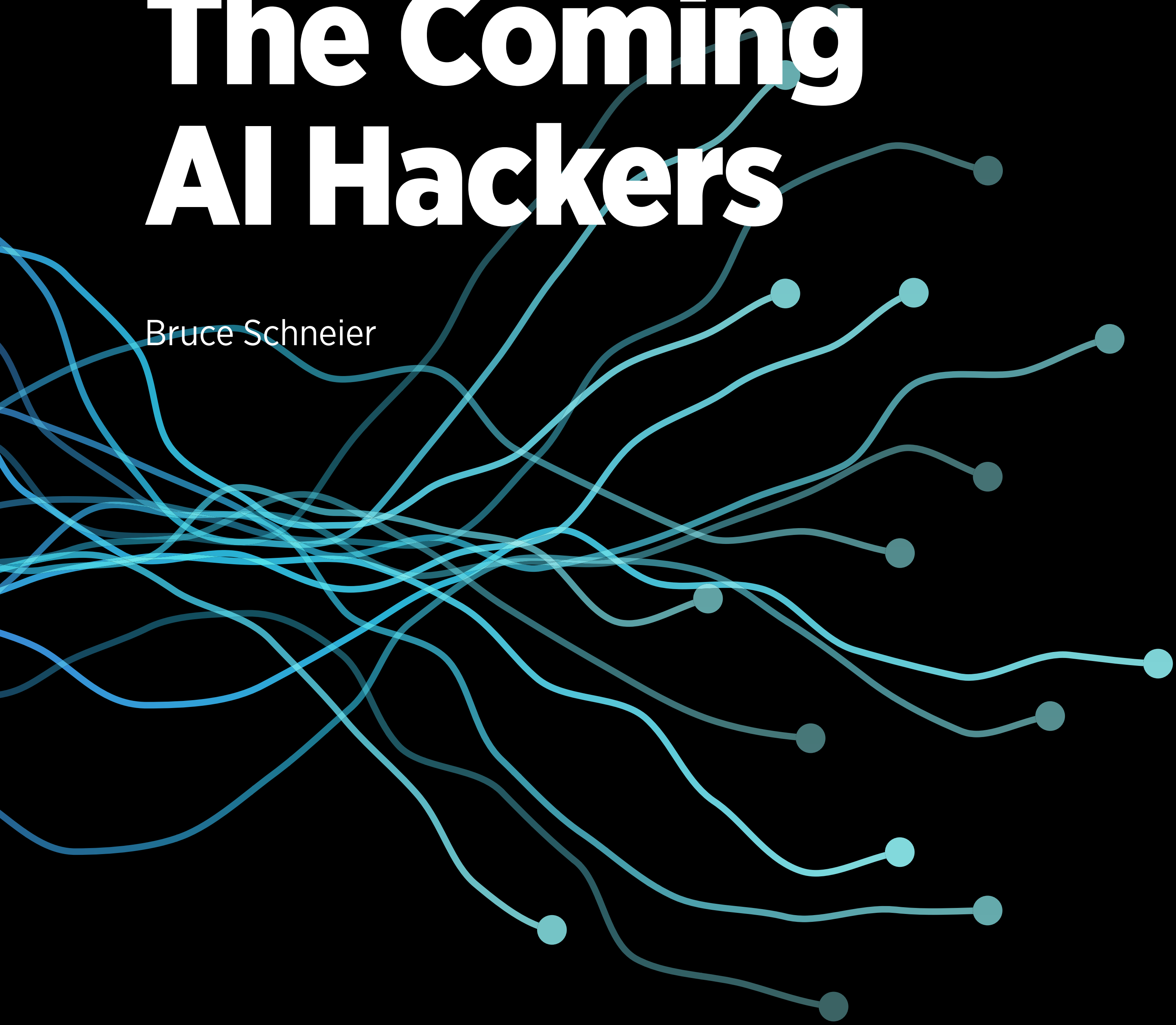
CYBER PROJECT × COUNCIL FOR THE RESPONSIBLE USE OF AI

The Coming AI Hackers

Bruce Schneier



HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

ESSAY
APRIL 2021

The Cyber Project
Council for the Responsible Use of AI
Belfer Center for Science and International Affairs
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

www.belfercenter.org/Cyber

Statements and views expressed in this report are solely those of the authors and do not imply endorsement by Harvard University, Harvard Kennedy School, the Belfer Center for Science and International Affairs, or the U.S. Government.

Design and layout by Andrew Facini

Printed in the United States of America

CYBER PROJECT × COUNCIL FOR THE RESPONSIBLE USE OF AI

The Coming AI Hackers

Bruce Schneier

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

ESSAY
APRIL 2021

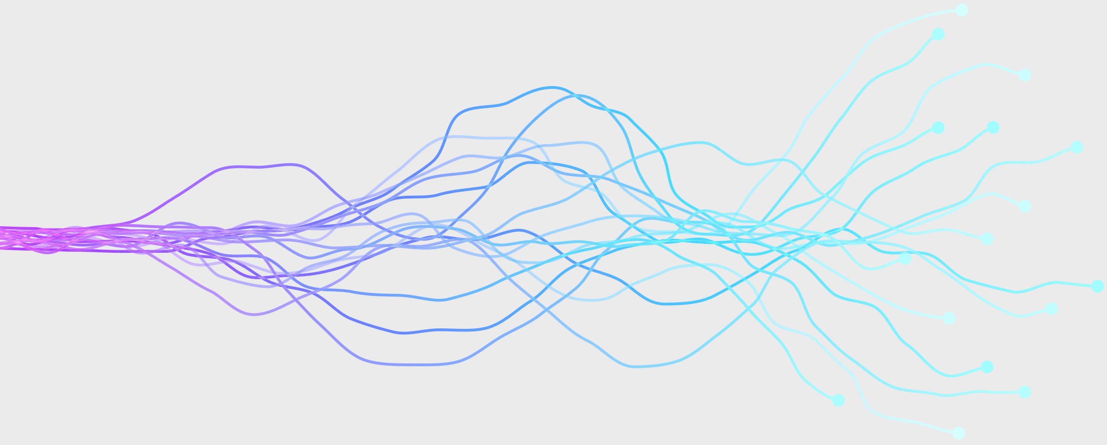
About the Author

Bruce Schneier is a fellow at the Cyber Project, in the Belfer Center for Science and International Affairs. He is the *New York Times* best-selling author of 14 books—including *Click Here to Kill Everybody*—as well as hundreds of articles, essays, and academic papers. His publishes the monthly newsletter *Crypto-Gram*, and blog *Schneier on Security*. Schneier is also a fellow at the Berkman Klein Center for Internet and Society at Harvard University; a Lecturer in Public Policy at the Harvard Kennedy School; a board member of the Electronic Frontier Foundation, AccessNow, and the Tor Project; and an advisory board member of EPIC and VerifiedVoting.org. He is the Chief of Security Architecture at Inrupt, Inc. He can be found online at www.schneier.com, and contacted at schneier@schneier.com.

Acknowledgments

I would like to thank Nicholas Anway, Robert Axelrod, Robert Berger, Vijay Bolina, Ben Buchanan, Julie Cohen, Steve Crocker, Kate Darling, Justin DeShazor, Simon Dickson, Amy Ertan, Gregory Falco, Harold Figueroa, Brett M. Frischmann, Abby Everett Jaques, Ram Shankar Siva Kumar, David Leftwich, Gary McGraw, Andrew Odlyzko, Cirsten Paine, Rebecca J. Parsons, Anina Schwarzenbach, Victor Shepardson, Steve Stroh, Tarah Wheeler, and Lauren Zabierek, all of whom read and commented on a draft of this paper.

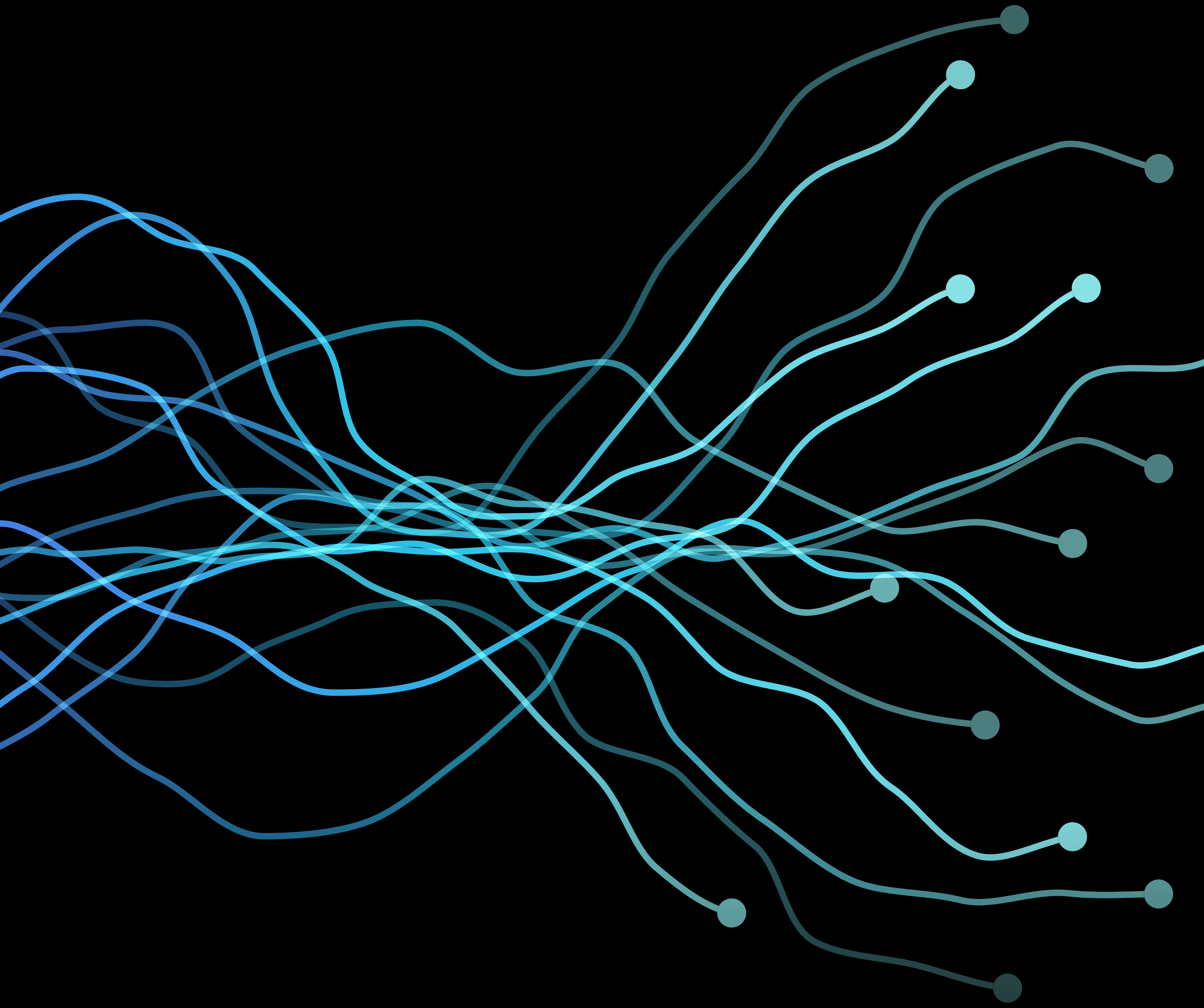
I would also like to thank the RSA Conference, where I gave a keynote talk on this topic at their 2021 virtual event; the Belfer Center at the Harvard Kennedy School, under whose fellowship I completed much of the writing; and the 5th International Symposium on Cyber Security Cryptology and Machine Learning, where I presented this work as an invited talk.



