O'REILLY®

Third Edition

Inside Cyber Warfare

Mapping the Cyber Underworld

Jeffrey Caruso
Foreword by Dan Geer

O'REILLY®

Inside Cyber Warfare

Get a fascinating and disturbing look into how state and nonstate actors throughout the world use cyber attacks to gain military, political, and economic advantages. In the third edition of this book, cyber warfare researcher Jeffrey Caruso explores the latest advances in cyber espionage and warfare that have emerged on the battlefields of Ukraine and the Middle East, including cyber attacks that result in the physical destruction of the target and the pairing of cognitive with maneuver warfare.

Inside Cyber Warfare features an exclusive deep dive into the wartime operations of an offensive cyber unit of Ukraine's Ministry of Defense as it works to defend the nation against Russian forces, particularly since the 2022 invasion:

- **See what happened when a Ukrainian cyber and special operations team worked together to destroy a secret missile laboratory**
- **Explore the legal status of cyber warfare and civilian hackers**
- **Discover how a cyber team with little money and limited resources learned to create fire from the manipulation of code in automated systems**
- **Distinguish reality from fiction regarding AI safety and existential risk**
- **Learn new strategies for keeping you and your loved ones safe in an increasingly complex and insecure world**

> "*Inside Cyber Warfare* explores the significance and vulnerabilities of our software-based economies. Jeff Caruso's fast-paced and far-ranging narrative explains the complexity behind cyber warfare in Ukraine, how cyber attacks can lead to real kinetic effects in the physical world, and why we will not be able to defend adequately against cyber attacks until the software industry is held to higher standards of accountability."
>
> —**Carmen A. Medina**
> Retired deputy director for intelligence, Central Intelligence Agency

Jeffrey Caruso is a US Coast Guard veteran and has worked in the cybersecurity and cyber warfare field since 2006. He has provided cyber intelligence briefings to the CIA's Open Source Center, the DIA, the FBI, and the Chief of Naval Operations Strategic Study Group. He has also been a frequent lecturer at the US Air Force Institute of Technology and the US Army War College.

SECURITY

linkedin.com/company/oreilly-media
youtube.com/oreillymedia

US $55.99 CAN $69.99
ISBN: 978-1-098-13851-6

THIRD EDITION

Inside Cyber Warfare

Mapping the Cyber Underworld

Jeffrey Caruso

Foreword by Dan Geer

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

Inside Cyber Warfare
by Jeffrey Caruso

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisitions Editor: Simina Calin
Development Editor: Virginia Wilson
Production Editor: Clare Laylock
Copyeditor: J.M. Olejarz
Proofreader: Krsta Technology Solutions

Indexer: BIM Creatives, LLC
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Kate Dullea

December 2009: First Edition
December 2011: Second Edition
September 2024: Third Edition

Revision History for the Third Edition
2024-09-16: First Release

See *http://oreilly.com/catalog/errata.csp?isbn=9781098138516* for release details.

978-1-098-13851-6

[LSI]

This book is dedicated to my brothers Max (aka Nomad) and Dima (aka Apostle) in Ukraine. Your ingenuity and bravery in war have been and remain a constant source of inspiration for me.

Table of Contents

Foreword

YOU ARE IN A MAZE OF TWISTY LITTLE PASSAGES, ALL ALIKE.
—Adventure (1976 video game)

Things are seldom what they seem, Skim milk masquerades as cream....
—Gilbert & Sullivan, H.M.S. Pinafore

This book is a guide—not a guide for those twisty little passages, not a guide for detecting skim milk in the crockery, not a tour guide, more like a wilderness guide.

We are near an inflection point, a place where a curve changes from one regime to another. A declaration of inflection is a claim more than an observation—it is very nearly impossible to confirm that you are at a moment of inflection except in retrospect. Any such announcement can only be early or late. Your informed risk tolerance determines your preference for being early or being late.

The inflection described in this book is that of moving from a regime where cybersecurity capabilities are contributory yet ancillary to a regime where they are primary, from a regime where cybersecurity technique operates at the perimeter without real power to determine outcomes to a regime where cybersecurity technique is the broadly dominant power, a metamorphosis where cybersecurity passes from "useful" through "necessary" on to "sufficient."

An inflection begins in specific edge cases first; Jeff covers that in the first chapter with the authority of a John von Neumann. Once begun, inflections spread ever faster and in every direction, the central feature of self-reinforcing processes. As the spread picks up speed, the inflection point is passed. The inflection point which Jeff declares is occurring in the here-and-now, what with the skies of the Donbas dense with drones, nearly every hospital under ransomware pressure, and everything in between. The sharpest bend in the curve may be just ahead, depending on whether the autonomy granted to AI artifacts includes their ability to reproduce. Regardless of whether it is just past, just now, or just over the horizon, it's proximal.

Cyber conflict up to and including cyber war is a reality. Ever more precise application of ever more compute power guarantees that every aspect of computing's inherent dual-use nature will be explored. The rate of advance will be heady—make that *is* heady. A world of connected endpoints, a swiftly declining fraction of which will be human, is the regime waiting on the downstream side of this inflection.

Just as militaries deploy massed artillery wars and psyops, cyber conflict comes in all grades. That some particular conflict is underway may be obvious or it may be hidden. Much of the weaponry will be repurposed civilian capabilities, it being somewhat irrelevant if the repurposing came from an authoritarian state's predilection for intervention or from a liberal state's unwillingness to put its foot down. If the explainability problem in AI is as unsolvable as it looks, it won't matter that there is no human in the loop.

Catastrophizing? No. We know so much now that there is no shortage of things to do or that we could be doing. This book is a guide to a way of thinking usefully, looking backward so that you see forward.

These are perilous times; to be a master of your fate you must study conflicts up to and including warfare if you are to deliver a lasting peace. Looking back is neither optional nor a new idea, and what John Adams wrote to Abigail Adams in 1780 rings true even yet:

> I must study Politicks and War that my sons may have liberty to study Mathematicks and Philosophy. My sons ought to study Mathematicks and Philosophy, Geography, natural History, Naval Architecture, navigation, Commerce and Agriculture, in order to give their Children a right to study Painting, Poetry, Musick, Architecture, Statuary, Tapestry and Porcelaine.

Study well.

— Daniel E. Geer, Jr., Sc.D.

Preface

The first edition of this book focused on the use of cyber warfare by Russia during its invasion of Georgia on August 1, 2008. The war lasted 12 days, but it colored what we thought we knew about cyber warfare and the Russian playbook for more than a decade—13 and a half years, to be exact—until February 24, 2022, when Russia-backed forces invaded Ukraine. This entire third edition was researched and written during wartime, and took two years and four months to complete. And as I write these last few words on June 20, 2024, there are no signs that the war will be ending anytime soon.

There are some major differences between the first and second editions and this one, the biggest being that this edition contains 100% fresh content. If you've read the book's prior versions, this will be an entirely new experience for you.

As a researcher in 2008, my access to details about Russia's cyber warfare operations was limited to the activities of the StopGeorgia.ru website and forum, plus various open sources. For this book, I had an exclusive over-the-shoulder look at a number of secretive offensive cyber operations (OCOs), aimed at critical Russian infrastructure, run by a special unit in Ukraine's military, which I cover in Chapter 6.[1]

Likewise, Chapter 5, "The New Enmeshed War Strategy", was largely informed by my efforts to support the work of Col. Andrew Milburn's volunteer rescue and training organization, the Mozart Group, which was operating on the front lines of the war in Ukraine. I had a bird's-eye view of the impact of information and kinetic warfare campaigns run by the late Wagner Group founder Yevgeny Prigozhin against Milburn and his team. I also witnessed the impact of local corrupt business leaders who appeared to be supporting the work of the Mozart Group while leveraging it for their own selfish interests. That latter aspect isn't addressed in this book; more about it is, however, available for reading at the Inside Cyber Warfare newsletter.

