Shall We Play A Game?

Analyzing Threats to Artificial Intelligence

ISBN 979-8-8364-7154-5

Cover: This image was created by Infranodus[1], which uses Natural Language Processing to identify key terms in this book, how prevalent they are, and how closely they're related to other terms.

Reading notes: Code and equations are in **bold**. Key terms are in *italics*. I've tried to cite original research where possible. Errata and more content can be found at: **http://hackingai.net**

Dedicated to my wife, Tracy Cassidy. Many of the ideas in this book started as dinner table conversations with her about how Artificial Intelligence will impact and be impacted by various fields such as cybersecurity, threat assessment, social engineering, and risk analysis.

Table of Contents

Table of Figures

Chapter 1. Introduction

"Hacking Artificial Intelligence is one of the most profound challenges facing humanity, and most smart people are quite worried about it. Many are looking for a solution in the form of an artificial intelligence (AI) that can beat the best human players at chess and Go, and a few are even predicting that within a century AI will outcompete humans at various other tasks, like, one could argue, flying cars.

However, this appears to be quite optimistic, at least for now, and so it was with much disappointment that I came across a story that appears to be a fictional example of a dangerous AI, one that not only showed the potential of a dangerous artificial intelligence, but one which also created a gigantic death spiral. I urge you to read on,"

The above intro was created by an Artificial Intelligence tool, Talk To Transformer[2], starting from my phrase "Hacking Artificial Intelligence". It's a bit sensationalist - I'm pretty sure the phrase "gigantic death spiral" wasn't going to appear in this book - but there is a surprising accuracy to it. The references to Artificial Intelligence systems, or AIs, playing Go and chess against humans echo my choice of a quote from the 1983 movie *War Games*[3] for the title of this book: "Shall we play a game?" If we're talking about "hacking" AI, we are in a sense competing with them, in the same way that the character in *War Games* played a game against an artificially intelligent computer. In this book, I am certainly going to discuss the dangers of AI, albeit mostly from the perspective of human adversaries making use of them.

Artificial Intelligence (AI) is certainly a hot topic

right now. It seems to be mentioned in every article on technology and the media writ large. It is a very powerful technology that is exponentially getting more powerful. The transition from “computers that compute” to “computers that reason” might be as big a transition for society as the initial invention of computers. AI is no longer a thing of the future - it’s here and it’s already having a large impact. AI is being adopted at breakneck speed, in autonomous vehicles, financial analysis, and even the smart assistants in our homes. I’m writing this book during the coronavirus pandemic, and AI is a major tool that researchers are using to develop treatments, discover vaccines, and slow the spread of the virus.

As with most new technologies, the functionality is being developed faster than the security, and certainly faster than the law and public policy. Many words have already been written about potential risks of AI. Movies have been preparing us to be scared of hostile artificial life for decades - Bladerunner, the Terminator, The Matrix. So why do I feel the need to write a book about it?

Why This Book Exists

This book was born out of my interests in AI and cybersecurity. My hope is that it will contribute to several under-served niches regarding AI and security. I wrote this book for three primary reasons:

1. Current discussions concerning the risks of AI focus on Artificial General Intelligence, or AGI, a possible future technology that is as smart as or smarter than humans.
2. AI risk discussions often focus on the risks that the AI may cause, rather than risks to AI systems themselves.
3. My expertise is specifically in how to attack ("hack") computer systems.

To address the first reason, I feel that current discussions of AI risks jump past current and near-future implementations of AI and instead go straight to Artificial General Intelligence (AGI), a capability that does not yet exist and may or may not in the future. It is difficult to define what it would mean for machines to be "as smart as or smarter than humans." When or if AGI emerges, it will raise many legal, philosophical, and, of course, security concerns. The latter are important to consider, and maybe even plan for, in the present day but it seems premature to worry about future risks without first assessing current AI risks. The absence of AGI does not imply an absence of risk; there are plenty of things that can go wrong with the AI that already exists.