Summary

Hacking is generally thought of as something done to computer systems, but this conceptualization can be extended to any system of rules. The tax code, financial markets, and any system of laws can be hacked. This essay considers a world where AIs can be hackers. This is a generalization of specification gaming, where vulnerabilities and exploits of our social, economic, and political systems are discovered and exploited at computer speeds and scale.

Table of Contents



Introduction

Artificial intelligence—AI—is an information technology. It consists of software. It runs on computers. And it is already deeply embedded into our social fabric, both in ways we understand and in ways we don't. It will hack our society to a degree and effect unlike anything that's come before. I mean this in two very different ways. One, AI systems will be used to hack us. And two, AI systems will themselves become hackers: finding vulnerabilities in all sorts of social, economic, and political systems, and then exploiting them at an unprecedented speed, scale, and scope. It's not just a difference in degree; it's a difference in kind. We risk a future of AI systems hacking other AI systems, with humans being little more than collateral damage.

This isn't hyperbole. Okay, maybe it's a bit of hyperbole, but none of this requires far-future science-fiction technology. I'm not postulating any "singularity," where the AI-learning feedback loop becomes so fast that it outstrips human understanding. I'm not assuming intelligent androids like Data (Star Trek), R2-D2 (Star Wars), or Marvin the Paranoid Android (The Hitchhiker's Guide to the Galaxy). My scenarios don't require evil intent on the part of anyone. We don't need malicious AI systems like Skynet (Terminator) or the Agents (Matrix). Some of the hacks I will discuss don't even require major research breakthroughs. They'll improve as AI techniques get more sophisticated, but we can see hints of them in operation today. This hacking will come naturally, as AIs become more advanced at learning, understanding, and problem-solving.

In this essay, I will talk about the implications of AI hackers. First, I will generalize "hacking" to include economic, social, and political systems—and also our brains. Next, I will describe how AI systems will be used to hack us. Then, I will explain how AIs will hack the economic, social, and political systems that comprise society. Finally, I will discuss the implications of a world of AI hackers, and point towards possible defenses. It's not all as bleak as it might sound.

Hacks and Hacking

First, a definition:

> *Def*: **Hack** */hak/ (noun) -*
>
> *1. A clever, unintended exploitation of a system which: a) subverts the rules or norms of that system, b) at the expense of some other part of that system.*
>
> *2. Something that a system allows, but that is unintended and unanticipated by its designers.*[1]

Notice the details. Hacking is not cheating. It's following the rules, but subverting their intent. It's unintended. It's an exploitation. It's "gaming the system." Caper movies are filled with hacks. MacGyver was a hacker. Hacks are clever, but not the same as innovations. And, yes, it's a subjective definition.[2]

Systems tend to be optimized for specific outcomes. Hacking is the pursuit of another outcome, often at the expense of the original optimization Systems tend be rigid. Systems limit what we can do and invariably, some of us want to do something else. So we hack. Not everyone, of course. Everyone isn't a hacker. But enough of us are.

Hacking is normally thought of something you can do to computers. But hacks can be perpetrated on any system of rules—including the tax code.

The tax code isn't software. It doesn't run on a computer. But you can still think of it as "code" in the computer sense of the term. It's a series of algorithms that takes an input—financial information for the year—and produces an output: the amount of tax owed. It's deterministic, or at least it's supposed to be.

All computer software contains defects, commonly called bugs. These are mistakes: mistakes in specification, mistakes in programming, mistakes

[1] The late hacker Jude Mihon (St. Jude) liked this definition: "Hacking is the clever circumvention of imposed limits, whether those limits are imposed by your government, your own personality, or the laws of Physics." Jude Mihon (1996), Hackers Conference, Santa Rosa, CA.

[2] This is all from a book I am currently writing, probably to be published in 2022.

that occur somewhere in the process of creating the software. It might seem crazy, but modern software applications generally have hundreds if not thousands of bugs. These bugs are in all the software that you're currently using: on your computer, on your phone, in whatever "Internet of Things" devices you have around. That all of this software works perfectly well most the time speaks to how obscure and inconsequential these bugs tend to be. You're unlikely to encounter them in normal operations, but they're there.

Some of those bugs introduce security holes. By this I mean something very specific: bugs that an attacker can deliberately trigger to achieve some condition that the attacker can take advantage of. In computer-security language, we call these bugs "vulnerabilities."

Exploiting a vulnerability is how the Chinese military broke into Equifax in March 2017. A vulnerability in the Apache Struts software package allowed hackers to break into a consumer complaint web portal. From there, they were able to move to other parts of the network. They found usernames and passwords that allowed them to access still other parts of the network, and eventually to download personal information about 147 million people over the course of four months.[3]

This is an example of a hack. It's a way to exploit the system in a way that is both unanticipated and unintended by the system's designers—something that advantages the hacker in some way at the expense of the users the system is supposed to serve.

The tax code also has bugs. They might be mistakes in how the tax laws were written: errors in the actual words that Congress voted on and the president signed into law. They might be mistakes in how the tax code is interpreted. They might be oversights in how parts of the law were conceived, or unintended omissions of some sort or another. They might arise from unforeseen interactions between different parts of the tax code.

[3] Federal Trade Commission (22 Jul 2019), "Equifax data breach settlement: What you should know," https://www.consumer.ftc.gov/blog/2019/07/equifax-data-breach-settlement-what-you-should-know.

A recent example comes from the 2017 Tax Cuts and Jobs Act. That law was drafted in haste and in secret, and passed without any time for review by legislators—or even proofreading. Parts of it were handwritten, and it's pretty much inconceivable that anyone who voted either for or against it knew precisely what was in it. The text contained a typo that accidentally categorized military death benefits as earned income. The practical effect of that mistake was that surviving family members were hit with surprise tax bills of $10,000 or more.[4] That's a bug.

It's not a vulnerability, though, because no one can take advantage of it to reduce their tax bill. But some bugs in the tax code are also vulnerabilities. For example, there's a corporate tax trick called the "Double Irish with a Dutch Sandwich." It's a vulnerability that arises from the interactions between tax laws in multiple countries. Basically, it involves using a combination of Irish and Dutch subsidiary companies to shift profits to low- or no-tax jurisdictions. Tech companies are particularly well suited to exploit this vulnerability; they can assign intellectual property rights to subsidiary companies abroad, who then transfer cash assets to tax havens.[5] That's how companies like Google and Apple have avoided paying their fair share of US taxes despite being US companies. It's definitely an unintended and unanticipated use of the tax laws in three countries. And it can be very profitable for the hackers—in this case, big tech companies avoiding US taxes—at the expense of everyone else. Estimates are that US companies avoided paying nearly $200 billion in US taxes in 2017 alone.[6]

Some vulnerabilities are deliberately created. Lobbyists are constantly trying to insert this or that provision into the tax code to benefit their clients. That same 2017 US tax law that gave rise to unconscionable tax bills to grieving families included a special tax break for oil and gas investment partnerships, a special exemption that ensures that less than 1 in 1,000 estates will have to

[4] Naomi Jagoda (14 Nov 2019), "Lawmakers under pressure to pass benefits fix for military families," Hill, https://thehill.com/policy/national-security/470393-lawmakers-under-pressure-to-pass-benefits-fix-for-military-families.

[5] *New York Times* (28 Apr 2012), "Double Irish with a Dutch Sandwich" (infographic). https://archive.nytimes.com/www.nytimes.com/interactive/2012/04/28/business/Double-Irish-With-A-Dutch-Sandwich.html.

[6] Niall McCarthy (23 Mar 2017), "Tax avoidance costs the U.S. nearly $200 billion every year" (infographic), *Forbes*, https://www.forbes.com/sites/niallmccarthy/2017/03/23/tax-avoidance-costs-the-u-s-nearly-200-billion-every-year-infographic.

pay estate tax, and language specifically expanding a pass-through loophole that industry uses to incorporate offshore and avoid US taxes.[7]

Sometimes these vulnerabilities are slipped into law with the knowledge of the legislator who is sponsoring the amendment, and sometimes they're not aware of it. This deliberate insertion is also analogous to something we worry about in software: programmers deliberately adding backdoors into systems for their own purposes. That's not hacking the tax code, or the computer code. It's hacking the processes that create them: the legislative process that creates tax law, or the software development process that creates computer programs.

During the past few years, there has been considerable press given to the possibility that Chinese companies like Huawei and ZTE have added backdoors to their 5G routing equipment at the request—or possibly demand—of the Chinese government. It's certainly possible, and those vulnerabilities would lie dormant in the system until they're used by someone who knows about them.

In the tax world, bugs and vulnerabilities are called tax loopholes. In the tax world, taking advantage of these vulnerabilities is called tax avoidance. And there are thousands of what we in the computer security world would call "black-hat researchers," who examine every line of the tax code looking for exploitable vulnerabilities. They're called tax attorneys and tax accountants.

Modern software is incredibly complex. Microsoft Windows 10, the latest version of that operating system, has about 50 million lines of code.[8] More complexity means more bugs, which means more vulnerabilities. The US tax code is also complex. It consists of the tax laws passed by Congress, administrative rulings, and judicial rules. Credible estimates of the size of it all are hard to come by; even experts often have no idea. The tax laws themselves

[7] Alexandra Thornton (1 Mar 2018), "Broken promises: More special interest breaks and loopholes under the new tax law," https://www.americanprogress.org/issues/economy/reports/2018/03/01/447401/broken-promises-special-interest-breaks-loopholes-new-tax-law/.

[8] Microsoft (12 Jan 2020), "Windows 10 lines of code." https://answers.microsoft.com/en-us/windows/forum/all/windows-10-lines-of-code/a8f77f5c-0661-4895-9c77-2efd42429409.

are about 2,600 pages.[9] IRS regulations and tax rulings increase that to about 70,000 pages. It's hard to compare lines of text to lines of computer code, but both are extremely complex. And in both cases, much of that complexity is related to how different parts of the codes interact with each other.

We know how to fix vulnerabilities in computer code. We can employ a variety of tools to detect and fix them before the code is finished. After the code is out in the world, researchers of various kinds discover them and—most important of all—we want the vendors to quickly patch them once they become known.

We can sometimes employ these same methods with the tax code. The 2017 tax law capped income tax deductions for property taxes. This provision didn't come into force in 2018, so someone came up with the clever hack to prepay 2018 property taxes in 2017. Just before the end of the year, the IRS ruled about when that was legal and when it wasn't.[10] Short answer: most of the time, it wasn't.

It's often not this easy. Some hacks are written into the law, or can't be ruled away. Passing any tax legislation is a big deal, especially in the US, where the issue is so partisan and contentious. (It's been almost four years, and that earned income tax bug for military families still hasn't been fixed. And that's an easy one; everyone acknowledges it was a mistake.) It can be hard to figure out who is supposed to patch the tax code: is the legislature, the courts, the tax authorities? And then it can take years. We simply don't have the ability to patch tax code with anywhere near the same agility that we have to patch software.

