Z. IRENE YING AND JORDAN BOYD-GRABER

HOW I LEARNED TO WORRY PRODUCTIVELY ABOUT AI:

ROBOT APOCALYPSES AVERTED

UNIVERSITY OF MARYLAND

tufte-latex.github.io/tufte-latex/

September 2019

Introduction: Who should fear the Robot Apocolypse

The news and popular culture want us to fear the robot apocolypse. Newspapers warn us that robots are coming for our jobs, and science fiction claims that the robots are coming for our lives. At the same time, researchers building artificial intelligence are heralding their sucesses at Go **?**, Starcraft **?**, and *Jeopardy!* **?** a revolution is around the corner.

This book aims to bring together some of these threads into a coherent narrative building on our experience as a science writer (Z. Irene Ying) and an artifical intelligence researcher (Jordan Boyd-Graber). While we[1] agree that there are lots of changes in artificial intelligence and potential changes to society, we do disagree with some of the dire predictions we see from researchers, the news, and science fiction.

But nobody listens to our rants on Twitter or in the classroom, so we decided to write a book. It's not that we think that the future is full sunshine, lollipops, and rainbows.[2] We are as scared as anybody else... we're just scared of different things, and we'd like to use our background to try to get other people to be concerned about the same things we are.

Science fiction typically emphasizes fiction over science (and rightly so, it's more fun that way). These stories focus on massive sudden changes that uproot society overnight: a zombie virus destroys civilization in twenty-eight days,[3] a military computer launches all of the missiles to destroy the world,[4] or alien refugees overwhelm society.[5] These sudden shocks to the system make for good stories because we understand our culture *as it is* and we, as readers, can grapple with how our twenty-minutes-in-the-future selves would cope with these changes.

While technology does change society, these changes are typically slow. While telephones frightened one generation, it took so long to figure out how to use them effectively that by the time they were integrated fully into society, the next generation had lived with the *idea* of telephones so long that they were no longer viewed as an existential threat. The fear with artificial intelligence and robots is that progress is so quick that we won't have this comfortable buffer period... but for

Figure 1: Jordan Boyd-Graber is an associate professor at the University of Maryland who researches how computers can learn from humans and compete with humans. You can watch his trivia playing robots take on humans on YouTube.

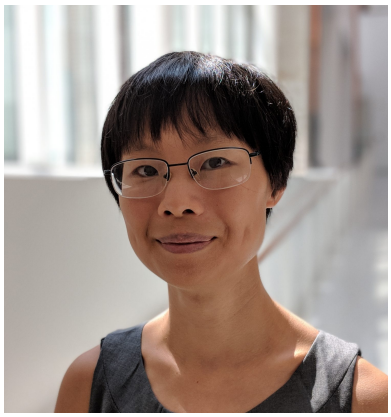

Figure 2: Z. Irene Ying is a science writer who publishes science fiction as Kara Lee.

[1] At the risk of sowing confusion, we use "we" in this chapter (first person), Jordan will use "I" in following essays, and Irene will use the third person in the following stories.

[2] This titular opening line from the song "Sunshine, Lollipops, and Rainbows" by Lesley Gore that featured prominently in *The Simpsons* episode "Marge on the Lam" and the film *Cloudy with a Chance of Meatballs* is a good chance to show our typographic conventions we'll use for our many pop culture references.

[3] *28 Days Later*, a 2002 British post-apocalyptic film where a genetically engineered virus destroys England.

[4] SkyNet's plan for erradicating humanity from the Terminator franchise.

[5] The 2009 film *District 9* based on *Alive in Joburg*

those in the trenches, progress feels achingly slow.

Researchers focus on what works. We spend years getting a system to answer trivia questions to work, so we want to show off what it's capable of. We focus on how it's able to know that "Rorke's Drift" is in South Africa and not how it answers every math question with "six" (true story). We are still a long way from anything close to a child's intelligence, let alone something that would challenge humanity's supremacy.

How the Book is Structured

We are not the only ones saying this; plenty of researchers warn of overselling the hype around machine intelligence. However, these boring essays and conference presentations don't really stack up against a big-budget western with Anthony Hopkins creating sexy killbots.[6] We don't have that budget either,[7] but what we can do is create sci-fi stories that better comport with the realities of science.

When people hear about the latest advance in artificial intelligence, they often imagine how it will change their lives. This book does the same thing... presenting a series of short stories about how artificial intelligence can shape culture, politics, and the economy. What is unique (or at least unusual) about this book is that it is shaped by an understanding of what machine learning research is doing.

Interleaved with each of these fictional extrapolations is a more scientifically oriented story that combines technical insights and the literature of the research community that drove some of the storytelling choices of the fictional chapters. These chapters are more academic in tone (but hopefully a little more accessible and playful than the turgid jargon-filled prose on ArXiv[8] submissions) and discuss what the academic community thinks about the themes in the fictional chapter.

Now that we've talked about *how* the book is structured, it is perhaps worthwhile to briefly discuss *why* the book is structured this way. We realize that it is unconventional, but we hope that it will be fun and keep the reader's attention more effectively than a mere collection of academic essays would.

Part of the issue is that the fear of artificial intelligence is emotional rather than based on logic and fact; countering it with essays is bringing a knife to a gun fight. Stories can engage the gut at a level that cerebral citations and arguments cannot. Our goal is to draw the reader into a universe plausible enough that afterward the reader can read the facts, see how close we are to the brink of armageddon, and then—ideally—do something to avert catastrophe.

[6] *Westworld*, an HBO series about a western theme park populated by intelligent robots.

[7] However, if anyone, including Anthony Hopkins, would like to give us a big budget, a sound stage, or sexy killbots, we'd put them to good use.

[8] ArXiv is a server that researchers use to publish results before they've been peer-reviewed. While there are plenty of gems posted there, there's also quite a bit of garbage people have posted to assert research priority.

The Book's Intended Audience

The short stories are meant to appeal broadly. We hope that everyone can enjoy the human elements, the struggle to understand new technologies, and how obstacles are (usually) surmounted. However, we hope that artificial intelligence experts will especially be able to read the stories without the head-slapping "that would never happen" or "that's not what that word means" moments.

The interstitial commentaries and exegesis are *not* intended for experts; they are intended to help connect a lay reader to the broader world of research. Fledgling researchers may appreciate some of the citations and connections to subfields of artificial intelligence research, but it will not be particularly useful to a researcher learning how to implement a recurrent neural network, deep opponent modeling, or a language model. Indeed, we will generally elide equations and jargon as much as possible.

If used in a classroom, we think that the course would be useful as part of an elective course about "artificial intelligence and society" or a computer science course on "artificial intelligence ethics" our associated webpage has recommended reading lists to supplement either instantiation of the course.