1 See my article for O'Reilly Radar, "D-Day in Kyiv", for more information about this.

Here's a quick overview of what's in the book.

This book covers the topic of war and includes some graphic descriptions of violence.

The Foreword was written by my friend Dan Geer, the longtime chief information security officer for the CIA-founded venture capital firm In-Q-Tel and a true icon in the world of information security. Back in 2003, Dan was fired by the consulting company where he was employed for cowriting a paper that called Microsoft a threat to national security. Twenty-one years later, DHS did exactly the same thing.[2]

In Chapter 1, "How Did We Get Here?", I begin with mathematician John von Neumann's prediction about high-speed computing being a "monster whose existence is going to change history, providing there is any history left." Almost every piece of critical infrastructure in the world is governed by software, a fundamentally flawed and unregulated system that is so awry that the more we spend on cybersecurity, the more incidents there are.

In Chapter 2, "Who Did It?", I address the risks of private-sector attribution of cyber attacks due to commercial incentives and a lack of accountability.

In Chapter 3, "Establishing Corporate Accountability", I show what happens historically when industries are left to regulate themselves (they don't do it), and what is typically required to bring regulation to an industry (the media and general public put relentless pressure on government to act due to an unacceptable level of human lives lost).

In Chapter 4, "The Legal Status of Cyber Warfare", I look at the potential repercussions for civilian hackers to engage in offensive operations against a nation-state at war, and provide a tool that will tell you if you're at risk of being considered an enemy combatant.

In Chapter 5, "The New Enmeshed War Strategy", I show how traditional warfare has changed to leverage our reliance on an internet-based information infrastructure that powers and tracks everything we do on a minute-to-minute basis.

2 "Cyber Safety Review Board Releases Report on Microsoft Online Exchange Incident from Summer 2023," April 20, 2024, The U.S. Department of Homeland Security

Chapter 6, "Cyber Attacks with Kinetic Effects", you'll read examples of recent real-world cyber attacks that have generated kinetic effects including explosions, fires, and in some cases loss of life, with nothing more than internet access to the target's automated control system(s).

In Chapter 7, "AI", I explore innovations in artificial intelligence and detail the most pressing present risks and the harms that have resulted from them. I also explore future risks and provide recommendations for prevention and mitigation.

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic
: Indicates new terms, URLs, email addresses, filenames, and file extensions.

This element signifies a general note.

This element indicates a warning or caution.

O'Reilly Online Learning

For more than 40 years, *O'Reilly Media* has provided technology and business training, knowledge, and insight to help companies succeed.

Our unique network of experts and innovators share their knowledge and expertise through books, articles, and our online learning platform. O'Reilly's online learning platform gives you on-demand access to live training courses, in-depth learning paths, interactive coding environments, and a vast collection of text and video from O'Reilly and 200+ other publishers. For more information, visit *https://oreilly.com*.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-889-8969 (in the United States or Canada)
707-827-7019 (international or local)
707-829-0104 (fax)
support@oreilly.com
https://oreilly.com/about/contact.html

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at *https://oreil.ly/InsideCyberWarfare3e*.

For news and information about our books and courses, visit *https://oreilly.com*.

Find us on LinkedIn: *https://linkedin.com/company/oreilly-media*.

Watch us on YouTube: *https://youtube.com/oreillymedia*.

Acknowledgments

I couldn't have written this book without the insights that I received from my brave friends in Ukraine and dozens of experts across multiple disciplines, including Dan Geer, Sc.D., who took time away from many more important things to write the Foreword for this edition, as well as Matt Georgy; Marcus Ranum; H. D. Moore; Hector Monsegur; David Thorstad, Ph.D.; Ellie Pavlick, Ph.D.; Suhail Balasinor; Olav Lysne, Ph.D.; Jorge Reyes; Alex Urbelis, J.D.; Drinor Selmanaj; Kathryn Ballentine Shepherd, J.D.; Emilio Iasiello; Anil Sood; Mukund Sarma; Boldizsar Bencsath, Ph.D.; and Col. Andrew Milburn (United States Marine Corps, retired). I'm sure there are others who I've forgotten to mention, but thank you so much for your time and assistance.

Also, thank you to my patient and amazingly insightful editor, Virginia Wilson, and the entire team at O'Reilly for their assistance and guidance as I struggled with winnowing down a mountain of material into a book that I hope people will find both interesting and informative.

Lastly, and most importantly, thank you to my astoundingly patient spouse, Lilly Andersen, for being so understanding and supportive in the midst of the constant uncertainty that comes with being married to a writer. You're living proof that angels walk among us.

CHAPTER 1
How Did We Get Here?

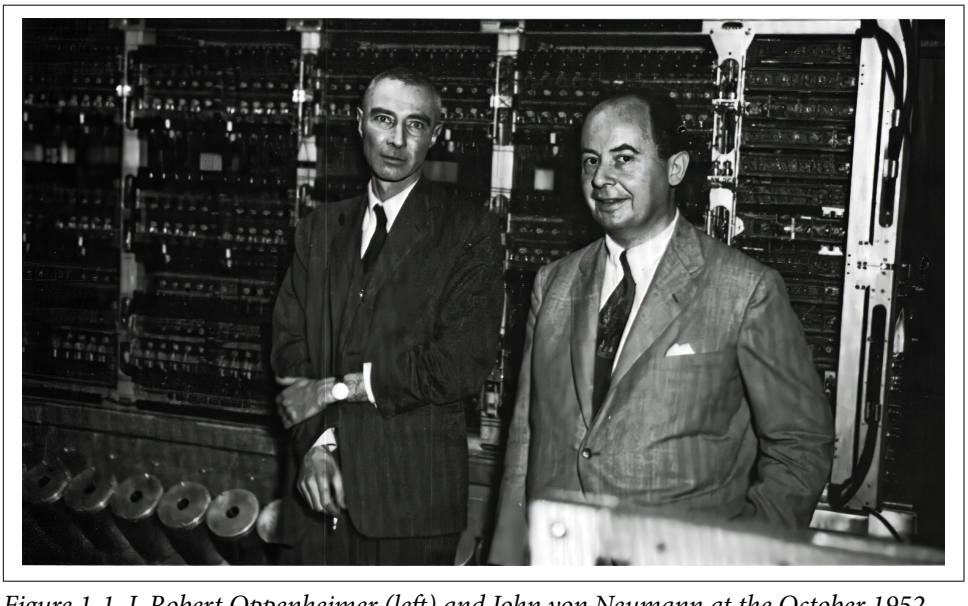


Figure 1-1. J. Robert Oppenheimer (left) and John von Neumann at the October 1952 dedication of the computer built for the Institute for Advanced Study. Oppenheimer, who was head of the Los Alamos National Laboratory during World War II, became the institute's director in 1947.[1]

1 A version of the computer in this image that was built for the Institute of Advanced Study was later built for Los Alamos. Both utilized the von Neumann architecture. The image is in the public domain.

What we are creating now is a monster whose existence is going to change history, provided there is any history left.

—John von Neumann[2]

In this chapter, I depict the cybersecurity industry as a super ouroboros: a snake that not only eats its own tail but also grows larger with every bite. Red (offensive) and blue (defensive) teams have been perpetually squaring off and creating new products and services for roughly 25 years while the customer (the technology enterprise, financial institution, hospital, power station, automobile manufacturer, etc.) pays the price.