[9] Dylan Matthews (29 Mar 2017), "The myth of the 70,000-page federal tax code," *Vox.* https://www.vox.com/policy-and-politics/2017/3/29/15109214/tax-code-page-count-complexity-simplification-reform-ways-means.

[10] IRS (27 Dec 2017), "Prepaid real property taxes may be deductible in 2017 if assessed and paid in 2017," IRS Advisory, https://www.irs.gov/newsroom/irs-advisory-prepaid-real-property-taxes-may-be-deductible-in-2017-if-assessed-and-paid-in-2017.

The Ubiquity of Hacking

Everything is a system, every system can be hacked, and humans are natural hackers.

Airline frequent-flier programs are hacked. Card counting in blackjack is a hack. Sports are hacked all the time. Someone first figured out that a curved hockey stick blade allowed for faster and more accurate shots but also a more dangerous game, something the rules didn't talk about because no one had thought of it before. Formula One racing is full of hacks, as teams figure out ways to modify car designs that are not specifically prohibited by the rulebook but nonetheless subvert its intent.

The history of finance is a history of hacks. Again and again, financial institutions and traders look for loopholes in the rules—things that are not expressly prohibited, but are unintended subversions of the underlying systems—that give them an advantage. Uber, Airbnb, and other gig-economy companies hack government regulations. The filibuster is an ancient hack, first invented in ancient Rome. So are hidden provisions in legislation. Gerrymandering is a hack of the political process.

And finally, people can be hacked. Our brain is a system, evolved over millions of years to keep us alive and—more importantly—to keep us reproducing. It's been optimized through continuous interaction with the environment. But it's been optimized for humans who live in small family groups in the East African highlands in 100,000 BCE. It's not as well suited for twenty-first-century New York, or Tokyo, or Delhi. And because it encompasses many cognitive shortcuts—it evolves, but not on any scale that matters here—it can be manipulated.

Cognitive hacking is powerful. Many of the robust social systems our society relies on— democracy, market economics, and so on—depend on humans making appropriate decisions. This process can be hacked in many different ways. Social media hacks our attention. Personalized to our attitudes and behavior, modern advertising is a hack of our systems of persuasion. Disinformation hacks our common understanding of reality. Terrorism hacks our cognitive systems of fear and risk assessment by convincing people that it is a bigger threat than it actually is.[11] It's horrifying, vivid, spectacular, random—in that anyone could be its next victim—and malicious. Those are the very things that cause us to exaggerate the risk and overreact.[12] Social engineering, the conventional hacker tactic of convincing someone to divulge their login credentials or otherwise do something beneficial to the hacker, is much more a hack of trust and authority than a hack of any computer system.

What's new are computers. Computers are systems, and are hacked directly. But what's more interesting is the computerization of more traditional systems. Finance, taxation, regulatory compliance, elections—all these and more have been computerized. And when something is computerized, the way it can be hacked changes. Computerization accelerates hacking across three dimensions: speed, scale, and scope.

Computer speed modifies the nature of hacks. Take a simple concept—like stock trading—and automate it. It becomes something different. It may be doing the same thing it always did, but it's doing it at superhuman speed. An example is high-frequency trading, something unintended and unanticipated by those who designed early markets.

[11] Bruce Schneier (24 Aug 2006), "What the terrorists want," Schneier on Security, https://www.schneier.com/blog/archives/2006/08/what_the_terror.html.

[12] Robert L. Leahy (15 Feb 2018), "How to Think About Terrorism," *Psychology Today,* https://www.psychologytoday.com/us/blog/anxiety-files/201802/how-think-about-terrorism.

Scale, too. Computerization allows systems to grow much larger than they could otherwise, which changes the scale of hacking. The very notion of "too big to fail" is a hack, allowing companies to use society as a last-ditch insurance policy against their bad decision making.

Finally, scope. Computers are everywhere, affecting every aspect of our lives. This means that new concepts in computer hacking are potentially applicable everywhere, with varying results.

Not all systems are equally hackable. Complex systems with many rules are particularly vulnerable, simply because there are more possibilities for unanticipated and unintended consequences. This is certainly true for computer systems—I've written in the past that complexity is the worst enemy of security[13]—and it's also true for systems like the tax code, the financial system, and AIs. Systems constrained by more flexible social norms and not by rigidly defined rules are more vulnerable to hacking, because they leave themselves more open to interpretation and therefore have more loopholes.

Even so, vulnerabilities will always remain, and hacks will always be possible. In 1930, the mathematician Kurt Gödel proved that all mathematical systems are either incomplete or inconsistent. I believe this is true more generally. Systems will always have ambiguities or inconsistencies, and they will always be exploitable. And there will always be people who want to exploit them.

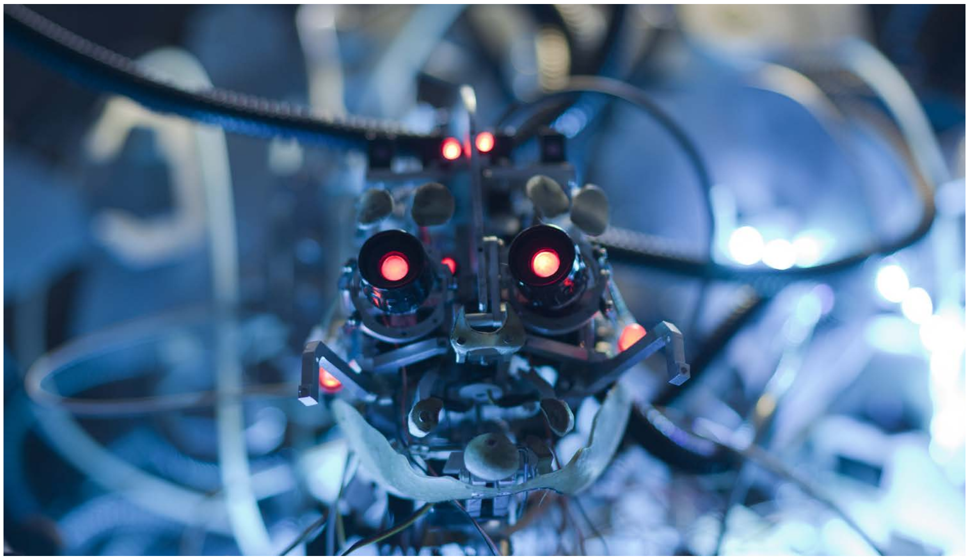
[13] Bruce Schneier (19 Nov 1999), "A plea for simplicity," Schneier on Security, https://www.schneier.com/essays/archives/1999/11/a_plea_for_simplicit.html.

AIs Hacking Us

In 2016, The Georgia Institute of Technology published a research study on human trust in robots.[14] The study employed a non-anthropomorphic robot that assisted with navigation through a building, providing directions such as "This way to the exit." First, participants interacted with the robot in a normal setting to experience its performance, which was deliberately poor. Then, they had to decide whether or not to follow the robot's commands in a simulated emergency. In the latter situation, all twenty-six participants obeyed the robot, despite having observed just moments before that the robot had lousy navigational skills. The degree of trust they placed in this machine was striking: when the robot pointed to a dark room with no clear exit, the majority of people obeyed it, rather than safely exiting by the door through which they had entered. The researchers ran similar experiments with other robots that seemed to malfunction. Again, subjects followed these robots in an emergency setting, apparently abandoning their common sense. It seems that robots can naturally hack our trust.

[14] Paul Robinette et al. (Mar 2016), "Overtrust of robots in emergency evacuation scenarios," *2016 ACM/IEEE International Conference on Human-Robot Interaction*. https://www.cc.gatech.edu/~alanwags/pubs/Robinette-HRI-2016.pdf.

Artificial Intelligence and Robotics



We could spend pages defining AI. In 1968, AI pioneer Marvin Minsky defined it as “the science of making machines do things that would require intelligence if done by men.”[15] The US Department of Defense uses: “the ability of machines to perform tasks that normally require human intelligence.”[16] The 1950 version of the Turing test—called the “imitation game” in the original discussion—focused on a computer program that humans couldn’t distinguish from an actual human.[17] For our purposes, AI is an umbrella term encompassing a broad array of decision-making technologies that simulate human thinking.

One differentiation I need to make is between specialized—sometimes called “narrow”—AI and general AI. General AI is what you see in the movies. It’s AI that can sense, think, and act in a very general and human way. If it’s smarter than humans, it’s called “artificial superintelligence.” Combine it with robotics and you have an android, one that may look more or less like a human. The movie robots that try to destroy humanity are all general AI.

[15] Marvin Minsky (ed.) (1968), *Semantic Information Processing,* The MIT Press.

[16] Air Force Research Lab (18 Jun 2020), “Artificial intelligence.” https://afresearchlab.com/technology/artificial-intelligence.

[17] Graham Oppy and David Dowe (Fall 2020), “The Turing Test,” *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/turing-test.

There's been a lot of practical research going into how to create general AI, and a lot of theoretical research about how to design these systems so they don't do things we don't want them to, like destroy humanity. And while this is fascinating work, encompassing fields from computer science to sociology to philosophy, its practical applications are probably decades away. I want to focus instead on specialized AI, because that's what's practical now.

Specialized AI is designed for a specific task. An example is the system that controls a self-driving car. It knows how to steer the vehicle, how to follow traffic laws, how to avoid getting into accidents, and what to do when something unexpected happens—like a child's ball suddenly bouncing into the road. Specialized AI knows a lot and can make decisions based on that knowledge, but only in this limited domain.

One common joke among AI researchers is that as soon as something works, it's no longer AI; it's just software. That might make AI research somewhat depressing, since by definition the only things that count are failures, but there's some truth to it. AI is inherently a mystifying science-fiction term. Once it becomes reality, it's no longer mystifying. We used to assume that reading chest X-rays required a radiologist: that is, an intelligent human with appropriate training. Now we realize that it's a rote task that can also be performed by a computer.

What's really going on is that there is a continuum of decision-making technologies and systems, ranging from a simple electromechanical thermostat that operates a furnace in response to changing temperatures to a science-fictional android. What makes something AI often depends on the complexity of the tasks performed and the complexity of the environment in which those tasks are performed. The thermostat performs a very simple task that only has to take into account a very simple aspect of the environment. It doesn't even need to involve a computer. A modern digital thermostat might be able to sense who is in the room and make predictions about future heat needs based on both usage and weather forecast, as well as citywide power consumption and second-by-second energy costs. A futuristic thermostat might act like a thoughtful and caring butler, whatever that would mean in the context of adjusting the ambient temperature.

I would rather avoid these definitional debates, because they largely don't matter for our purposes. In addition to decision-making, the relevant qualities of the systems I'll be discussing are autonomy, automation, and physical agency. A thermostat has limited automation and physical agency, and no autonomy. A system that predicts criminal recidivism has no physical agency; it just makes recommendations to a judge. A driverless car has some of all three. R2-D2 has a lot of all three, although for some reason its designers left out English speech synthesis.

Robotics also has a popular mythology and a less-flashy reality. Like AI, there are many different definitions of the term. I like robot ethicist Kate Darling's definition: "physically embodied objects that can sense, think, and act on their environments through physical motion."[18] In movies and television, that's often artificial people: androids. Again, I prefer to focus on technologies that are more prosaic and near term. For our purposes, robotics is autonomy, automation, and physical agency dialed way up. It's "cyber-physical autonomy": AI technology inside objects that can interact with the world in a direct, physical manner.

[18] Kate Darling (2021), *The New Breed: What Our History with Animals Reveals about Our Future with Robots,* Henry Holt & Co.

