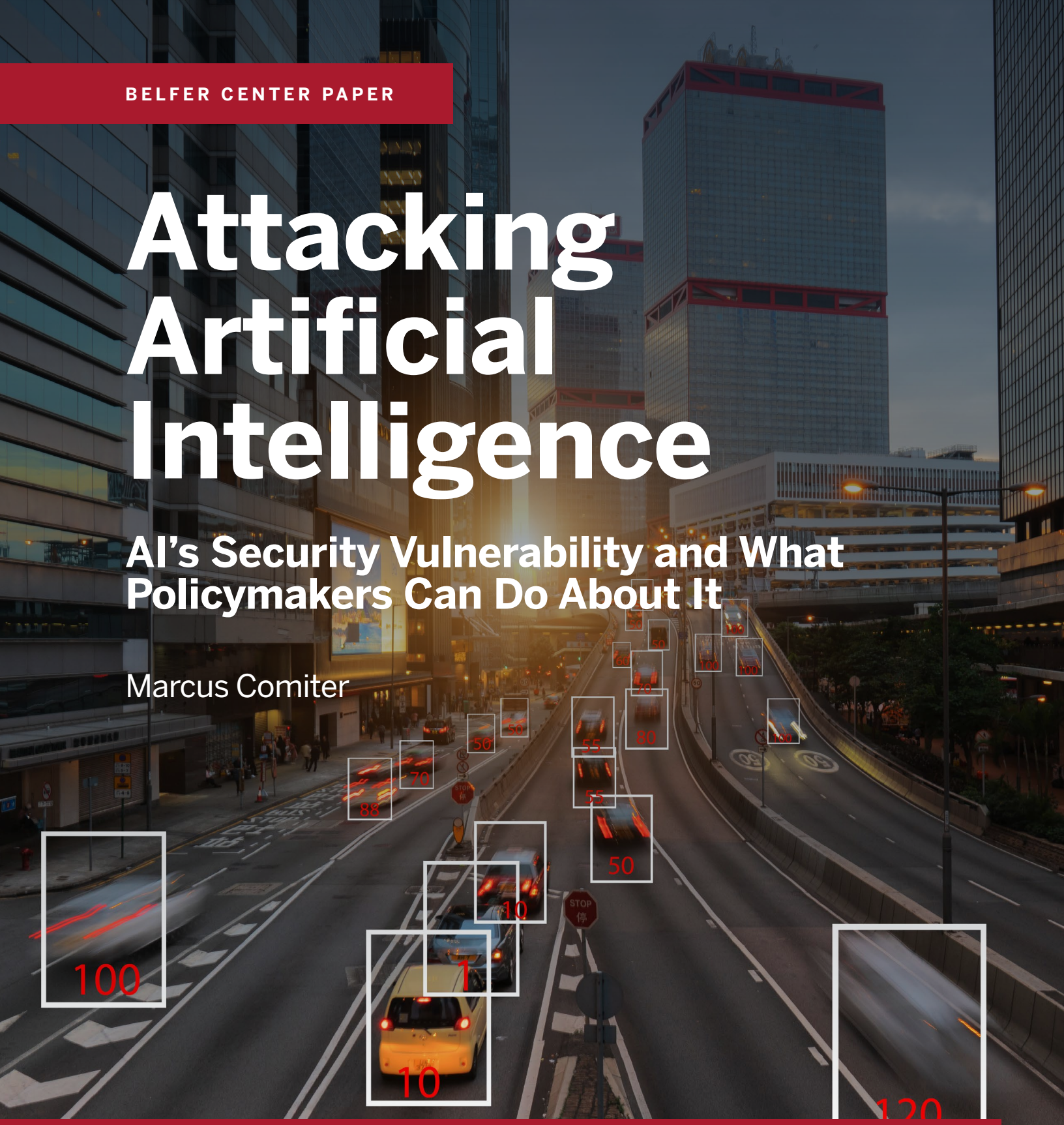
BELFER CENTER PAPER

Attacking Artificial Intelligence

AI's Security Vulnerability and What Policymakers Can Do About It

Marcus Comiter



HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

PAPER
AUGUST 2019

Belfer Center for Science and International Affairs
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

www.belfercenter.org

Design and Layout by Andrew Facini

Cover photo: Adobe Stock.

Printed in the United States of America

BELFER CENTER PAPER

Attacking Artificial Intelligence

AI's Security Vulnerability and What Policymakers Can Do About It

Marcus Comiter

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

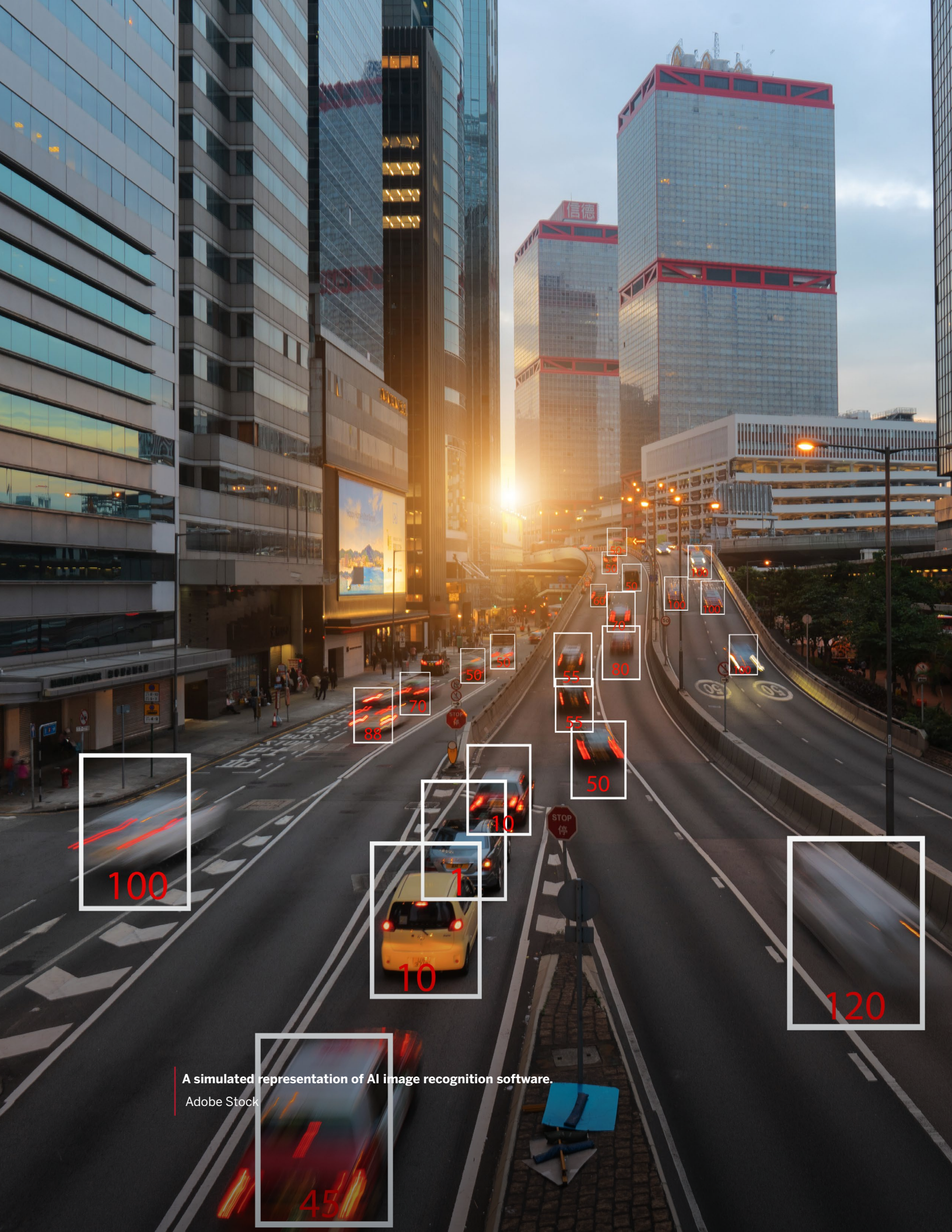
PAPER
AUGUST 2019

About the Author

Marcus Comiter is a Ph.D. Candidate in Computer Science at Harvard University and a Non-Resident Fellow at the Belfer Center for Science and International Affairs. Comiter's computer science research focuses on machine learning, computer and wireless networking (including 5G wireless networks), and security. Comiter's policy research leverages knowledge of the latest computer science research to study the public policy and cybersecurity implications of new technologies, such as machine learning/artificial intelligence and big data. His research has been published in both top computer science and public policy venues, and awarded a best paper award.

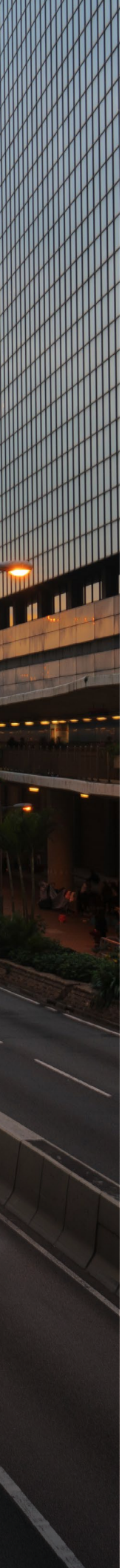
At Harvard, Comiter has helped design Harvard's first courses in blockchain/cryptocurrencies, 5G wireless networking, software defined networking, and the Internet of Things. Comiter's industry experience includes the Security and Microarchitecture groups at Intel Research Labs, where his work resulted in a patent, and serving as a Fellow at US Ignite, a Washington D.C. nonprofit catalyzing the development of next generation network applications. Comiter received his A.B. magna cum laude with highest honors in Computer Science and Statistics from Harvard University.

Table of Contents



A simulated representation of AI image recognition software.

Adobe Stock



Executive Summary

Artificial intelligence systems can be attacked.

The methods underpinning the state-of-the-art artificial intelligence systems are systematically vulnerable to a new type of cybersecurity attack called an "artificial intelligence attack." Using this attack, adversaries can manipulate these systems in order to alter their behavior to serve a malicious end goal. As artificial intelligence systems are further integrated into critical components of society, these artificial intelligence attacks represent an emerging and systematic vulnerability with the potential to have significant effects on the security of the country.

These "AI attacks" are fundamentally different from traditional cyberattacks.

Unlike traditional cyberattacks that are caused by "bugs" or human mistakes in code, AI attacks are enabled by inherent limitations in the underlying AI algorithms that currently cannot be fixed. Further, AI attacks fundamentally expand the set of entities that can be used to execute cyberattacks. For the first time, physical objects can be now used for cyberattacks (e.g., an AI attack can transform a stop sign into a green light in the eyes of a self-driving car by simply placing a few pieces of tape on the stop sign itself). Data can also be weaponized in new ways using these attacks, requiring changes in the way data is collected, stored, and used.

Critical parts of society are already vulnerable.

There are five areas most immediately affected by artificial intelligence attacks: content filters, the military, law enforcement, traditionally human-based tasks being replaced by AI, and civil society. These areas are attractive targets for attack, and are growing more vulnerable due to their increasing adoption of artificial intelligence for critical tasks.

This report proposes "AI Security Compliance" programs to protect against AI attacks.

Public policy creating "AI Security Compliance" programs will reduce the risk of attacks on AI systems and lower the impact of successful attacks. Compliance programs would accomplish this by encouraging stakeholders to adopt a set of best practices in securing systems against AI attacks, including considering attack risks and surfaces when deploying AI systems, adopting IT-reforms to make attacks difficult to execute, and creating attack response plans. This program is modeled on existing compliance programs in other industries, such as PCI compliance for securing payment transactions, and would be implemented by appropriate regulatory bodies for their relevant constituents.

Regulators should mandate compliance for governmental and high-risk uses of AI.

Regulators should require compliance both for government use of AI systems and as a pre-condition for selling AI systems to the government. In the private sector, regulators should make compliance mandatory for high-risk uses of AI where attacks would have severe societal consequences, and optional for lower-risk uses in order to avoid disrupting innovation.

Introduction

> " Artificial intelligence algorithms can be attacked and controlled by an adversary."

The terrorist of the 21st century will not necessarily need bombs, uranium, or biological weapons. He will need only electrical tape and a good pair of walking shoes. Placing a few small pieces of tape inconspicuously on a stop sign at an intersection, he can magically transform the stop sign into a green light in the eyes of a self-driving car. Done at one sleepy intersection, this would cause an accident. Done at the largest intersections in leading metropolitan areas, it would bring the transportation system to its knees. It's hard to argue with that type of return on a $1.50 investment in tape.

This is a study of how an obscure problem within artificial intelligence—currently the concern of a tiny subfield of yet another subfield of computer science—is on a dangerous collision course with the economic, military, and societal security of the future, and what can be done about it. The artificial intelligence algorithms that are being called upon to deliver this future have a problem: by virtue of the way they learn, they can be attacked and controlled by an adversary. What we see as a slightly vandalized stop sign, a compromised artificial intelligence system sees as a green light. *Call it an "artificial intelligence attack" (AI attack).*

This vulnerability is due to inherent limitations in the state-of-the-art AI methods that leave them open to a devastating set of attacks that are as insidious as they are dangerous. Under one type of attack, adversaries can gain control over a state-of-the-art AI system with a small but carefully chosen manipulation, ranging from a piece of tape on a stop sign[1] to a sprinkling of digital dust invisible to the human eye on a digital image.[2] Under another, adversaries can poison AI systems, installing backdoors that can be used at a time and place of their choosing to destroy the system. Whether it's causing a car to careen through a red light, deceiving

1 Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

2 Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014)

a drone searching for enemy activity on a reconnaissance mission, or subverting content filters to post terrorist recruiting propaganda on social networks, the danger is serious, widespread, and already here.

However, just as not all applications of AI are "good," not all AI attacks are necessarily "bad." As autocratic regimes turn to AI as a tool to monitor and control their populations, AI "attacks" may be used as a protective measure against government oppression, much like technologies such as Tor and VPNs are.

Regardless of their use, AI attacks are different from the cybersecurity problems that have dominated recent headlines. These attacks are not bugs in code that can be fixed—they are inherent in the heart of the AI algorithms. As a result, exploiting these AI vulnerabilities requires no "hacking" of the targeted system. In fact, attacking these critical systems does not even always require a computer. This is a new set of cybersecurity problems, and cannot be solved with the existing cybersecurity and policy toolkits governments and businesses have assembled. Instead, addressing this problem will require new approaches and solutions.

" Addressing this problem will require new approaches and solutions."

Given time, researchers may discover a technical silver bullet to some of these problems. But time ran out yesterday. For a technology that was nascent a decade ago, AI is now being used as a key ingredient across every industry, from Main Street to Wall Street, from the baseball diamond to the battlefield. And on cue, as with every other recent technological development—the Internet, social media, and the Internet of Things—in our haste we are turning a blind eye to fundamental problems that exist.

This report seeks to provide policymakers, politicians, industry leaders, and the cybersecurity community an understanding of this emerging problem, identify what areas of society are most immediately vulnerable, and set forth policies that can be adopted to find security in this important new era.

The report is split into four sections. First, it begins by giving an accessible yet comprehensive description of how current AI systems can be attacked, the forms of these attacks, and a taxonomy for categorizing them.

Second, the report identifies the most critical areas affected by this new class of vulnerabilities. While the number of systems affected by this new threat will only grow as AI increases its penetration into the modern world, this report focuses on five high priority areas that require immediate attention: content filters, military, law enforcement, human tasks being replaced with AI, and civil society.

Third, the report contextualizes AI vulnerabilities within the larger cybersecurity landscape. It argues that AI attacks constitute a new vertical of attacks distinct in nature and required response from existing cybersecurity vulnerabilities. This section also discusses the use of AI attacks as an offensive cyber weapon.

Fourth, the report proposes the idea of "AI Security Compliance" programs to protect against AI attacks. These compliance programs will reduce the risk of attacks on AI systems and lower the impact of successful attacks. They will accomplish this by encouraging stakeholders to adopt a set of best practices in securing systems against AI attacks, including considering attack risks and surfaces when deploying AI systems, adopting IT-reforms that will make attacks more difficult to execute, and creating attack response plans to mitigate attack damage.

The report further suggests regulators should mandate compliance in portions of both the public and private sectors. In the public sector, compliance should be mandated for governmental uses of AI and be a pre-condition for private firms selling AI systems to the government. In the private sector, compliance should be mandated for high-risk private sector AI applications, but should be optional for lower-risk uses in order to avoid disrupting innovation in this rapidly changing field.

This policy will improve the security of the community, military, and economy in the face of AI attacks. But for policymakers and stakeholders alike, the first step towards realizing this security begins with understanding the problem, which we turn our attention to now.

Part I: Technical Problem

Understanding the Problem through a Historical Analogy

General George Patton may have won the D-Day campaign for the Allies without ever firing a shot. In support of the future D-Day landings, Patton was given charge of the First United States Army Group (FUSAG). Rather than fighting in arms, the FUSAG fought in deception. To convince the German command that the invasion point would be Pas de Calais rather than Normandy, the FUSAG orchestrated a major force deployment—including hundreds of tanks and other vehicles—directly across the English Channel from it.

These tanks, however, were not what they seemed. Unable to spare the vehicles needed for this show of force from the actual war effort, the Allies instead used inflatable balloons painted to look like tanks. Although more characteristic of a technique employed by Bugs Bunny against Elmer Fudd than George Patton against Nazis, it did the trick. German reconnaissance was fooled. The images captured by the Luftwaffe planes were interpreted as a major buildup of forces in anticipation of an invasion of Pas de Calais, leaving the beaches of Normandy under-fortified.[3]

Given access to the site, we would not expect a human to mistake what was essentially a painted balloon for a multi-ton metal machine. But German reconnaissance worked by recognizing patterns: the shapes and markings representing tanks and other military assets in images. Relegated to pattern matching, German reconnaissance was easy to fool with a few strategic markings placed on the inflatable balloons. Although surprising, this is the same flaw that dooms AI algorithms, allowing them to be fooled in similar and even more pernicious manners.