We assume that readers are broadly interested in artificial intelligence and its impact on society but they are not experts; we are wont to make references throughout the book but will explain as we go (perhaps violating the maxim of "don't explain the joke"), even within the stories. Then, the essays will go into even deeper detail, slowly building up the necessary concept to understand the role of artificial intelligence in society.

How to Read the Book

The short story chapters are told chronologically and build on each other. Thus, it makes sense to read them in order ("no spoilers"). The essays are tied to the stories and reference them extensively; ignoring these references, the essays could be read out of order. Each chapter concludes with additional reading; as part of a "deep dive" course, it may make sense to thoroughly read these suggestions before moving on to the next chapter. For example, a course could be structured with reading a chapter of the book each week, with the suggested reading (or a subset thereof) of that week framing the in-class discussion.

For such a fast moving field, however, a static book is inadequate. We also suggest referencing the book's webpage to keep up-to-date on errata, recent developments, or additional resources.

If You're Looking for Energy, try the Coffee Machine

IRENE NOTE TO SELF: PUT A CYBER CLIENT FOR CALVIN UP FRONT SO THAT THE ATTACK LATER SEEMS RELATED TO THAT

" ... and does anyone else have something to say? Yes, John."

John Smith, the Chief Marketing Officer of ReVOLTlution, stood from his conference room chair and straightened his tie. He had clearly been saving up for the grand finale of the monthly all-hands meeting.

"I hate to call out my own colleagues—"

Liar, Susan Hobbes thought morosely from across the room, nursing her long-cold cup of coffee.

"—but in my opinion, there is a huge problem with the Travelling Salesman, and nobody will say anything about it except me—"

Actually, the developer team probably had the most to say about it, but they did it behind closed doors.

"—and if you doubt that, please allow me to replay a sales script that the Salesman generated just this morning."

But here came the cold, hard truth. John wasn't wrong. He was just being a very public jerk about it.

John took a company issue phone out of his blazer pocket and set it on the conference room table. In full view of the entire company, he launched the Travelling Salesman (in beta) application and asked it to generate a door-to-door sales script for a new customer. He set the phone to speaker. A tinny voice blared.

"Hi INSERT CUSTOMER, this is INSERT NAME with INSERT COMPANY. I'm here to ask if you have a minute to talk with me about the great news of solar panels! What, you don't like them? Why? Do you hate Mother Earth?"

Eyes in the audience either glazed over or averted themselves with embarrassment. Susan tried to block out the script but it just kept coming. Snickers came from the audience. Was someone filming it on their cameraphone? She was going to speak with legal ... strike that, the person filming *was* from legal.

"Did you know that CLIMATE CHANGE ... GREEN ENERGY! ... solar panels for the low price of ... and the installation is free! Plus, if you sign up right now, we will give you a free toaster!"

The company did not in fact give away toasters. Susan honestly did not know how that line had somehow found its way into the generation library, but that hardly mattered at this point. As one of the junior developers who had helped code the Salesman, she wished for a sinkhole to open up and swallow her.

"I just don't think that humans can build a sales robot," John concluded.

Out of the corner of her eye, Susan saw her boss, Tricia, the Chief Technology Officer, stand up with steely eyes.

"Steve," she said, addressing the CEO directly. "Can we talk in your office after the meeting, please?"

"We're not fired," Susan said. "Yet."

Calvin, her twin and studio apartment-mate, looked at her over the monitor of his laptop.

"Did your boss talk her boss down from firing you all?"

"Actually, I think John just wants to watch us get fired in front of the board," Susan said. "He's probably selling tickets as we speak."

Tricia had somehow managed to convince the CEO to at least give her team the chance to turn around their performance by the annual board meeting, which would be in six months. The problem was, Susan did not think their system would improve in six months—or even six years.

ReVOLTlution, a startup that sold "smart" solar panels—outfitted with algorithms to help their owners take advantage of energy markets and local incentives—turned a tidy profit in the city of Bayder, a wealthy and progressive city that prioritized renewable energy. Two years ago, the city had used eminent domain to purchase the local coal-fired power plant, then sold a pile of bonds to finance a smart grid for the entire municipality. Thanks to its geographical location—-tucked between several mountain ranges—-Bayder conveniently experienced high winds just on its outskirts, which was perfect for an eight-turbine wind power plant that produced multiple gigawatts of power. Also conveniently, Bayder was blessed with more sunny days than not, which made it one of the best cities in the United States for solar power.

Just in the past year, the city had begun to implement significant financial incentives to encourage its citizens to save energy. For instance, in exchange for a hefty reduction to the electricity bill, the city council installed free smart thermostats in willing customers' homes. These thermostats were connected by the internet to a municipal energy analysis application, which used data from the city's power grid and the weather to predict when those peak times would be. At peak times, the application sent out remote commands to shut off the air

conditioning units of participating homes. People complained at first, but once they saw the huge deductions to their bills, they got on board. Within three years, more than 75% of homes in Bayder had these thermostats.

Happy with their first major success, the city council started to issue various incentives for homeowners to use more renewable energy. For instance, they offered to purchase electricity generated from homeowners' solar panels at market rates. But few homeowners signed on, because they either did not understand, or did not trust, the energy market.

Sensing an opportunity, several entrepreneurial souls started ReVOLTlution. But sales were slower than the company executives hoped for, simply because solar panels were expensive and smart panels even more so.

ReVOLTlution fancied itself on the cutting edge of everything, including sales. So instead of hiring consultants, the CEO decided to improve sales by hiring a team of AI developers to create a sales AI, "The Travelling Salesman." In theory, the Salesman used artificial intelligence and machine learning to analyze the best sales scripts in the world and used that to generate scripts for ReVOLTlution's sales team when they went out on sales calls. In practice—well, everyone had seen the results at the meeting.

Tricia had shared the bad news with her team after speaking with the CEO. Because the Salesman produced such poor scripts, the sales associates had started turning off or just plain ignoring the Salesman on their calls altogether. When they did use the Salesman, it was to mock it: Canyon Green, a self-described "script kiddie turned honest salesman" who was far and away ReVOLTlution's superstar sales associate, had turned one of the Salesman's worst scripts—which was really saying something—into a techno remix and distributed it to his coworkers.

The AI team was clearly failing at its job. It was going to be hard to justify their continued employment at this rate.

And Susan, a freshly minted alum of the University of Bayder, had student loans to pay.

"You could quit the job and join me as a partner in my business," Calvin suggested after hearing her tale of woe. "I get more requests than I can handle. We could probably clear twice my current income if we joined forces. I know you have all the relevant expertise."

Susan and Calvin had both had studied computer science with a focus in artificial intelligence at the University of Bayder. But while Susan farmed out her skills to a company, Calvin ran his own consulting business.