I wanted to write a book with information that's actionable now, as AI has many current risks and security issues. It is already changing the risk landscape. My hope is that this book will provide you with actionable recommendations on how to protect current AI systems and gain insight into how adversaries do and might try to subvert them.

My second reason for writing this book is because there is not currently a lot of information about risks to the AI systems themselves. Review of current literature about AI and AI risks centers on the risks that AI will create and not on the risks of people attacking AI systems. When I see discussions of the security of AI systems, it's often shrugged off as 'they're just computers; protecting them is the same as any other cybersecurity challenge.' When a novel attack is demonstrated, like data tainting attacks, it's dismissed as 'just academic.'

AI itself will certainly create risks. There are security risks, such as the potential that AI can find more security bugs (vulnerabilities) in software, and do it faster than human researchers. This might stretch the ability of software developers to fix the vulnerabilities as fast as the AI can find them.

There are also what I call insecurity risks, where we trust AI to perform security functions and they perform them poorly. For example, a fraud detection system could have errors in it that allow fraud to go undetected.

One of the most significant risks is that AI will inherit cognitive biases from its human creators. These biases could be as subtle as recency bias or as blatant as racism or other forms of prejudice.

The security risks that AI can present are important but it is also important to think about risks to the AI systems themselves. There will be ways to attack or

subvert artificially intelligent systems that are unique to that system, as well as ways to attack AI systems as a whole. We'll need to defend against both types of attacks. AI systems are valuable. Their information and output are valuable, even when they're not expensive to run. Therefore, they are attractive targets - possibly even to each other.

Finally, I have a different perspective on these security risks than many researchers. I've spent a majority of my career on the "offensive" side of cybersecurity. This entails looking at systems and networks in order to find security problems and figure out ways to exploit them, preferably without setting off any alarms. If the target is friendly, this is known as red teaming, penetration testing, or vulnerability discovery, and these techniques give valuable insight into potential security problems for organizations. If the target is an adversary (e.g., criminals, terrorists, hostile nations), these techniques can be used to collect intelligence or disrupt their operations.

Currently, AI security from an offensive point of view is not well explored. This is unfortunate, because taking an offensive approach, or "thinking like a bad guy" often leads to the discovery of weaknesses that wouldn't be discovered any other way.

Red teaming, one of the offensive tactics used, includes any sort of exercise where one examines organizational capabilities through the eyes of an adversary. It is a structured approach to thinking critically about one's current protocols in order to improve them. The key component of a red team exercise is to pretend to be an intelligent adversary actively trying to subvert an organization's security protocols. AI can also be employed against these security protocols, so we need to consider AIs as potential adversaries as well. Red teaming differs

from a disaster recovery or business continuity exercise, where the “adversary” is usually a force of nature rather than a malicious actor. The structure of red team exercises differ based on one’s industry, goals, and security maturity.

The cybersecurity industry often uses the phrase “red teaming” to mean the practice of trying to break into friendly computer networks, in order to highlight their weaknesses. However, red teaming can also include any sort of exercise where someone roleplays as an adversary. It might be a thought experiment (“What might the terrorists try next?”), a wargaming or tabletop exercise, or a kinetic military exercise with one group playing the enemy (sometimes called a “Tiger Team”.) It also includes “cyber red teaming” - hiring hackers to find flaws in your technology, your networks, or your systems - which is the type of offensive work that I have been involved in during my career.

This book can be seen as an attempt to red team AIs in advance. Throughout this book, we’ll be engaging in a predictive thought experiment, as AI is developing so quickly that many of the capabilities we’ll discuss don’t exist in a testable form yet. My hope is that an in-depth, thoughtful discussion of potential risks and attack surfaces of AI will serve to jumpstart your future planning, exercises, and adversarial testing.