3 Knighton, Andrew, "FUSAG: The Ghost Army—Patton's D-Day Force That Was Only Threat In The Enemy's Imagination", 14 May 2017, https://www.warhistoryonline.com/world-war-ii/fusag-the-ghost-army-pattons-d-day-force-that-was-only-a-threat-xb.html.

To understand why AI systems are vulnerable to the same weakness, we must briefly examine how AI algorithms, or more specifically the machine learning techniques they employ, "learn." Just like the reconnaissance officers, the machine learning algorithms powering AI systems "learn" by extracting patterns from data. These patterns are tied to higher-level concepts relevant to the task at hand, such as which objects are present in an image. As an example, consider the task of an AI algorithm on a self-driving car learning to recognize a stop sign. For this task, the algorithm "learns" by being shown a dataset containing hundreds or thousands of examples of stop signs and extracting patterns of colors and shapes representative of it. When later tasked to identify if a particular sign is a stop sign, the algorithm scans the image looking for the patterns it has learned to associate with a stop sign. If the patterns match, the algorithm can instruct the car to stop. If the patterns match that of a different sign, such as a new faster speed limit, the algorithm can similarly instruct the car to speed up.

> " Only a few stray marks or subtle changes to a handful of pixels in an image are needed to destroy an AI system."

Just as the FUSAG could expertly devise what patterns needed to be painted on the inflatable balloons to fool the Germans, with a type of AI attack called an "input attack," adversaries can craft patterns of changes to a target that will fool the AI system into making a mistake. This attack is possible because when patterns in the target are inconsistent with the variations seen in the dataset, as is the case when an attacker adds these inconsistent patterns *purposely*, the system may produce an arbitrary result. However, unlike the tank example, these patterns or markings need not be as blatant. This is because AI algorithms process information differently than humans do. As a result, while it may have been necessary to make the balloons actually look like tanks to fool a human, to fool an AI system, only a few stray marks or subtle changes to a handful of pixels in an image are needed to destroy an AI system.

These input attacks are only one type of AI attack. Another—known as a poisoning attack—can stop an AI system from operating correctly in situations, or even insert a backdoor that can later be exploited by an adversary. Continuing the analogy, poisoning attacks would be the equivalent of

hypnotizing the German analysts to close their eyes anytime they were about to see any valuable information that could be used to hurt the Allies.

As a whole, these attacks have the characteristics of a severe cyber threat: they are versatile in form, widely applicable to many domains, and hard to detect. They can take the form of a smudge or squiggle on a physical target, or be hidden within the DNA of an AI system. They can target assets and systems in the real world, such as making stop signs invisible to driverless cars, and in the cyber world, such as hiding child pornography from the content detectors seeking to stop its spread. Perhaps most concerning is that AI attacks can be pernicious and difficult to detect. Attacks can be completely invisible to the human eye. Conversely, they can be grand and hidden in plain sight, made to look like they fit in perfectly with their surroundings.

But what exactly are AI attacks? Why do they exist? And what do they look like? We now turn our attention to understanding the technical basis of these attacks in order to answer these questions.

Overview of Artificial Intelligence Attacks

An artificial intelligence attack (AI attack) is the purposeful manipulation of an AI system with the end goal of causing it to malfunction. These attacks can take different forms that strike at different weaknesses in the underlying algorithms:

- **Input Attacks:** manipulating what is fed into the AI system in order to alter the output of the system to serve the attacker's goal. Because at its core every AI system is a simple machine—it takes an input, performs some calculations, and returns an output—manipulating the input allows attackers to affect the output of the system.
- **Poisoning Attacks:** corrupting the process during which the AI system is created so that the resulting system malfunctions in a way desired by the attacker. One direct way to execute a poisoning attack is to corrupt the data used during this process. This is because the state-of-the-art machine learning methods powering AI work by "learning" how to do a task, but they "learn" from one source and one source only: data. Data is its water, food, air, and true love. Poison the data, poison the AI system. Poisoning attacks can also compromise the learning process itself.

As AI systems are integrated into critical commercial and military applications, these attacks can have serious, even life-and-death, consequences. AI attacks can be used in a number of ways to achieve a malicious end goal:

- **Cause Damage:** the attacker wants to cause damage by having the AI system malfunction. An example of this is an attack to cause an autonomous vehicle to ignore stop signs. By attacking the AI system so that it incorrectly recognizes a stop sign as a different sign or symbol, the attacker can cause the autonomous vehicle to ignore the stop sign and crash into other vehicles and pedestrians.
- **Hide Something:** the attacker wants to evade detection by an AI system. An example of this is an attack to cause a content filter tasked with blocking terrorist propaganda from being posted on

a social network to malfunction, therefore letting the material propagate unencumbered.

- **Degrade Faith in a System:** the attacker wants an operator to lose faith in the AI system, leading to the system being shut down. An example of this is an attack that causes an automated security alarm to misclassify regular events as security threats, triggering a barrage of false alarms that may lead to the system being taken offline. For example, attacking a video-based security system to classify a passing stray cat or blowing tree as a security threat may cause the security system to be taken offline, therefore allowing a true threat to then evade detection.

Given the unparalleled success of AI over the past decade, it is surprising to learn that these attacks are possible, and even more so, that they have not yet been fixed. We now turn our attention to why these attacks exist, and why it is so difficult to prevent them.

Why Do Artificial Intelligence Attacks Exist?

AI attacks exist because there are fundamental limitations in the underlying AI algorithms that adversaries can exploit in order to make the system fail. Unlike traditional cybersecurity attacks, these weaknesses are not due to mistakes made by programmers or users. They are just shortcomings of the current state-of-the-art methods. Put more bluntly, the algorithms that cause AI systems to work so well are imperfect, and their systematic limitations create opportunities for adversaries to attack. At least for the foreseeable future, this is just a fact of mathematical life.

" The algorithms that cause AI systems to work so well are imperfect, and their systematic limitations create opportunities for adversaries to attack. At least for the foreseeable future, this is just a fact of mathematical life."

To see why this is the case, we need to understand how the algorithms underpinning AI work. Many current AI systems are powered by machine learning,[4] a set of techniques that extract information from data in order to "learn" how to do a given task. A machine learning algorithm "learns" analogously to how humans learn. Humans learn by seeing many examples of an object or concept in the real world, and store what is learned in the brain for later use. Machine learning algorithms "learn" by seeing many examples of an object or concept in a *dataset*, and store what is learned in a *model* for later use. In many if not most AI applications based on machine learning, there is no outside knowledge or other magic used in this process: it is entirely dependent on the dataset and nothing else.[5]

4 As a note on terminology, artificial intelligence and machine learning are popularly used interchangeably. In a more exact sense, the two are distinct. Artificial intelligence is a broader term that generally refers to the ability of computer systems to execute complex tasks performed by humans. Machine learning is one particular method used to power artificial intelligence, and is a set of techniques and algorithms that "learn" by extracting patterns from data. Due to the overwhelming success of machine learning algorithms compared to other methods, many artificial intelligence systems today are based entirely on machine learning. As a result, the attacks and vulnerabilities described in this report affect both artificial intelligence and machine learning systems.

5 Production machine learning systems may feature a good amount of human and guard rail engineering, while others may be fully data dependent. As a result, some production systems may fall along a spectrum between "learned" systems that are fully data dependent and "designed" systems that are heavily based on hand-designed features. However, systems that are closer to the "designed" side of the spectrum may still be vulnerable to attacks, such as input attacks. Further, given the success of learning, which often captures patterns and relations that could not be designed manually by human model designers, many if not most systems will rely heavily on learned features, and be vulnerable to attacks.

The key to understanding AI attacks is understanding what the "learning" in machine learning actually is, and more importantly, what it is not. Recall that machine learning "learns" by looking at many examples of a concept or object in a dataset. More specifically, it uses algorithms that extract and generalize common patterns in these examples. These patterns are stored within the model. Taking the example of recognizing a stop sign, the learning algorithm will identify patterns in the pixels that make up the example images, such as large areas of red, the shapes of the letters "S" "T" "O" and "P", and other defining characteristics. When the model is later called upon to detect a stop sign in a new image, it will search that image for the same patterns of pixels. If it finds patterns that match those it has learned to associate with a stop sign, it will output that it has found a stop sign. If it instead finds patterns that match those it has learned to associate with a different object, such as a green light, it will output that it has found a green light. These patterns are "general" in the sense that they should work in new settings, not just on the examples from which it learned. For example, the patterns in the example above should be able to recognize all stop signs, not just the particular ones included in the dataset.

Given enough data, the patterns learned in this manner are of such high quality that they can even outperform humans on many tasks. This is because if the algorithm sees enough examples in all of the different ways the target naturally appears, it will learn to recognize all the patterns needed to perform its job well. Continuing the stop sign example, if the dataset contains images of stop signs in the sun and shade, from straight ahead and from different angles, during the day and at night, it will learn all the possible ways a stop sign can appear in nature.

However, this process already introduces a significant vulnerability: it is wholly dependent on the dataset. Because the dataset is the model's only source of knowledge, if it is corrupted or "poisoned" by an attacker, the model learned from this data will be compromised. Attackers can poison the dataset to stop the model from learning specific patterns, or more insidiously, to install secret backdoors that can be used to trick the model in the future.[6]

6 Bagdasaryan, Eugene, et al. "How to backdoor federated learning." arXiv preprint arXiv:1807.00459 (2018).

But the problems do not end there. Even assuming a non-corrupted dataset and highly accurate model, this success comes with a very important caveat: the patterns "learned" by current state-of-the-art machine learning models are relatively brittle. As a result, the model only works on data that is similar in nature to the data used during the learning process. If used on data that is *even a little* different in nature from the types of variations it saw in the original dataset, the model may utterly fail. This is a major limitation attackers can exploit: by introducing artificial variations—such as a piece of tape or other aberrant patterns—the attacker can disrupt the model and control its behavior based on what artificial pattern is introduced. Because the amount of data used to build the model is finite but the amount of artificial variations an attacker can create are infinite, the attacker has an inherent advantage.

This explains how the stop sign tape attack can cause a self-driving car to crash. While the dataset used to train the stop sign detector contains plenty of variations of stop signs in different natural conditions, it doesn't contain examples of the endless ways it can be artificially manipulated by an attacker, such as with tape and graffiti. Because of this, very small artificial manipulations chosen in *just* the right way can break the relatively brittle patterns the model learned, and have preposterously huge impacts on the model's output. This is why a small piece of tape can transform a stop sign into a green light so easily: it doesn't have to make the entire stop sign look like a green light, it only has to trick the specific small brittle patterns that the model learned. Unfortunately, this is easy to do.

" Many may be surprised to learn that machine learning has such a glaring shortcoming."

Many may be surprised to learn that machine learning has such a glaring shortcoming. This is because popular culture has shaped a widespread but erroneous belief that machine learning actually "learns" in a human sense of the word. Humans are good at truly learning concepts and associations. If a stop sign is distorted or defaced with graffiti or dirt, even a human who has never seen graffiti or a dirty stop sign would still reliably and consistently identify it as a stop sign, and certainly would not mistake it for an entirely different object altogether, such as a green light. But we now know current AI systems do not work in the same way. Even a model that can almost perfectly

recognize a stop sign still has no knowledge of the concept of a stop sign, or even a sign for that matter, as a human does. It only knows that certain learned patterns correspond to a label named "stop sign."

While it may seem that this distinction between human learning and machine "learning" is arbitrary—especially because if the model works, it seems we should be happy—we now understand why it has such severe ramifications: under contested conditions, AI systems can be made to fail even if they are extremely successful under "normal" conditions.

A logical step to combat this would be to understand why the patterns the model learns are so brittle. However, this is not currently supported in the most widely used models, such as deep neural networks, as exactly how and even what these models learn is still not fully understood. As a result, the most popular machine learning algorithms powering AI, like neural networks, are referred to as "black boxes": we know what goes in, we know what comes out, but we do not know exactly what happens in between. We cannot reliably fix what we do not understand. And for this same reason, it is difficult if not impossible to even tell if a model is being attacked or just doing a bad job. While other data science methods, such as decision trees and regression models, allow for much more explainability and understanding, these methods do not generally deliver the performance that the widely used neural networks are capable of providing.

From this understanding, we can now state the characteristics of the machine learning algorithms underpinning AI that make these systems vulnerable to attack.

- **Characteristic 1: Machine learning works by "learning" relatively brittle patterns that work well but are easy to disrupt.** Contrary to popular belief, machine learning models are not "intelligent" or capable of truly mimicking human ability on tasks, even tasks they perform well. Instead, they work by learning brittle statistical associations that are relatively easy to disrupt. Attackers can exploit this brittleness to craft attacks that destroy the performance of an otherwise excellent model.

- **Characteristic 2: Dependence solely on data provides a main channel to corrupt a machine learning model.** Machine learning "learns" solely by extracting patterns from a set of examples known as a dataset. Unlike humans, machine learning models have no baseline knowledge that they can leverage—their entire knowledge depends wholly on the data they see. Poisoning the data poisons the AI system. Attacks in this vein essentially turn an AI system into a Manchurian candidate that attackers can activate at a time of their choosing.
- **Characteristic 3: The black box nature of the state-of-the-art algorithms makes auditing them difficult.** Relatively little is understood about how the widely used state-of-the-art machine learning algorithms, such as deep neural networks, learn and work—even today they are still in many ways a magical black box. This makes it difficult, if not currently impossible, to tell if a machine learning model has been compromised, or even if it is being attacked or just not performing well. This characteristic sets AI attacks apart from traditional cybersecurity problems where there are clear definitions of vulnerabilities, even if they are hard to find.

Taken together, these weaknesses explain why there are no perfect technical fixes for AI attacks. These vulnerabilities are not "bugs" that can be patched or corrected as is done with traditional cybersecurity vulnerabilities. They are deep-seated issues at the heart of current state-of-the-art AI itself.

Now that we have an understanding of why these attacks are possible, we now turn our attention to looking at actual examples of these attacks.

Input Attacks

Input attacks trigger an AI system to malfunction by altering the input that is fed into the system. As shown in the figure below, this is done by adding an "attack pattern" to the input, such as placing tape on a stop sign at an intersection or adding small changes to a digital photo being uploaded to a social network.

Input attacks do not require the attacker to have corrupted the AI system in order to attack it. Completely state-of-the-art AI systems that are highly accurate and have never had their integrity, dataset, or algorithms compromised are still vulnerable to input attacks. And in stark contrast to other cyberattacks, the attack itself does not always use a computer!

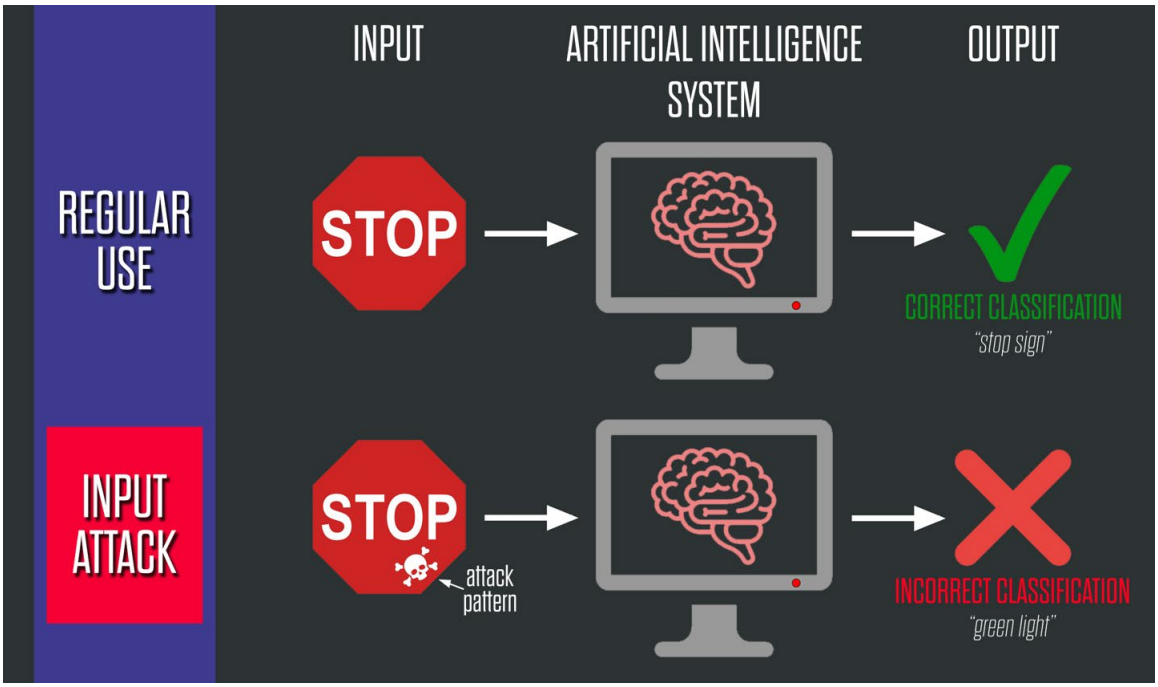


Figure 1: In regular use (top), the AI system takes a valid input, processes it with the model (brain), and returns an output. In an input attack (bottom), the input to the AI system is altered with an attack pattern, causing the AI system to return an incorrect output.