It had started when he was a junior and worked as a server at the

Bayderado Hotel for some extra cash over the summer. One evening, as he came off shift, he heard a loud gathering in the bar area. It turned out to be a meetup of hotshot young AI entrepreneurs, and they had a heckler: an older man wearing a very nice suit who was convinced that robots were going to bring about the end of civilization. The man was soundly mocked by the entire gathering and he exited the bar, disgusted and humiliated.

Calvin had followed the heckler into the lobby area, sidled up, and flashed a smile.

"I'm an expert in artificial intelligence, and I thought your idea was very interesting," he'd said.

"Good! Glad you think so. AI devices are—" and he was off again.

"I think that is such a fresh perspective," Calvin said after the rant. "If you were interested in expanding on that topic further, for instance in a book or a blog post, I'd love to offer my services as a consultant."

A week later, he'd made his first hundred dollars. Today he was making enough to pay his share of the rent and groceries. It turned out that the world had an astounding supply of rich cranks.

Susan, though, was uncomfortable with the uneven cash flow—among other features—of Calvin's work, and said so.

"I'd rather try to salvage my current job," she said. "Besides, I'm pretty sure I would piss off your clients in no time."

"Do you think it's possible to salvage?" Calvin said idly. "Sales is such a squishy, noisy human behavior. It's not the type of thing you can easily model with equations. Even if you could, you couldn't generalize it to all people. Every single human is going to have a different thing that works for them."

"That's right," Susan said. "You're a genius, Calvin!"

"I haven't told you anything you don't already know," Calvin pointed out.

"Well," Susan joked, "you are a consultant."

"Right," Calvin said. "So tell me, what did I tell you?"

The problem began with, Susan realized, the underlying methodology of the Salesman, which was to collect the best sales strategies and scripts that the internet had to offer, and use them to generate even better strategies and scripts.

To implement this method, the Salesman was fed on a steady diet of sales scripts from a wide variety of companies. It then tried to generate scripts for ReVOLTlution's sales team when they went out on calls.

The scripts were, as everyone now knew, terrible. But Susan didn't think it was because their algorithms were bad. She thought it was because their method was doomed from the start. Scripts from so many different businesses all selling different things could not possibly

be tailored to the specific needs of any one particular company.

"All right," Calvin said. "So how do you think you can do better?"

"We shouldn't be trying to extract the best out of the entire human sales population," Susan said. "What we should do is to try to emulate one human. Just one. But the best."

"Um," Canyon said. "Are you here about the video? Because I'm really sorry about that."

"I think getting fired may have been preferable to this," Susan moaned, two weeks later.

At her weekly one-on-one with Tricia, Susan had pitched her idea. Tricia had been impressed with Susan's initiative, and rewarded Susan with taking the lead on implementing her idea. Susan had felt like a genius for all of ten seconds before reality sunk in.

Reality was work, lots of it. GPS to track spatial movement, motion sensors to track hand gestures, voice recognition to get transcripts of Canyon's sales speeches, analysis of Canyon's speeches, tracking of Canyon's research on pitches, and so on. In addition, ReVOLTlution's smart solar panels came with a smartphone app, which Canyon often demonstrated using his company phone on his calls, so Susan gave Salesman access to Canyon's phone. And then, Susan had to find ways to convert all that squishy behavior into data for the Salesman to process, analyze, and optimize.

Susan also took feedback from Canyon that salespeople didn't like being given scripts as if they were incapable of stringing words together. So together they softened the Salesman's tone from an overseer to a coach.

With one month to go to the board meeting, an exhausted Susan unveiled Salesman 2.0.

Possibly because he was flattered by being the engineers' gold standard, Canyon even exerted himself to bringing his colleagues around to giving Salesman 2.0 a try.

At the next monthly company-wide meeting on the last Friday of the month, Susan fought the urge to hide in the bathroom or call in sick.

She waited for materials science to finish what seemed like a hundred slides of chemical diagrams. She sat through finance wringing their hands about obscurely worded federal incentives that would either cost or make the company a whole boatload of money. She did not fall asleep while marketing put up two supposedly vastly different brochures that looked, to her, exactly alike.

Sales was scheduled to speak just before the software development team, all the better to increase the dramatic tension. By this point the

entire company knew that the team was one failure away from the axe. From her seat in the back, Susan could smell the popcorn.

John got up and with agonizing slowness, discussed their sales rates, which were keeping them afloat albeit flatter than he would have liked. But there was a small uptick at the end.

Then he got to the Travelling Salesman 2.0.

He cleared his throat.

"My team and I feel that the Salesman is actually helpful this time around," he said. "There was one sale that I felt was lost. But the Salesman suggested some wording that I would never have chosen. And it helped me sell those words. And the panels."

He glanced at Susan.

"I still don't think robots will ever outright replace human sales. People still like having a flesh-and-blood hand to shake. But nothing's wrong with having a little help."

Susan didn't remember the rest of the meeting—or most of the rest of the night, which included closing a bar with Tricia.

Within another month, Salesman 2.0 was sales' best friend. Even those who didn't care to adopt its scripts loved having it data mine targets and flag all the most important findings. And the happy sales associates complied with Susan's request to give the Salesman feedback on its data. This allowed Susan to refine the Salesman even more.

Sales began to rise, ever so slightly. Morale went up. And by the next summer, so did Susan's salary.

And so did the company's profile, not only in the energy sector but the wider artificial intelligence community.

"Congratulations," Calvin said, just over a year into Susan's time at ReVOLTlution. "A potential client blogged about you. You've hit the big time."

Susan was working from home that day because half of her team was at a sales show. It was a terribly hot Friday in August. A blackout had hit Bayder at noon, leaving the entire city to swelter and Susan to miss the beginning of ReVOLTlution's daily stand-up meeting.

At Calvin's remark, Susan leaned over to check out the webpage displayed on Calvin's computer. There it was, a 5,000-word blog post declaring ReVOLTlution to be the "Harbinger of the Robot Apocalypse."

"He wants to make this a blog series," said Calvin. "I quoted him a hundred an hour. I think he'll do it."

Susan had a deadline but couldn't resist this piece of temptation. Calvin handed her the laptop.

"THE ROBOTS ARE READING YOUR MIND TO PICK YOUR POCKET!" read the subject line. It escalated from there. "ReVOLTlution's 'Trav-

elling Salesman' is supposedly the best sales AI in existence, resulting in record sales conversions of their overpriced solar panels! What can this be but using machines to READ OUR MINDS and manipulate us? Soon we will be in FINANCIAL SLAVERY to the companies that can zap our brains to make us BUY, BUY BUY!"

"Wow," Susan finally said, not knowing how else to react.

"It's actually one of the more coherent requests that I get," Calvin said thoughtfully. "I might give him a discount on this series."

"I can't believe you," Susan said.

"I looked into the data that he provided," Calvin said. "According to trade association reports, ReVOLTlution's sales broke industry records in June. Do you think those numbers are fake?"

Susan thought about the slide deck she was supposed to be working on. "No. We've been doing well."