This chapter will show you that there has never been such a thing as a "secure" or "healthy" network, from the first high-speed computer, known as MANIAC, to the present time; that the business of exposing vulnerabilities only makes the attacker's job easier; and that, when used in medicine, the model of finding new ways to attack a network, advising the company about it, and then publishing your findings, which lets bad actors use that information, would be not only illegal but a crime against humanity.

By the time you're finished you'll have learned that software programming is inherently insecure, that the multibillion-dollar cybersecurity industry exploits that fact, and that it pays much better to play offense than defense.

von Neumann's Monster

There is a marker in the history of civilization at which our future security became more perilous than ever before. Logic and math combined to form a new type of computing that enabled the creation of a thermonuclear weapon, a weapon so powerful that if used today it would result in an estimated two billion people dying if a nuclear war happened between India and Pakistan, and five billion people dying if the war was between the United States and Russia, due to the global effects that radiation would have on crops, marine fisheries, and livestock.[3]

2 William Poundstone, "Unleashing the Power", review of *Turing's Cathedral,* by George Dyson, *New York Times,* May 4, 2012.

3 The two atomic bombs dropped on Japan during World War II were fission bombs. The thermonuclear device that von Neumann was helping to create was a fusion bomb, which is a more complex nuclear reaction resulting in a much more powerful explosion. More information on the differences can be found here; Lili Xia et al., "Global Food Insecurity and Famine from Reduced Crop, Marine Fishery and Livestock Production Due to Climate Disruption from Nuclear War Soot Injection", *Nature Food* 3 (2022): 586–596.

Many would argue that the successful detonation of the first ever thermonuclear device in 1952 would certainly qualify as that marker, but the risk of such a war happening is extremely low thanks to the doctrine of mutually assured destruction (MAD).[4] MAD relies on the theory of rational deterrence, which says that if two opponents each have the capability of using nuclear weapons, and that both players would die if either player used it, then neither will use it.

However, John von Neumann wasn't nearly as worried about the bomb that he helped build as he was about the high-speed computer that he invented in order to mathematically prove that such a bomb was possible. The "monster" in von Neumann's quote at the start of this chapter wasn't the bomb. It was his stored program architecture code that ran the MANIAC computer at Los Alamos National Laboratory, an architecture that was inspired by his former student Alan Turing's paper "On Computable Numbers."[5] Stored-program architecture went on to become the basis for digital computing worldwide.

With thermonuclear war, while the potential for harm was astronomical, the risk of it happening was very low thanks to MAD. Computing, on the other hand, was a seductive charmer that only a select few understood fully in the beginning. As computing became more pervasive and complex, no one knew more than their specialty. The average person can assess risk when it comes to things that they understand, but no one completely grasps how computing works, even the experts, and so our collective risk has grown to the point where online sabotage, extortion, theft, and espionage are unstoppable. Cyber insurance companies are now worried about claims so large that they could result in bankrupting the industry.

In order to understand just how unsafe the world is today because of the perils inherent in software and hardware, we need to return to Los Alamos and the MANIAC computer (see Figure 1-2).

4 Alan J. Parrington, "Mutually Assured Destruction Revisited: Strategic Doctrine in Question," *Airpower Journal* 11, no. 4 (1997), 4–19.

5 Lily Rothman, "How *Time* Explained the Way Computers Work", *Time*, May 28, 2015.

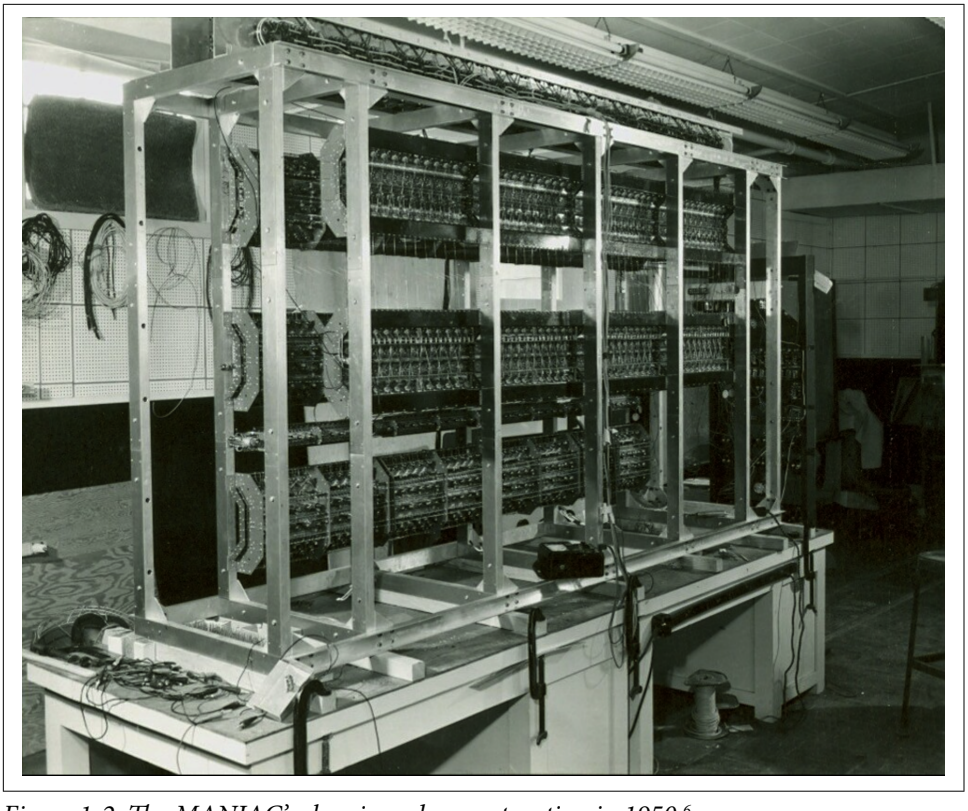


Figure 1-2. The MANIAC's chassis under construction in 1950.[6]

MANIAC's primary purpose was to run mathematical models to test the thermonuclear process of a hydrogen bomb explosion. It successfully achieved that with a single mathematical operation that ran nonstop for sixty days. Then on November 1, 1952, IVY MIKE, the code name for the world's first thermonuclear device, had a successful detonation on Elugelab, an island that was part of the Enewetak Atoll of the Marshall Islands.

6 Taken at Los Alamos National Laboratory 1950. Unless otherwise indicated, this information has been authored by an employee or employees of the Los Alamos National Security, LLC (LANS), operator of the Los Alamos National Laboratory under Contract No. DE-AC52-06NA25396 with the U.S. Department of Energy. The U.S. Government has rights to use, reproduce, and distribute this information. The public may copy and use this information without charge, provided that this Notice and any statement of authorship are reproduced on all copies. Neither the Government nor LANS makes any warranty, express or implied, or assumes any liability or responsibility for the use of this information.

MANIAC also went on to become the first computer to beat a human in a modified game of chess; there were no bishops because of the limitations of the machine. MANIAC's entire memory storage was five kilobytes (about the size of a short email), sitting within a six-foot-by-eight-foot beast weighing one thousand pounds. MANIAC went through three iterations between 1952 and 1965.

New challenges swiftly arose in the area of software because programs were haphazardly written without any formal rules or structure, and it all came to a head in 1968 at the NATO Software Engineering Conferences at Garmisch-Partenkirchen, a resort in the Bavarian Alps (see Figure 1-3). It should come as no surprise that there were serious differences of opinion among the attendees, all of whom came from the elite universities in their respective countries as well as Bell Labs and IBM.