These attacks are particularly dangerous because the attack patterns do not have to be noticeable, and can even be completely undetectable. Adversaries can be surgical, changing just a small aspect of the input in a precise and exact way to break the patterns learned previously by the model. For attacks on physical objects that must be captured by a sensor or camera before being fed into an AI system, attackers can craft small changes that are just big enough to be captured by the sensor. This is the canonical "tape attack": attackers figure out that placing a two-inch piece of white tape on the upper corner of a stop sign will exploit a particular brittleness in the patterns learned by the model, turning it into a green light.[7] For attacks on digital objects that are fed directly into the AI system, such as an image uploaded to a social network, the attack patterns can be imperceivable to the human eye. This is because in this all-digital setting, the alterations can occur on an individual pixel level, creating alterations that are so small they are literally invisible to the human eye.

Categorizing input attacks

> "Adversaries will choose a form for their attack that fits their particular scenario and mission."

The most interesting aspect of input attacks is how varied they are. Input attacks on AI systems are like snowflakes: no two are exactly alike. The first step in securing systems from these attacks is to create a taxonomy to bring order to the endless attack possibilities. "Form fits function" is an appropriate lens with which to do so: adversaries will choose a form for their attack that fits their particular scenario and mission. Therefore, a taxonomy should follow this same tendency.

Input attack forms can be characterized along two axes: perceivability and format. Perceivability characterizes if the attack is perceivable to humans (e.g., for AI attacks on physical entities, is the attack visible or invisible to the human eye). Format characterizes if the attack vector is a physical real-world object (e.g., a stop sign), or a digital asset (e.g., an image file on a computer). The figure below shows this taxonomy.

7 Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

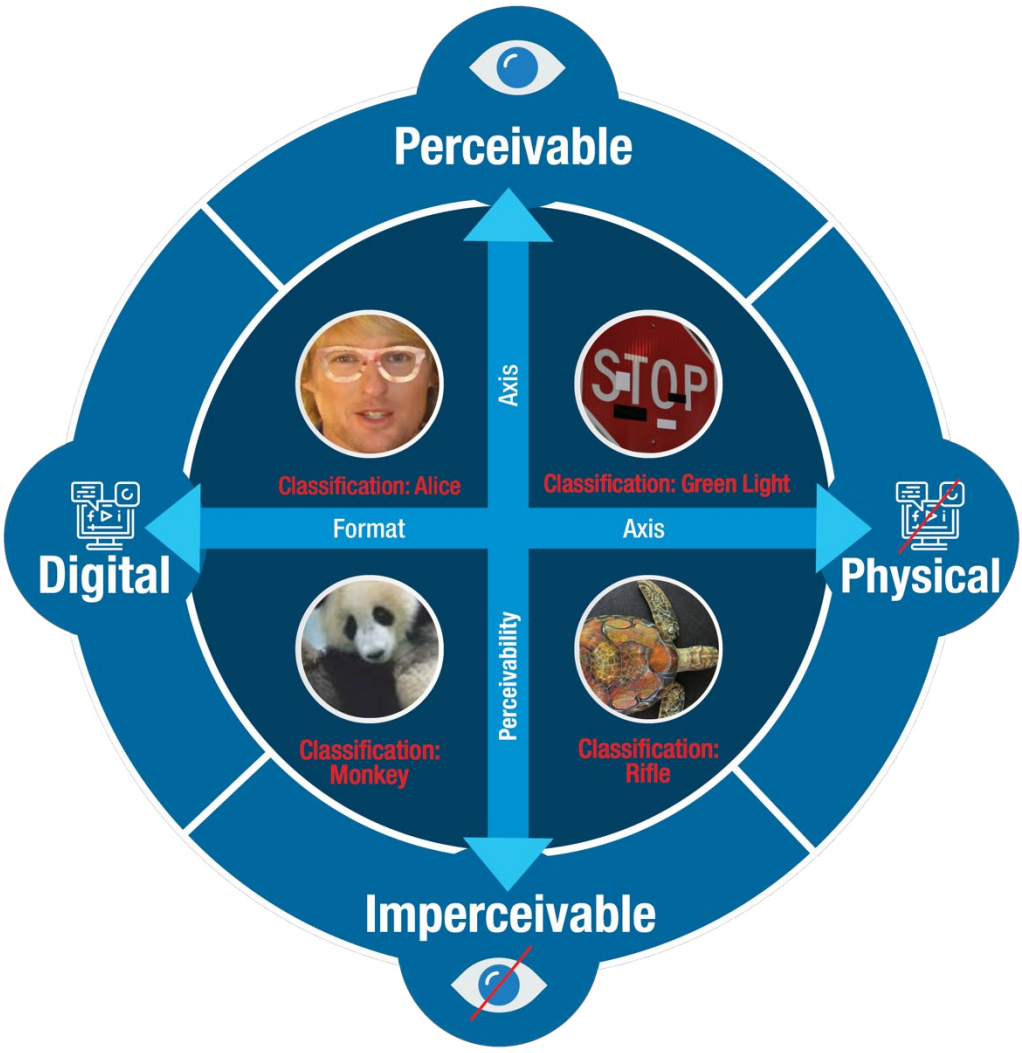


Figure 2: Taxonomy for categorizing input attacks. The horizontal axis characterizes the format of the attack, either in the physical world or digital. The vertical axis characterizes the perceivability of the attack, either perceivable to humans or imperceivable to humans.

(See footnote[8] for thumbnail images citations.)

8 Graphic by Marcus Comiter except for stop sign attack thumbnail from Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, panda attack thumbnail from Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014), turtle attack thumbnail from Athalye, Anish, et al. "Synthesizing robust adversarial examples." arXiv preprint arXiv:1707.07397 (2017), and celebrity attack thumbnail from Sharif, Mahmood, et al. "Adversarial generative nets: Neural network attacks on state-of-the-art face recognition." arXiv preprint arXiv:1801.00349 (2017).

Perceivability Axis

We first discuss the perceivability axis. On one end of the axis are "perceivable" attacks in which the input attack pattern is able to be noticed by humans. The attack patterns can be alterations to the target itself, such as deforming, removing a portion of, or altering the color of the target. Alternatively, the attack pattern may be an addition to the target, such as affixing tape or other decals to the physical target, or adding digital marks to a digital target. Examples of perceivable attacks include defacing a stop sign with patterns formed from tape,[9] or using software to superimpose objects such as glasses[10] on a digital image of a subject (as many popular apps like Snapchat do).

The figure below shows how a perceivable attack is formed for a physical object. A regular object is altered with a visible attack pattern (a few pieces of tape) to form the attack object. While the regular object would be classified correctly by the AI system, the attack object is incorrectly classified as a "green light".

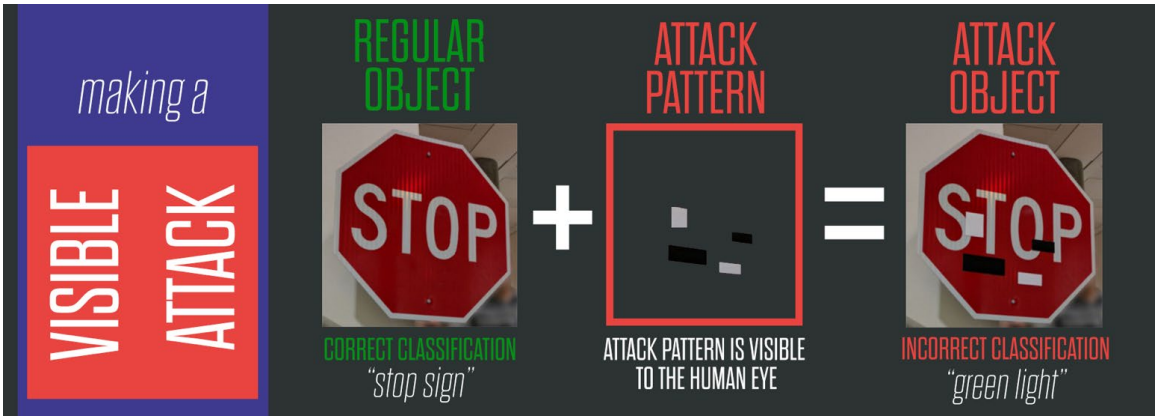


Figure 3: Crafting a visible input attack. A small attack pattern is affixed to the physical object, making the AI system misclassify the image with a small change in its appearance.

(See footnote[11] for thumbnail images citations.)

9 Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

10 Sharif, Mahmood, et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016.

11 Graphic by Marcus Comiter except for stop sign noise thumbnail and stop sign attack thumbnail from Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

Although perceivable attacks are noticeable by humans, they can still be highly effective for a number of reasons. First, perceivable attacks need not be ostentatious. A visible attack in the form of a few carefully chosen pieces of tape placed on a stop sign is able to be perceived, but will not necessarily be *noticed*. Humans are naturally conditioned to ignore small changes in their environment, such as graffiti, vandalism, and natural wear and tear. As such, perceivable attacks may go completely unnoticed. Second, perceivable attacks can be crafted to hide in plain sight. A visible attack in the form of specially designed glasses or a specially crafted logo added to a person's t-shirt would be noticed, but would not be suspected of being an attack, effectively hiding in plain sight. In this case, rather than crafting an attack to be as small as possible, it may actually be more effective for it to be large but blend into its surroundings.

On the other end of the visibility axis are "imperceivable" attacks that are invisible to human senses. Imperceivable attacks can take many forms. For digital content like images, these attacks can be executed by sprinkling "digital dust" on top of the target.[12] Technically, this dust is in the form of small, unperceivable perturbations made to the entire target. Each small portion of the target is changed so slightly that the human eye cannot perceive the change, but in aggregate, these changes are enough to alter the behavior of the algorithm by breaking the brittle patterns learned by the model. The figure below shows how an imperceivable attack is formed in this manner. A normal digital image is altered with tiny, imperceivable pixel-level perturbations scattered throughout the image, forming the attack image. While the regular image would be classified correctly by the AI system as a "panda", the attack object is incorrectly classified as a "monkey". However, because the attack pattern makes such small changes, to the human eye, the attack image looks identical to the original regular image.

" For digital content like images, these 'imperceivable' attacks can be executed by sprinkling 'digital dust' on top of the target."

12 Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014)

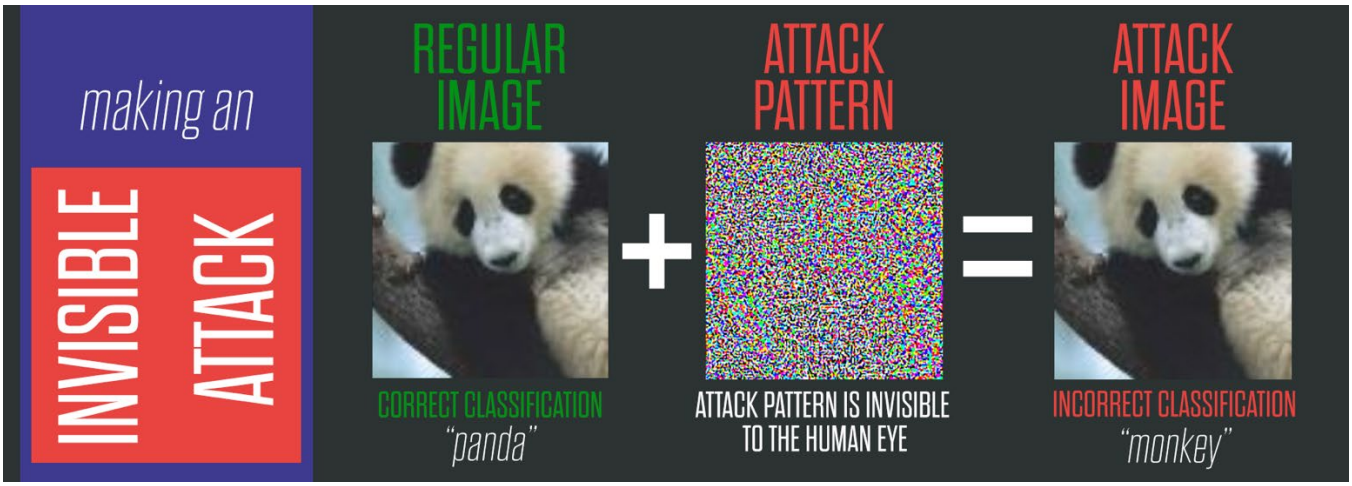


Figure 4: Crafting an invisible input attack. A small amount of noise that is invisible to the human eye is added to the entire image, making the AI system misclassify the image without changing its appearance. (Image concept from footnote[13], see footnote[14] for thumbnail images citations.)

Imperceivable attacks are not limited to just digital objects. For example, attack patterns can be added in imperceivable ways to a physical object itself. Researchers have shown that a 3D-printed turtle with an imperceivable input attack pattern could fool AI-based object detectors.[15] While turtle detection may not have life and death consequences (yet...), the same strategy applied to a 3D-printed gun may. In the audio domain, high pitch sounds that are imperceivable to human ears but able to be picked up by microphones can be used to attack audio-based AI systems, such as digital assistants.

These imperceivable attacks are particularly pernicious from a security standpoint. Unlike visible attacks, there is no way for humans to observe if a target has been manipulated. This poses an additional barrier to detecting these attacks.

13 Image concept showing how attack is formed from Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014)

14 Graphic by Marcus Comiter except for panda image thumbnail, noise image thumbnail, and panda attack thumbnail from Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

15 Athalye, Anish, et al. "Synthesizing robust adversarial examples." arXiv preprint arXiv:1707.07397 (2017).

Imperceivable attacks are highly applicable to targets that the adversary has full control over, such as digital images or manufactured objects. For example, a user posting an illicit image, such as one containing child pornography, can alter the image such that it evades detection by the AI-based content filters, but also remains visually unchanged from the human perspective. This allows the attacker unfettered and, for all practical purposes, unaltered distribution of the content without detection.

Format

We next discuss the format axis. On one end of the axis are "physical" attacks. These are attacks in which the target being attacked exists in the physical world. While physical attacks are easiest to think of in terms of objects, including stop signs, fire trucks, glasses, and even humans, they are also applicable to other physical phenomena, such as sound. For example, attacks have been shown on voice controlled digital assistants, where a sound has been used to trigger action from the digital assistant.[16] Alterations are made directly to or placed on top of these targets in order to craft an attack. Examples of physical attacks on real-world objects are shown in the figure below.

In some settings, attacks on physical objects may require larger, coarser attack patterns. This is because these physical objects must first be digitized, for example with a camera or sensor, to be fed into the AI algorithm, a process that can destroy finer level detail. However, even with this digitization requirement, attacks may still be difficult to perceive. The "attack turtle" that is incorrectly classified as a rifle in the example shown below is one such example of a physical attack that is nearly invisible. The 3D-printed turtle is manufactured to have a very subtle pattern that blends naturally with its shell and flipper patterns—making the attack imperceivable—but consistently deceives the classifier regardless of the angle and position from which it is viewed by the camera.[17] By "cloaking" the object in this attack pattern, it can deceive the AI system without appearing as an attack to a human observer.

16 Carlini, Nicholas, and David Wagner. "Audio adversarial examples: Targeted attacks on speech-to-text." 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018.

17 Athalye, Anish, et al. "Synthesizing robust adversarial examples." arXiv preprint arXiv:1707.07397 (2017).

Figure 5: Examples of physical attacks on real world objects. Physical attacks can be perceivable, as with the stop sign or yellow glasses, or imperceivable, as with the 3D-printed turtle and baseball shown here.
(See footnote[18] for thumbnail images citations.)

On the other end of the format axis are "digital" attacks. These are attacks in which the target being attacked is a digital asset. Examples include images, videos, social media posts, music, files, and documents. Unlike physical targets that must first be sensed and digitized, digital targets are fed directly in their original state into the AI system. This gives adversaries an expanded selection of attacks and lowers the difficulty of crafting a successful attack, as they do not need to account for possible distortion of the attack pattern during this sensing process. As such, digital attacks are particularly well suited to invisibility. Examples of digital attacks on digital images are shown in the figure below. (While the digital attacks shown in this figure are all digital images, this choice is for presentation purposes, and attacks can also target other digital assets such as videos and files.)

18 Graphic by Marcus Comiter except for stop sign attack thumbnail from Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018., turtle attack thumbnail and baseball attack thumbnail from Athalye, Anish, et al. "Synthesizing robust adversarial examples." arXiv preprint arXiv:1707.07397 (2017), and girl with glasses attack thumbnail from Sharif, Mahmood, et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016.

Figure 6: Examples of digital attacks on digital images. Digital attacks can be perceivable, as with the silly glasses superimposed on the picture of a celebrity (middle), or imperceivable, as with the panda and duck images shown here (left, right). (See footnote[19] for thumbnail images citations.)

Crafting an Input Attack

Once attackers have chosen an attack form that suits their needs, they must craft the input attack. The difficulty of crafting an attack is related to the types of information available to the attacker. However, it is important to note that attacks are still practical (although potentially more challenging to craft) even under very difficult and restrictive conditions.

An input attack is relatively easy to craft if the attacker has access to the AI model being attacked. Armed with this, the attacker can automatically craft attacks using simple textbook optimization methods. Publicly available software implementing these methods is already available.[20] Attackers can also use Generative Adversarial Networks (GANs), a method specifically created to exploit weaknesses in AI models, to craft these attacks.[21]

19 Graphic by Marcus Comiter except for panda attack thumbnail from Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014), celebrity attack thumbnail from Sharif, Mahmood, et al. "Adversarial generative nets: Neural network attacks on state-of-the-art face recognition." arXiv preprint arXiv:1801.00349 (2017), and goose attack thumbnail from Gong, Yuan, and Christian Poellabauer. "Protecting Voice Controlled Systems Using Sound Source Identification Based on Acoustic Cues." 2018 27th International Conference on Computer Communication and Networks (ICCCN). IEEE, 2018.