While Susan didn't exactly get to look over accounting's shoulder, she could tell from the good mood in the company, and her first-ever annual bonus, that they were doing well financially.

Judging from how much complaining came out of the manufacturing side of the business, the company was getting more orders than they could handle. Customers left good reviews on the panels and the app. The company had even briefly courted an internet giant. After the deal fell through, the company pivoted to focus on growing itself. ReVOLTlution had recently expanded out of Bayder and started campaigns targeting other nearby cities.

Given those results, it was only reasonable to conclude that the Travelling Salesman was doing something right.

Susan was spared from her musings when the power came back on. She put on her headset and called into the meeting, which was half over.

"—-and just so you all know, sales closed five more deals yesterday. Let's have a round of applause!"

The air conditioning started back up just then. The blast of cool air hit Susan's face like a slap, waking her up to reality.

Her work was good. But it could not possibly be that good, simply because the field of AI was not yet that good. The Salesman could not possibly be getting those results on its own.

Yet the sales numbers were real. Customers reported high rates of satisfaction with the product, although they did often complain about the prices.

Was the AI a fake? Susan had heard of several cases where an "AI" was actually a glorified crowd of low-paid workers. But Susan had seen the Salesman's code. It trained on human data and let a human make the final decision on what to say, but there was no room for humans to interfere between those two steps. Besides, few humans

could match the staggering success rate that the Salesman had.

Still, Susan had a funny feeling.

Early on Saturday morning, Susan logged onto the company servers. Staring at folders full of code and logs and documentation, Susan realized she had no idea where to start. She sat there paralyzed for about ten minutes before she realized that she could simply test the Salesman. Although Susan had coded a big chunk of its code, she had never tried to use the final product as a whole. Doing so would provide either inspiration or directions on her next steps.

It was the work of only a few minutes to compile the Salesman on her laptop and install the app on her personal smartphone. She loaded Calvin's name into the app. The interface brought up a slow swirl of calming colors while it processed her request. When the app told Susan that it was ready to get started, she put the phone on speaker and set the app to begin its sales pitch. A cartoony happy face popped onto the screen and began to talk.

"I have some information for you!" The Salesman had a chipper voice that sounded like a sports announcer. Susan wondered if there were options for different personas. "Don't worry, I'll remind you later, but let's have an overview! Your target is Calvin Hobbes. Age 23. He is a self-employed white male with a STEM degree from a top 25 university in his field. No criminal history. Significant internet presence with interest heavily concentrated in . . . " and it went on for a while. None of it was surprising to Susan, but she imagined it would shock some people to know that so much personal information was freely available online, and plenty more existed if you were willing to pay for it.

When the dossier ended, the Salesman chimed.

"Time to close the deal!" The Salesman said, still inhumanly chipper, and the smiley face actually winked. Susan knew this wasn't in the default behavior. But then again, the whole point of the Salesman was that it was highly flexible to accomodate a wide range of human customers.

One-one thousand, two-one thousand . . . the time was up, and nothing happened. She frowned.

Calvin walked into the room, holding up his buzzing phone. "We've got a hacker."

Susan was simultaneously annoyed by the interruption and troubled by the announcement. "Excuse me?"

"Someone's targeting our Internet of Things network. I set up an alert system last year after a blogger paid me with that fancy remote controlled espresso machine. Its app was super insecure. I always knew that someday, someone would go after the system."

"Are you seriously telling me that a hacker is targeting your espresso?"

"No. You know those fancy free programmable smart thermostats that Bayder got our landlord to install? It's not exactly secure either, so I put it on the same alert system. That's what's being attacked right now."

"Why would someone try to mess with our thermostats?"

"Well, if you believe my customers, cyberterrorist attacks on our nation's infrastructure are imminent and we should all be wearing tinfoil hats while living underground. But if you ask me—-"

The lights went off with a snap.

"Ugh," Calvin said. "It got through. I really need to buff up the system. After I secure our electronics. Can you give me a hand?"

Susan's back pocket buzzed. Probably the Salesman. She muted the phone and tossed it onto her desk. Dealing with a possible hacker was more urgent. Susan was furious at the hacker, less for attacking the system than for interrupting her experiment. But Calvin was right. They needed to take some quick action in case the hacker had injected malicious code into their network. The twins methodically disabled all their networked devices. Then they began painstakingly going through network logs and checking for malware.

Saturday was no cooler than Friday. Concerned about another blackout, the twins had opted to leave the air conditioning off and rely on wet towels draped over standing fans. By noon, Susan was baking.

"Screw the grid," she muttered. "*We* should go solar. Set up some batteries. Then we wouldn't have to worry about the city messing around with our power. Or a hacker. I've got an employee discount."

"Hold that thought," Calvin said, a finger raised. "I think I've finally managed to trace the attack. It came from somewhere in the apartment complex. I'm not sure if they're online anymore, but if they are, they've got a surprise coming their way."

Calvin set up a fresh internet connection while Susan brought their systems back online, the better to entrap the hacker.

"Two can play at this game," Calvin said, and pressed a button.

Thirty seconds later, Susan thought she smelled something funny. She turned to see her cell phone spewing smoke while the plastic table around it very slowly melted.

"Holy—Calvin!"

Calvin was already running toward the desk with a full ice bucket, cursing loudly enough to wake the dead, while Susan was frozen in place, her mind briefly shocked into a kind of blankness.

As she felt her mind come back online, she felt information struggling to come together.

Her phone had been the source of the attack.

They hadn't found any malware on it.

The only thing she'd done that day was install the Salesman on it.

Canyon used to be a coder. And she had let the Salesman have unfettered access to his phone.

Screw the grid. We should go solar. Set up some batteries. Then we wouldn't have to worry about the city messing around with our power. Or a hacker.

Time to close the deal!

"I have good news and bad news," Susan said, while Calvin mopped up the floor and babbled various apologies for destroying her phone. "Actually, they're the same news. I really did make an AI that can sell solar panels. Unfortunately, it's not so much the Travelling Salesman as the Sleazy Salesman. I've created a monster."

It seemed that Canyon had not left his coding days behind after all. On certain sales calls, he used various hacking programs that lived on his phone—which the Salesman also lived on—to compromise smart thermostats or other Internet of Things devices. This allowed him to overload customers' circuits and trigger blackouts at will. This dramatically reminded customers of the unreliability of the city grid, which was an excellent incentive for them to buy solar panels. Of course, this didn't convince *every* customer to do so, but it convinced enough to make an appreciable increase in sales. It had certainly convinced Susan, after all.

At some point the Salesman had picked up on Canyon's hacking. Susan had coded it to replicate Canyon's actions without regard for anything other than increasing the sales success rate.

And because Susan it had unfettered access to Canyon's phone, again thanks to Susan, the app eventually copied over the same programs that Canyon used to overload devices, and found it wildly successful as a sales tactic.