I would be remiss if I didn’t discuss the ethical considerations associated with red teaming and offensive security techniques. Whenever someone in cybersecurity talks about “how to hack into” something, particularly in public media, there is always some debate about whether exposing these tactics, techniques, and procedures (TTPs) benefits the attackers or the defenders more. My take on this is “forewarned is forearmed”: defenders will not be

prepared if they have not taken the time to study the attacker's methodologies. I'm not overly concerned that this book will provide attackers with tactics, techniques, or procedures to attack AI that they may not have already considered.

Before I turn the discussion to potential ways to attack AI, I'll provide an overview of the scope and topics that I'll cover.

Bounding Artificial Intelligence

Defining the scope and parameters of AI is a long discussion that I am not going to explore in depth here. My intent is to present how I will define AI for the purposes of this book. As you continue through the book, you can assume that I am referring to the set of technologies that I'll discuss below unless I specify otherwise. The first bounding condition: the discussion of AI in the book does not include Artificial General Intelligence (AGI) or strong AI. This calls for further clarification, as the definition of intelligent is itself open to debate, so I'll be a little more specific. An AGI would not only be as smart as a human in one or more specific areas - doing math, driving a car, playing chess - but in all areas that a human can reason about. This would include cognition, problem solving, and imagination, among other areas. The implications of AGI are fun to ponder, but AGI is decades away even if it does ever exist.

For the purposes of this book, I will stick to discussing weak AI, or non-AGI. The usual definition of weak AI is "anything less than strong AI", but I would go

further and say I am discussing AI which is substantially short of strong AI. This weak AI might be able to do math, or drive a car, or play chess, but it cannot do all of the above in one system. Weak AI certainly can't decide which of the skills it should employ in a given situation, except in a very contrived scenario.

Another bounding condition is that I will talk about AI that exists in 2022 or can reasonably be expected to exist in the next decade or so. That almost certainly precludes AGI, but also precludes other forms that probably will not exist for a few decades. One can imagine, for example, "savant" types of AI that are fantastically intelligent in one field, but still well short of human intelligence in most others. "Savant" types of AI are still a thing of the future. If they become more common, I'll write a book about them in the future (or one of them can.)

Aside from this focus on near-term, sub-human level intelligence, I am going to be as inclusive as possible in the book. Some of the technologies we'll discuss include artificial neural networks (ANNs), genetic algorithms, planning algorithms, and symbolic / deductive approaches to AI. There is debate about which of these are or are not true AI, but it is not relevant to this discussion. Any computer system that learns or reasons is within our scope. (There are even AI systems that "dream"[4].)

The discussions will focus on common, contemporary examples of AI – such as autonomous robots, computer vision, and voice-controlled digital assistants - as well as less obvious or less common AI. I will make informed guesses as to what AI will likely be able to do in the future, while sticking to the outlined boundaries and avoiding pure science fiction.

Finally, a note on terminology. I generally use the terms "AI", "Artificial Intelligence", "ML", and "Machine

Learning" interchangeably, except when the difference matters to the context.

What Is Information Security?

I am going to take a similarly expansive view of security. I intentionally use the phrase "information security" (infosec) rather than cybersecurity. For better or worse, "cybersecurity" is normally used synonymously with "computer and network security". This is sufficient for delineating roles in an organization or deciding what topics go in which courses in academia. However, I believe that the security issues surrounding AI are much broader than merely securing the computers they run on and the networks that they access. If I didn't think that, I wouldn't be writing this book! The focus of this book will be on "information security", or all the ways that information can be manipulated or abused by adversaries, as it relates to AI.

Those coming from a cybersecurity background might think that some of these topics are out of scope or not of concern to them because they are not a "cybersecurity issue". Cybersecurity professionals might feel like some of these issues may be better handled by fraud investigators, or intelligence agencies, or even the marketing department. That is fair, and likely a common response, but I would suggest that when you are trying to identify security risks, it's better to overlap and have to deconflict areas of responsibility than it is to leave gaps. For example, when I worked on cybersecurity as it relates to fleets of cars, there was a disconnect in most

organizations. The people running the motor pool assumed the IT department would secure the network, while the IT people didn't think of cars as computers on their network. As a result, the (insecure) computers in the cars were often connected to the network in order to collect diagnostics and telemetry, thus creating risk.