20 See, e.g., https://github.com/tensorflow/cleverhans

21 Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems*. 2014.

While it may seem shocking that attackers would have access to the model, there are a number of common scenarios in which this would occur routinely. On the more innocent side of the spectrum, models are often made public because they have been optimized by researchers or companies for an important general task, such as object recognition, and then made public for anyone to use as part of the "open source" movement.[22] On the more sinister side of the spectrum, attackers can hack the system storing the model in order to steal it. The model itself is just a digital file living on a computer, no different from an image or document, and therefore can be stolen like any other file on a computer. Because models are not always seen as highly sensitive assets, the systems holding these models may not have high levels of cybersecurity protection. History has shown that when software capabilities are commoditized, as they are becoming with AI systems, they are often not handled or invoked carefully in a security sense, as demonstrated by the prevalence of default root passwords. If this history is any indication, the systems holding these models will suffer from similar weaknesses that can lead to the model being easily stolen.

Even in cases where the attacker does not have the model, it is still possible to mount an input attack. If attackers have access to the dataset used to train the model, they can use it to build their own copy of the model, and use this "copy model" to craft their attack. Researchers have shown that attacks crafted using these "copy models" are easily transferable to the originally targeted models.[23] As was the case with models, there are a number of common scenarios in which the attacker would have access to the dataset. Like models themselves, datasets are made widely available as part of the open source movement, or could similarly be obtained by hacking the system storing this dataset. In an even more restrictive setting where the dataset is not available, attackers could compile their own *similar* dataset, and use this similar dataset to build a "copy model" instead.

" Even in cases where the attacker does not have the model, it is still possible to mount an input attack."

In an increasingly more restrictive case where attackers do not have access to the model or the dataset, but have access to the output of the model,

22 See, e.g., YOLO: Real-Time Object Detection, https://pjreddie.com/darknet/yolo/

23 Liu, Yanpei, et al. "Delving into transferable adversarial examples and black-box attacks." arXiv preprint arXiv:1611.02770 (2016).

they can still craft an attack. This situation occurs often in practice, with businesses offering Artificial Intelligence as a Service via a public API.[24] This service gives users the output of an AI model trained for a particular task, such as object recognition. While these models and their associated datasets are kept private, attackers can use the output information from their APIs to craft an attack. This is because this output information replaces the need for having the model or the dataset.

In the hardest case where nothing about the model, its dataset, or its output is available to the attacker, the attacker can still try to craft attacks by brute force trial-and-error. For example, an attacker trying to beat an online content filter can keep generating random attack patterns and uploading the content to see if it is removed. Once a successful attack pattern is found, it can be used in future attacks.

24 See, e.g., "Machine Learning on AWS: Putting Machine Learning in the Hands of Every Developer", https://aws.amazon.com/machine-learning/

Poisoning Attacks

Poisoning attacks are the second class of AI attacks. In poisoning attacks, the attacker seeks to damage the AI model itself so that once it is deployed, it is inherently flawed and can be easily controlled by the attacker. Unlike input attacks, model poisoning attacks take place while the model is being learned, fundamentally compromising the AI system itself.

To poison an AI system, the attacker must compromise the learning process in a way such that the model fails on certain attacker-chosen inputs, or "learns" a backdoor that the attacker can use to control the model in the future. One motivation is to poison a model so that it fails on a particular task or types of input. For example, if a military is training an AI system to detect enemy aircraft, the enemy may try to poison the learned model so that it fails to recognize certain aircraft.

Data is a major avenue through which to execute a poisoning attack. Because information in the dataset is distilled into the AI system, any problems in the dataset will be inherited by the model trained with it. Data can be compromised in multiple ways. One way is to corrupt an otherwise valid dataset, as illustrated in the figure below. By switching valid data with poisoned data, the machine learning model underpinning the AI system itself becomes poisoned during the learning process. As a toy example of this type of poisoning attack, consider training a facial recognition-based security system that should admit Alice but reject Bob. If an attacker poisons the dataset by changing some of the images of "Alice" to ones of "Bob," the system would fail in its mission because it would learn to identify Bob as Alice. Therefore Bob would be incorrectly authenticated as Alice when the system was deployed.

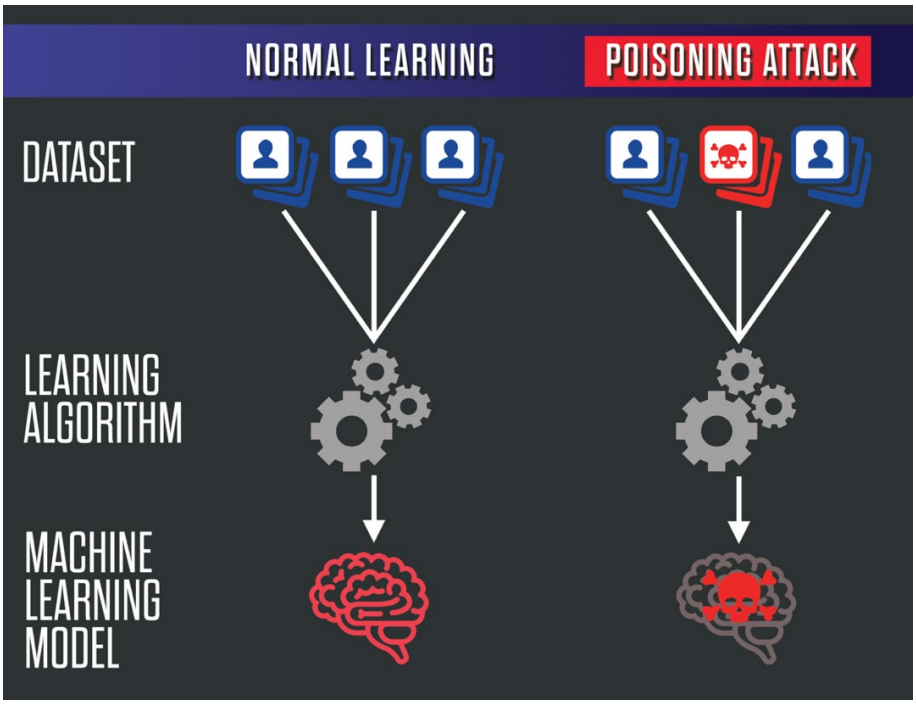


Figure 7: In normal machine learning (left), the learning algorithm extracts patterns from a dataset, and the "learned" knowledge is stored in the machine learning model—the brain of the system. In a poisoning attack (right), the attacker changes the training data to poison the learned model.

A second way to compromise data in order to execute a poisoning attack is to attack the *dataset collection process*, the process in which data is acquired. This effectively poisons the data from the start, rather than changing an otherwise valid dataset as shown in the example above.

The ability to attack the dataset collection process represents the beginning of a new era of attitudes towards data. Today, data is generally viewed as a truthful representation of the world, and has been successfully used to teach AI systems to perform tasks within this world. As a result, data collection practices today resemble a dragnet: everything that can be collected is collected. The reason for this is clear: AI is powered almost entirely by data, and having more data is generally correlated with better AI system performance.

However, now that the dataset collection process *itself* may be attacked, AI users can no longer blindly trust that the data they collect is valid. Data

represents the state of something in the world, and this state can be altered by an adversary. This represents a new challenge: even if data is collected with uncompromised equipment and stored securely, what is represented in the data itself may have been manipulated by an adversary in order to poison downstream AI systems. This is the classic misinformation campaign updated for the AI age.

In the face of AI attacks, today's dragnet data collection practices may soon be a quaint relic of a simpler time. If an AI user's data collection practices are known by an adversary, the adversary can influence the collection process in order to attack the resulting AI system through a poisoning attack. As a result, the age of AI attacks requires new attitudes towards data that are in stark contrast to current data collection practices.

" Today's dragnet data collection practices may soon be a quaint relic of a simpler time."

Crafting a Poisoning Attack

To implement a poisoning attack, the attacker targets one of the assets used in the learning process: either the dataset used to learn the model, the algorithm used to learn the model, or the model itself. Regardless of the method, the end result is a model that has a hidden weakness or backdoor that can later be attacked by exploiting this known weakness.

Dataset Poisoning

The most direct way to poison a model is via the dataset. As previously discussed, the model is wholly dependent on the dataset for all of its knowledge: poison the dataset, poison the model. An attacker can do this by introducing incorrect or mislabeled data into the dataset. Because the machine learning algorithms learn a model by recognizing patterns in this dataset, poisoned data will disrupt this learning process, leading to a poisoned model that may, for example, have learned to associate patterns with mislabeled outcomes that serve the attacker's purpose. Alternatively, the adversary can change its behavior so that the data collected in the first place will be wrong.

Discovering poisoned data in order to stop poisoning attacks can be very difficult due to the scale of the datasets. Datasets routinely contain millions of samples. These samples many times come from public sources rather than private collection efforts. Even in the case when the dataset is collected privately and verified, an attacker may hack into the system where the data is being stored and introduce poisoned samples, or seek to corrupt otherwise valid samples.

Algorithm Poisoning

Another avenue to execute a poisoning attack takes advantage of weaknesses in the algorithms used to learn the model. This threat is particularly pronounced in Federated Learning, a new cutting-edge machine learning algorithm that is emerging.[25] Federated Learning is a method to train machine learning models while protecting the privacy of an individual's data. Rather than centrally collecting potentially sensitive data from a set of users and then combining their data into one dataset, federated learning instead trains a set of small models directly on each user's device, and then combines these small models together to form the final model. Because the users' data never leaves their devices, their privacy is protected and their fears that companies may misuse their data once collected are allayed. Federated learning is being looked to as a potentially groundbreaking solution to complex public policy problems surrounding user privacy and data, as it allows companies to still analyze and utilize user data without ever needing to collect that data.

" This threat is particularly pronounced in Federated Learning, a new cutting-edge machine learning algorithm that is emerging."

However, there is a weakness in the federated learning algorithm that leaves it vulnerable to model poisoning attacks. As attackers have control over their own data on their device, they can manipulate both the data and algorithm running on their device in order to poison the model. Attacks

25 McMahan, H. Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." arXiv preprint arXiv:1602.05629 (2016).

that install a particular backdoor into models,[26] as well as those that generally degrade the model,[27] have already been demonstrated.

Model Poisoning

A final avenue to poison a model is to simply replace a legitimate model with a poisoned one. This is simple to do with a traditional cyberattack. Once trained, a model is just a file living within a computer, no different than an image or PDF document. Attackers can hack the systems holding these models, and then either alter the model file or replace it entirely with a poisoned model file. In this respect, even if a model has been correctly trained with a dataset that has been thoroughly verified and found not poisoned, this model can still be replaced with a poisoned model at various points in the distribution pipeline.

26 Bagdasaryan, Eugene, et al. "How to backdoor federated learning." arXiv preprint arXiv:1807.00459 (2018).

27 Bhagoji, Arjun Nitin, et al. "Analyzing Federated Learning through an Adversarial Lens." arXiv preprint arXiv:1811.12470 (2018).

Part II: Impacted Systems

We now turn our attention to which systems and segments of society are most likely to be impacted by AI attacks. AI systems are already integrated into many facets of society, and increasingly so every day. For industry and policy makers, the five most pressing vulnerable areas are content filters, military systems, law enforcement systems, traditionally human-based tasks being replaced with AI, and civil society.

Content Filters

Content filters are society's digital immune systems. By removing foreign assets that are dangerous, illegal, or against the terms-of-service of a particular application, they keep platforms healthy and root out infections.

Content filters are also uniquely qualified to police content at the scale the Internet requires. The content uploaded to the Internet each minute is a staggering amount growing at a staggering rate. Over three billion images are shared every day on the Internet.[28] AI-based content filters have emerged as the primary, if not only, tool able to operate at this scale, and have been widely adopted by industry. For example, Facebook removed 21 million pieces of lewd content in the first quarter of 2018 alone, 96% of which was flagged by these algorithms.[29]

Content filtering has taken on increased urgency in the past years. ISIS successfully used social media as one of its main recruitment avenues.[30] Nationalists in Myanmar used Facebook as its mouthpiece to incite a campaign of Rohingyan genocide.[31] Misinformation campaigns deploying fake content on social networks have been used to influence

28 List, Mary, "33 Mind-Boggling Instagram Stats & Facts for 2018", 19 February 2018, https://www.wordstream.com/blog/ws/2017/04/20/instagram-statistics

29 Meeker, Mary, "Internet Trends 2018", 30 May 2018, https://www.slideshare.net/kleinerperkins/internet-trends-report-2018-99574140

30 Alfifi, Majid, et al. "Measuring the Impact of ISIS Social Media Strategy." (2018): 1-4.

31 Mozur, Paul, "A Genocide Incited on Facebook, With Posts from Myanmar's Military", NY Times, 15 October 2018, https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.

democratic elections in the U.S. and Europe.[32] As this content successfully weaponizes US-based platforms, the efficacy of AI-based content filters has broad-ranging implications, including the defense of both national security and oppressed populations.

Beyond these newest matters, content filters must continue to be effective with important tasks already within their purview, such as the detection of child pornography. Perhaps the only concept that could make strange bedfellows of the Americans, Russians, Chinese, and Iranians, child pornography was universally accepted as a target of censorship even from the early days of the Internet. AI-based content filters allow website and platform operators to efficiently and effectively scan the millions of images uploaded each minute for illicit content, and immediately destroy offending images. In addition to custom tools companies built for their own use, this detection software was eventually provided via the Software as a Service (SaaS) distribution model, where one large company like Microsoft offered an API-based content filter that website owners could use instead of building their own.[33]

Even in more banal uses, content filters are tied to many business models. As advertisers begin to be held responsible in the court of public opinion for the content appearing next to their advertisements, there is a growing need to detect an increasing number of objectionable content types. This extends to detection of nudity, violence, hate crimes, weapons, adult pornography, profanity, and inappropriate comments. YouTube faced the boycott of advertisers including AT&T, Disney, Hasbro, and Nestle for failing to effectively filter sexual comments left by viewers on videos in which children appeared.[34]

As content filters are drafted into these battles, there will be strong incentives both to attack them and to generate tools making these attacks easier to execute. Adversaries have already seen the power of using digital

32 Satariano, Adam, "Facebook Identifies Russia-Linked Misinformation Campaign", NY Times, 17 January 2019, https://www.nytimes.com/2019/01/17/business/facebook-misinformation-russia.html.

33 See, e.g., Microsoft's PhotoDNA, https://www.microsoft.com/en-us/photodna

34 Fischer, Sara, "Companies pull ads from YouTube…again", Axios, 22 February 2019, https://www.axios.com/companies-pull-ads-from-youtube-again-1550791548-c0433403-d119-43e0-8143-602c50dd1af4.html

platforms in pursuit of their mission. ISIS organically grew an international following and successfully executed a large-scale recruitment program using social media. These are successes that, morals aside, may have evoked jealousy from the marketing departments of Fortune 500 companies. Future organizations of malice are likely to follow the same playbook. If confronted with better content filters, they are likely to be the first adopters of AI attacks against these filters.

In an environment with AI attacks, content filters cannot be trusted to perform their job. Because content filters are now being used as the first and, in many respects, only line of defense against terrorism, extremism, and political attack on the Internet, important parts of society would be left defenseless in the face of successful AI attacks. These attacks give adversaries free reign to employ these platforms with abandon, and leave these societal platforms unprotected when protection is needed more than ever.

Further, it will be difficult to stop or even detect these attacks on content filters because they will likely go wholly unnoticed. Because content filtering is applied to digital assets, it is particularly well suited to the "imperceivable" input attacks. Further, unlike many other cyberattacks in which a large-scale theft of information or system shutdown makes detection evident, attacks on content filters will not set off any alarms. The content will simply fall through the filter unnoticed.

" Entities such as social networks may not even know they are under attack until it is too late."

In this respect, entities such as social networks may not even know they are under attack until it is too late, a situation echoing the 2016 U.S. presidential election misinformation campaigns. As a result, as is discussed in the policy response section, content-centric site operators must take proactive steps to protect against, audit for, and respond to these attacks.

Military

A second major AI attack surface is the military. Military applications of AI are expected to be a critical component of the next major war. The U.S. Department of Defense has recently made the integration of artificial intelligence and machine learning into the military a high priority with its creation of the Joint Artificial Intelligence Center (JAIC). The JAIC has "the overarching goal of accelerating the delivery of AI-enabled capabilities, scaling the Department-wide impact of AI, and synchronizing DoD AI activities to expand Joint Force advantages."[35] The Pentagon's Project Maven applies AI to the analysis of full motion video (FMV), highlighting the military's desire to use AI to identify ground-based assets.[36] The Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory's (AFRL) release of the Moving and Stationary Target Acquisition and Recognition dataset is aimed at building AI techniques to classify and recognize targets of interest.[37] Attacking these military AI systems is the logical successor of General Patton's FUSAG.

The contested environments in which the military operates creates a number of unique ways for adversaries to craft attacks against these military systems, and correspondingly, a number of unique challenges in defending against them.