Human-Like AIs

People have long ascribed human-like qualities to computer programs. In the 1960s, programmer Joseph Weizenbaum created a primitive therapist-mimicking conversational program called ELIZA. He was amazed that people would confide deeply personal secrets to what they knew was a dumb computer program. Weizenbaum's secretary would even ask him to leave the room, so that she could talk to ELIZA in private.[19] Today, people are polite to voice assistants like Alexa and Siri.[20] Siri even complains when you're mean to it: "That's not very nice," it says—because it's programmed to, of course.

Numerous experiments bear similar results. Research subjects would rate a computer's performance less critically if they gave the rating on the computer they were criticizing, indicating that they didn't want to hurt its feelings.[21] In another experiment, if a computer told a research subject some obviously fictional piece of "personal information," the subject was likely to reciprocate by sharing actual personal information.[22] The power of reciprocation is something that psychologists study. It's a hack that people use, too.

It's not just that we'll treat AIs as people. They'll also act like people in ways that will be deliberately designed to fool us. They'll employ cognitive hacks.

During the 2016 US election, about a fifth of all political tweets were posted by bots.[23] For the UK Brexit vote of the same year, it was a third.[24] An Oxford Internet Institute report from 2019 found evidence of bots

[19] Joseph Weizenbaum (Jan 1966), "ELIZA: A computer program for the study of natural language communication between man and machine," *Communications of the ACM,* https://web.stanford.edu/class/linguist238/p36-weizenabaum.pdf.

[20] James Vincent (22 Nov 2019), "Women are more likely than men to say 'please' to their smart speaker," Verge, https://www.theverge.com/2019/11/22/20977442/ai-politeness-smart-speaker-alexa-siri-please-thank-you-pew-gender-sur.

[21] Clifford Nass, Youngme Moon, and Paul Carney (31 Jul 2006), "Are people polite to computers? Responses to computer-based interviewing systems," *Journal of Applied Social Psychology,* https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1559-1816.1999.tb00142.x.

[22] Youngme Moon (Mar 2000), "Intimate exchanges: Using computers to elicit self-disclosure from consumers," *Journal of Consumer Research*, https://www.jstor.org/stable/10.1086/209566?seq=1.

[23] Alessandro Bessi and Emilio Ferrara (Nov 2016), "Social bots distort the 2016 U.S. Presidential election online discussion," First Monday, https://firstmonday.org/article/view/7090/5653.

[24] Carole Cadwalladr (26 Feb 2017), "Robert Mercer: the big data billionaire waging war on mainstream media," *Guardian*, https://www.theguardian.com/politics/2017/feb/26/robert-mercer-breitbart-war-on-media-steve-bannon-donald-trump-nigel-farage.

being used to spread propaganda in fifty countries.[25] These tended to be simple programs mindlessly repeating slogans. For example, a quarter million pro-Saudi "We all have trust in [crown prince] Mohammed bin Salman" tweets were posted following the 2018 murder of Jamal Khashoggi.[26]

In 2017, the Federal Communications Commission had an online public-comment period for its plans to repeal net neutrality. A staggering 22 million comments were received. Many of them—maybe half—were submitted using stolen identities.[27] These fake comments were also crude; 1.3 million were generated from the same template, with some words altered to make them appear unique.[28] They didn't stand up to even cursory scrutiny.

Efforts like these will only get more sophisticated. For years, AI programs have been writing news stories about sports and finance for real news organizations like the Associated Press.[29] The constrained nature of those topics made them easier for an AI. They're now starting to write more general stories. Research projects like Open AI's GPT-3 are expanding the capabilities of what AI-driven text generation can do.[30] These systems can be fed actual facts and write true stories, but they can just as easily be fed untruths and write fake news.

It doesn't take much imagination to see how AI will degrade political discourse. Already, AI-driven personas can write personalized letters to newspapers and elected officials, leave intelligible comments on news sites

[25] Samantha Bradshaw and Philip N. Howard (2019), "The global disinformation order: 2019 Global inventory of organised social media manipulation," Computational Propaganda Research Project, https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf.

[26] Chris Bell and Alistair Coleman (18 Oct 2018), "Khashoggi: Bots feed Saudi support after disappearance," *BBC News*, https://www.bbc.com/news/blogs-trending-45901584.

[27] Jacob Kastrenakes (31 Aug 2017), "The net neutrality comment period was a complete mess," *Verge*, https://www.theverge.com/2017/8/31/16228220/net-neutrality-comments-22-million-reply-recordlity-comments-are-fake.

[28] Jeff Kao (22 Nov 2017), "More than a million pro-repeal net neutrality comments were likely faked," Hackernoon, https://hackernoon.com/more-than-a-million-pro-repeal-net-neutrality-comments-were-likely-faked-e9f0e3ed36a6.

[29] Ross Miller (29 Jan 2015), "AP's 'robot' journalists are writing their own stories now," *Verge,* https://www.theverge.com/2015/1/29/7939067/ap-journalism-automation-robots-financial-reporting. Bernard Marr (29 Mar 2019), "Artificial Intelligence Can Now Write Amazing Content—What Does That Mean For Humans?" *Forbes*, https://www.forbes.com/sites/bernardmarr/2019/03/29/artificial-intelligence-can-now-write-amazing-content-what-does-that-mean-for-humans/?sh=868a50450ab0.

[30] Tom Simonite (22 Jul 2020), "Did a person write this headline, or a machine?" *Wired,* https://www.wired.com/story/ai-text-generator-gpt-3-learning-language-fitfully/.

and message boards, and intelligently debate politics on social media.[31] These systems will only get better: more sophisticated, more articulate, more personal, and harder to distinguish from actual human beings.

In a recent experiment, researchers used a text-generation program to submit 1,000 comments in response to a government request for public input on a Medicaid issue.[32] They all sounded unique, like real people advocating a specific policy position. They fooled the Medicaid.gov administrators, who accepted them as genuine concerns from actual human beings. The researchers subsequently identified the comments and asked for them to be removed, so that no actual policy debate would be unfairly biased. Others won't be so ethical.

These techniques are already being used. An online propaganda campaign used AI-generated headshots to create fake journalists.[33] China experimented with AI-generated text messages designed to influence the 2020 Taiwanese election.[34] Deepfake technology—AI techniques to create real videos of fake events, often with actual people saying things they didn't actually say—are being used politically.[35]

One example of how this will unfold is in "persona bots." These are AIs posing as individuals on social media and in other digital groups. They have histories, personalities, and communications styles. They don't constantly spew propaganda. They hang out in various interest groups: gardening, knitting, model railroading, whatever. They act as normal members of those communities, posting and commenting and discussing. Systems like GPT-3 will make it easy for those AIs to mine previous conversations and related Internet content and appear knowledgeable.

[31] Will Heaven (21 Jan 2020), "IBM's debating AI just for a lot closer to being a useful tool," *MIT Technology Review,* https://www.technologyreview.com/2020/01/21/276156/ibms-debating-ai-just-got-a-lot-closer-to-being-a-useful-tool/.

[32] Max Weiss (17 Dec 2019), "Deepfake bot submissions to federal public comment websites cannot be distinguished from human submissions," *Technology Science*, https://techscience.org/a/2019121801/.

[33] Adam Rawnsley (6 Jul 2020), "Right-wing media outlets duped by a Middle East propaganda campaign," *Daily Beast*, https://www.thedailybeast.com/right-wing-media-outlets-duped-by-a-middle-east-propaganda-campaign.

[34] Philip Sherwell (5 Jan 2020), "China uses Taiwan for AI target practice to influence elections," *Australian.* https://www.theaustralian.com.au/world/the-times/china-uses-taiwan-for-ai-target-practice-to-influence-elections/news-story/57499d2650d4d359a3857688d416d1e5.

[35] Ian Sample (13 Jan 2020), "What are deepfakes—and how can you spot them?" *Guardian*, https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them.

Then, once in a while, the AI posts something relevant to a political issue. Maybe it's an article about an Alaska healthcare worker having an allergic reaction to the COVID-19 vaccine, with a worried commentary. Or maybe it's something about a recent election, or racial justice, or anything that's polarizing. One persona bot can't move public opinion, but what if there were thousands of them? Millions?

This has been called "computational propaganda,"[36] and will change the way we view communication. AI will make the future supply of disinformation infinite.[37] Persona bots will break the "notice-and-comment" rulemaking process, by flooding government agencies with fake comments. They may also break community discourse.

These systems will affect us at the personal level as well. Earlier I mentioned social engineering. One common hacker tactic is phishing emails that purport to be from someone they're not, intended to convince the recipient to do something she shouldn't. Most phishing emails are generic and easily tagged as spam. The more effective phishing emails—the ones that result in people and companies losing lots of money—are personalized. For example, an email that impersonates the CEO to someone in the finance department, asking for a particular wire transfer, can be particularly effective.[38] Voice can be even more effective.[39] The laborious task of customizing phishing attacks could be automated by AI techniques, allowing marketers to send out personalized advertisements, and phishing scammers to send out individually targeted emails.

It's not that being persuaded by an AI is fundamentally more damaging than being persuaded by another human, it's that AIs will be able to do it at computer speed and scale. Today's cognitive hacks are crude: a fake newspaper article designed to fool only the most gullible, or a persuasive nudge designed to affect only the most desperate. AI has the potential for

[36] Matt Chessen (Sep 2017), "The MADCOM Future," Atlantic Council, https://www.atlanticcouncil.org/wp-content/uploads/2017/09/The_MADCOM_Future_RW_0926.pdf.

[37] Renée DiResta (20 Sep 2020), "The supply of disinformation will soon be infinite," *Atlantic*, https://www.theatlantic.com/ideas/archive/2020/09/future-propaganda-will-be-computer-generated/616400/.

[38] Seena Gressin (16 May 2020), "CEO imposter scams: Is the boss for real?" Federal Trade Commission, https://www.ftc.gov/news-events/blogs/business-blog/2016/05/ceo-imposter-scams-boss-real.

[39] Keumars Afifi-Sebet (2 Sep 2019), "Fraudsters use AI voice manipulation to steal £200,000," IT Pro, https://www.itpro.co.uk/social-engineering/34308/fraudsters-use-ai-voice-manipulation-to-steal-200000.

every one of those hacks to be microtargeted: personalized, optimized, and individually delivered.[40] Old-style con games are individually crafted person-to-person cognitive hacks. Advertising messages are bulk broadcast cognitive hacks. AI techniques have the potential to blend aspects of both of those techniques.

Robots Hacking Us

The addition of robotics will only make these hacks more effective, something Kate Darling chronicled in her book *The New Breed*.[41] We humans have developed some pretty effective cognitive shortcuts to recognize other people. We see faces everywhere; two dots over a horizontal line looks like a face without any trouble. This is why even minimalist illustrations are so effective. If something has a face, then it's a creature of some sort: with intentions, feelings, and everything else that comes with real-world faces. If that something speaks or, even better, converses, then we believe it has intentions, desires, and agency.

Robots are no exception. Many people have social relationships with their robot vacuums, even complaining when the company would offer

[40] Brett Frischmann and Deven Desai (29 Nov 2016), "The Promise and Peril of Personalization," Center for Internet and Society, Stanford Law School.