Figure 1-3. An unidentified photographer captured this image at the NATO Software Engineering Conference in 1968.[7]

One side viewed the use of the phrase "software crisis" as unwarranted and unnecessarily dramatic. Sure, there were some supply problems, and there were "certain classes of systems that are beyond our capabilities," but—their view went—we can handle payroll and sort routines perfectly well!

7 "The Birth of Software Engineering" posted on Github.

Douglas Ross, a pioneer in computer-aided design at MIT, and one who believed there was a serious software programming crisis at hand, had a perfectly succinct response to the critics: "It makes no difference if my legs, arms, brain, and digestive tract are in fine working condition if I am at the moment suffering from a heart attack. I am still very much in a crisis."[8]

Around that same time, the US Department of Defense was struggling with how to do secure programming on a shared server provided by IBM. A task force, chaired by Willis Howard Ware from the RAND Corporation and including representatives from the US's National Security Agency, Central Intelligence Agency, and Department of Defense, as well as academia, spent two years on the problem. The Ware Task Force produced a report in 1970 that advocated for a system of document flags representing the four levels of clearance: Unclassified, Confidential, Secret, and Top Secret. They would also program a rule that they called "No Read Up," basically what we call "least privilege access" today.

Unfortunately, the Ware report concluded that it would be very difficult to secure such a system; one would have to delineate each method whereby the No Read Up rule could be defeated, and then create a flowchart that solves the security problem for each method of compromise. And that process would be next to impossible because, according to Ware, "the operating systems were Swiss cheese in terms of security loopholes."[9]

Another two years went by with little to no action, and then Major Roger Schell of the US Air Force's Electronic Systems division commissioned a new study. According to Schell, the Ware report was all doom and gloom with no solutions: "You've got all these problems. They offer almost nothing by way of solutions. The US Air Force wants to know solutions."[10]

Is Software Killing People?

As the years went by, and computers were called upon to perform more and more complex tasks, the question of software's dependability became more and more critical. In 1969, J. C. R. Licklider of the Defense Department's Advanced Research Project Agency (today known as DARPA) was one of the contrarian voices decrying the use of software to run missile defense systems that protected American cities. It was a "potentially hideous folly," Licklider said, because all software contains bugs.[11]

8 Donald A. MacKenzie, *Mechanizing Proof* (Cambridge, MA: MIT Press, 2001), 36.

9 Ibid, 160.

10 Ibid, 160.

11 Ibid, 299.

In 1994, Donald MacKenzie, author of *Mechanizing Proof* (published in 2001 by MIT Press), was determined to investigate just how big a problem software with programming flaws was. Were people actually dying because of it? Were the fears overblown or justified? He set about to determine the number of deaths in computer-related accidents worldwide up until 1992. As he found:

> The resultant data set contained 1,100 deaths. Over 90 percent of those deaths were caused by faulty human-computer interaction (often the result of poorly designed interfaces or of organizational failings as much as mistakes by individuals).[12] Physical faults such as electromagnetic interference were implicated in a further 4 percent of deaths. Software bugs caused no more than 3 percent, or thirty, deaths: two from a radiation-therapy machine whose software control system contained design faults, and twenty-eight from faulty software in the Patriot antimissile system that caused a failed interception in the 1991 Gulf War.[13]

MacKenzie's research paper on computer-related deaths was published separately in the journal *Science and Public Safety* in 1994. In that paper he raised the problem of underreporting of less catastrophic accidents, "such as industrial accidents involving robots or other forms of computer-controlled automated plants."[14]

For example, in 2001, five patients of Panama's National Oncological Institute died due to radiation overexposure that resulted from software flaws in the radiation treatment planning software.[15]

The acceleration of the transition to the electronic health record (EHR) in health care from 2011 to the present brought about a shocking number of computer-related injuries and deaths in spite of being a notoriously underreported sector. This doesn't diminish the fact that EHRs have also had their successes; however, unless and until hospitals and clinics are required to report the cases where EHRs caused harm, patients should be advised of the risks, known and unknown, associated with them.

The adoption of EHRs accelerated under US Presidents Bush and Obama. President Bush wrote an executive order (EO) that created the Office of the National Coordinator for Health Information Technology with the goal of achieving nationwide adoption of EHRs by 2014. President Obama funded the rollout by incorporating EHRs into his American Recovery and Reinvestment Act of 2009, which paid health care

12 Computer-related accidents may include bugs, malware, a poorly designed user interface, and failure.

13 *Mechanizing Proof*, 300.

14 Donald MacKenzie, "Computer-Related Accidental Death: An Empirical Exploration", *Science and Public Policy* 21, 4 (August 1994): 233–248; possibly the worst example of software flaws resulting in fatalities were the two Boeing 737 crashes (in October 2018 and February 2019, which) that resulted in the deaths of 346 people.

15 "FDA Statement on Radiation Overexposures in Panama", US Food and Drug Administration, content current as of June 13, 2019.

providers more money if they implemented EHR systems and associated requirements.[16] It also incentivized providers with higher payments for "meaningful use," which goes beyond simple record-keeping functions to improving quality of care. The Centers for Medicare and Medicaid Services had a three-stage process to encourage hospitals to expand their use of EHR technology and receive incentive payments for doing so.[17]

Stage one
: Capture basic health information and provide printed copies to patients after a visit.

Stage two
: Improve clinical processes and quality of care; improve information sharing.

Stage three
: Improve health outcomes.

These incentives prompted a gold-rush mentality among software companies that wanted to take advantage of the federal dollars available to support this initiative. And hospitals that took Medicare and Medicaid payments were mandated to transition to EHR by a certain deadline if they wanted to keep getting paid by the federal government. The combination of greed and speed led to terrible results.

Researchers at the University of Illinois at Urbana-Champaign, MIT, and Rush University Medical Center tackled the subject of accidents and deaths by robotic surgeries between 2000 and 2013 as reported to the FDA MAUDE database, which houses reports on medical device usage. The researchers' goal was "to determine the frequency, causes, and patient impact of adverse events in robotic procedures across different surgical specialties."[18] Their results showed that "144 deaths (1.4% of the 10,624 reports), 1,391 patient injuries (13.1%), and 8,061 device malfunctions (75.9%) occurred during the study period." The researchers believed that these numbers were low due to underreporting.

16 Jim Atheron, "Development of the Electronic Health Record", Virtual Mentor, *American Medical Association Journal of Ethics* 13, no. 3 (March 2011): 186–189.

17 "Stages of Promoting Interoperability Programs: First Year Demonstrating Meaningful Use", Department of Health and Human Services, accessed 2022.

18 Homa Alemzadeh et al., "Adverse Events in Robotic Surgery: A Retrospective Study of 14 Years of FDA Data," *PLoS One* 11, no. 4 (2016): e0151470.

A report by Kaiser Health News and *Fortune* magazine entitled "Death by a Thousand Clicks" interviewed more than one hundred physicians, patients, IT experts and administrators, health policy leaders, attorneys, top government officials, and representatives at more than a half-dozen EHR vendors, including the CEOs of two of the companies.[19] Its result:

> Our investigation found alarming reports of patient deaths, serious injuries and near misses—thousands of them—tied to software glitches, user errors or other flaws that have piled up, largely unseen, in various government-funded and private repositories.
>
> Compounding the problem are entrenched secrecy policies that continue to keep software failures out of public view. EHR vendors often impose contractual "gag clauses" that discourage buyers from speaking out about safety issues and disastrous software installations—though some customers have taken to the courts to air their grievances. Plaintiffs, moreover, say hospitals often fight to withhold records from injured patients or their families. Indeed, two doctors who spoke candidly about the problems they faced with EHRs later asked that their names not be used, adding that they were forbidden by their health care organizations to talk. Says Assistant U.S. Attorney Foster, the EHR vendors "are protected by a shield of silence."