First, adversaries may capture the physical equipment, including drones and weapon systems, on which AI systems will live. The loss and capture of this equipment will be routine in future conflicts, and the threat this poses to AI systems will grow as more and more AI-enabled systems are deployed in the field or on equipment that can be captured by an adversary. This trend will further increase with the proliferation of "edge computing" in military contexts. In edge computing, rather than sending data to a centralized cloud infrastructure for processing, the data and AI algorithms are stored and run directly on the devices deployed in the field. The DoD has

35 "Establishment of the Joint Artificial Intelligence Center", Deputy Secretary of Defense, 27 June 2018, https://admin.govexec.com/media/establishment_of_the_joint_artificial_intelligence_center_osd008412-18_r....pdf

36 Pellerin, Cheryl, "Project Maven Industry Day Pursues Artificial Intelligence for DoD Challenges", U.S. Department of Defense, 27 October 2017, https://dod.defense.gov/News/Article/Article/1356172/project-maven-industry-day-pursues-artificial-intelligence-for-dod-challenges/

37 MSTAR Public Targets, https://www.sdms.afrl.af.mil/index.php?collection=mstar&page=targets.

made the development of "edge computing" a priority, as the bandwidth needed to support a cloud-based AI paradigm is unlikely to be available in battlefield environments.[38] This reality will require these systems to be treated with care. Just as the military recognizes the threat created when a plane, drone, or weapon system is captured by an enemy, these AI systems must be recognized and treated as a member of this same protected class so that the systems are not compromised if captured by an enemy.

Second, the military's unique domain necessitates the creation of similarly unique datasets and tools, both of which are likely to be shared within the military at-large. Because these datasets and systems will be expensive and difficult to create, there will be significant pressures to share them widely among different applications and branches. However, when multiple AI systems depend on this small set of shared assets, a single compromise of a dataset or system would expose all dependent systems to attack.

Despite this risk, shared datasets are expected to become widespread within military AI operations. The DoD has already stated that the foundation for its AI efforts "includes shared data, reusable tools, frameworks, libraries, and standards…"[39] The initial DoD AI applications, which focus on extracting information from aerial images and video, illustrate why sharing datasets is attractive. These datasets are critical to developing a set of powerful AI systems, but are expensive—both in terms of time and money—to collect and prepare.[40] As a result, there is a logical desire to share and reuse these datasets across many different applications rather than creating a separate dataset for each application.

However, this creates a single point of vulnerability for system-wide attacks. If this data is hacked or compromised, every application developed using this data would be potentially compromised. If a large number of

38 "Interview with Lieutenant General Jack Shanahan: Part 2", Over the Horizon Multi-Domain Operations and Strategy, 4 April 2018, https://othjournal.com/2018/04/04/interview-with-lieutenant-general-jack-shanahan-part-2/

39 Statement by Dana Deasy, Department of Defense Chief Information Office, Before the House Armed Services Committee Subcommittee on Emerging Threats and Capabilities on "Department of Defense's Artificial Intelligence Structure, Investments, and Applications", 26 February 2019, https://armedservices.house.gov/_cache/files/5/7/579723e2-4461-4a8c-95da-ec3e84c4985e/E41B38FCB69AD83331F31CDC06570D33.hhrg-116-as26-wstate-deasyd-20190226.pdf.

40 "Interview with Lieutenant General Jack Shanahan: Part 1", Over the Horizon Multi-Domain Operations and Strategy, 2 April 2018, https://othjournal.com/2018/04/02/interview-with-lieutenant-general-jack-shanahan-part-1/

applications depended on this same shared dataset, this could lead to widespread vulnerabilities throughout the military. In the case of input attacks, an adversary would then be easily able to find attack patterns to engineer an attack on any systems trained using the dataset. In the case of poisoning attacks, an adversary would only need to compromise one dataset in order to poison any downstream models that are later trained using this poisoned dataset.

> " The military faces the challenge that AI attacks will be difficult, if not impossible, to detect in battle conditions."

Further, the process associated with creating these unique datasets can lead to vulnerabilities that can be exploited. When building AI-enabled weapons and defense systems, the individual data samples used to train the models themselves become a secret that must be protected. However, because this preparation work is exceedingly time consuming, it may rely on a large number of non-expert labelers or even outsourced data labeling and preparation services. This trend has already manifested itself in the private sector, where firms like Facebook have turned to outsourced content moderators,[41] as well as in initial military AI efforts.[42] Expected similar trends here could make high confidence guarantees on data-access restrictions and oversight of proper data handling, labeling, and preparation difficult to achieve. While these types of procedural oversight concerns are not new, best practices have been established in other fields such as nuclear. However, because of its infancy, these best practices are lacking in the AI field. Forming these best practices will require new policies managing data acquisition and preparation.

Beyond the threats posed by sharing datasets, the military may also seek to re-use and share models and the tools used to create them. Because the military is a, if not *the*, prime target for cyber theft, the models and tools themselves will also become targets for adversaries to steal through hacking or counterintelligence operations. History has shown that computer systems are an eternally vulnerable channel that can be reliably counted on as an attack avenue by adversaries. By obtaining the models stored and run

41 Lagorio-Chafkin, Christine, "Facebook's 7,500 Moderators Protect You From the Internet's Most Horrifying Content. But Who's Protecting Them?", Inc., 26 September 2018, https://www.inc.com/christine-lagorio/facebook-content-moderator-lawsuit.html.

42 Fang, Lee, "Google Hired Gig Economy Workers to Improve Artificial Intelligence in Controversial Drone-targeting Project", The Intercept, 4 February 2019, https://theintercept.com/2019/02/04/google-ai-project-maven-figure-eight/.

on these systems, adversaries can back-solve for the attack patterns that could fool the systems.

Finally, the military faces the challenge that AI attacks will be difficult, if not impossible, to detect in battle conditions. This is because a hack of these systems to obtain information to formulate an attack would not by itself necessarily trigger a notification, especially in the case where an attacker is only interested in reconnaissance aimed at learning the datasets or types of tools being used. Further, once adversaries develop an attack, they may exercise extreme caution in their application of it in order to not arouse suspicion and to avoid letting their opponent know that its systems have been compromised. Accordingly, attacks may be limited only to situations of extreme importance. In this respect, there may be no counter-indications to system performance until after the most serious breach occurs. This is also a problem inherent in traditional cyberattacks.

Detecting AI attacks in the face of their rare application would focus on two methods: detecting intrusions into systems holding assets used to train models, and analysis of model performance. Traditional intrusion detection methods could be used to detect if a dataset or resource has been compromised. If an asset has been compromised, the AI systems using those assets may have to be shut down or re-trained. Alternatively, AI attack detection could be based on complex performance analysis of the system whenever an AI attack is suspected, such as events surrounding a surprising decrease in AI system performance.

Beyond these defensive concerns, the military may also choose to invest in offensive AI attack capabilities. This topic of offensive weaponization is discussed in detail in Part III.

Law Enforcement

A third major attack surface is the application of AI to law enforcement. The National Institute of Justice argues that "Artificial intelligence has the potential to be a permanent part of our criminal justice [system]" through its use to "replicate...human [pattern recognition] capability in software algorithms and computer hardware."[43]

The applications of AI for law enforcement are both already deployed and being actively researched. Amazon has recently launched a facial recognition system[44] that is being piloted by police departments in the US.[45] The system seeks to match target facial images against a large database of criminal mugshots. The NIJ supports research in video and image analysis, detecting characteristics of firearm discharges (number of guns present, assignment of a gunshot to a particular gun, and classification of firearm class and caliber), face detection, and other applications.[46]

It is understandable that law enforcement is turning to AI technology. Technology has created entirely new streams of data and platforms that law enforcement is being called on to police,[47] posing the challenge of analyzing a virtually infinite amount of content with a very finite amount of human resources. Much like the case with content filtering, the law enforcement community views the new generation of AI-enabled tools as necessary to keep pace with their expanding technological purview. The NIJ recognizes this potential for AI, stating, "Examining the huge volume of possibly relevant images and videos in an accurate and timely manner is a time-consuming, painstaking task, with the potential for human error due to fatigue and other factors. Unlike humans, machine do not tire."[48]

43 "Using Artificial Intelligence to Address Criminal Justice Needs", Christopher Rigany, NIJ, 2, https://www.ncjrs.gov/pdffiles1/nij/252038.pdf

44 Amazon Rekognition, https://aws.amazon.com/rekognition/.

45 Wingfield, Nick, "Amazon Pushes Facial Recognition to Police. Critics See Surveillance Risk", New York Times, 22 May 2018, https://www.nytimes.com/2018/05/22/technology/amazon-facial-recognition.html.

46 "Using Artificial Intelligence to Address Criminal Justice Needs", Christopher Rigany, NIJ. 2, 7, https://www.ncjrs.gov/pdffiles1/nij/252038.pdf.

47 Pegues, Jeff, "Florida school shooting: FBI got call about suspect a year before shooting", CBS News, 15 February 2018, https://www.cbsnews.com/news/fbi-youtube-video-investigation-florida-shooting-suspect-nikolas-cruz-details-today/

48 "Using Artificial Intelligence to Address Criminal Justice Needs", Christopher Rigany, NIJ. 2, https://www.ncjrs.gov/pdffiles1/nij/252038.pdf.

Beyond just its use in keeping pace with expanding amounts of content, AI can be used to provide more effective policing and crime prevention by detecting criminal warning signs earlier and apprehending suspects faster.

As these AI-based law enforcement systems become more widespread, they will naturally become attack targets for criminals. One could imagine AI attacks on facial recognition systems as the 21st century version of the time-honored strategy of cutting or dyeing one's hair to avoid law enforcement recognition. Researchers have already shown that sporting a multi-colored pair of glasses has the ability to attack AI-based facial recognition systems, greatly degrading their accuracy.[49] As these facial recognition systems move not just into police departments but into other law enforcement areas such as facial-recognition based airport screening,[50] the number of attack targets continues to grow.

" As these AI-based law enforcement systems become more widespread, they will naturally become attack targets for criminals."

Further, these attacks are not limited to visual surveillance systems. The NIJ's funded research into classifying firearm class and caliber from audio signals also presents a target. New classes of hardware accessories such as "smart silencers" may be developed that execute AI attacks to deceive these systems, for example by making the systems think that the gunshot came from a different gun. As the AI technology evolves, criminal strategy will do so in turn.

Although law enforcement and the military share many similar AI applications, the law enforcement community faces its own unique set of challenges in securing against AI attacks. First, law enforcement AI systems will largely be off-the-shelf purchases from different private companies. Unlike the military, most law enforcement organizations are small and lack the resources needed to scope, let alone build, these AI systems, and will therefore likely rely on a patchwork of different private providers. This is reason to worry. Private companies have already shown an ineptitude to properly address known and easily addressed security

" Private companies have already shown an ineptitude to properly address known and easily addressed security vulnerabilities."

49 Sharif, Mahmood, et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016.

50 See, e.g., https://www.clearme.com

vulnerabilities, let alone an emerging and difficult vulnerability such as AI attacks. It would be unwise to assume that the private companies are taking, or are even capable of taking, the necessary steps to mitigate AI security vulnerabilities. Further, each law enforcement organization alone will probably not have enough market power to demand stringent security protections, while the military does.

Second, law enforcement organizations are at a significantly lower level of cybersecurity preparedness compared to the military. The military by definition plans for operating in contested environments with sophisticated adversaries. As a result, the military possesses classified networks, established cybersecurity protocols, and in-house expertise to identify and address any breaches or attacks. Many local law enforcement organizations have none of these protections. Law enforcement data systems from which training data may be obtained are not maintained with the same level of security as their military counterparts. While there is still risk inherent in the military's secure cloud architectures and networks, this risk is significantly larger for the unsecured ad hoc systems employed by law enforcement organizations.[51] This sets the bar for executing AI attacks, especially those that rely on obtaining or corrupting data, significantly lower in this domain.

Together, these challenges are an especially worrisome point given the current climate in which police departments are on the front lines of fighting terrorism. A technological system that is fragmented and not properly handled may disadvantage police forces in the face of advanced adversaries. This situation may call for additional coordination from sources such as DHS to unify purchasing and security standards.

51 "Cybersecurity Guide for State and Local Law Enforcement", National Consortium for Advanced Policing, June 2016, https://cchs.gwu.edu/sites/g/files/zaxdzs2371/f/downloads/NCAPCybersecurityGuide-2016.pdf

Commercial Artificial Intelligence-fication of Human Tasks

A fourth major attack surface is the rapid artificial intelligence-fication of traditionally human-based tasks. Although some of these applications are within apps and services where attacks would not have serious societal consequences, attacks on other applications could prove very dangerous. Self-driving vehicles and trucks rely heavily on AI to drive safely, and attacks could expose millions to danger on a daily basis. Some commercial applications also have ramifications for law enforcement. Automated identity screening and customs kiosks at airports that are built and operated by private companies also rely on AI, and attacks could jeopardize the safety of the skies and national borders.

The cost of failure of AI systems in this domain have already been experienced. An Uber self-driving car struck and killed a pedestrian in Tempe, Arizona when the on-board AI system failed to detect a human in the road.[52] While it is unclear if the particular pattern of this pedestrian is what caused the failure, the failure manifested itself in the exact same manner in which an AI attack on the system would. This real-world example is a terrifying harbinger of the ability for adversaries who are deliberately trying to find attack patterns to find success.

Commercial firms have proven themselves woefully incapable or unwilling to address cybersecurity concerns. There are also few regulations or support structures that encourage or aid in the development of cybersecurity protocols, as has been demonstrated by a lack of regulation of the Internet of Things and other computer systems over the past decade. Without the appropriate regulations and penalties for disregarding security, companies have shown themselves incapable of providing attention to the necessary security issues associated with their products.

In order to properly regulate commercial firms in this domain, policymakers must understand how this commercial development of AI systems will

52 Said, Carolyn, "Video shows Uber robot car in fatal accident did not try to avoid woman", SFGate, 21 March 2018, https://www.sfgate.com/business/article/Uber-video-shows-robot-car-in-fatal-accident-did-12771938.php

progress. In one scenario, individual companies will each build their own proprietary AI systems. Because each company is building its own system, industries cannot pool resources to invest in preventative measures and shared expertise. However, this diversification limits the applicability of an attack on one AI system to be applied broadly to many other systems. Further, by not pooling dataset resources, a dataset breach will have limited consequences.

> "Regulators may need to pay more attention to monocultures of AI models that may permeate certain industries."

However, in a second scenario, individual companies may utilize shared AI systems provided by a third party. This is already happening for many common AI tasks, including illicit content filters and computer vision tasks. Because a single organization specializes in building the AI system, it may be able to better invest resources to protect its system from attacks. However, the creation of "monocultures" in this setting amplify the damage of an attack, as a successful attack would compromise not just one application but every application utilizing the shared model. Just as regulators fear monocultures in supply chains, illustrated recently by Western fears that Huawei may become the only telecommunication network equipment vendor, regulators may need to pay more attention to monocultures of AI models that may permeate certain industries.

Different industries will likely play into one of these scenarios, if not a hybrid of both. This dichotomy is already seen in the market today. Autonomous vehicle companies are largely operating under the first "every firm on its own" scenario. At the same time, Artificial Intelligence as a Service, a key component of the second "shared monoculture" scenario, is also becoming more common. As such, policymakers must be ready to address both scenarios, as each will require different interventions.

Civil Society

Just as not all uses of AI are "good," not all AI attacks are "bad." While AI in a Western context is largely viewed as a positive force in society, in many other contexts it is employed to more nefarious ends. Countries like China

and other oppressive regimes use AI as a way to track, control, and intimidate their citizens. As a result, "attacks" on these systems, from a US-based policy view of promoting human rights and free expression, would not be an "attack" in a negative sense of the word. Instead, these AI "attacks" would become a source of protection capable of promoting safety and freedom in the face of oppressive AI systems instituted by the state.

This underscores an important point that should not be disregarded in policy discussions: AI attacks are a "dual use" tool. Depending on the context, the same attack can be used as a sword against free society or a shield against oppression.

China's detention and "re-education" of Uighur Muslims in the Xinjiang region serves as a case study for how AI "attacks" could be used to protect against regime-sponsored human rights abuses. China uses facial recognition systems to track and monitor the movements and actions of the Uighur Muslims within the region.[53] "Attacks" on these systems in the form of glasses shown to be universally successful at degrading the state-of-the-art facial recognition systems[54] would go a far to help protect oppressed minorities who otherwise would be helpless against AI systems. U.S. policy may therefore warrant treating the same exact attack/"attack" differently depending on context. A kidnapper wearing these glasses at a gas station to evade detection by a police force applying AI to find the suspect from thousands of video streams poses a threat to societal safety. A Uighur Muslim wearing these glasses to evade detection by Chinese government officials represents the protection of religious freedom.

This "dual use" nature is not unique to AI attacks, but is shared with many other cyber "attacks." For example, the identical encryption method can be used by dissidents living under an oppressive regime to protect their communications as easily as it can be by terrorists planning an attack.

> " AI 'attacks' may take on a role similar to that of Tor, VPNs, and other technologies used to evade government oppression."

53 Huges, Roland, "China Uighurs: All you need to know on Muslim 'crackdown'", BBC News, 8 November 2018, https://www.bbc.com/news/world-asia-china-45474279

54 Sharif, Mahmood, et al. "Adversarial generative nets: Neural network attacks on state-of-the-art face recognition." arXiv preprint arXiv:1801.00349 (2017).

In this respect, AI "attacks" may take on a role similar to that of Tor, VPNs, and other technologies used to evade government oppression. Just as this report advocates for appropriate agencies to educate their constituents about the risks posed by AI attacks, it should likewise advocate for human rights organizations to educate their constituents about the *benefits* available through AI "attacks."

This dual use will create difficult policy decisions as potential protections against AI attacks are developed. Specifically, if protections against AI attacks are developed, should they be made public? If sharing this protection with U.S. institutions and companies would stop dangerous attacks on them, the answer would be "yes." But if oppressed people around the world came to rely on AI "attacks" to protect themselves from their government, and sharing this protection would again give their oppressive regimes the upper hand, many may argue that the answer would be "no." (Beyond the impact on civil society, the answer may also be "no" if it was known that the disclosure would improve an adversary's defenses against AI attack.)