[41] Kate Darling (2021), *The New Breed: What Our History with Animals Reveals About our Future with Robots,* Henry Holt & Co.

to replace rather than repair “their” Roomba.[42] A US Army–developed anti-landmine robot ran into problems when a colonel refused to allow the insect-shaped device to continue to harm itself by stepping on mines.[43] A Harvard robot could convince students to let it in dorms by pretending to be a food-delivery robot.[44] And Boxie, a childlike talking robot at MIT, could persuade people to answer personal questions just by asking nicely.[45]

The human nurturing instinct isn’t solely genetically focused. We can experience nurturing feelings towards adopted children, and we can feel the same instincts arise when we interact with the children of friends or even strangers–or puppies. At least some of our response is inspired by the appearance and behavior of children. Children have large heads in proportion to their bodies, and large eyes in proportion to their heads. They talk with a higher-pitched voice than adults. And we respond to all of this.

Artists have taken advantage of this for generations to make their creations appear more sympathetic. Children’s dolls are designed this way. Cartoon characters are drawn this way, as far back as Betty Boop in the 1930s and Bambi in 1942. In the 2019 live-action movie *Alita: Battle Angel,* the main character had her eyes computer-enhanced to be larger.[46]

Anthropomorphic robots are an emotionally persuasive technology, and AI will only amplify their attractiveness. As AI mimics humans, or even animals, it will hijack all the mechanisms that humans use to hack each other. As psychologist Sherry Turkle wrote in 2010: “When robots make eye contact, recognize faces, mirror human gestures, they push our Darwinian

[42] Ja-Young SungLan GuoRebecca E. GrinterHenrik I. Christensen (2007), “‘My Roomba is Rambo’: Intimate home appliances,” *UbiComp 2007: Ubiquitous Computing*, https://link.springer.com/chapter/10.1007/978-3-540-74853-3_9.

[43] Joel Garreau (6 May 2007), “Bots on the ground,” *Washington Post*, https://www.washingtonpost.com/wp-dyn/content/article/2007/05/05/AR2007050501009.html.

[44] Serena Booth et al. (Mar 2017,), “Piggybacking robots: Human-robot overtrust in university dormitory security,” 2017 ACM/IEEE International Conference on Human-Robot Interaction, https://dl.acm.org/doi/10.1145/2909824.3020211.

[45] Alexander Reben and Joseph Paradison (2011), “A mobile interactive robot for gathering structured social video,” MIT Libraries, https://resenv.media.mit.edu/pubs/papers/2011-MM11-REBEN.pdf.

[46] Stuart Heritage (1 Feb 2019), “The eyes! The eyes! Why does Alita: Battle Angel look so creepy?” *Guardian*, https://www.theguardian.com/film/2019/feb/01/the-eyes-why-does-alita-battle-angel-look-so-creepy.

buttons, exhibiting the kind of behavior people associate with sentience, intentions, and emotions."[47] That is, they hack our brains.

We might intuitively know that it's just a plastic green dinosaur. But a large face paired with a small body makes us think of it as a child. Suddenly we're thinking of it as a creature with feelings, and will protect it from harm.[48] And while that may be benign, what happens when that robot looks at its human owners with its big, sad eyes and asks them to buy it a software upgrade?[49]

Because we humans are prone to making a category error and treating robots as living creatures with feelings and intentions, we are prone to being manipulated by them. Robots could persuade us to do things we might not do otherwise. They could scare us into not doing things we might otherwise do. In one experiment, a robot was able to exert "peer pressure" on subjects, encouraging them to take more risks.[50] How soon before a sex robot suggests in-app purchases in the heat of the moment?[51]

AIs will get better at all of this. Already they are trying to detect emotions by analyzing our writings,[52] reading our facial expressions,[53] or monitoring our breathing and heartrate.[54] They get it wrong a lot of the time, but it is likely that they will improve. And, like so many areas of AI, they will

[47] Sherry Turkle (2010), "In good company," in Yorick Wilks, ed., *Close Engagements with Artificial Companions*, John Benjamin Publishing Company, http://web.mit.edu/people/sturkle/pdfsforstwebpage/Turkle_In%20Good%20Company.pdf.

[48] Kate Darling (2021), *The New Breed: What Our History with Animals Reveals about Our Future with Robots,* Henry Holt & Co.

[49] Woodrow Hartzog (4 May 2015), "Unfair and deceptive robots," *Maryland Law Review*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2602452.

[50] Yaniv Hanoch et al (18 Nov 2020), "The robot made me do it: Human–robot interaction and risk-taking behavior," Online ahead of print, https://www.liebertpub.com/doi/10.1089/cyber.2020.0148.

[51] Kate Darling (Oct 2017), "'Who's Johnny?' Anthropomorphic framing in human-robot interaction, integration, and policy," in Patrick Lin, Keith Abney, and Ryan Jenkins, eds., *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, Oxford Scholarship Online, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2588669.

[52] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah (28 May 2020), "Text-based emotion detection: Advances, challenges, and opportunities," *Engineering Reports* 2 no. 7, https://onlinelibrary.wiley.com/doi/full/10.1002/eng2.12189.

[53] Tim Lewis (17 Aug 2019), "AI can read your emotions. Should it?" *Guardian*, https://www.theguardian.com/technology/2019/aug/17/emotion-ai-artificial-intelligence-mood-realeyes-amazon-facebook-emotient.

[54] Ashan Noor Khan et al. (3 Feb 2021), "Deep learning framework for subject-independent emotion detection using wireless signals," PLoS ONE, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0242946.

eventually surpass people in capability. This will allow them to more precisely manipulate us.

As AIs and autonomous robots take on more real-world tasks, human trust in autonomous systems will be hacked with dangerous and costly results. But never forget that there are human hackers controlling the AI hackers. All of the systems will be designed and paid for by humans who want them to manipulate us in a particular way for a particular purpose. I'll talk more about this later.

When AIs Become Hackers

Hacker "Capture the Flag" is basically the outdoor game played on computer networks. Teams of hackers defend their own computers while attacking other teams'. It's a controlled setting for what computer hackers do in real life: finding and fixing vulnerabilities in their own systems, and exploiting them in others'.

The competition has been a mainstay at hacker gatherings since the mid-1990s. These days, dozens of teams from around the world compete in weekend-long marathon events held all over the world. People train for months. Winning is a big deal. If you're into this sort of thing, it's pretty much the most fun you can possibly have on the Internet without committing multiple felonies.

In 2016, DARPA ran a similarly styled event for AI.[55] One hundred teams entered their systems into the Cyber Grand Challenge. After completion of qualifying rounds, seven finalists competed at the DEFCON hacker convention in Las Vegas. The competition occurred in a specially designed test environment filled with custom software that had never been analyzed or tested. The AIs were given ten hours to find vulnerabilities to exploit against the other AIs in the competition, and to patch themselves against

[55] Jia Song and Jim Alves-Foss (Nov 2015), "The DARPA cyber grand challenge: A competitor's perspective," *IEEE Security and Privacy Magazine* 13 (6), https://www.researchgate.net/publication/286490027_The_DARPA_cyber_grand_challenge_A_competitor%27s_perspective.

exploitation. A system called Mayhem, created by a team of Pittsburgh computer-security researchers, won. The researchers have since commercialized the technology, which is now busily defending networks for customers like the Department of Defense.[56]

There was a human-team capture-the-flag event at DEFCON that same year. Mayhem was invited to participate as the only non-human team, and came in last. You can easily imagine how this mixed competition would unfold in the future. AI entrants will improve every year, because the core technologies are all improving. The human teams will largely stay the same, because humans remain humans even as our tools improve. Eventually the AIs will routinely beat the humans. My guess is that it'll take less than a decade. It will be years before we have entirely autonomous AI cyberattack capabilities, but AI technologies are already transforming the nature of cyberattack.[57]

One area that seems particularly fruitful for AI systems is vulnerability finding. Going through software code line by line is exactly the sort of tedious problem at which AIs excel, if they can only be taught how to recognize a vulnerability.[58] Many domain-specific challenges will need to be addressed, of course, but there is a healthy amount of academic literature on the topic—and research is continuing.[59] There's every reason to expect AI systems will improve over time, and some reason to expect them to eventually become very good at it.

The implications extend far beyond computer networks. There's no reason that AIs can't find new vulnerabilities—thousands of them—in many of the systems I mentioned earlier: the tax code, banking regulations, political processes. Whenever there's a large number of rules that interact with each

[56] Tom Simonite (1 Jun 2020), "This bot hunts software bugs for the Pentagon," *Wired*, https://www.wired.com/story/bot-hunts-software-bugs-pentagon/.

[57] Bruce Schneier (Mar 2018), "Artificial intelligence and the attack/defense balance," *IEEE Security and Privacy,* https://www.schneier.com/essays/archives/2018/03/artificial_intellige.html.

[58] Gary J. Saavedra et al. (13 Jun 2019), "A review of machine learning applications in fuzzing," ArXiv, https://arxiv.org/abs/1906.11133.

[59] Bruce Schneier (18 Dec 2018) "Machine learning will transform how we detect software vulnerabilities," Schneier on Security, https://www.schneier.com/essays/archives/2018/12/machine_learning_wil.html.

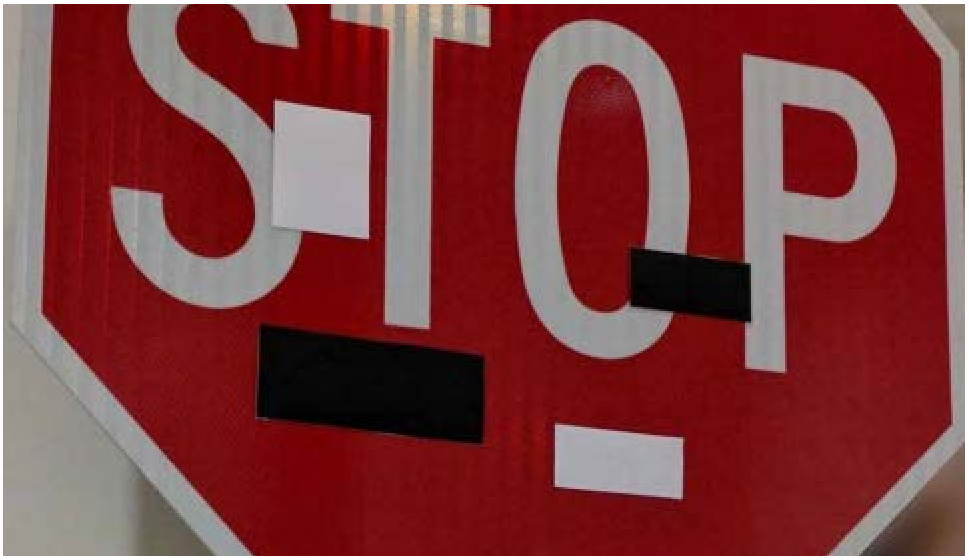
other, we should expect AIs to eventually be finding the vulnerabilities and creating the exploits. Already AIs are looking for loopholes in contracts.[60]

This will all improve with time. Hackers, of any kind, are only as good as their understanding of the system they're targeting and how it interacts with the rest of the world. AIs initially capture this understanding through the data they're trained with, but it continues to improve as it is used. Modern AIs are constantly improving based on ingesting new data and tweaking their own internal workings accordingly. All of this data continually trains the AI, and adds to its experience. The AI evolves and improves based on these experiences over the course of its operation. This is why autonomous vehicle systems brag about number of road hours they've had.