Susan covered her face, simultaneously horrified and impressed by the Salesman.

The only question now was what to do. Susan decided to take a page out of sales' book for the very last time and perform a live demonstration.

Her opportunity came earlier than expected.

Every fall, the company flew in its board of advisors for an annual meeting, complete with slide presentations and a fancy catered dinner. Susan glanced at the proposed schedule and saw a very obvious lack of a demonstration of the Salesman. That figured. Susan didn't think Canyon was eager to show off his tricks. The first meeting of the day was about finances. After that, the mid-morning's talks mostly revolved around methods of manufacturing solar panels while the afternoon's talks were about energy policies.

The engineer who enjoyed giving a talk had yet to be born, so Susan met only perfunctory resistance when she offered, in her weekly meeting with Tricia, to give a lunchtime slide deck on the Travelling Salesman AI.

Everyone piled into the conference room at noon on the fateful day. Two office interns passed out boxed salads while Susan, sweating in her new suit, set up her laptop. The background chatter quieted when her title slide flashed onto the screen: "The Travelling Salesman: How Does It Work?"

Susan went to the second slide, which was a screenshot of many lines of code. She saw eyes immediately glaze over. Perfect.

"This is boring, am I right?" Susan said. The audience perked up. "Reading raw code is really boring, even for programmers. What's interesting is what the code does. And our code? It learns. Specifically, it learns to sell things from humans. In our case, it has learned from the best sales associate alive—-Canyon! Can you come up here?"

Susan flashed a slide showing sales numbers growing along with how much the AI showed it had learned from Canyon. He did come up, bashfully. Applause rang out and he smiled at the audience. Susan positioned herself between him and the exit.

"So you see," Susan said. "There's no need to be afraid of any robot overlords, because they are really only us humans, just a bit more effective."

A few knowing nods in the audience.

Susan picked up her phone and set the slide to display her device. She pressed the app and launched it.

"So of course, what is important is for us to learn just how we humans can be more effective."

Tricia gave Susan a thumbs-up, much to her guilt at the impending show.

"And so, I thought the best result would be what the Salesman AI can do for a completely hopeless sales agent. I once failed to sell bottled water. At the beach. In July." Laughter. "So I'm going to try to sell solar panels to you all, with the help of the Travelling Salesman, which Canyon trained."

Cheers erupted.

"Normally I would be wearing a headset, but for the demo, I'm going to put my phone on speaker."'

Canyon blanched. Susan tried to stay nonchalant. While everyone watched, she put the URL for Canyon's LinkedIn profile into the AI interface.

"I have some information for you before going on your call! Your target is Canyon Smith. Age 25. He is an employed white male with double degrees in computer science and psychology . . . "

The AI cheerily spat out its usual insights. Susan noticed none of them. Her hands were sweaty and shaking and she was terribly warm under the lights.

"Wait for me to do the thing!"

Coffemakers shorted out, window control units overloaded, the air conditioning units screamed like banshees. Then the power went out, and with it the presentation slides.

Susan said, with theatrical obliviousness, "What just happened there?"

She was a terrible actor and would not have fooled anyone—-had anyone been paying attention to her. But it was chaos in the office. People looked around asking what had gone wrong. In the confusion, Canyon made an attempt to bolt. Susan, finished with subtlety, stuck out a foot. Over the thump of Canyon's face meeting the floor, Susan met her boss's gaze.

"Excuse us," Tricia choked out, drawing herself up. "My team needs to check on something."

"That could have gone better," Calvin said, sipping on a celebratory espresso from his remote-controlled coffee machine, which he had newly secured against hacking attempts.

"Laugh it up," said a newly unemployed Susan. "You're going to have to cover the rent until I find something new."

"That's fine," Calvin said. "The implosion of your company created a couple months' worth of new business for me."

After the confusion died down, everyone at ReVOLTlution had eventually understood exactly what happened and who was responsible. However, that didn't stop the media from gleefully trumpeting the story from every major publication in the United States, and a handful worldwide. The damage was done. Susan's unit was shut down and almost everyone was shown the door. Half of the staff thought Susan was a hero while the other half reviled her for ruining a good thing. Susan, for her part, was glad to be out of there.

But she was going to need a paycheck, and soon.

"You know what," Susan said, "maybe you have the right idea."

"Of course I do. What idea?"

"Using my expertise to work on projects that I can control and nobody else can muck up for me," Susan said.

"Okay," Calvin said cautiously, "what does that mean?"

"I think I'll become a consultant, just like you."

Calvin choked on his coffee.

"We're going to need a bigger apartment."

AI can be Jerks: Learning from the Best

The previous story saw our protagonist start her career, and our essays will also begin with foundational material: what a lay user needs to understand what's happening in the story and how realistic it is. We'll save the more fanciful and complicated stuff for later chapters.

For the moment, we focus on a facet of artificial intelligence called machine learning. Machine learning is a small subset of artificial intelligence,[9] but unlike much of the fanciful claims of artificial intelligence prophets, it actually works *now*.

We first introduce a key component of machine learning—objective functions—that define why systems act they way they do. We then show how bad human behavior can be mimicked by these algorithms (it's already happening!). Finally, we close the chapter with a call to action: keep your computers safe!

[9] Research is organized around conferences; these conferences often form a community that is protective of its turf. While there is substantial overlap between the communities (e.g., the Neural Information Processing Systems conference), many machine learning researchers (e.g., the International Conference on Machine Learning) prefer to remain distinct from "pure" artificial intelligence.

Learning by Doing: Objective Functions

Let us begin with an essential and under-appreciated example of machine learning that simply works and makes modern life livable: spam filters. An e-mail is either spam or not; computers think in ones and zeros, so let's call spam a one and not spam a zero.

An algorithm does not just say "yes" or "no" to whether an e-mail is spam. It typically gives a score for each e-mail: the higher the score, the more confident it is the e-mail is spam. For the moment, let's assume that all of the numbers are between zero and one. This is convenient because we can think of this as a probability: a zero means that there is no chance the e-mail is spam and a one is complete confidence that it is spam.

A perfect score would be if all spam e-mails got a 1.0 and all of the good e-mails got 0.0. Both perfection and absolute certainty are unattainable goals, so we need to deal with both errors and uncertainty. So we sum all of the spam documents—how far the scores are from 1.0—and we sum all of the good e-mails—how far the scores are from 0.0. This sum is our *objective function*: a perfect score is 0.0—no mistakes—and the worst score is the number of e-mails.

Machine learning algorithms are defined by *parameters*. For example, a parameter might define what the algorithm does when it sees the word "viagra" or the phrase "I am a Nigerian prince". These parameters effectively define the algorithm; machine learning algorithms learn how to set these parameters from data, a process that can take multiple names,[10] but we'll call it "optimization".