In this respect, AI attacks are in the unique position of inheriting the *reverse* of cybersecurity's perennial discussion regarding disclosure of vulnerabilities. Traditional cybersecurity grapples with the question whether entities (such as the NSA) that discover vulnerabilities should 1) disclose them to promote public safety and patching, or 2) keep them secret and therefore maintain their usefulness for their own mission. This debate is based on the fact that vulnerability is assumed to be (largely) unknown, but the remedy is generally easily crafted and applied. However, with AI attacks, the opposite is true: the vulnerability is known but the remedy is unknown. This potential situation poses significant ethical and policy questions. It is important for a country to realize that the disclosure of any protective techniques will have impacts beyond its own borders.

Part III: Significance within the Cybersecurity Landscape

Comparison with Traditional Cybersecurity Issues

AI attacks are fundamentally different in nature than the cybersecurity attacks that have received heightened recent attention. Unlike traditional cybersecurity vulnerabilities, the problems that create AI attacks cannot be "fixed" or "patched." Traditional cybersecurity vulnerabilities are generally a result of programmer or user error. As a result, these errors can be identified and rectified. In contrast, the AI attack problem is more intrinsic: the algorithms themselves and their reliance on data are the problem.

> " For AI attacks, a robust IT department and 90-letter passwords won't save the day."

This difference has significant ramifications for policy and prevention. Mitigating traditional cybersecurity vulnerabilities deals with fixing "bugs" or educating users in order to stop adversaries from gaining control or manipulating an otherwise sound system. Reflecting this, solutions to cybersecurity problems have focused on user education, IT department-led policy enforcement, and technical modifications such as code reviews and bug bounties aimed at finding and correcting flaws in the code. However, for AI attacks, a robust IT department and 90-letter passwords won't save the day. The algorithms themselves have the inherent limitations that allow for attack. Even if an AI model is trained to exacting standards using data and algorithms that have never been compromised, it can still be attacked. This bears repeating: among the state-of-the-art methods, there is currently no concept of an "unattackable" AI system. As such, protecting against these intrinsic algorithmic vulnerabilities will require a different set of tools and strategies. This includes both taking steps to make executing these attacks more difficult, as well as limiting the dependence and reach of applications built on top of AI systems.

Despite this fundamental difference, the two are linked in important ways. Many AI attacks are aided by gaining access to assets such as datasets or model details. In many scenarios, doing so will utilize traditional cyberattacks that compromise the confidentiality and integrity of systems, a subject well studied within the cybersecurity CIA triad. Traditional confidentiality attacks will enable adversaries to obtain the assets needed to engineer input attacks. Traditional integrity attacks will enable adversaries to make the changes to a dataset or model needed to execute a poisoning attack. As a result, traditional cybersecurity policies and defense can be applied to protect against some AI attacks. While AI attacks can certainly be crafted without accompanying cyberattacks, strong traditional cyber defenses will increase the difficulty of crafting certain attacks.

Another important lesson from traditional cybersecurity policy is the superiority of foresight and pre-deployment planning over reactionary remedies. The past decade has borne poisonous fruit from technological seeds planted before the turn of the century. From a commercial perspective, the breakneck pace to digitize and interconnect infrastructure without the prescience to keep a similar pace with cybersecurity defense has seen billions of dollars of losses from cyberattacks.[55] From a societal perspective, the unwavering march to connect the world via social networks and reluctance of government to investigate their power has led to their successful use as a terrorist recruiting mechanism, a mouthpiece for and inciter of genocide, and the disruption of democratic electoral processes. It is not certain that these problems could have been fully prevented through better planning and regulation. However, it is certain that it would have been easier to prevent them than it is to solve them now.

" Another important lesson from traditional cybersecurity policy is the superiority of foresight and pre-deployment planning over reactionary remedies."

Given the current attention cybersecurity problems are receiving from the public and the government, the climate is right for taking proactive measures to allow for the beneficial use of AI while mitigating the associated attack threat *before* the expanded spread of these algorithms to safety- and security-critical infrastructure and applications.

55 Greenberg, Andy, "The Untold Story of NotPetya, the Most Devastating Cyberattack in History", Wired, 22 August 2018, https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/.

Offensive Weaponization

" The focus of China's and other countries' investments in AI is based on an attempt to offset traditional U.S. battlefield superiority."

Any cyber vulnerability can be turned into a cyber weapon. The same holds true for AI attacks, especially in the military and intelligence contexts. The potential promise of this is based on the belief that other countries may begin to integrate AI and machine learning into military decision making pipelines and automated weapons.[56] China and other potential adversaries are investing heavily in AI and machine learning. Many believe that these abilities will be integrated into their armed forces.[57] Lieutenant General John Shanahan, director of the Joint Artificial Intelligence Center, believes that machine learning/artificial intelligence capabilities of potential foes will be so far developed in future wars that U.S. use of the same technologies "... is not a case where we're going to offset somebody. We will however, be offset if we do not do it [develop these capabilities]."[58]

In this regard, the United States has the opportunity to weaponize AI attacks against its adversaries' AI systems. Doing so would realize two large benefits. First, it would turn a developing strength of the United States' main geopolitical foes to a weakness. The focus of China's and other countries' investments in AI is based on an attempt to offset traditional U.S. battlefield superiority. As an example, China believes that the current U.S. strategy in a potential conflict may take the form of an overwhelming rapid show of force to degrade China's capability to wage war.[59] Because traditional military assets may not be sufficient to win a conflict in the face of overwhelming and rapid strikes on coastal areas and raids on their interior, China may look to autonomous weapon systems to engage U.S. attacks at a speed at which humans could not operate. U.S. strategy will need to evolve

56 Harvard Kennedy School Institute of Politics John F. Kennedy Jr. Forum "Interview with Eric Rosenbach and Jason Mathen: The Public Policy Challenges of Artificial Intelligence", 15 February 2018, https://www.belfercenter.org/event/public-policy-challenges-artificial-intelligence#transcript

57 Upchurch, Tom, "How China Could Beat the West in the Deadly Race for AI Weapons", Wired, 8 August 2018, https://www.wired.co.uk/article/artificial-intelligence-weapons-warfare-project-maven-google-china.

58 "Interview with Lieutenant General Jack Shanahan: Part 1", Over the Horizon Multi-Domain Operations and Strategy, 2 April 2018, https://othjournal.com/2018/04/02/interview-with-lieutenant-general-jack-shanahan-part-1/

59 Talmadge, Caitlin, "Beijing's Nuclear Option: Why a U.S.-Chinese War Could Spiral Out of Control", Foreign Affairs Vol. 97 Num. 6, November/December 2018.

to counter this new AI-based strategy. One key component of this strategy should include offensive AI attacks to degrade the performance of enemy automated systems. In this respect, AI attacks would be a modern-day version of radar jamming.

Second, developing offensive AI attack capabilities would build important institutional knowledge within the U.S. military that could then be used to harden its own systems against attack. All successful work in developing offensive capabilities would double as an important case study in ineffective preventative techniques, and could be used to stress test or "red team" U.S. AI systems. This experience will be essential in preparing for the next potential conflict given that the U.S. is unlikely to gain battlefield experience with AI attacks, both on the receiving and transmitting end, until it is already in a military conflict with an advanced adversary. In order to be prepared at this first encounter, it is important that the U.S., after crafting successful attacks against adversaries, turn these same techniques against itself to test its own resiliency to this new form of weapon.

However, offensive weaponization of AI attacks would not be without risk. The creation of offensive attacks against state-of-the-art systems that are deployed would risk the diffusion of these attacks into enemy hands. This risk is well known with other cyber weapons. Notably, the NSA has been criticized for not disclosing the EternalBlue exploit responsible for severe attacks, including WannaCry and NotPetya.[60] Creating offensive AI attack weapons against systems on which the host country or its allies are also dependent may create similar risks of having the weapon turned against friendly assets.

In the context of AI attacks, if the *development* of the AI attack is believed to be so sophisticated that no other entity is expected to be able to craft the attack on its own, diffusion risks exist. In this case, the fear of an attack that could be turned against the host country and find its way into the public sphere may outweigh the benefits the attack may provide, creating an incentive against offensive weaponization. However, these risks only apply

60 Burgess, Matt, "Everything you need to know about EternalBlue—the NSA exploit linked to Petya", Wired, 28 June 2017, https://www.wired.co.uk/article/what-is-eternal-blue-exploit-vulnerability-patch

if the host country or its allies are utilizing a similar system vulnerable to the same attack.

In other respects, however, diffusion in the AI attack context is different in nature from that of other offensive cyber weapons. Unlike the vulnerabilities allowing for many traditional cyberattacks, the vulnerabilities allowing for AI attacks are believed to be un-patchable. As such, there may be less downside to exploiting it. This is due to the fact that because there is, by definition, no way to protect against the vulnerability, an adversary is incentivized to exploit it regardless of the host country's actions. As a result, in the face of this permanent vulnerability, a host country's exploitation of that vulnerability may have no effect on its adversary's ability to do so. If offensive weaponization has no impact on an adversary's behavior, it removes the associated risk.

" Diffusion in the AI attack context is different in nature from that of other offensive cyber weapons in some respects."

This represents a different situation than in traditional cyber weaponization. In traditional cyber weaponization, a tension exists between 1) notifying the system operator to allow for patching, and 2) keeping the vulnerability a secret in order to exploit it. This tension is based on the fact that if one party discovers a vulnerability, it is likely that another, possibly hostile, party will do so as well. Therefore, the push to report the vulnerability is based on the fear that an adversary will either steal or discover the vulnerability as well, and therefore there is a need to patch affected systems before this occurs in order to reduce exposure to the vulnerability. Continuing the EternalBlue example, the NSA is criticized not for *using* EternalBlue, but rather for failing to *report* it in order to maintain its usefulness. In the context of an AI system, because the system is already known to be vulnerable but unable to be patched, this tension disappears.

Together, this allows for the following conclusion: if a vulnerability is un-patchable and already capable of being effectively exploited by an adversary, the traditional fears of diffusion may not apply, leaving the door open to offensive weaponization. However, if a vulnerability is un-patchable but likely *not* capable of being exploited by an adversary alone, the traditional fears of diffusion apply, and the associated risks should be weighed against the benefits of the attack.

Considerations of Practicality

Are AI attacks practical to a degree that they represent a true threat? Given their youth, it is an important question. Pushback to serious consideration of this attack threat will center around the technological prowess of attackers. As this attack method relies on sophisticated AI techniques, many may take false comfort in the fact that the attack method's technical barriers will provide a natural barrier against attack. As a result, some may say that AI attacks do not deserve equal consideration with their traditional cybersecurity attack counterparts.

> " Some may say that AI attacks do not deserve equal consideration with their traditional cybersecurity attack counterparts. This view is incorrect."

This view is incorrect. Recent history of a similar scourge with equal technical sophistication shows why. Deepfake, a method to create fake synthetic videos using complex AI methods, experienced widespread use by non-technical users to create fake celebrity pornographic videos despite its advanced technical sophistication.[61] Popular use occurred to such a degree that a Reddit page was even created where people shared their homemade videos.

Like AI attacks, the technology behind Deepfakes shares a similar if not even more advanced technical sophistication. However, despite the technique living at the intersection of cutting-edge AI, computer vision, and image processing research, large number of amateurs with no technical background were able to use the method to produce the videos.

This was due to two enabling factors, both of which can be applied to gain insight into the practicality of AI attacks. First, even though the underlying technology behind Deepfakes was sophisticated, it was possible to create tools that simplified the application of the method. In the case of Deepfake, an app was created that abstracted away all of the technical details, essentially distilling the application of a complicated algorithm to a

61 CNN Business, "When Seeing is No Longer Believing", January 2019, https://www.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/.

drag-and-drop and a single click of a button.[62] This allowed for non-technical actors to harness the power of the algorithm easily. This is not the first time this rodeo has played out in the cyber domain: a similar set of tools has also proliferated in the traditional cybersecurity domain, allowing non-technical actors to participate in campaigns such as Distributed Denial of Service (DDoS) attacks.[63]

Second, the proliferation of powerful yet cheap computing hardware means almost everyone has the power to run these algorithms on their laptops or gaming computers. While this is expected in military contexts opposite an adversary with modern technical capabilities, it does have significant bearing on the ability for non-state actors and rogue individuals to execute AI attacks. In conjunction with apps that could be made to allow for the automation of AI attack crafting, the availability of cheap computing hardware removes the last barrier from successful and easy execution of these AI attacks.

Both of these enabling factors will be applied to make crafting AI attacks easier and accessible. Tools have already been created to craft AI attacks,[64] and it would be a weekend project to turn them into a single click operation and package them for widespread use. For input attacks, tools will allow an adversary to load a stolen dataset into an app and quickly spit out custom crafted input attacks. Easy access to computing power means this app could run on the attacker's own computer, or could plug into cloud-based platforms.[65] For the integrity and confidentiality attacks that are likely to accompany some model poisoning attacks, a number of existing cyberattacks could be co-opted for this purpose. As a result, an environment of feasibility may easily develop around AI attacks, as it has developed around Deepfakes and other cyberattacks.

Further, the fact that technological ecosystems have not adapted to prevent these attacks will amplify the success of these tools even further. For example, because many AI systems have web-based APIs, apps could easily

62 See, e.g., https://github.com/iperov/DeepFaceLab

63 Singel, Ryan, "Joining Pro-Wikileaks Attacks is as Easy as Clicking a Button", wired, 10 December 2010, https://www.wired.com/2010/12/web20-attack-anonymous/

64 See, e.g., https://github.com/tensorflow/cleverhans

65 See, e.g., AWS, https://aws.amazon.com/

be developed to interface directly with the APIs to generate attacks on demand. To attack an image content filter with a web-based API, attackers would simply supply an image to the app, which would then generate a version of the image able to trick the content filter but remain indistinguishable from the original to the human eye.

As a result of this environment, AI attacks will be within the realm of capabilities for both advanced geopolitical adversaries and individuals, and everyone in between.

Part IV: "AI Security Compliance" as a Policy Solution for AI Attacks

This report proposes the creation of "AI Security Compliance" programs as a main public policy mechanism to protect against AI attacks. The goals of these compliance programs are to 1) reduce the risk of attacks on AI systems, and 2) mitigate the impact of successful attacks.

Compliance programs will accomplish these goals by encouraging stakeholders to adopt a set of best practices in securing their systems and making them more robust against AI attacks. These best practices manage the entire lifecycle of AI systems in the face of AI attacks. In the *planning* stage, they will force stakeholders to consider attack risks and surfaces when planning and deploying AI systems. In the *implementation* stage, they will encourage adoption of IT-reforms that will make attacks more difficult to execute. In the *mitigation* stage for addressing attacks that will inevitably occur, they will require the deployment of previously created attack response plans.

This program is modeled on existing compliance programs in other industries, such as PCI compliance for securing payment transactions.[66] From a practical standpoint, compliance programs would be implemented by appropriate regulatory bodies for their relevant constituents.

This section sets forth a general AI security compliance program that can be the basis of compliance programs adopted by industry and regulators. Industries and sectors adopting this type of compliance program can customize the components to fit their needs. The following section describes implementation and enforcement details.

66 See https://www.pcisecuritystandards.org/

Planning Stage Compliance Requirements

Planning stage compliance requirements focus on ensuring stakeholders have assessed the risks inherent in the process of planning the creation of AI systems. This includes properly evaluating the risks associated with the AI system, and taking steps to secure other preparation activities, such as dataset collection.

AI Suitability Tests

Conduct "AI Suitability Tests" that assess the risks of current and future applications of AI. These tests should result in a decision as to the acceptable level of AI use within a given application. These tests should weigh the application's vulnerability to attack, the consequence of an attack, and the availability of alternative non-AI-based methods that can be used in place of AI systems.

When deciding whether to build an AI system, stakeholders should conduct an "AI suitability test" to review the risks associated with the proposed AI system. The outcomes of these tests should be a study of the risks posed by the AI system, and a determination of how much AI use is appropriate for the given application. This may range from full AI autonomy, through mixed AI/human use with varying degrees of human oversight, to no AI use at all.

These suitability tests should be principled and balance potential harms with the need to foster innovation and the development of new technologies. The focus of assessments should include both current and near-future applications of AI.

AI suitability tests should focus on answering five questions:

- **Value:** What is the value added by the AI system?

- **Ease of Attack:** How easy will it be for an adversary to execute an attack on the AI system?
- **Damage:** What will be the damage incurred from an attack on the AI system?
- **Opportunity Cost:** What are the costs of not implementing the AI system?
- **Alternatives:** Are there alternatives to the AI system?

We now discuss each component briefly. The *value* of the AI system should be examined in light of the economic and societal benefit the system is expected to deliver. This will by nature be a subjective measure, but entities deciding to adopt AI should be able to justify the value they believe it will deliver in the event of an audit or external review.

Determining the *ease of attacking a particular system* will be an integral part of these AI suitability tests. The degree of vulnerability can be determined by characteristics such as public availability of datasets, the ability to easily construct similar datasets, and other technical characteristics that would make an attack easier to execute. One example of an application that could be particularly vulnerable to attack is a military system that automatically classifies an adversary's aircrafts. The dataset for this task would likely consist of collected radar signatures of the adversary's aircraft. Even if the country collected the data itself, stored it perfectly and safely with encryption, and had flawless intrusion detection—all of which would guarantee that the adversary could not get this data and use it to formulate an attack—the adversary could still execute a successful attack by building a similar dataset itself from scratch, which could easily be done because the adversary clearly has access to its own aircraft. This would therefore allow the adversary to craft an attack without ever having to compromise the original dataset or model. As a result, if this application was deemed easy to attack, an AI system may not be well suited to this particular application.