I spoke with a patient safety researcher in Switzerland who confirmed that it wasn't just a US problem. He told me that there's no formal tracking of EHR's safety record because (a) it's not required by law, and (b) at least one hospital administrator told him, "we don't want to know the safety performance of our system because what would that mean?" The administrator further went on to say that safety researchers should be more diplomatic about what they say because it might result in vendors leaving the market. "Isn't that what we want for unsafe systems?" the researcher asked him somewhat incredulously.

At a 2017 meeting with health care leaders in Washington, former Vice President Joe Biden railed against the infuriating challenge of getting his son Beau's medical records from one hospital to another. "I was stunned when my son for a year was battling stage 4 glioblastoma," said Biden. "I couldn't get his records. I'm the vice president of the United States of America…It was an absolute nightmare. It was ridiculous, absolutely ridiculous, that we're in that circumstance.[20]

19 Fred Schulte and Erika Fry, "Death by 1,000 Clicks: Where Electronic Health Records Went Wrong", *Fortune*, March 18, 2019.

20 Ibid.

The impact of poorly designed, fault-prone EHR software has predictably struck our most vulnerable population—children. A study published in the medical journal *Patient Safety* in November 2018 analyzed nine thousand pediatric patient safety reports, made in the period 2012–2017, from three different health care institutions that were likely related to EHR use. "Of the 9,000 reports," it found, "3,243 (36 percent) had a usability issue that contributed to the medication event, and 609 (18.8 percent) of the 3,243 might have resulted in patient harm."[21]

The averseness of the health care profession to tracking usability and patient harm (potential and realized) is evident by the fact that only one state—Pennsylvania—has required acute health care facilities to report patient safety events that either could have harmed or did harm the patient.[22] While other states have reporting requirements, they're narrowly focused and no one really wants to track that data. In that respect, it's very similar to the reluctance of companies to report a cybersecurity incident—and if they do report it, it's to underplay the seriousness. The same corporate reluctance applies to fixing vulnerabilities that have been discovered by security researchers.

To Disclose, or Not to Disclose, or to Responsibly Disclose

A vulnerability is a flaw, glitch, or weakness in the coding of a software product that may be exploited by an attacker.[23]

Vulnerability disclosure was once hotly debated with almost religious ferocity. On the one side were the skeptics who painstakingly tried to explain that when you build on top of a fundamentally insecure base, you are perpetuating insecurity. Or to frame it another way, if you build a perfect house on sand, its collapse is inevitable. If you publicly announce that the house is built on sand, and a third party announces that there is one point of the structure where, if pushed at just the right angle with *n* amount of pressure, you can greatly accelerate that process, you're making it really easy for others to cause problems that otherwise they would have to discover on their own.

On the other side were the proponents who, while acknowledging that nothing will ever be 100% secure, say that public disclosure is a must because back when we didn't have it, companies that were informed about vulnerabilities in their software didn't bother fixing them. It wasn't until researchers went public with their findings, with

21 Raj M. Ratwani et al., "Identifying Electronic Health Record Usability and Safety Challenges in Pediatric Settings," *Health Affairs (Project Hope)* 37, no. 11 (2018): 1752–1759, doi:10.1377/hlthaff.2018.0699.

22 Shawn Kepner and Rebecca Jones, "2020 Pennsylvania Patient Safety Reporting: An Analysis of Serious Events and Incidents from the Nation's Largest Event Reporting Database", *Patient Safety* 3, no. 2 (2021): 6–21.

23 Kelley Dempsey et al., "Automation Support for Security Control Assessments: Software Vulnerability Management", National Institute of Standards and Technology, report 8011, no. 4 (April 2020).

the help of journalists who gave them headlines, that companies would spend the money on developing patches. Technologist Bruce Schneier summed it up in his essay "Damned Good Idea" when he wrote: "Public scrutiny is the only reliable way to improve security, while secrecy only makes us less secure."[24]

There are numerous examples of companies that were notified about problems in their software products, didn't patch or update them, and suffered a breach. A few of those appear below.

Sony PlayStation Network

The Sony PlayStation Network had a month-long series of data breaches starting around April 17, 2011, that resulted in the compromise of personal data for 77 million Sony PlayStation Network users. It was large enough to attract the attention of Representative Mary Bono Mack (R-CA), chair of the congressional Subcommittee on Commerce, Manufacturing, and Trade, who sent a formal letter to Kazuo Hirai, Sony's executive deputy president, with a request to appear before the committee and answer 13 questions. Hirai declined to appear personally but responded to the questions via a letter.

The committee did take testimony from four individuals about the matter, one of whom was Dr. Gene Spafford, a professor of computer science at Purdue University. In his testimony, Spafford said that Sony knew that it was using outdated Apache web server software and no firewall because those issues had been reported several months earlier in an online forum that Sony employees monitored.

Equifax

> *Any complex software contains flaws.*

That's an excerpt from the official statement of the Apache Software Foundation on the 2017 Equifax data breach, and while it seems simplistic, it's a necessary reminder that there's no such thing as 100% secure. The only question defenders have to answer is how easy they want to make it for the attackers to get in.

In the case of the Equifax breach, the company made it really easy by not immediately installing a software vulnerability patch that Apache had written and pushed out the same day that Apache had been notified of it—March 7, 2017. Because Equifax never installed the patch, the personal credit information of 143 million people had been compromised.

24 Bruce Schneier, "Schneier: Full Disclosure of Security Vulnerabilities a 'Damned Good Idea'", Schneier on Security, January 2007.

If the vulnerability had not been publicly disclosed, attackers would have had to discover the software flaw on their own. But that's not how the system of vulnerability disclosure works.

The Apache Software Foundation issued a press release that included the following paragraph:

> Since vulnerability detection and exploitation has become a professional business, it is and always will be likely that attacks will occur even before we fully disclose the attack vectors, by reverse engineering the code that fixes the vulnerability in question or by scanning for yet unknown vulnerabilities.

Twitter

The worst example of corporate negligence when it comes to poor cybersecurity practices has to be Twitter.

An explosive 2022 whistleblower complaint that was leaked to the *Washington Post* revealed the egregious state of Twitter's cybersecurity:

- More than half of Twitter's 500,000 servers were running operating systems so out of date that many did not support basic privacy and security features and lacked vendor support.
- More than a quarter of the around 10,000 employee computers had software updates disabled.
- More than one major security incident occurred every week, involving millions of user accounts.
- Every engineer had a full copy of Twitter's proprietary source code on their laptop instead of in the cloud or in a data center; laptops had poor to no security configurations including running endpoint security software that had been discontinued by the vendor.

This wasn't the first time Twitter had cybersecurity issues. After the company suffered a serious breach and compromise of user data in 2009, the Federal Trade Commission put Twitter, then a private company, under a consent decree in March 2011 with a list of requirements that it must perform to maintain an acceptable level of security. One of those requirements specified the appointment of a "qualified, objective, independent third-party professional" to do several tasks:

- Set forth the specific administrative, technical, and physical safeguards that respondent has implemented and maintained during the reporting period.
- Explain how such safeguards are appropriate to respondent's size and complexity, the nature and scope of respondent's activities, and the sensitivity of the nonpublic personal information collected from or about consumers.