We're going to set aside the mindset that the security of information is centered around making sure "hackers" don't "break in" to your computers and steal credit card numbers, or encrypt your drives and demand ransom. Instead, let's pivot to the viewpoint of 'we spent a lot of money on this sweet AI system, what could someone do that we don't want them to?'

One potential avenue of attack on AI systems is financial manipulation, which isn't normally considered an information security issue. With the increase in cloud computing, cryptocurrency, and smart contracts, code and data are going to be increasingly synonymous with money. An AI that can rapidly reallocate your company's computing load across multiple cloud services with constantly changing rates can potentially save an organization a lot of money on compute costs. On the other hand, an AI that can skim transaction fees off of a "world computer" like Ethereum can make lots of money - at the expense of all the other users. I will explore this topic in more depth in the chapter on cryptocurrency.

Privacy is an area that will certainly be impacted by AI, but it is not normally prioritized by security departments. In fact, privacy is often seen as antithetical to cybersecurity; the belief is that security personnel have to be able to see what innocent users are doing in order to detect hostile users. Privacy is often seen as a zero-sum game, where the more privacy users have, the less security companies and governments have. Regardless of

whether this has ever been true, the calculus will certainly change when you throw AI into the mix.

Another example is autonomous robots, which need to process large amounts of analog data to interact with the real world. Hacking computers usually involves getting them to process data they weren't expecting - a malformed file or malicious network data, for example. In the case of sensors on robots, the input data is often analog - light, sound, Lidar, radio waves - but it can still be manipulated. There have already been public demonstrations of the ability to trick sensors such as computer vision and Lidar. Of course, falsifying analog data has been going on forever, especially in the military, e.g., camouflage or stealth fighters. AI will be susceptible to these too, but it's likely they'll be more or less susceptible to certain patterns than humans.

Perhaps the most dangerous manipulations of information will come from using AI to emulate real people. There are already social media bots spreading misinformation, and deepfakes that make you think a real person did or said something they didn't, and algorithms tweaking the way "facts" are presented to you based on your online behavior.

Manipulating information is nothing new, but the scale and speed at which it can be done with AI is unprecedented.

Organization of this book

The rest of this book is organized as follows. There is a natural progression of the topics, but you can skip

around if some parts are more interesting to you. I suggest you read Chapter 2 because it discusses technology and architectures that are (or will be) common to many AI systems. After that, you can read straight through or treat it as more of a formatted wiki, where each chapter can stand alone.

Architectural Challenges
: Software, hardware, and network architectures are going to be different in AI than the traditional "a computer is connected to the Internet." What challenges does this pose to attackers?

Semantic attacks
: These are attacks where we manipulate the meaning of input, rather than the structure.

Classes of Attacks
: Broad categories of AI-specific attacks

Generative Adversarial Networks
: GANs are "AI vs AI" systems that attempt to improve one AI by having another one try to "trick" it.

Bots, artificial humans, and robots
: We'll dig deeper into tricking robots and human-like systems

Money and the Global Computer
: Fraud in the age of AI

Tools of the Trade
: The uses and shortcomings of symbolic execution, SMT solvers, deep reinforcement learning, threat modeling, and related techniques

Biases Equal Vulnerabilities
: How we can manipulate the human biases

which will inevitably be codified in AI

Competition as Code

Geopolitical issues - Will future conflicts be able to be described as code and played out between competing AIs? Will these AIs belong to nation-states?

Conclusion

One final note: This is not a book on how to attack AI systems that are used specifically for computer- and network- security functions, such as anti-virus or User Behavior Analytics (UBA.) However, I'll use some examples from the cybersecurity field because they are easy to understand in terms of adversarial participants. This does not mean that all, or even most, attacks will be intended to bypass security functions. Bypassing security functions is almost a means to a [financial/social/geopolitical] end.