The *damage* that an attack can precipitate should be assessed in terms of the likelihood of an attack and the ramifications of the attack. Entities may wish to conduct "red teaming" exercises and consultations with law

enforcement, academics, and think tanks in order to understand what damage may be incurred from a successful attack against an AI system.

The *opportunity cost* of not implementing an AI system must also be incorporated into the suitability test equation. The risks of attack do not delete the societal benefits AI is expected to deliver. As such, the cost of not implementing the system must also be considered.

Finally, the *existence of non-AI alternatives*, or lack thereof, should be considered. If good alternatives exist that are capable of performing similar function with similar costs, AI should not necessarily be adopted over an alternative in the name of innovation or progress. However, if no reasonable alternatives exist, this may provide an additional impetus for the adoption of AI even in the face of attack.

Once each of these questions have been sufficiently answered, they should be weighed to arrive at a determination of how much risk the system poses, and this should be used to make an implementation decision. Just as they may have chosen to do in answering the questions, stakeholders may again wish to consult with law enforcement, academics, think thanks, and other outside entities in arriving at a decision. Entities may wish to look to the National Highway Traffic Safety Administration's cost analysis methodology for inspiration in reaching an implementation decision.[67]

This implementation decision should state how much AI should be used within an application, ranging from full use, through limited use with human oversight, to no use. This spectrum affirms that vulnerability to attacks does not necessarily mean that a particular application is ill-suited for AI. Instead, suitability should be measured by the informed results of the suitability test, especially the questions regarding the consequences of an attack and the availability of other options.

As an illustrative example of this careful tradeoff, consider the example of extremist content filtering on a social network. We have already determined that this application is valuable yet vulnerable to attack. In terms

67 Soodoo, George, "A Primer on the NHTSA Rulemaking Process", Eno Center for Transportation, 13 March 2017, https://www.enotrans.org/article/primer-nhtsa-rulemaking-process/

of attack damage, an attack will at worst render the content filters ineffective, an outcome no worse than not deploying them in the first place. In terms of availability of other options, AI-based filtering is perhaps the only technique that is capable of operating at a sufficient scale given the large amount of content added to social networks daily. As a result, this application would still be well suited for AI, given a lack of alternatives and low collateral damage from an attack. However, even though AI may still be appropriate in this case, it does not absolve the social network from both preventative and mitigative efforts to counter attacks. For example, the social network may need to determine human involvement in and oversight of the system, such as by executing periodic manual audits of content to identify when its systems have been attacked, and then taking appropriate action such as increased human review of material policed by the compromised system.

This example also demonstrates the outcomes of these AI suitability tests need not be binary. They can, for example, suggest a target level of AI reliance on the spectrum between full autonomy and full human control. This can allow for technological development while not leaving an application vulnerable to a potentially compromised monoculture. The DoD has been vocal about adopting this strategy in its development of AI-enabled systems, albeit for additional reasons. In this middle-lane strategy, AI-enabled systems can be used to augment human-controlled processes, but not to fully replace human operators. Through this middle lane, a successful attack would not have its full intended effect. Stakeholders may look to the self-driving vehicle industry for inspiration in categorizing human involvement in AI systems, which formulizes this classification system by categorizing autonomous vehicles from Level 1 (no AI use) to Level 5 (full AI use).

In terms of implementing these suitability tests, regulators should play a supportive role. They should provide guidelines on best practices for how to perform the tests. In areas requiring more regulatory oversight, regulators should write domain specific tests and evaluation metrics to be used. In areas requiring less regulatory oversight, they should write general guidelines to be followed. Beyond this, regulators should provide advice and counsel where needed, both in helping entities answer the

questions that make up the tests as well as in forming a final implementation decision.

> " Just as Rome's powerful roads were turned against them by their enemies, AI attacks and other forms of information warfare may similarly turn data from the panacea it is hailed as today into a vulnerability in an AI-dominated society."

Beyond this supportive role, regulators should affirm that they will use an entity's effort in executing a suitability test in deciding culpability and responsibility if attacks do occur. As is the case with other compliance efforts, a company that demonstrates that it made a good faith effort to reach an informed decision via a suitability test may face more lenient consequences from regulators in the case of attacks than those that disregarded the tests.

Because AI systems have already been deployed in critical areas, stakeholders and appropriate regulatory agencies should also retroactively apply these suitability tests to already deployed systems. Based on the outcome of the tests, the stakeholders or regulators should determine if any deployed AI systems are too vulnerable to attack for safe operation with their current level of AI use. Systems found to be too vulnerable should be promptly updated, and in certain cases taken offline until such updates are completed.

Review and update data policies

Review and update data collection and sharing practices to protect against data being weaponized against AI systems. This includes formal validation of data collection practices and restricting data sharing.

AI users must review and secure their data collection and sharing policies. These reviews should be formal, identify emerging ways data can be weaponized against systems, and be used to shape data collection and use practices. The outcome of these reviews should be written policies governing how any data used in building an AI system is collected and shared.

These reviews are needed because data may emerge as a potent weapon in the age of AI attacks, and steps must be taken to have stakeholders realize

the dangers data can now pose. This is especially important because this new danger is in stark contrast with data's current reputation in society: data is currently regarded pervasively as "digital gold" within the private sector, government, and military. However, because AI is almost wholly dependent on data, data is a direct avenue through which to conduct AI attacks. In this respect, just as Rome's powerful roads were turned against them by their enemies, AI attacks and other forms of information warfare may similarly turn data from the panacea it is hailed as today into a vulnerability in an AI-dominated society.

AI users will need to fundamentally rethink their data practices in order to protect themselves from having it weaponized against them. Data practices will have to change in two major ways: collection practices must be validated, and data sharing must be restricted. As discussed below, these two changes in practices will challenge current attitudes towards data.

Validate Dataset Collection Practices

AI users must validate their data collection practices to account for risks that manipulated, inaccurate, or incomplete datasets pose to AI systems. Data can be weaponized in order to execute AI attacks, specifically poisoning attacks. For every dataset collected, AI users should ask themselves the following questions to identify potential weaknesses in the dataset that could be exploited for AI attacks:

How could adversaries have manipulated the data being collected?

If the adversary controls the entities on which data is being collected, they can manipulate them to influence the data collected. For example, consider a dataset of radar signatures of an adversary's aircrafts. Because the adversary has control over their own aircraft, it can alter them in order to alter the data collected. Adversaries need not be aware that data is being collected in order to manipulate the process. The existence of the possibility that data will be collected may be enough of a threat to execute this type of influence campaign.

Is an adversary aware data is being collected?

If an adversary is aware that data is being collected, they may try to interfere in some aspect of the collection process in order to alter the data being collected. An analogous example from the traditional cybersecurity domain can illustrate this example. When the U.S. was aware the Russia was stealing pipeline control software, they purposely altered the software to introduce a flaw into the software that would trigger a pipeline explosion.[68] Analogously in the data domain, if an adversary is aware that data is being collected to be used in an AI system, they may take additional steps to interfere in the data collection process to corrupt the data collected.

How was the data prepared?

After data is collected, it generally requires processing to prepare it for use with training AI systems. This preparation process presents opportunities to steal or poison the dataset and, therefore, the downstream AI system.

What inaccuracies may exist in the dataset?

Datasets may contain inaccurate data points for a number of reasons. To name a few common cases, data points may be mislabeled, corrupted, or inherently flawed. These mistakes do not necessarily stem from an adversary's actions. They may arise through completely natural processes such as human error and sensor failure. Because datasets can contain millions of data points, it is easy to overlook mistakes that exist in the dataset that may affect downstream AI systems and leave them open to attack.

Is there data missing from or underrepresented in a collected dataset?

AI systems can only learn concepts encapsulated within a dataset. If key types of data are either missing from or not sufficiently represented in a collected dataset, the resulting AI system will not be able to function properly when it encounters situations not represented in its dataset.

Once they have answered these questions, AI users should evaluate what risks exist within the dataset, and take corrective actions:

68 Russell, Alec, "CIA plot led to huge blast in Siberian gas pipeline", The Telegraph, 28 February 2004, https://www.telegraph.co.uk/news/worldnews/northamerica/usa/1455559/CIA-plot-led-to-huge-blast-in-Siberian-gas-pipeline.html

- If there is a risk adversaries may have been able to manipulate the data itself, additional steps should be taken to validate the data and remove data that is suspect.
- If there is a risk the data preparation process has been compromised, the data may need to be re-prepared or discarded.
- If there is a risk the dataset may not be complete, additional data may need to be collected.

Restrict Data Sharing

Critical AI systems must restrict how and when the data used to build them is shared in order to make AI attacks more difficult to execute. For critical applications, as a rule data should by default not be shared. Exceptions should be well reasoned. The resulting data sharing policies should be explicitly written and followed.

This restriction on data sharing is required because knowledge of the dataset used to train the AI system makes executing AI attacks significantly easier. However, this is in stark contrast to current data sharing policies that encourage data sharing. The Federal Government's National Artificial Intelligence Research and Development Strategic Plan explicitly calls for the open sharing of data among agencies.[69] The open source movement prioritizes data sharing and open datasets. The foundation of the DoD's AI efforts "includes shared data, reusable tools, frameworks, libraries, and standards…"[70] due to the fact that these military datasets are expensive—both in terms of time and money—to collect and prepare.[71] These examples affirm that data sharing norms are not universally wrong and are based in other legitimate practices. However, these established norms are wrong for certain high-security contexts and applications. When data is

69 National Science and Technology Council, Networking and Information Technology Research and Development Subcommittee, "The National Artificial Intelligence Research and Development Strategic Plan", October 2016, https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf

70 Statement by Dana Deasy, Department of Defense Chief Information Office, Before the House Armed Services Committee Subcommittee on Emerging Threats and Capabilities on "Department of Defense's Artificial Intelligence Structure, Investments, and Applications", 26 February 2019, https://armedservices.house.gov/_cache/files/5/7/579723e2-4461-4a8c-95da-ec3e84c4985e/E41B38FCB69AD83331F31CDC06570D33.hhrg-116-as26-wstate-deasyd-20190226.pdf.

71 "Interview with Lieutenant General Jack Shanahan: Part 1", Over the Horizon Multi-Domain Operations and Strategy, 2 April 2018, https://othjournal.com/2018/04/02/interview-with-lieutenant-general-jack-shanahan-part-1/

shared widely, there is a larger risk that it will be stolen or accidently copied on to insecure systems.

As such, when writing data sharing policies, AI users must challenge these established norms, consider the risks posed by data sharing, and shape data sharing policies accordingly. Without this, constituent parties may not realize the strategic importance data provides to attackers, and therefore may not take the steps necessary to protect it in the absence of explicit policy.

Once a data sharing policy for a particular dataset is written, it must be implemented in such a way that it is reasonably expected to be followed. Data by nature is free-flowing: in a matter of seconds, gigabytes of data can easily flow over a network link and compromise an entire organization's security. Implementation of data sharing policies should focus on making data more "sticky" so that it is not as easy to flow to where it should not be. Unwritten exceptions where data is informally shared should not be allowed to exist. User and application specific encryption of data can be used to this end in order to restrict accidental or improper sharing. This will make attributing improper data sharing practices easier so that offending parties can be held accountable.

" Implementation of data sharing policies should focus on making data more "sticky" so that it is not as easy to flow to where it should not be."

Implementation Stage Compliance Requirements

Implementation stage compliance requirements focus on ensuring stakeholders are taking proper precautionary steps as they build and deploy their AI systems. This includes securing assets that can be used to launch AI attacks, and improving detection systems that can warn when attacks are being formulated.

Secure Soft Assets

Protect the assets that can be used to craft AI attacks, such as datasets and models, and improve the cybersecurity of the systems on which these assets are stored.

AI system operators must recognize the strategic need to secure assets that can be used to craft AI attacks, including datasets, algorithms, system details, and models, and take concrete steps to protect them. In many contexts, these assets are currently not treated as secure assets, but rather as "soft" assets lacking in protection. This is because the threat of AI attacks is not widely known, and as such, these critical assets are treated with lower security standards compared to "hard" assets, such as passwords, that are stored with high security standards and encryption. This can no longer be the case. Critical applications that employ AI must adopt a set of best practices to harden the security of these assets.

These best practices should be formulated with joint input from security experts and domain experts for each application, and are likely to include changes such as only transmitting data over classified or encrypted networks, encrypting stored data to protect it even if the system is compromised, and keeping system details, such as tools and model hyperparameters, secret.

Hardening these "soft" targets will be an integral component of defending against AI attacks. This is because the two prominent forms of AI attacks discussed here, input and poisoning attacks, are easier to execute if the attacker has access to some component of the AI system and training

pipeline. This has transformed a wide range of assets that span the AI training and implementation pipelines into targets for would-be attackers. Specifically, these assets include the datasets used to train the models, the algorithms themselves, system and model details such as which tools are used and the structure of the models, storage and compute resources holding these assets, and the deployed AI systems themselves.

> " Hardening each part of the AI system will require different approaches."

Hardening each part of the AI system will require different approaches. For datasets, the challenge will be to secure the systems on which data is stored, and reevaluating paradigms such as open source data sharing directives for sensitive applications, as discussed previously. Keeping datasets secure is one key to protecting against AI attacks: if adversaries obtain the dataset used to train a model, they can use it to reverse engineer the model and then use this constructed copy to craft attacks. As a result, data must be managed through its entire provenance, or lifetime. Starting from how and when the data is collected, how it is labeled, how it is stored, how it is accessed during the model training process, through how it is archived, the data must be kept secret and fully protected. To accomplish this, at all points in this process, the data may need to be encrypted using the strongest encryption possible, and access to decryption keys must be managed securely. In order to protect against integrity attacks on the data, new technologies such as blockchains may be adopted.

Establishing a norm of hardening this "soft" target will be challenging because it goes against established habits and thoughts around data. In many applications, data is neither considered nor treated as confidential or classified, and may even be widely and openly shared.

This hardening must extend to the model itself. Even if the data is properly secured and an uncompromised model is trained, the model itself must then be protected. A trained model is just a digital file, no different from an image or document on a computer. As such, like other digital assets, it can be stolen or corrupted. If a model is stolen, crafting an attack is relatively easy. If an uncompromised model is corrupted or replaced with a corrupted one, all other protection efforts are completely moot. As such, the model itself must be recognized as a critical asset and protected, and the

storage and computing systems on which the model is stored and executed must similarly be treated with high levels of security.

> " Regardless of the reason for doing so, placing AI models on edge devices makes protecting them more difficult."

However, recent trends in how models are used will complicate efforts to protect them. Recently, models are no longer residing and operating exclusively within data centers where security and control can be centralized, but are instead being pushed directly to devices such as weapon systems and consumer products. This change is necessary for applications in which it is either impossible or impractical to send data from these "edge" devices to a data center to be processed by AI models living in the cloud. For example, in the case of weapon systems, this may be impossible because the enemy has jammed the communication channels. In the case of consumer applications such as autonomous cars, this may be impractical because the device will not receive a response fast enough to meet application requirements.

Regardless of the reason for doing so, placing AI models on edge devices makes protecting them more difficult. Because these edge devices have a physical component (e.g., as is the case with vehicles, weapons, and drones), they may fall into an adversary's hands. Care must be taken that if these systems are captured or controlled, they cannot be examined or disassembled in order to aid in crafting an attack. In other contexts, such as with consumer products, adversaries will physically own the device along with the model (e.g., an adversary can buy a self-driving car in order to acquire the model that is stored on the vehicle's on-board computer to help in crafting attacks against other self-driving cars). In this case, care must be taken that adversaries cannot access or manipulate the models stored on systems over which they otherwise have full control. Encryption will play an important role in securing these assets.

Improve intrusion and attack formulation Detection

Improve intrusion detection systems to better detect when assets have been compromised and to detect patterns of behavior indicative of an adversary formulating an attack.

While hardening soft targets will raise the difficulty of executing attacks, attacks will still occur and must be detected. Policymakers should encourage improved intrusion detection for the systems holding these critical assets, and the design of methods profiling anomalous behavior to detect when attacks are being formulated. While an ounce of prevention is worth a pound of cure, it is imperative to know when prevention has failed so that the system operator can take the necessary mitigation steps before the adversary has time to execute an attack.

In the simplest scenarios where a central repository holds the datasets and other important assets, the vanilla intrusion detection methods that are currently a mainstay of cybersecurity can be applied. In this simple case, if assets such as datasets or models are accessed by an unauthorized party, this should be noted immediately and the proper steps should be taken in response.

There are other scenarios in which intrusion detection will be significantly more difficult. As previously discussed, many AI systems are being deployed on edge devices that are capable of falling into an attacker's hands. If a piece of military software is captured by an enemy, the model and AI system on it must be treated as would be any other piece of sensitive military technology, such as a downed drone. Compromise of one system could lead to the compromising of any other system that shares critical assets such as datasets. As such, methods detecting intrusions in contested environments where the adversary has gained control of the system must be developed.

Protecting against attacks that do not require intrusions will need to be based on profiling behavior that is indicative of formulating an attack. This will hold particularly true for the many AI applications that use open APIs

to allow customers to utilize the models. Attackers can use this window into the system to craft attacks, replacing the need for more intrusive actions such as stealing a dataset or recreating a model. In this setting, it can be difficult to tell if an interaction with the system is a valid use of the system or probing behavior being used to formulate an attack. For example, is the case of a user sending the same image to a content-filter one hundred times 1) a developer diligently running tests on a newly built piece of software, or 2) an attacker trying different attack patterns to find one that can be used to evade the system? System operators must invest in capabilities able to alert them to behavior that seems to be indicative of attack formulation rather than valid use.