- Explain how the safeguards that have been implemented meet or exceed the protections required.
- Certify that respondent's security program is operating with sufficient effectiveness to provide reasonable assurance to protect the security, privacy, confidentiality, and integrity of nonpublic consumer information and that the program has so operated throughout the reporting period.

The worst part of the Twitter (now X) fiasco is that no one at the company, from former CEO Jack Dorsey on down, seemed to care about the fact that a global telecommunications platform was so vulnerable to compromise, disinformation, and data theft. It takes a certain amount of skill, time, and money to break into a well-defended network even with the inherent weaknesses found in computer programming. But it becomes exponentially easier when security controls are carelessly applied or missing altogether.

Problematic Reporting of Exploits and Vulnerabilities

Speaking of making things easier for an attacker, that's precisely what the controversial and firmly entrenched practice of vulnerability disclosure has done by incentivizing security researchers to discover and reveal not only the flaws they found in a product's programming but also how to exploit them as well. The path to exploitation, also called the proof of concept, needs to be demonstrated so that the owner of the code can see that the flaw is genuine and how hard or easy it is to exploit. The worst-case scenario is an easy-to-exploit vulnerability that could cause serious harm to the users of the product.

The Cybersecurity and Infrastructure Security Agency (CISA) has created a catalog of known vulnerabilities that threat actors have been exploiting to make it easier for the federal government's cyber defenders to conduct vulnerability management. Prioritize the vulnerability by how exploitable it is, test the patch, then execute the fix.

CISA has authority over every federal branch, department, and agency that isn't part of the intelligence community or the Department of Defense and has issued a binding operational directive (BOD 22-01, "Reducing the Significant Risk of Known Exploitable Vulnerabilities") that requires those departments and agencies to do the following:[25]

25 "BOD 22-01: Reducing the Significant Risk of Known Exploited Vulnerabilities", Cybersecurity and Infrastructure Security Agency, November 3, 2021.

1. Within 60 days of issuance, agencies shall review and update agency internal vulnerability management procedures.
2. Remediate each vulnerability according to the timelines set forth in the CISA-managed vulnerability catalog.
3. Report on the status of vulnerabilities listed in the repository.

Part of CISA's responsibilities under that directive is to maintain a catalog of Known Exploited Vulnerabilities (KEVs) and regularly review it based on changes in the cybersecurity landscape.[26]

The Exploit Database

It's important to note that there is no correlation between the CVE database run by MITRE, the National Vulnerability Database run by the National Institute of Standards and Technology, and the Exploit Database run by Offensive Security.[27] In fact, a study by Unit 42 discovered that of the 11,079 exploits that were mapped to CVEs at the time of the study, 80% of them were published in the Exploit Database an average of 23 days before their respective CVE was published. That's the exact opposite of what should happen, assuming that you agree that public disclosure should happen at all. The entire point of responsible disclosure is that the vulnerability and proof of concept isn't made public until after the patch is ready and the CVE released. Even then, it's still a race for most companies to get the patch tested and installed before the bad guys move against you.

The minus numbers in Figure 1-4 show the vast majority of exploits were published in the Exploit Database before their respective CVEs were published. That not only doesn't comply with the spirit of responsible disclosure, but it also begs the question of where the line is drawn between serving the needs of vulnerability researchers versus harming the companies, organizations, and agencies that are relying on cybersecurity products and services for protection.

26 See the CISA's Known Exploited Vulnerabilities catalog.

27 Offensive Security, which maintains the Exploit Database, is a for-profit company with over 800 employees that provides training and certification in penetration testing as well as offering penetration testing services to companies.

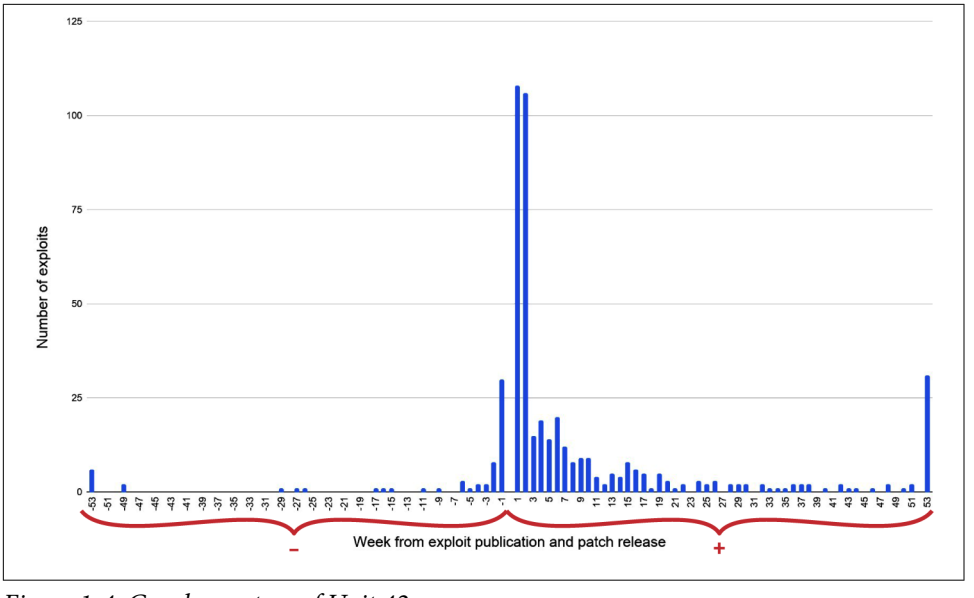


Figure 1-4. Graph courtesy of Unit 42.

A Protection Racket?

The cybersecurity industry has a reverse financial incentive. Systems built with software are inherently vulnerable, and many of those vulnerabilities are unknown. Money and professional reputations are made by discovering new vulnerabilities, of which there is an unlimited supply, and announcing them to the world. Then more money is made by creating products to protect those systems that are put at risk every time a new exploitable vulnerability is born. There's a name for a business that's responsible for both making you vulnerable to attack and defending you from that attack. The business is organized crime, the racket is called "protection," and it's illegal.[28]

Imagine if this model also applied to medicine. Medical hackers would make their living finding ways that the human body was vulnerable to illness and then announce those new vulnerabilities at medical conferences, which would then start a race between pharmaceutical companies to create a new drug ("patch") and bad actors to create malicious organisms (bioweapons) to exploit those vulnerabilities.

28 Paolo Campana, "Organized Crime and Protection Rackets", in Wim Bernasco, Jean-Louis van Gelder, and Henk Elffers (eds.), *The Oxford Handbook of Offender Decision Making*, Oxford Handbooks (2017; online ed., Oxford Academic, 6 June 2017).

What a horror show that would be! In fact, we call that scenario *biological warfare,* and it's been prohibited worldwide since 1928 under the United Nations' Biological Weapons Convention.

No other industry is allowed to do what the cybersecurity industry has been doing since its inception. It's time for a new model with different incentives and accompanying regulations.

Summary

My goal with this chapter was to demonstrate how cybersecurity is an industry driven by profit incentives that are at cross-purposes to its stated mission. Startups are reliant upon venture capital (VC) firms and private equity that have a different "why" than the founders do, a why that is entirely driven by how big of a multiple the VC firm can exit for rather than how effective the product or service is at preventing losses by unauthorized network access.

It is an industry divided between offense and defense, but one where only those engaged in offense are rewarded with speaking engagements, VC funding, and celebrity hacker status, which isn't surprising as this chapter also demonstrated how there has never been such a thing as a "secure" or "healthy" network since the birth of the first high-speed computer—MANIAC.