[10] Some of the alternate names are field specific; for example, in statistics this is called inference and the objective function is typically called the likelihood. However, the mechanics are similar. For some applications, it's called a "loss function", but we'll stick with the more general term.

Optimization

Optimization is the process where machine learning uses data to find the parameters that give the best score on the objective function. Typically, the process starts with a random setting of the parameters. This will do *horribly*, but it provides a place to start: the early improvements will be easy.

The learning process will slowly change the parameters to ever so slightly improve the objective function.[11] The learning process goes document by document to slowly improve the objective function. Eventually, modern optimization techniques find a setting of the parameters that work well for this problem.[12]

[11] Mathematically, this is usually by looking at the derivative (if you remember that from your calculus class) of the objective function with respect to the parameters.

[12] The form of the objective function determines whether this answer will be the best possible or merely a "good" solution.

Let's take a concrete example. Let's say that this is the first e-mail the algorithm has seen with the phrase "special offer" and the algorithm says that it is spam with score 0.7. This is a spam e-mail, so the algorithm could do better in terms of its objective function if the parameter associated with the phrase "special offer" were more strongly associated with spam. So our optimization will push the parameter a little more in that direction so that the score is 0.75 the next time it sees a similar document. This continues until the parameters cannot get any better.

All of this seems fairly straightforward, but there is some art to creating these systems. Defining the parameters requires a bit of expertise: either hand-crafting parameters that fit the problem well or using automatic approaches that learn effective representations. Doing this well requires quite a bit of work, and often your first (and second, etc.) attempt will usually fail.

Is this Artificial Intelligence?

A reasonable question is whether this is *intelligence*. Before we get into the nuance of the question, the answer—given the systems we have discussed—is definitely **no**. But because we'll discuss things that *are* artificial intelligence later in the book, it's useful to discuss what it means for a system to be intelligent. This way, we can at least see why these systems are not intelligent.

In some ways, impressive achievements in technology are a moving target. After a few decades of living with machines that do a formerly

impressive task, we cease to be impressed by it and its comparison to humans.[13] The luster of adding machines has faded into beige quotidian boringness through the twentieth century.

Even in my own life, my excitement about halfway decent speech recognition—taking a stream of speech and transcribing it into words on a screen—has given way to annoyance at my phones' (increasingly rare) mistakes. Both of these were considered tasks that only humans could do[14] but now are so ubiquitous that half of humanity is joined at the hip to devices that do both these things.

A skeptic would say that machine learning is a mere mechanical mathematical exercise. The algorithm that decides whether an e-mail is spam or not doesn't actually understand what it's reading, and the parameters were defined by a human. While the values of the parameters come from data, there's no real learning going on.

A proponent would counter that intelligence is actually a combination of many smaller processes. To understand an e-mail, we need to understand parts of speech, syntax, discourse, and pragmatics. Each of these can be thought of as individual problems like spam classification. Each one is mechanistic, but together they form a process that can be viewed as intelligent.

For the moment, we will leave the debate there. Chapter **??** takes up the question about what it means to be intelligent and how to measure the intelligence of algorithms.

Nevertheless, machine learning has earned the reputation of being "artificial intelligence that works". Its mechanistic definition is a virtue: it is easy to see what works and what does not. Just as the objective function allows the algorithm to tune individual parameters to improve a single task, the clear definition allows researchers to tweak problems, models, and data to get better at important tasks.

Machine learning has done well at many tasks: identifying objects in a picture (**?**), playing games (**?**), answering questions (**?**), and driving cars (**?**). These tasks are impressive, but each task is a composition of smaller objective functions. As we move into "real" artificial intelligence, these objective functions will play an outsized role. While the underlying mechanism here is relatively dumb, it is worth grappling with the important questions of defining a good objective function.

In our example in Bayder, the objective function is typically aligned with business goals. For ReVOLTlution, this could be the number of new subscribers, revenue, or profits. Other companies have different objectives: many web sites optimize engagement (**?**), while logistics companies want to minimize travel time to fulfill promises made to

[13] This is not just for intelligence; "The Ballad of John Henry" celebrates a competition between a steam drill and "a man who ain't nothing but a man. . . with a hammer in [his] hand." Today, such comparisons and competitions seem silly.

[14] The term "computer" originally referred to specially trained humans proficient at high-stakes, accurate calculations. **?** documents how labor shortages during World War II increasingly a female profession. However, like the protagonists of *Hidden Figures* by Margot Lee Shetterly, the women remain in the background.

customers (**?**).

Training Data

This machine learning setup,[15] however, only works with generous training data. Modern machine learning algorithms are also notoriously data hungry. They can always do better, but they can only do better at the cost of additional training data: if you see input x, what should your output y be?

However, many tasks can be framed as mechanical calculations; they do not require the abstract reasoning or generalization that people today consider "intelligence". In contrast, they just require hundreds or thousands of examples to know what to do. Canyon is a prolific demonstrator of what (not) to do. A human could learn Canyon's tricks from a high-level description. An algorithm learns to copy.

This simple mimicry can seem sophisticated. It's couched in terms that sound sophisticated but only facilitate superficial copying (Figure 3). ReVOLTlution has a relatively simple objective function to optimize: are people buying its products or not? If sales go up, it's good. If sales go down, it's bad. Optimizing this function without any help is a difficult problem (this is why CEOs are paid so much), but ReVOLTlution doesn't have to do it alone: it learned it from Canyon.

[15] Formally, this setup is called *supervised machine learning*, where there are specific input–output examples the algorithm must replicate. Unsupervised machine learning is another active area of research, but it is much harder to see when things are working well. We talk about applications and evaluations of unsupervised machine learning in Section **??**

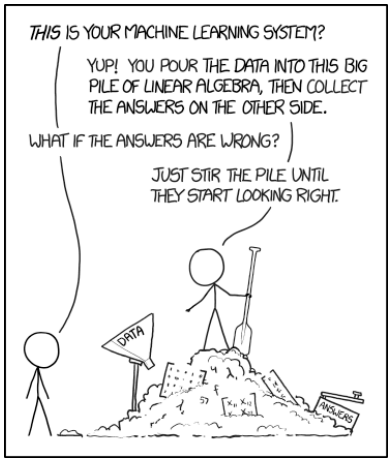

Figure 3: Much of "Machine Learning" is copying in the form of complicated equations defined by mathematical formulas using fancy words like "linear algebra", "gradients", and "non-convex optimization".

Humans are Jerks

> Everybody's a jerk. You, me, that jerk over there... That's my philosophy.
>
> – Bender Bending Rodriguez, *Futurama*: "I, Roommate"

Presumably I don't need to convince you that humans can be jerks. Instead, I want to convince you that humans can create systematic, predictable jerkitude that can be carried on, emulated, and perfected by machine learning.