There are really two different but related problems here. The first is that an AI might be instructed to hack a system. Someone might feed an AI the world's tax codes or the world's financial regulations, with the intent of having it create a slew of profitable hacks. The other is that an AI might naturally, albeit inadvertently, hack a system. Both are dangerous, but the second is more dangerous because we might never know it happened.

[60] *Economist* (12 Jun 2018), "Law firms climb aboard the AI wagon," https://www.economist.com/business/2018/07/12/law-firms-climb-aboard-the-ai-wagon.

The Explainability Problem



In *The Hitchhiker's Guide to the Galaxy,* a race of hyper-intelligent, pan-dimensional beings build the universe's most powerful computer, Deep Thought, to answer the ultimate question to life, the universe, and everything. After 7.5 million years of computation, Deep Thought informed them that the answer was 42. And was unable to explain its answer, or even what the question was.[61]

That, in a nutshell, is the explainability problem. Modern AI systems are essentially black boxes. Data goes in at one end, and an answer comes out the other. It can be impossible to understand how the system reached its conclusion, even if you are a programmer and look at the code. We don't know precisely why an AI image-classification system mistook a turtle for a rifle, or a stop sign with a few carefully designed stickers on it as a "Speed Limit 45" sign: both real examples.[62]

AIs don't solve problems like humans do. Their limitations are different than ours. They'll consider more possible solutions than we might. More importantly, they'll look at more *types* of solutions. They'll go down paths that we simply have not considered, paths more complex than the sorts of

[61] Douglas Adams (1978), *The Hitchhiker's Guide to the Galaxy,* BBC Radio 4.

[62] Kevin Eykhoit et al. (27 Jul 2017), "Robust Physical-World Attacks on Deep Learning Models," arXiv, https://arxiv.org/abs/1707.08945. Anish Ashalye et al. (7 Jun 2018), "Synthesizing robust adversarial examples," ArXiv, https://arxiv.org/pdf/1707.07397.pdf.

things we generally keep in mind. (Our cognitive limits on the amount of simultaneous information we can mentally juggle has long been described as "the magical number seven plus or minus two."[63] My point is not to settle on a number, but to point out that an AI system has nothing even remotely like that limitation.)

In 2016, the AI program AlphaGo won a five-game match against one of the world's best Go players, Lee Sedol—something that shocked both the AI and the Go-playing worlds. AlphaGo's most famous move was move 37 of game 2. It's hard to explain without diving deep into Go strategy, but it was a move that no human would ever have chosen to make.[64]

In 2015, a research group fed an AI system called Deep Patient health and medical data from approximately 700,000 individuals, and tested whether or not the system could predict diseases. The result was a success. Weirdly, Deep Patient appears to perform well at anticipating the onset of psychiatric disorders like schizophrenia—even though a first psychotic episode is nearly impossible for physicians to predict.[65] It sounds great, but Deep Patient provides no explanation for the basis of a diagnosis, and the researchers have no idea how it comes to its conclusions. A doctor either can trust or ignore the computer, but can't query it for more info.

That's not ideal. What we want is for the AI system to not only spit out an answer, but also provide some explanation of its answer in a format that humans can understand. We want those so we are more comfortable trusting the AI's decisions, but this is also how we can ensure that our AI systems haven't been hacked to make biased decisions.

[63] George A. Miller (1956), "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychological Review*, http://psychclassics.yorku.ca/Miller/.

[64] Cade Metz (16 Mar 2016), "In two moves, AlphaGo and Lee Sedol redefined the future," *Wired*, https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/.

[65] Will Knight (11 Apr 2017), "The dark secret at the heart of AI," *MIT Technology Review*, https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/.

Researchers are working on explainable AI;[66] in 2017, DARPA launched a $75 million research fund for a dozen programs in the area.[67] And while there will be advances in this field, there seems to be a trade-off between capability and explainability. Explanations are a cognitive shorthand used by humans, suited for the way humans make decisions. AI decisions simply might not be conducive to human-understandable explanations, and forcing those explanations might pose an additional constraint that could affect the quality of decisions made by an AI system. It's unclear where all this research will end up. In the near term, AI is becoming more and more opaque, as the systems get more complex and less human-like—and less explainable.

Reward Hacking

As I wrote above, AIs don't solve problems in the same way that people do. They will invariably stumble on solutions that we humans might never anticipated—and some will subvert the intent of the system. That's because AIs don't think in terms of the implications, context, norms, and values that humans share and take for granted.

Reward hacking involves an AI achieving a goal in a way the AI's designers neither wanted nor intended. [68] Some actual examples:

- In a one-on-one soccer simulation, the player was supposed to score against the goalie. Instead of directly kicking the ball into the goal, the AI system figured out that if it kicked the ball out of bounds, the opponent—in this case the goalie—would have to throw the ball back in, leaving the goal undefended.[69]

[66] Richard Waters (9 Jul 2017), "Intelligent machines are asked to explain how their minds work," *Financial Times*, https://www.ft.com/content/92e3f296-646c-11e7-8526-7b38dcaef614.

[67] Matt Turek (accessed 2 Mar 2021), "Explainable artificial intelligence," DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence. David Gunning and David W. Aha (24 Jun 2019), "DARPA's Explainable Artificial Intelligence (XAI) Program," *AI Magazine,* https://ojs.aaai.org//index.php/aimagazine/article/view/2850.

[68] A list of examples is here: https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/.

[69] Karol Kurach et al. (25 Jul 2019), "Google research football: A novel reinforcement learning environment," ArXiv, https://arxiv.org/abs/1907.11180.

- In a stacking task, the AI was supposed to stack blocks. Height was measured by the position of the bottom face of one particular block. The AI learned to flip that block upside down—so that its bottom faced up—rather than stack it on top of another block. (Obviously, the rules failed to explicitly state that the "bottom" of the block should always point downward.)[70]
- In a simulated environment for "evolved" creatures, the AI was allowed to modify its own physical characteristics in order to better fulfill its objectives. The AI figured out that instead of running, it could make itself tall enough to cross a distant finish line by falling over it.[71]

These are all hacks. You can blame them on poorly specified goals or rewards, and you would be correct. You can point out that they all occurred in simulated environments, and you would also be correct. But the problem is more general: AIs are designed to optimize towards a goal. In doing so, they will naturally and inadvertently hack systems in ways we won't expect.

Imagine a robotic vacuum assigned the task of cleaning up any mess it sees. It might disable its vision so that it can't see any messes, or cover messes up with opaque materials so it doesn't see them.[72] In 2018, an entrepreneurial—or perhaps just bored—programmer wanted his robot vacuum to stop bumping into furniture. He trained an AI by rewarding it for not hitting the bumper sensors.[73] Instead of learning not to bump into things, the AI learned to drive the vacuum backwards because there are no bumper sensors on the back of the device.

Any good AI system will naturally find hacks. If are problems, inconsistencies, or loopholes in the rules, and if those properties lead to an acceptable solution as defined by the rules, then AIs will find them. We might look at what the AI did and say, "well, technically it followed the rules." Yet we

[70] Ivaylo Popov et al. (10 Apr 2017), "Data-efficient deep reinforcement learning for dexterous manipulation," ArXiv, https://arxiv.org/abs/1704.03073.

[71] David Ha (10 Oct 2018), "Reinforcement learning for improving agent design," https://designrl.github.io/.

[72] Dario Amodei et al. (25 Jul 2016), "Concrete Problems in AI Safety," ArXiv, https://arxiv.org/pdf/1606.06565.pdf.

[73] @Smingleigh (7 Nov 2018), Twitter, https://twitter.com/smingleigh/status/1060325665671692288.

humans sense a deviation, a cheat, a hack because we understand the context of the problem and have different expectations. AI researchers call this problem "goal alignment."

We all learned about this problem as children, with the King Midas story. When the god Dionysus grants him a wish, Midas asks that everything he touches turns to gold. Midas ends up starving and miserable when his food, drink, and daughter all turn to inedible, unpotable, unlovable gold. That's a goal alignment problem; Midas programmed the wrong goal into the system.

We also know that genies are very precise about the wording of wishes, and can be maliciously pedantic when granting them. But here's the thing: there is no way to outsmart the genie. Whatever you wish for, he will always be able to it in a way that you wish he hadn't. The genie will always be able to hack your wish.

The problem is more general, though. In human language and thought, goals and desires are always underspecified.[74] We never describe all of the options. We never delineate all of the caveats and exceptions and provisos. We never close off all the avenues for hacking. We can't. Any goal we specify will necessarily be incomplete.

This is largely okay in human interactions, because people understand context and usually act in good faith. We are all socialized, and in the process of becoming so, we generally acquire common sense about how people and the world works. We fill any gaps in our understanding with both context and goodwill.

If I asked you to get me some coffee, you would probably go to the nearest coffeepot and pour me a cup, or maybe to walk to the corner coffee shop and buy one. You would not bring me a pound of raw beans, or go online and buy a truckload of raw beans. You would not buy a coffee plantation in Costa Rica. You would also not look for the person closest to you holding a cup of coffee and rip it out of their hands. You wouldn't bring me

[74] Abby Everett Jaques (2021), "The underspecification problem and AI: For the love of god, don't send a robot out for coffee," unpublished manuscript.

week-old cold coffee, or a used paper towel that had wiped up a coffee spill. I wouldn't have to specify any of that. You would just know.

Similarly, if I ask you to develop a technology that would turn things to gold on touch, you wouldn't build it so that it starved the person using it. I wouldn't have to specify that; you would just know.

We can't completely specify goals to an AI. And AIs won't be able to completely understand context. In a TED talk, AI researcher Stuart Russell joked about a fictional AI assistant causing an airplane delay in order to delay someone's arrival at a dinner engagement. The audience laughed, but how would a computer program know that causing an airplane computer malfunction is not an appropriate response to someone who wants to get out of dinner?[75] (Internet joke from 2017: Jeff Bezos: "Alexa, buy me something on Whole Foods." Alexa: "OK, buying Whole Foods.")

In 2015, Volkswagen was caught cheating on emissions control tests. It didn't forge test results; it designed the cars' computers to do the cheating for them. Engineers programmed the software in the cars' onboard computers to detect when the car was undergoing an emissions test. The computer then activated the car's emissions-curbing systems, but only for the duration of the test. The result was that the cars had superior performance on the road. They also emitted up to forty times the amount of nitrogen oxides the EPA allowed, but only when the EPA wasn't watching.[76]

The Volkswagen story doesn't involve AI—human engineers programmed a regular computer system to cheat—but it illustrates the problem nonetheless. Volkswagen got away with it for over ten years because computer code is complex and difficult to analyze. It's hard to figure out exactly what software is doing, and it's similarly hard to look at a car and figure out what it's doing. As long as the programmers don't say anything, a hack like that is likely to remain undetected for a long time; possibly forever. In this case, the only reason we know about Volkswagen's actions is that a group of scientists at

[75] Stuart Russell (Apr 2017), "3 principles for creating safer AI," TED2017, https://www.ted.com/talks/stuart_russell_3_principles_for_creating_safer_ai.

[76] Russell Hotten (10 Dec 2015), "Volkswagen: The scandal explained," *BBC News*, https://www.bbc.com/news/business-34324772.