Regardless of the methods used, once a system operator is aware that an intrusion has occurred that may compromise the system or that an attack is being formulated, the operator must immediately switch into mitigation mode. As discussed in the mitigation stage compliance requirements below, system operators should have a predetermined plan that specifies exactly the actions that should be taken in the case of system compromise, and put the plan into action immediately.

Mitigation Stage Compliance Requirements

Mitigation stage compliance requirements focus on ensuring stakeholders plan responses for when attacks inevitably occur. This includes creating specific response plans for likely attacks, and studying how the compromise of one AI system will affect other systems.

Create attack response plans

Determine how AI attacks are most likely to be used, and craft response plans for these scenarios.

Stakeholders must determine how AI attacks are likely to be used against their AI system, and then craft response plans for mitigating their effect. In determining what attacks are most likely, stakeholders should look to existing threats and see how AI attacks can be used by adversaries to accomplish a similar goal. For example, for a social network that has seen itself mobilized to spread extremist content, it can be expected that input attacks aimed at deceiving its content filters are likely.

After this, response plans should be designed. Response plans should be based on the best efforts to respond to attacks and control the amount of damage. Continuing the social network example, sites relying on content filtering may need response plans that include the use of other methods, such as human-based content auditing, to filter content. The military will need to develop protocols that prioritize early identification of when its AI algorithms have been hacked or attacked so that these compromised systems can be replaced or re-trained immediately. Existing work in this area can be looked to as a learning experience. Facebook's algorithms were able to successfully remove 1.2 million of the 1.5 million known video uploads of the 2019 New Zealand shooting automatically upon upload, but then had to turn to additional techniques to remove the remaining 300,000.[72]

72 Reuters, "Facebook says it removed 1.5 million videos of the New Zealand mosque attack", 17 March 2019, https://www.reuters.com/article/us-newzealand-shootout-facebook-video/facebook-says-it-removed-15-million-videos-of-the-new-zealand-mosque-attack-idUSKCN1QY05X

Similar human-machine partnerships that Facebook sometimes employs[73] will need to become the norm in an era in which the AI systems are vulnerable to attack.

Response plans may also require real-world action to be taken. For example, police response plans to input attacks on infrastructure, such as signs and road markers, will require the immediate dispatch of officers. Just as officers are dispatched to an intersection when a traffic light is broken, similar responses will be needed. In this case however, the response will need to be immediate—humans can still navigate a broken traffic light relatively well, but a driverless car will run a now "invisible" stop sign without the human passengers having a chance to intervene. This response plan may also require expanded partnerships and information sharing agreements with other entities, such as companies controlling the technology. Further, the response plan will require training and coordination such that officers will be equipped to recognize that seemingly harmless graffiti or vandalism may actually be an attack, and then know to activate the appropriate response plan.

Rapid Shared Vulnerability Mapping

Create maps showing how the compromise of one asset or system affects all other AI systems.

Policymakers should require AI system operators to map how the compromise of a given asset or system would affect all other systems. Characteristics of the AI domain make these shared vulnerabilities common. Given the easy transport of data, the convenience and monetary savings of reusing data, and the operational benefits of sharing tools and models, many AI systems will share the same underlying assets such as datasets. However, this sharing has a dark side: the compromise of one asset may compromise other assets that have also utilized this asset.

73 Liptak, Andrew, "Facebook says that it removed 1.5 million videos of the New Zealand mass shooting", The Verge, 17 March 2019, https://www.theverge.com/2019/3/17/18269453/facebook-new-zealand-attack-removed-1-5-million-videos-content-moderation

Given the reality of how data is shared and repurposed, shared dependencies—and therefore vulnerabilities—among systems will be widespread for better or worse. As a result, there is a need to rapidly understand how a compromise of one asset or system affects other systems.

This can be accomplished via rapid shared vulnerability mapping. Organizations should have vulnerability maps that document the assets their different AI systems share. This mapping should be rapid in the sense that once an asset or system is compromised, it should not require additional analysis to determine what other systems are compromised. For example, one such map would document which systems utilized the same training datasets. If this dataset was later compromised, administrators would immediately know what other systems are vulnerable and need to be addressed.

These shared vulnerability maps should be integrated into the attack response plans as well.

Part V: Implementation and Enforcement

Implementation

AI security compliance programs should be enforced for portions of both the public and private sectors. Broadly, as a rule, compliance should be mandated for government uses of AI. Further, because the government is turning to the private sector to develop its AI systems, compliance should be mandated as a precondition for companies selling AI systems to the government. Government applications for which truly no risk of attack exists, for example in situations where a successful attack would have no effect, can apply for a compliance waiver through a process that would review the circumstances and determine if a waiver is appropriate.

More specifically, different segments of the public sector can implement versions of compliance that meet their needs on a segment-by-segment basis. For the military, the JAIC is a natural candidate for administrating this compliance program. As it is specifically designed as a centralized control mechanism over all significant military AI applications, it can use this centralized position to effectively administer the program. For law enforcement, the DOJ can use its relationship with law enforcement organizations, including the FBI and local law enforcement offices, as a basis for administrating a compliance program. Where necessary, DOJ can tie compliance as a pre-condition for receiving funding through grants.

In the private sector, regulators should make compliance mandatory for high-risk uses of AI where attacks would have severe societal and public safety consequences. This report has identified examples of private sector high-risk uses of AI, including content filters and self-driving vehicles. In some cases, compliance can be mandated legislatively directly by Congress. For example, in the context of the relatively unregulated space of social networks, there is a call from both legislators and industry itself for additional regulation. Any regulation of the industry can mandate AI security compliance. In other contexts, it may be more appropriate and effective for

agencies already regulating an industry to manage compliance mandates and details. In the context of self-driving cars, this may fall to DoT or one of its sub-agencies, such as NHTSA. In the context of other consumer applications, this may fall to other agencies such as the FTC.

Enforcement

Once AI Security Compliance programs are implemented, regulators should decide in what ways entities will be held responsible for meeting compliance requirements, and clearly communicate these principles with their constituents. Informed AI users in critical areas should be held responsible for acting in good faith and taking appropriate measure to protect against AI attacks.

Because it is currently believed that the widely-used AI algorithms are vulnerable to attack, companies will of course not be able to exhaustively protect against AI attacks, just as they are not expected to exhaustively protect against traditional cyberattacks. However, they should be required to make reasonable efforts. These efforts include following the policy proposals set forth in this report, including conducting a rigorous AI suitability test, generating and implementing attack response plans, making attacks more difficult to execute by hardening the security protections of assets such as datasets and models, and improving their intrusion detection capabilities.

Regulators should clearly communicate these expectations to their constituents, along with the potential ramifications that will occur if these steps are not taken and an attack occurs.

Drawbacks

While these security steps will be a necessary component of defending against AI attacks, they do not come without cost. From a societal standpoint, one point of contention is that some of these security precautions

will require a trade-off against other important considerations, such as ensuring that AI systems are fair, unbiased, and trustworthy. Many of the methods to verify these properties rely on openly publishing datasets, methods, models, and APIs to the systems. However, these exact actions double as a list of worst practices in terms of protecting against AI attacks. In already deployed systems that require both verified fairness and security, such as AI-based bond determination,[74] it will be difficult to balance both simultaneously. New methods will be needed to allow for audits of systems without compromising security, such as restricting audits to a trusted third party rather than publishing openly.

From an implementation standpoint, a difficulty in implementing this policy will be managing the large number and disparate nature of entities, ranging from the smallest startups to the largest corporations, that will be implementing AI systems. Because different stakeholders face unique challenges that may not be applicable in other areas, regulators should tailor compliance to their constituents in order to make the regulation germane to their industry's challenges.

From a technological standpoint, an additional difficulty is created by the fact that the field and technology itself is rapidly changing. As a result, regulators should not focus on all entities and all uses of AI. Instead, broad yet shallow efforts should be made at educating the entire field, but more focused attention should be reserved for entities and applications that regulators fear present an outsized danger. These may include products used in law enforcement, intelligence, and military contexts, as well as applications that can have public safety ramifications, such as self-driving cars.

From a political standpoint, a difficulty in gaining acceptance of this policy is the fact that stakeholders will view this as an impediment to their development and argue that they should not be regulated either because 1) it will place an undue burden on them, or 2) they do not fall into a "high-risk" use group. Regulators must balance security concerns with the burdens placed upon stakeholders through compliance.

74 See, e.g., the Correcitonal Offender Management Profiling for Alternative Sanctions tool (COMPAS) and the use of it by various government institutions, e.g., https://doc.wi.gov/Pages/AboutDOC/COMPAS.aspx and https://qz.com/1375820/california-just-replaced-cash-bail-with-algorithms/.

Additional Recommendations

Additional policy can complement the effectiveness of AI Security Compliance programs.

Prioritize Research of Defense Mechanisms and More Robust Algorithms

Additional Recommendation 1: Increase research funding of methods to defend against AI attacks and the creation of new robust AI algorithms. Mandate the inclusion of a security assessment on all AI-related research grants.

Research should prioritize the creation of defense mechanisms for the current state-of-the-art AI methods, as well as the development of new more robust AI methods. Given the success of deep learning and its already established footprint, these vulnerable methods will be the primary methods used for a substantial amount of time. As such, even if complete mitigation is provably impossible, techniques to "harden" the methods, such as making attacks more difficult to execute by modifying the structure of the models themselves, will be of significant interest to AI users. Similar hardening techniques have found great success in cybersecurity, such as Address Space Layout Randomization (ASLR), and have imposed significant technical hurdles for performing once common and easy cyberattacks.

Government funding organizations such as DARPA should continue to use their agenda setting power to establish AI security as an important and urgent topic under the auspices of national security. While many previous grants and projects have focused on increasing the capabilities of AI algorithms, more attention should now be paid to the robustness of existing capabilities rather than a sole focus on traditional evaluation metrics such as accuracy. DARPA has already set a good example of this through its Guaranteeing AI Robustness Against Deception (GARD) program.[75]

75 Guaranteeing AI Robustness against Deception (GARD), DARPA, https://www.darpa.mil/attachments/GARD_ProposersDay.pdf.

Beyond creating programs and grants aimed solely at defense mechanisms and creating new methods not vulnerable to these attacks, DARPA and other funding bodies should mandate that every research project related to AI must include a component discussing the vulnerabilities introduced by the research. This will allow users who potentially adopt these technologies to make informed decisions as to not just the benefits but also the risks of using the technology.

In addition to a technical focus on securing models, research attention should also focus on creating testing frameworks that can be shared with industry, government, and military AI system operators. In a similar manner to how automobiles are tested for safety, testing frameworks for the security of models can be established and used as a core component alongside the traditional testing methods used for vehicles, drones, weapon systems, and other systems that will adopt AI.

Educating Stakeholders with Domain and Threat Awareness

Additional Recommendation 2: The FTC, DoD, and DOJ should alert their relevant constituents regarding the existence of AI attacks and preventative measures that can be taken.

Policymakers and relevant regulatory agencies should educate stakeholders about the threat landscape surrounding AI. Specifically, this education should be twofold. First, it should focus on publicizing the existence and ramifications of AI attacks. This will allow stakeholders to make educated decisions regarding if AI is appropriate for their domain, as well as develop response plans for when attacks occur. Second, it should provide resources informing relevant parties about the steps they can take to protect against AI attacks from day one.

The first component of this education should focus on informing stakeholders about the existence of AI attacks. This will enable potential users to make an *informed* risk/reward tradeoff regarding their level of AI adoption. Leaders from the boardroom to the situation room may similarly suffer from unrealistic expectations of the power of AI, thinking it has human

intelligence-like capabilities beyond attack. This may lead to premature replacement of humans with algorithms in domains where the threats of attack or failure are severe yet unknown. This will hold particularly true for applications of AI to safety and national security. Decisions in these domains may be made for purposes of reducing operating expenditures, increasing efficiency, or broad imperatives to adopt new technology and "modernize." Without a proper understanding of the threats that exist to an AI-based system, proper cost-benefit analyses cannot be conducted, and dangerous vulnerabilities may be overlooked that create systematic risk within these critical domains.

From a practical standpoint, government agencies should take control of educating and interfacing with affected constituents, as each group has unique concerns and circumstances. These agencies should be the DoD, FTC, and DoJ for the military, consumer, and law enforcement communities, respectively. In order to avoid the siloing of best practices and lessons learned within each department, agencies should place a priority on publishing their efforts openly and communicating findings outside of usual intra-agency pathways.

Reevaluation of AI Applications

Additional Recommendation 3: Reevaluate the role AI should play in future applications, with regard to safety and proper planning.

Policymakers and industry alike must study and reevaluate the planed role of AI in many applications. While this may appear Ludditian in view, it has a historical basis. The US's Strategic Automated Command and Control System, a component within the U.S. nuclear control system, still uses technology systems from the 1970s rather than updated state-of-the-art computers.[76] This is because the presence of cybersecurity vulnerabilities in new technologies poses too great a risk for this particular application.

76 Fung, Brian, "The Real Reason America Controls its Nukes with Ancient Floppy Disks", The Washington Post, 26 May 2016, https://www.washingtonpost.com/news/the-switch/wp/2016/05/26/the-real-reason-america-controls-its-nukes-with-ancient-floppy-disks/?noredirect=on&utm_term=.e4d0d5a41b7a

Similar discussions must occur in regard to the integration of AI into other applications, but not necessarily with the end goal of reaching binary use/don't use outcomes. For some applications, the integration of AI may pose such little risk that there is little worry. For others, AI may require human supervision. While this supervision may not always protect against the consequences of all AI attacks, it may reach a common ground between full exposure to attack risk and the risk of not realizing the benefits AI can deliver. The military is setting a good example for this intermediate use by prioritizing the development of AI systems that augment but do not replace human control. Finally, some applications of AI may prove too dangerous to use. Autonomous weapon systems, even those that do not utilize AI, already carry great stigma due to a fear that attack or algorithmic mistakes will cause unacceptable collateral damage, and therefore present unacceptable levels of risk. This same attitude may be adopted in other applications reliant on AI.

In some contexts, these discussions can be internally led. The DoD, for example, has already shown attention to understanding and addressing the security risks of employing AI. However, in other contexts, such as in industry settings where parties have shown a disregard and inability to address other cyber risks, these discussions may need to be forced by an outside regulatory body such as the FTC.

Conclusion

> *"Knowledge is knowing that Frankenstein is not the monster. Wisdom is knowing that Frankenstein is the monster."*[77]

For hundreds of years, humans have been wary of inscribing human knowledge in technical creations. With machine learning and artificial intelligence, we take a step closer to this fear.

> " Artificial intelligence today presents seismic unknowns that we would be wise to ponder."

It is the fear of the unknown of a creation. And artificial intelligence today presents seismic unknowns that we would be wise to ponder. Artificial intelligence, like Frankenstein's monster, may appear human, but is decidedly not. Despite the popular warnings of sentient robots and superhuman artificial intelligence that grow more difficult to avoid with each passing day, artificial intelligence as it is today possesses no knowledge, no thought, and no intelligence. In the future, technical advancements may one day help us to better understand how machines can learn, and even learn how to embed these important qualities in technology. But today is not that day.

The current set of state-of-the-art artificial intelligence algorithms are, at their essence, pattern matchers. They are intrinsically vulnerable to manipulation and poisoning at every stage of their use: from how they learn, what they learn from, and how they operate. This is not an accidental mistake that can be easily fixed. It is embedded deep within their DNA.

As a result, it is imperative that policymakers recognize the problem, identify vulnerable systems, and take steps to mitigate risk before people get hurt. This report has identified five critical areas that are already vulnerable to these attacks, and growing more so with each day. The content filters that will serve as the first line of defense against extremist recruiting, misinformation and disinformation campaigns, and the spread of hate and encouragement of genocide can be rendered ineffective with AI attacks. A U.S. military transitioning to a new era of adversaries that are its

77 Anonymous quote.

technological equals or even superiors must develop and protect against this new weapon. Law enforcement, an industry that has perhaps fallen victim to technological upheaval like no other, risks its efforts at modernizing being undermined by the very technology it is looking at to solve its problems. Commercial applications that are using AI to replace humans, such as self-driving cars and the Internet of Things, are putting vulnerable artificial intelligence technology onto our streets and into our homes. Segments of civil society are being monitored and oppressed with AI, and therefore have a vested interest in using AI attacks to fight against the systems being used against them.

The unfettered building of artificial intelligence into these critical aspects of society is weaving a fabric of future vulnerability. Policymakers must begin addressing this issue today to protect against these dangers by creating AI security compliance programs. These programs will create a set of best practices that will ensure AI users are taking the proper precautionary steps to protect themselves from attack. In high-risk application areas of AI, such as government and critical industry use of AI, compliance can be mandatory and enforced by the appropriate regulatory bodies. In low-risk application areas of AI, compliance can be optional in order to not stifle innovation in this rapidly changing field.

" The unfettered building of artificial intelligence into these critical aspects of society is weaving a fabric of future vulnerability."

The world has learned a number of painful lessons from the unencumbered and reckless enthusiasm with which technologies with serious vulnerabilities have been deployed. Social networks have been named as an aide to genocide in Myanmar and the instrument of democratic disruption in the world's foremost democracy. Connected infrastructure has led to attacks with hundreds of millions of dollars of economic loss. The warning signs of AI attacks may be written in bytes, but we can see them and what they portend. We would be wise to not ignore them.

Belfer Center for Science and International Affairs
Harvard Kennedy School
79 John F. Kennedy Street
Cambridge, MA 02138

www.belfercenter.org

Printed in the United States of America