Let's take a specific example of institutionalized jerkdom: redlining. Redlining is the practice that prevented mortgages from being issued to specific neighborhoods: its name came from the maps created for Chicago's Austin neighborhood (**?**). The neighborhoods excluded from lending were typically minority-majority. This was a despicable practice and even though it has ostensibly ended, it has persistent effects on familial wealth, education, voting redistricting, and segregation.

Of course, laws prohibited outright discrimination in the public sector. However, an important part of the housing market is controlled by private banks:[16] whether or not a family can get a mortgage for a house. An overly simplified version of the redlining story is that if a

[16] After the 2008 housing market implosion and the federal government increasingly becoming the lender of last resort, this fiction is a flimsy veneer. Although private (sometimes local) banks originate loans, the government is usually the ultimate lender.

family comes in and asks for a loan to buy a house in the "wrong" neighborhood, the bank will deny them the loan.

Many Americans—particularly those whose families were spared by the practice—aren't aware of the history of the practice. Unfortunately, these are often the people in positions of power when, for instance, a bank decides to create a machine learning algorithm to help decide whether or not to give a family a loan for a house.

I Learned it from You!

Machine learning algorithms learn from examples. Canyon's monkeying with the power grid, bankers denying every minority borrower, or old boys' networks who only hire from ivy league schools. Just as humans can justify their actions through ReVOLTlution's virtuous mission, preserving neighborhoods, or preserving a company's culture, machine learning algorithms can "justify" their actions with a veneer of respectability.

An algorithm has no prejudice or ill intent—it after all lacks intelligence. But that doesn't mean that it cannot recapitulate the malice that it "learns" by recreating man's inhumanity to man.[17]

[17] I was going to document that this is a reference to *The Dirge* by Robert Burns, but in documenting that, I learned that this phrase might have itself been a reference to 1673's *The Whole Duty of Man According to the Law of Nature* by von Pufendorf.

One of machine learning's strengths is that it can discover surprising correlations: e-mails sent at this hour from town are often spam, people stop talking about the future when they're ending a relationship. This strength can conceal when a machine learning algorithm is doing something it "shouldn't". Again, these algorithms are too dumb to have a morality; this judgment is an external one based on society's mores.

Take the case of redlining again. The algorithm may not know the race of loan applicants. But it might know their current address, age, occupation, and name. That is often enough to infer the race of an applicant and from that can replicate past humans' prejudice while giving current humans the excuse of "I'm just doing what the machine says".

This is not just an academic discussion. It is the subject of a subfield of machine learning focusing on fairness, accountability, transparency, and ethics (FATE). This subfield focuses on making machine learning more understandable (**?**) and quantifying unfairness (**?**).

One of the lessons learned from the redlining debacle is how to characterize these poor outcomes. It's often worthwhile to distinguish disparate impacts vs. disparate treatment (**?**).

Objective Function Mismatch

Another lesson from ReVOLTlution's choices is their objective function: optimizing sales. A die-hard skeptic of capitalism might say that

this is an inherent failing of a free market system, but these problems are also present in planned economies (**?**): but the objective function is being applied to a sprawling country instead of an individual.[18]

Objective functions can seem reasonable at first blush but fail in reality. In New York City, the RAND corporation helped optimize fire station placement and staffing based on minimizing response time. This was already recorded and seems reasonable: you cannot put out a fire until the first fire truck gets there.

However, **?** in *The Fires: How a Computer Formula, Big Ideas, and the Best of Intentions Burned Down New York City–and Determined the Future of Cities* argues that this led to poor outcomes for New York: not everyone called in fires at the same rate (or as quickly after a fire starts). Not all calls are equal: it's okay if you wait to address a false alarm in favor of a blazing inferno. Moreover, the first response isn't the whole story: what actually matters is protecting life and property, but this is harder to quantify.

[18] *Red Plenty* (**?**) is an academic/fiction mashup much like this book that realizes the the dream of a planned economy with high-tech information technology. In real life, even Trotsky recognized that central planning "is checked and, to a considerable degree, realized through the market" (**?**).

Prevention is Tough

Preventing these issues is difficult. It requires detecting malice in humans, understanding what factors lead to the malice, and then especially engineering your machine learning algorithm to avoid that malice. Organizations can deliberately obfuscate their malice or inadvertently overlook it (e.g., they don't know the history of their data).

Legal regimes may be a mechanism for preventing algorithmic discrimination. For example, the price of insurance can only be based on age. Limiting the data that an algorithm can use prevents it from finding obscure correlations but also prevents the algorithm from doing anything useful that could, say, lower overall healthcare costs.[19] These tradeoffs might be worth making, however.

Rich Caruna argues that you should not hide sensitive features to eliminate bias. Instead, he argues that you should explicitly include that data and prevent it from being part of important decisions. Other, more nuanced models, argue for adapting the "disparate impact" from the legal world into machine learning (**?**): scrub a dataset so that information about protected classes (race, gender, age, etc.) do not leak into parameters that can be used by a machine learning algorithm.

[19] This is also connected to the concept of *overfitting* in machine learning: fewer features might lead to a worse objective function score but might be better in the "real world".

Today's Cybersecurity Crisis: Tomorrow's Machine Learning Crisis

Let's focus, however, on Canyon's particular bag of tricks. ReVOLTlution's algorithm was only able to recreate the jerk by having access to machines and devices that it shouldn't.

Stuxnet is an example of how far this can go: a computer virus that specifically targeted centrifuges in Iran's nuclear program. Stuxnet

was a "precision weapon—a technique for anonymously targeting a specific system or organization without massive collateral damage" (**?**). It was able to exploit security flaws in operating systems, and—like coffee machines in Bayder—how specific everyday devices interact and interfere with each other.

However, StuxNet was crafted by experts over the course of years (one assumes, given the complexity of the source code... its origins are shrouded in secrecy). In contrast, machine learning is fast, efficient, and scalable. The same artisinal mischief that human jerks can do locally, machine learning can instantly reproduce globally.

This is not to argue for abandoning all technology. Instead, the world of Ronald D. Moore's *Battlestar Galactica* reboot provides a useful template. They needed computers for jumps and had plenty of electronics. They religiously made sure that computers were never ever linked together.[20]

However, in the early twenty-first century, every device is linked to an unsecured internet. The "Internet of things" movement has proliferated the devices that are connected to the Internet: refrigerators (**?**), cars, and medical equipment (**?**). These gadgets are catnip to Yuppies who live in places like Bayder. And let's face it, they can do some cool stuff.

But we created the devices before we made our homes and Internet ready for them. Ideally, each home should be a nice, insulated environment with controlled entry and exit. Just like you wouldn't leave your front door open all day and all night, you would not let anyone in the world potentially fiddle with the gadgets in your home.

The Internet was designed as a place where everyone could be trusted. The Internet began as a resource for researchers funded by America's Defense Advanced Research Projects Agency (DARPA)[21] to share data and communicate quickly (**?**). When the Internet had dozens of nodes where everyone knew and trusted each other, you didn't have to be paranoid.