Finally, it is an industry desperately in need of a moral imperative similar to the Hippocratic Oath, to help the sick, and abstain from intentional wrongdoing.[29]

If you're looking for additional reading, I encourage you to seek out *Mechanizing Proof*, referenced in the footnotes of this chapter, and Dr. Olav Lysne's *The Huawei and Snowden Questions*, published by Springer and available for free download.

29 The principle of "First, do no harm," although widely attributed to the Hippocratic Oath, is not a correct translation, nor is it even possible to do. A more accurate translation is "to help the sick, and abstain from intentional wrongdoing." If you substitute "vulnerable" for "sick," I think you have the makings of a good oath for cybersecurity practitioners to take.

CHAPTER 2
Who Did It?

> *The required level of attribution needed…is not proved beyond a reasonable doubt in a court of law, but good enough for CNN.*
>
> —Dmitri Alperovitch[1]

In this chapter I'll examine the main underlying assumptions upon which cybersecurity analysts in government and the private sector build a case for the attribution of blame.

I'll address the risks of private-sector analysis due to commercial incentives and a lack of accountability.

Most importantly, I'll explain why attribution should be a transparent process and why evidence needs to be shared due to a lack of trust that has built up over the years, especially trust in classified sources because no one, not even the NSA, can tell you who sat behind the keyboard of an attack.

Finally, I'll close with a recommendation from international law professors Michael Schmitt and Yuval Shaney for an international attribution mechanism, even though the United States, the United Kingdom, and Israel most likely will not support it.

By the time you're finished you'll understand how the incentives for attribution work for cybersecurity companies as well as government agencies, and how those incentives are different. And I hope that in the future you'll treat every assessment of attribution with a critical eye knowing how those incentives may influence outcomes.

1 "Deterrence in Cyberspace: Debating the Right Strategy with Ralph Langner and Dmitri Alperovitch," Brookings Institution, Washington, D.C., September 20, 2011.

Attribution Is Inferred, Not Deduced

You might not know it by reading the work of journalists who cover cybersecurity, viewing threat reports that attribute cyber attacks to a nation-state, or listening to the many cybersecurity experts who have embraced attribution as something that they have more and more confidence in, but attribution at its core is built upon a series of assumptions, not facts. And those assumptions, unlike the axioms or postulates found in math and science, aren't known to be true without proof. There is no cybersecurity equivalent to Euclid's postulate "the whole is greater than the part," for example.

Further, the knowledge base upon which attribution is derived, as described by cybersecurity researchers Florian J. Egloff and Myriam Dunn Cavelty, "is neither value-free nor purely objective but built on assumptions and choices that make certain outcomes more or less likely."[2]

Timo Steffens, the author of *Advanced Persistent Threats: How to Identify the Actors Behind Cyber Espionage*, writes that attribution relies on abductive reasoning rather than deductive reasoning, but you can substitute "abductive" for "inductive" if you aren't already familiar with the term. Inductive reasoning is based upon one making an inference from an observation.[3]

For example, as you leave the house in the morning, you notice that the grass is wet. You infer from that observation that it must have rained last night. Or as you drive to work, you notice that there's a police car parked in front of the bank and the bank isn't open yet. You infer that there was a break-in attempt or a robbery.

The biggest problem with inductive or abductive reasoning is that we are merely arriving at a hypothesis with incomplete information and therefore our conclusion may very well be wrong. The grass could be wet because the automatic sprinkler system is malfunctioning. The police car may be sitting in front of the bank because the police officer is making an early-morning deposit into their account via the ATM or night deposit box.

Historically, two of the biggest assumptions in attribution up until around 2014 were that hackers who engaged in financial crime were from Russia and that those who stole trade secrets or other valuable intellectual property (IP) were from China. Therefore, if a company reported a breach, and it was a bank or a retail giant, then that would be a financial crime and therefore Russian hackers were responsible.

2 Florian J. Egloff and Myriam Dunn Cavelty, "Attribution and Knowledge Creation Assemblages in Cybersecurity Politics," *Journal of Cybersecurity* 7, no.1 (February 2021): tyab002.

3 Timo Steffens, *Attribution of Advanced Persistent Threats: How to Identify the Actors Behind Cyber-Espionage* (Berlin: Springer, 2020).

For example, when an aerospace company discovered a breach and hired a cybersecurity company like Mandiant to investigate, Mandiant's incident responders' first assumption would be that the hackers responsible worked for the Chinese government. Other countries that were also engaging in economic espionage—like France, Israel, Germany, and so on—weren't a consideration within the US cybersecurity industry at all. I frequently observed during those years that the industry suffered from a form of target fixation, and it only saw one culprit—China.

That was, of course, completely wrong. Russian hackers were stealing IP as far back as 1996, as evidenced by Moonlight Maze, the granddaddy of advanced persistent threats as we know them today. Moonlight Maze was discovered in 1999 and involved massive amounts of classified data stolen from the Pentagon, NASA, the US Department of Energy, defense industrial base companies, and other sources. It was attributed to Russia for the following reasons:

- The hackers' working hours corresponded to Russian working hours.
- The hackers didn't work during Russian orthodox holidays.
- Russian internet service providers were used.

However, the attack wasn't attributed definitively to any particular military or intelligence unit. The closest anyone got was Ben Macintyre at the *Times of London*, who reported that it was believed to be hackers associated with the Russian Academy of Sciences, which, in turn, did work for the Russian military. (Clearly, we were a lot more cautious about attribution back then than we are today.) Michael Vatis, the director of the FBI's National Infrastructure Protection Center, had this to say about attribution in his testimony before Congress on March 1, 2000:

> One major difficulty that distinguishes cyber threats from physical threats is determining who is attacking your system, why, how, and from where. This difficulty stems from the ease with which individuals can hide or disguise their tracks by manipulating logs and directing their attacks through networks in many countries before hitting their ultimate target. The "Solar Sunrise" case illustrates this point. This will continue to pose a problem as long as the internet remains rife with vulnerabilities and allows easy anonymity and concealment.

The Solar Sunrise case Vatis referred to happened during the month of February 1998, when a series of attacks were directed at unclassified Defense Department computers in an attempt to exploit a known vulnerability in the Sun Solaris operating system. Since the attacks happened at the same time that the US military was preparing for possible military operations against Iraq, it was assumed that the attacks were orchestrated by the Iraqi government. A multiagency investigation was launched, and within a matter of days it was determined that the hackers weren't from the Iraqi government at all. They were a group of teens from California and Israel.

Vatis pointed out in his testimony that the lesson learned from Solar Sunrise was that it was necessary to "gather information from victims and other sites within the US pursuant to applicable legal authorities before making conclusions about the likely identity of the attacker and determining what response to take."

In other words, be cautious about assuming who the attacker is until you've gathered enough evidence to rule out other possibilities.

A series of attacks known as Titan Rain followed in 2003 (though they weren't reported until 2005), and unlike Moonlight Maze, whose attribution was fuzzy and considered a one-off, Titan Rain (whose name later changed to Byzantine Hades) was directly attributed by the US government to the Third Department of China's People's Liberation Army (PLA), also known by its military unit cover designator 61398.

On January 31, 2013, the *New York Times* announced that it had been breached by hackers from China. Mandiant was named in the article as the company that was hired to do incident response.

On February 18, 2013, the *New York Times* ran this headline: "Chinese Army Unit Is Seen as Tied to Hacking Against U.S." The article described the findings from Mandiant's now-famous APT1 report, with APT1 being Mandiant's code name for PLA unit 61398. By the end of the year, Mandiant had been acquired by FireEye for $1 billion. The lesson for everyone in the cybersecurity industry was clear—attribution to China meant headlines, and headlines brought new business and higher valuations.