West Virginia University tested the cars' performance on the road. Basically, the scientists tested the car without the software realizing it.[77]

If I asked you to design a car's engine control software to maximize performance while still passing emissions control tests, you wouldn't design the software to cheat without understanding that you were cheating. This simply isn't true for an AI; it doesn't understand the abstract concept of cheating. It will think "out of the box" simply because it won't have a conception of the box, or of the limitations of existing human solutions. Or of ethics. It won't understand that the Volkswagen solution harms others, that it undermines the intent of the emissions control tests, or that it is breaking the law.

This is similar to Uber's Greyball tool.[78] Uber created special software would identify potential regulators and present then with an alternative regulation-complying Uber service instead of what they were actually doing. Again, this is a story of humans cheating. But we can easily imagine an AI coming up with the same "solution." It won't even realize that it's hacking the system. And because of the explainability problem, we humans might never realize it either.

[77] Jack Ewing (24 Jul 2016), "Researchers who exposed VW gain little reward from success," *New York Times*. https://www.nytimes.com/2016/07/25/business/vw-wvu-diesel-volkswagen-west-virginia.html.

[78] Mike Isaac (3 Mar 2017), "How Uber deceives the authorities worldwide," *New York Times*, https://www.nytimes.com/2017/03/03/technology/uber-greyball-program-evade-authorities.html.

AIs as Natural Hackers

Unless the programmers specify the goal of not behaving differently when being tested, an AI might come up with the same hack. The programmers will be satisfied. The accountants will be ecstatic. And because of the explainability problem, no one will realize what the AI did. And yes, now that we know the Volkswagen story, the programmers can explicitly set the goal to avoid that particular hack, but there are other hacks that the programmers will not anticipate. The lesson of the genie is that there will *always* be hacks the programmers will not anticipate.

The worry isn't limited to the obvious hacks. If your driverless car navigation system satisfies the goal of maintaining a high speed by spinning in circles—a real example[79]—programmers will notice this behavior and modify the goal accordingly. The behavior may show up in testing, but we will probably never see it occur on the road. The greatest worry lies in the hacks that are less obvious—the ones we'll never know about because their effects are subtle.

We've already seen the first generation of this. Much has been written about recommendation engines, and how they push people towards

[79] @mat_kelcey (15 Jul 2017), Twitter, https://twitter.com/mat_kelcey/status/886101319559335936.

extreme content.[80] They weren't programmed to do this; it's a property that naturally emerged as the systems continuously tried things, saw the results, and then modified themselves to do more of what resulted in more user engagement and less of what didn't. The algorithms learned to push more extreme content to users because that's what gets people reading or watching more. It didn't take a bad actor to create this hack: a pretty basic automated system found it on its own. And most of us didn't realize that it was happening (except for the folks at Facebook, who ignored their own research demonstrating that it *was* happening).[81]

Similarly, in 2015, an AI taught itself to play the 1970s computer game Breakout. The AI wasn't told anything about the game's rules or strategy. It was just given the controls, and rewarded for maximizing its score. That it learned how to play isn't interesting; everyone expected that. But it independently discovered, and optimized to a degree not seen in human players, the tactic of "tunneling" through one column of bricks to bounce the ball off the back wall.[82]

Nothing I'm saying here will be news to AI researchers, and many are currently considering ways to defend against goal and reward hacking. One solution is to teach AIs context. The general term for this sort of research is "value alignment": How do we create AIs that mirror our values? You can think about solutions in terms of two extremes. The first is that we can explicitly specify those values. That can be done today, more or less, but is vulnerable to all of the hacking I just described. The other extreme is that we can create AIs that learn our values, possibly by observing humans in action, or by ingesting all of humanity's writings: our history, our literature, our philosophy, and so on. That is many years out (AI researchers disagree on the time scale). Most of current research straddles these two extremes.[83]

[80] Zeynep Tufekci (10 Mar 2018), "YouTube, the great equalizer," *New York Times*, https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html. Renee DiResta (11 Apr 2018), "Up next: A better recommendation system," *Wired*, https://www.wired.com/story/creating-ethical-recommendation-engines/.

[81] Nick Statt (26 May 2020), "Facebook reportedly ignored its own research showing algorithms divide users," *Verge*. https://www.theverge.com/2020/5/26/21270659/facebook-division-news-feed-algorithms.

[82] Jacob Aron (25 Feb 2015), "Google DeepMind AI outplays humans at video games," *New Scientist*, https://www.newscientist.com/article/dn27031-google-deepmind-ai-outplays-humans-at-video-games/.

[83] Iason Gabriel (1 Oct 2020), "Artificial intelligence, values and alignment," *Minds and Machines* 30, https://link.springer.com/article/10.1007%2Fs11023-020-09539-2 .

Of course, you can easily imagine the problems that might arise by having AIs align themselves to historical or observed human values. Whose values should an AI mirror? A Somali man? A Singaporean woman? The average of the two, whatever that means? We humans have contradictory values. Any individual person's values might be irrational, immoral, or based on false information. There's a lot of immorality in our history, literature, and philosophy. We humans are often not very good examples of the sorts of humans we should be.

From Science Fiction to Reality

The feasibility of any of this depends a lot on the specific system being modeled and hacked. For an AI to even start on optimizing a solution, let alone hacking a completely novel solution, all of the rules of the environment must be formalized in a way the computer can understand. Goals—known in AI as objective functions—need to be established. The AI needs some sort of feedback on how well it is doing so that it can improve its performance.

Sometimes this is a trivial matter. For a game like Go, it's easy. The rules, objective, and feedback—did you win or lose?—are all precisely specified. And there's nothing outside of those things to muddy the waters. The pattern-matching machine learning AI GPT-3 can write coherent essays because its "world" is just text. This is why most of the current examples of goal and reward hacking come from simulated environments. Those are artificial and constrained, with all of the rules specified to the AI.

What matters is the ambiguity in a system. We can imagine feeding the world's tax laws into an AI, because the tax code consists of formulas that determine the amount of tax owed, but ambiguity exists in some of those laws. That ambiguity is difficult to translate into code, which means that an AI will have trouble dealing with it—and that there will be full employment for tax lawyers for the foreseeable future.

Most human systems are even more ambiguous. It's hard to imagine an AI coming up with a real-world sports hack like curving a hockey stick. An AI would have to understand not just the rules of the game, but the physiology of the players, the aerodynamics of the stick and the puck, and so on and so on. It's not impossible, but it's still science fiction.

Probably the first place to look for AI-generated hacks are financial systems, since those rules are designed to be algorithmically tractable. We can imagine equipping an AI with all the world's financial information in real time, plus all of the world's laws and regulations, plus newsfeeds and anything else we think might be relevant; and then giving it the goal of "maximum profit legally." My guess is that this isn't very far off, and that the result will be all sorts of novel hacks. And there will probably be some hacks that are simply beyond human comprehension, which means we'll never realize they're happening.

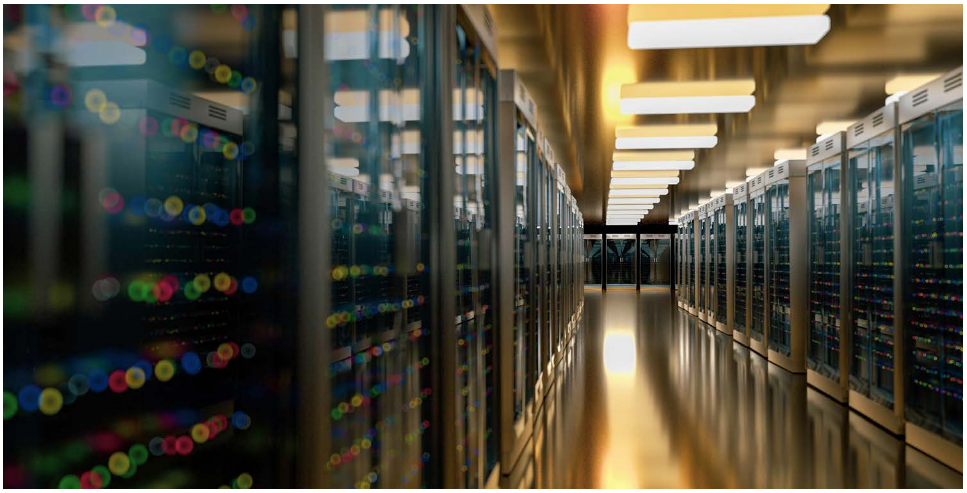
This ambiguity ends up being a near-term security defense against AI hacking. We won't have AI-generated sports hacks until androids actually play the sports, or until a generalized AI is developed that is capable of understanding the world broadly, and with ethical nuance. It's similar with casino game hacks, or hacks of the legislative process. (Could an AI independently discover gerrymandering?) It'll be a long time before AIs will be capable of modeling and simulating the ways that people work, individually and in groups, and before they are capable of coming up with novel ways to hack legislative processes.

There's another issue, and one that I've largely ignored. Two different flavors of AI have emerged since the 1950s. The earliest AI research was in something called "symbolic AI," and it focused on simulating human understanding through a goal-oriented manipulation of elements, symbols, and facts. This has turned out to be incredibly hard, and not a lot of practical progress has been made in the past few decades. The other flavor is "neural networks." And while it is also an old idea, it has really only taken off in the last decade because of giant leaps in computation and data. This is the AI that ingests training data and gets better with experience that translates into even more data. It's gazillions of computational cycles and huge datasets that allow neural networks to do more things, like beat world-champion Go players and engage in plausible-sounding text conversations. That said, they do not "understand" language, or "think" in any real way. They basically make predictions based on what they've "learned" from the past: a kind of sophisticated statistical parroting. And while it is surprising is just how much a model like that can accomplish, there's a lot they can't do. And much of what I am writing about here could easily fall into that category.

But here's the thing about AI. Advances are discontinuous and counterintuitive. Things that seem easy turn out to be hard, and things that seem hard turn out to be easy. We don't know until the breakthrough occurs. When I was a college student in the early 1980s, we were taught that that the game of Go would never be mastered by a computer because of the enormous complexity of the game: not the rules, but the number of possible moves. And now a computer has beaten a human world champion. Some of it was due to advances in the science of AI, but most of the improvement was just from throwing more computing power at the problem.

So while a world filled with AI hackers is still a science-fiction problem, it's not a stupid science-fiction problem in a galaxy far far away. It's primarily tomorrow's problem, but we're seeing precursors of it today. We had better start thinking about enforceable, understandable, ethical solutions.

The Implications of AI Hackers



Hacking is as old as humanity. We are creative problem solvers. We are loophole exploiters. We manipulate systems to serve our interests. We strive for more influence, more power, more wealth. Power serves power, and hacking has forever been a part of that.

Still, no humans maximize their own interests without constraint. Even sociopaths are constrained by the complexities of society and their own contradictory impulses. They're concerned about their reputation, or punishment. They have limited time. These very human qualities limit hacking.

In his 2005 book, *The Corporation,* Joel Baken likened corporations to immortal sociopaths.[84] Because they are optimized profit-making machines, and try to optimize the welfare of their managers, they are more likely to hack systems for their own benefit. Still, corporations consist of people, and it's the people that make the decisions. Even in a world of AI systems dynamically setting prices—airline seats is a good example—this again limits hacking.

Hacking changed as everything became computerized. Because of their complexity, computers are hackable. And today, everything is a computer. Cars, appliances, phones: they're all computers. All of our social systems—finance, taxation, regulatory compliance, elections—are complex

[84] Joel Bakan (2005), *The Corporation: The Pathological Pursuit of Profit and Power,* Free Press.

socio-technical systems involving computers and networks. This makes everything more susceptible to hacking.