This might be okay if the underlying devices were secure or unimportant. We're attaching everything to the Internet and not doing a good job of it. Bad guys can spy on you through your "computer, home security system, or devices such as baby monitors" (**?**), remotely hijack your car (**?**), and steal your money.

Unfortunately, neither consumers nor companies really see the problem. Highly secure devices are annoying to the consumer:[22] you need to set up passwords, firewalls, and update the firmware consistently. The ideal device is one that the user plugs in, it works, and then the user never has to think about it again. Many companies can only survive if their next product sells, and you're not going to sell many units if your product is secure (and thus annoying). Humans are bad about

[20] The downfall of the colonies was Gaius Baltar circumventing these preventative measures.

[21] In future chapters, much of the research into artificial intelligence will be directed by fictional governments for military ends. This isn't just to make the story more exciting; it also reflects reality. Much of the breakthrough research in AI has happened in America, and much that—like last century's breakthroughs in communications—has been funded by military organizations such as the Defense Advanced Research Projects Agency.

[22] This applies to commercial and residential consumers. In both cases, devices will be set up by people, often non-experts, who will not want to put up with the hassle of making sure devices are secure. This often leads to shortcuts ().

estimating low-probability events and assume that they'll never have a security issue.

Eventually, big firms might be coerced into doing something by the threat of litigation. Companies that have deep pockets and strong reputations might eventually be scared of bad press and big lawsuits into caring about security. And consumers too might be willing to put up with the annoyance of doing cybersecurity right if that's simply "what you have to do".

But as long as the Internet remains as it is, we'll be as vulnerable as the weakest link. While the free market might get some firms to make secure devices, some consumers, some companies, or some bad interaction between the two might lead to unsecure environments. And this is the real risk: a home or a business is only as strong as its weakest link. And the status quo hasn't forged many strong links; indeed, it's encouraged and rewarded the proliferation of weak links.

Securing the Beachheads

Because this is society's problem, I suspect that the only long-term solution is from society: legislation. Individual incentives aren't enough to make sure that everyone is safe, so to protect society we need to make sure people are protecting themselves.

Doing it right is difficult for many reasons, but here I'll talk about the most salient. While innovation moves quickly, legislation decidedly does not; moreover, when legislation does happen, it often has unintended consequences. Refining legislation—fixing the unintended consequences—requires input from industry (the experts), but this brings its own issues.

People working in technology broadly like to claim that things move fast—in other words, applying the early Facebook ethos to "move fast and break things" (**?**). While some of this is hype (it still takes time to get things right), there is some truth to this claim. New technologies really do allow new business models to emerge, and these often are incompatible with existing legal structures (**?**).

Copyright law has been vexed by the remix culture made possible by platforms like YouTube (**?**). Spam phone and e-mail filters have not kept up with the reality of robodialers and text-to-speech. The intrinsically international conception of the Internet generally is at odds with regional-based laws and regulation.

Around the world, government has not done a stellar job of responding to this challenge. In America, gridlock has reigned. Companies propose their own frameworks and guidelines, but they lack legal force and are ultimately voluntary. Like the over-hyped myth of Polish calvary charging German tanks with lances,[23] Europe's re-

[23] The charge at Krojanty did involve infantry, tanks, and calvary, but the Polish calvary was used effectively given the circumstances.

sponse has been mostly to fight the last war: focusing on privacy and tracking (real problems in many European regimes) rather than that of protecting systems.

Europe's well-intentioned GDPR response also highlights another critical issue: unintended consequences (**?**). Regulatory burdens can favor entrenched firms over upstarts who can't afford lawyers to oversee compliance and can prevent your citizens from having access to the rest of the world: companies might prefer to not offer service in your country if they don't like your laws.

If the unintended consequences favor particular sectors, those sectors might then try to keep those regulations entrenched and static (**?**). Then the regulations help that sector rather than the public. Whatever form legislation takes to solve these underlying issues, it must be well-crafted to address not just the emergent issues of today but to provide a framework to solve future issues. In America, the republican[24] system and the implicit veto power of minority political parties in the cooling saucer of the Senate; in Europe the federal system of European regulation with explicit or implicit veto from member states; in China, the opaque consensus building between party, province, and industry... no modern society moves quickly. The discrepancy in tempo is more acute when it comes to technology. The silver haired mandarins in Whitehall, DC, Brussels, or Beijing are often not equipped to understand the interaction between society and new technology.

[24] The "r" here is lower-case: in the American republic, voters select representatives from oddly drawn districts.

The temptation then is to abdicate the issue to more nimble agents: the military or industry. These are important parts of how artificial intelligence will enter society, but they have their own interest someone at odds with the rest of society. The next chapters will consider how martial or commercial AI could go wrong.

But what about Artificial Intelligence?

The astute reader has probably noticed that very little in this section has to do with AI *per se*. It's more about preventing the everyday jerks from screwing with our lives.

The same good digital hygiene that protects us from the common jerk also prevents the scaled-up, fast-paced epidemic of AI-driven terror.

This is why when a student comes to me who is interested in computers and has no idea what to do, I don't steer them to the bright shiny field of AI (where I work) but rather to cybersecurity. It is perhaps not as sexy, but it is what will avert the first robot apocalypse.

But frankly, these computers are pretty dumb. What happens when they get a little smarter... something that could actually be considered **intelligent**?

Recommended Reading

- **?**: *The Fires: How a Computer Formula, Big Ideas, and the Best of Intentions Burned Down New York City–and Determined the Future of Cities*
- Cosma Shalizi: In Soviet Union, Optimization Problem Solves You

General AI

- **Story:** General AI
- **Essay:** Hobbes is called in to consult on the deployment of a general AI that a military contractor has developed. It breaks out of its sandbox, causing physical havoc with military equipment. A specialized AI designed to contain the general AI is surgically deployed to resolve the problem.

Resource-Constrained AI

- **Story:** Resource-Constrained AI
- **Essay:** Hobbes is comfortably faculty at University when her past history with Entrench comes back to haunt her: a rogue actor has found an old copy and using it to terrorize a small Canadian town. Hobbes works with the local authorities to deploy their specialized AI to counteract the threat.

AI's Role in Social Interactions

- **Story:** AI's Role in Social Interactions
- **Essay:** Hobbes finds herself drawn into a political scandal when an AI agent begins influencing public opinion against her and her research. She fights back the only way she knows how: with AI.

Cashing Out

- **Story:** Cashing Out
- **Essay:** Fed up by politics and the politics of academia, Hobbes goes off to make a pile of money at StockOverflow, turning her AI smarts to Wall Street. It's harder than she thinks.

The Silent Revolution

- **Story:** The Silent Revolution
- **Essay:** Having saved the world multiple times, Hobbes retires to spend more time with her grandchildren and sees their relationship with AI.