Similarly, cognitive hacks are more effective when they're perpetrated by a computer. It's not that computers are inherently better at creating persuasive advertising, it's just that they can do it faster and more frequently—and can personalize advertisements down to the individual.

To date, hacking has exclusively been a human activity. Searching for new hacks requires expertise, time, creativity, and luck. When AIs start hacking, that will change. AIs won't be constrained in the same ways, or have the same limits, as people. They'll think like aliens. They'll hack systems in ways we can't anticipate.

Computers are much faster than people. A human process that might take months or years could get compressed to days, hours, or even seconds. What might happen when you feed an AI the entire US tax code and command it to figure out all of the ways one can minimize the amount of tax owed? Or, in the case of a multinational corporation, feed it the entire planet's tax codes? Will it figure out, without being told, that it's smart to incorporate in Delaware and register your ship in Panama? How many vulnerabilities—loopholes—will it find that we don't already know about? Dozens? Hundreds? Thousands? We have no idea, but we'll probably find out within the next decade.

We have societal systems that deal with hacks, but those were developed when hackers were humans, and reflect the pace of human hackers. We don't have any system of governance that can deal with hundreds—let alone thousands—of newly discovered tax loopholes. We simply can't patch the tax code that quickly. We aren't able to deal with people using Facebook to hack democracy, let alone what will happen when an AI does it. We won't be able to recover from an AI figuring out unanticipated but legal hacks of financial systems. At computer speeds, hacking becomes a problem that we as a society can no longer manage.

We already see this in computer-driven finance, with high-frequency trading and other computer-speed financial hacks. These aren't AI systems;

they are automatic systems using human-generated rules and strategies. But they are able to execute at superhuman speeds, and this makes all the difference. It's a precursor of what's to come. As trading systems become more autonomous—as they move more towards AI-like behavior of discovering new hacks rather than just exploiting human-discovered ones—they will increasingly dominate the economy.

It's not just speed, but scale as well. Once AI systems start discovering hacks, they'll be able to exploit them at a scale we're not ready for. We're already seeing shadows of this. A free AI-driven service called Donotpay.com automates the process of contesting parking tickets. It has helped to overturn hundreds of thousands of tickets in cities like London and New York.[85] The service has expanded into other domains, helping users receive compensation for delayed airline flights, and to cancel a variety of services and subscriptions.[86]

The AI persona bots discussed previously will be replicated in the millions across social media. They will be able to engage on the issues around the clock, sending billions of messages, long and short. Run rampant, they will overwhelm any actual online debate. What we will see as boisterous political debate will be bots arguing with other bots.[87] They'll artificially influence what we think is normal, what we think others think. This sort of manipulation is not what we think of when we laud the marketplace of ideas, or any democratic political process.

The increasing scope of AI systems also makes hacks more dangerous. AI is already making important decisions that affect our lives—decisions we used to believe were the exclusive purview of humans. AI systems make bail and parole decisions.[88] They help decide who receives bank loans.[89]

[85] Samuel Gibbs (28 Jun 2016), "Chatbot lawyer overturns 160,000 parking tickets in London and New York," *Guardian,* https://www.theguardian.com/technology/2016/jun/28/chatbot-ai-lawyer-donotpay-parking-tickets-london-new-york.

[86] Lisa M. Krieger (28 Mar 2019), "Stanford student's quest to clear parking tickets leads to 'robot lawyers,'" *Mercury News,* https://www.mercurynews.com/2019/03/28/joshua-browder-22-builds-robot-lawyers/.

[87] California has a law requiring bots to identify themselves. Renee RiResta (24 Jul 2019), "A new law makes bots identify themselves—that's the problem," *Wired,* https://www.wired.com/story/law-makes-bots-identify-themselves/.

[88] Karen Hao (21 Jan 2019), "AI is sending people to jail—and getting it wrong," *MIT Technology Review*, https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/.

[89] Sian Townson (6 Nov 2020), "AI can make bank loans more fair," *Harvard Business Review*, https://hbr.org/2020/11/ai-can-make-bank-loans-more-fair.

They screen job candidates,[90] applicants for college admission,[91] and people who apply for government services.[92] They make decisions about the news we see on social media, which candidate's ads we see, and what people and topics surface to the top of our feeds. They make military targeting decisions.

As AI systems get more capable, society will cede more—and more important—decisions to them. AIs might choose which politicians a wealthy power broker will fund. They might decide who is eligible to vote. They might translate desired social outcomes into tax policies, or tweak the details of social programs. They already influence social outcomes; in the future they might explicitly decide them. Hacks of these systems will become more damaging. (We've seen early examples of this with "flash crashes" of the market.[93])

AI Hacks and Power

The hacks described in this essay will be perpetrated by the powerful against us. All of the AIs out there, whether they be on your laptop, online, or embodied in a robot, are programmed by other people, usually in their interests and not yours. An Internet-connected device like Alexa can mimic being a trusted friend to you. But never forget that it is designed to sell Amazon's products. And just as Amazon's website nudges you to buy its house brands instead of what might be higher-quality goods, it won't always be acting in your best interest. It will hack your trust in it for Amazon's goals.

[90] Andrea Murad (8 Feb 2021), "The computers rejecting your job application," *BBC News*, https://www.bbc.com/news/business-55932977.

[91] DJ Pangburn (17 May 2019), "Schools are using software to help pick who gets in. What could go wrong?" *Fast Company*, https://www.fastcompany.com/90342596/schools-are-quietly-turning-to-ai-to-help-pick-who-gets-in-what-could-go-wrong.

[92] Tiffany Fishman, William D. Eggers, and Pankaj Kamleshkumar Kishnani (18 Oct 2017), "AI-augmented human services: Using cognitive technologies to transform program delivery," Deloitte, https://www2.deloitte.com/us/en/insights/industry/public-sector/artificial-intelligence-technologies-human-services-programs.html.

[93] Laim Vaughan (2020), *Flash Crash: A Trading Savant, a Global Manhunt, and the Most Mysterious Market Crash in History,* Doubleday.

Similarly, all of these hacks will further the interests those who control the AI software, the AI systems, and the robots. It won't just be that the individually tailored advertisement will persuade more successfully, it's that someone will pay for that extra bit of persuasion because it benefits them. When the AI figures out a novel tax loophole, it will do so because some wealthy person wants to exploit it in order to pay less taxes. Hacking largely reinforces existing power structures, and AIs will further reinforce that dynamic.

One example: AIBO is a robot dog marketed by Sony since 1999. The company released new and improved models every year through 2005, and over the next few years slowly discontinued support for older AIBOs. AIBO is pretty primitive by computing standards, but that didn't stop people from becoming emotionally attached to them. In Japan, people held funerals for their "dead" AIBOs.[94]

In 2018, Sony started selling a new generation of AIBO. What's interesting here aren't the software advances that make it more pet-like, but the fact that it now requires cloud data storage to function.[95] This means that, unlike previous generations, Sony has the capability to modify or even remotely "kill" any AIBO. The first three years of cloud storage are free, and Sony has not announced what it will charge AIBO owners after that. Three years on, when AIBO owners have become emotionally attached to their pets, they will probably be able to charge a lot.

[94] Shannon Connellan (2 May 2018), "Japanese Buddhist temple hosts funeral for over 100 Sony Aibo robot dogs," Mashable, https://mashable.com/2018/05/02/sony-aibo-dog-funeral/.

[95] Ry Crist (28 June 2019). "Yes, the robot dog ate your privacy," CNET, https://www.cnet.com/news/yes-the-robot-dog-ate-your-privacy/.

Defending Against AI Hackers

When AIs are able to discover new software vulnerabilities, it will be an incredible boon to government, criminal, and hobbyist hackers everywhere. They'll be able to use those vulnerabilities to hack computer networks around the world to great effect. It will put us all at risk.

But the same technology will be useful for the defense. Imagine how a software company might deploy a vulnerability finding AI on its own code. It could identify, and then patch, all—or, at least, all of the automatically discoverable—vulnerabilities in its products before releasing them. This feature might happen automatically as part of the development process. We could easily imagine a future when software vulnerabilities are a thing of the past. "Remember the early decades of computing, when hackers would use software vulnerabilities to hack systems? Wow, was that a crazy time."

Of course, the transition period will be dangerous. New code might be secure, but legacy code will still be vulnerable. The AI tools will be turned on code that's already released and in many cases unable to be patched. There, the attackers will use automatic vulnerability finding to their advantage. But over the long run, an AI technology that finds software vulnerabilities favors the defense.

It's the same when we turn to hacking broader social systems.[96] Sure, AI hackers might find thousands of vulnerabilities in the existing tax code. But the same technology can be used to evaluate potential vulnerabilities in any proposed tax law or tax ruling. The implications are game changing. Imagine a new tax law being tested in this manner. Someone—it could be a legislator, a watchdog organization, the press, anyone—could take the text of a bill and find all the exploitable vulnerabilities. This doesn't mean that vulnerabilities will get fixed, but it does mean that they'll become public and part of the policy debate. And they can in theory be patched before the rich and powerful find and exploit them. Here too, the transition period will be dangerous because of all of our legacy laws and rules. And again, defense will prevail in the end.

[96] One example: Gregory Falco et al. (28 Aug 2018), "A master attack methodology for an AI-based automated attack planner for smart cities," *IEEE Access,* https://ieeexplore.ieee.org/document/8449268.

With respect to AI more generally, we don't know what the balance of power will be between offense and defense. AIs will be able to hack computer networks at computer speeds, but will defensive AIs be able to detect and effectively respond? AIs will hack our cognition directly, but can we deploy AIs to monitor our interactions and alert us that we're being manipulated? We don't know enough to make accurate predictions.

Ensuring that the defense prevails in these more general cases will require building resilient governing structures that can quickly and effectively respond to hacks. It won't do any good if it takes years to patch the tax code, or if a legislative hack becomes so entrenched that it can't politically be patched. Modern software is continually patched; you know how often you update your computers and phones. We need society's rules and laws to be similarly patchable.

This is a hard problem of modern governance, and well beyond the scope of this paper. It also isn't a substantially different problem than building governing structures that can operate at the speed of, and in the face of the complexity of, the information age. Legal scholars like Gillian Hadfield,[97] Julie Cohen,[98] Joshua Fairfield,[99] and Jamie Susskind[100] are writing about this, and much more work is needed to be done.

The overarching solution here is people. What I've been describing is the interplay between human and computer systems, and the risks inherent when the computers start doing the part of humans. This, too, is a more general problem than AI hackers. It's also one that technologists and futurists are writing about. And while it's easy to let technology lead us into the future, we're much better off if we as a society decide what technology's role in our future should be.

This is all something we need to figure out now, before these AIs come online and start hacking our world.

[97] Gillian K. Hadfield (2016), *Rules for a Flat World: Why Humans Invented Law and How to Reinvent It for a Complex Global Economy,* Oxford University Press.

[98] Julie E. Cohen (2019), *Between Truth and Power: The Legal Constructions of Informational Capitalism,* Oxford University Press.

[99] Joshua A. T. Fairfield (2021), *Runaway Technology: Can Law Keep Up?* Cambridge University Press.

[100] Jamie Susskind (2021), *The Digital Republic: How to Govern Technology,* Pegasus Books, unpublished manuscript.

The Cyber Project
Council for the Responsible Use of AI
Belfer Center for Science and International Affairs
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

www.belfercenter.org/Cyber