On June 9, 2014, CrowdStrike's blog seemingly announced a breakthrough. The company claimed to have found evidence that a Chinese hacker group that it was tracking, known as Putter Panda, was actually PLA military unit 61486. The evidence was a hat in a dorm room (see Figure 2-1).

CrowdStrike said of its own report, "While there are no 'smoking keyboards' in the unclassified intelligence CrowdStrike has collected on PUTTER PANDA, the balance of evidence available points to an extensive operation conducted by a PLA unit with a nexus to space-based communication systems. The alleged location and imagery associated with Chen Ping further corroborates the likelihood that this actor is affiliated with the PLA 12th Bureau of the 3rd Department of the GSD."

That's speculation, not evidence. CrowdStrike could have moved away from speculation and closer to evidence if it proposed alternative explanations and then weighed each of them for pros and cons, arriving at a best guess or estimate.

Figure 2-1. From the CrowdStrike blog post "Hat-tribution to PLA Unit 61486," June 9, 2014.

CrowdStrike also claimed to have been the first to uncover the address of the headquarters of the PLA's General Staff Dept. 12th Bureau in Shanghai. It was not. Project 2049 did that in 2011.

The company further claimed to have uncovered a secret hacker code disguised as a forum about cars that distributed assignments to Chinese hackers. In fact, it was nothing more than a forum about cars. The analyst who wrote the report used machine translation instead of having a native Chinese speaker translate it for him.[4]

CrowdStrike did not prove that the person it identified as Chen Ping—aka "cpyy," a handle used in Chinese forums and email addresses—was actually named Chen Ping, was an employee of PLA unit 61486, or was even a hacker. All of that is speculation on the part of the researchers. In fact, the name "Chen Ping" came from a WHOIS registration where the phone number, postal code, and email address were all fictitious, but because the name Chen Ping corresponds with the "cp" of cpyy used in email addresses associated with the WHOIS registration, the name, the thinking went, must be real. Or was it that cpyy, while providing invented data for every spot in the WHOIS registration of cppy.net, just made up his name, too?

4 This was shared in private emails between a native Chinese speaker, an information security professional, and myself on June 18, 2014.

Apart from CrowdStrike's poor analytic practices, the fact is that no company, especially Mandiant, was looking to attribute these types of attacks to any other nation-state than China. As a result, China was blamed for the ones that it did, and the ones that it didn't do. And because of bad assumptions, no one could tell the difference.

Examining Our Assumptions

When you read that a nation-state has been blamed for a cyber attack that had newsworthy fallout like a large power outage, or a hospital losing access to its medical records, or the utter destruction of Sony Pictures servers, the conclusion can be both convincing and intimidating. So it's important to begin with understanding the core assumptions upon which all attribution is built.

These assumptions aren't necessarily shared when claims of attribution are made. You're expected to trust the source (i.e., whoever is issuing the finding of attribution) to know what they're doing, to have the right motivations for doing it, and to strive to be as objective as possible in their analysis prior to coming to a finding. If you take nothing else away from this book, I hope it's that you will always ask questions and never accept anything without stress-testing the assumptions that it rests upon.

Let's set cybersecurity aside for a moment and look at the field of medical diagnosis. When a doctor provides you with a diagnosis, you want to believe that the doctor arrived at that diagnosis through a rigorous process of truth-finding. That's our assumption as patients. Is the assumption true or false? Consider this finding from a study:

> In reviewing 25 years of U.S. malpractice claim payouts, Johns Hopkins researchers found that diagnostic errors—not surgical mistakes or medication overdoses—accounted for the largest fraction of claims, the most severe patient harm, and the highest total of penalty payouts. Diagnosis-related payments amounted to $38.8 billion between 1986 and 2010, they found.
>
> "This is more evidence that diagnostic errors could easily be the biggest patient safety and medical malpractice problem in the United States," says David E. Newman-Toker, M.D., Ph.D., an associate professor of neurology at the Johns Hopkins University School of Medicine and leader of the study published online in *BMJ Quality and Safety*. "There's a lot more harm associated with diagnostic errors than we imagined."

Studies like this one by Johns Hopkins are possible because in medicine there's a clear path to arriving at the truth of a diagnosis. If the diagnosis and treatment was successful, the patient recovered. If it wasn't, the patient didn't, and then there were repercussions for the doctor and the hospital in the form of fines or worse. It would be impossible to perform such a study on the hundreds of claims of nation-state attribution made over the past 20 years because 98% of the time, no one knows the truth.

Let's take that model from medicine and layer it on top of a cyber attack.

Medical diagnosis	Cyber attack attribution
The patient is in distress. Cause unknown.	A company has suffered a breach. Culprit unknown.
The physician attempts to match the patient's symptoms with the disease or condition responsible.	The digital forensics and incident response (DFIR) company investigates the breach attempts to determine who's responsible.
The hospital that employs the physician may order a review of the physician's diagnosis if the patient's condition worsens.	The cybersecurity company that employs the DFIR person or team isn't held accountable for a wrong attribution because one rarely knows if an attribution is right or wrong.

So what's the problem, you may ask?

Well, in medicine, you learn from missed diagnoses because there's a clear outcome to the treatment. That clear outcome proceeds from a correct assumption to a validated outcome, better known as a fact, or it proceeds from an incorrect assumption to a postmortem to a matter of record so that, hopefully, it isn't repeated by other doctors down the road.

That is not the case in the world of cyber attribution because we almost never know who actually performed an attack. Instead, a series of assumptions are made, but unlike medicine, those assumptions are almost never validated with facts, because no one, not even the NSA, can tell you who was operating the keyboard of a computer used in an attack. The agency, according to former employees that I've spoken with for this chapter, is quite good at analyzing netflow and tracing traffic through proxies and VPNs, and it may even get to the exact computer that was used by the attacking operator, but that's as far as it can go. Hence, assumptions come into play.

Following are the main assumptions that Steffens describes in his book *Attributions of Advanced Persistent Threats.*

You may find the use of the term "premise" versus "assumption" a little confusing. When I'm quoting a section from the book by Steffens, I use "premise" because that's the term he used. Otherwise, I use "assumption." I do that because premises are usually found in mathematical or logical arguments, and those premises are statements that are *assumed* to be true until tested, after which they're found to be either true or false.

In cybersecurity, premises aren't tested. They're simply assumed to be true. It falls to the critic to test them; however, there's no incentive to do that. So I prefer to use the term assumption, meaning unproven statement, which I believe is more easily understood by people who aren't trained in math or philosophy.

Premises about resources and relations to governments

> "A fundamental assumption for attribution is that attacks that were done by the same actor will be similar in certain ways."
>
> "Criminals do not engage in espionage—or if they do, they do it at the behest of a government."
>
> "According to commonly accepted belief, most APT [advanced persistent threat] groups work for exactly one government, not for several."

Premises about the setup of teams

> "Cyber-espionage is generally assumed to be a full-time job. Basically all analysis methods about patterns-of-life and timestamp statistics rely on the premise that the attackers work regular office hours—roughly from 8 or 9 am to 5 pm."
>
> "Another premise is that malware source code is owned by the APT group or its sponsor, and not by individual developers. This is an important requirement for clustering attacks into intrusion sets by the malware that was used."

Premises about data

> "An important premise is that code similarity does not happen by chance, but is more likely due to the same developer having written it."
>
> "If a malware that has been attributed to an APT group is found on a computer, for simplicity it is usually assumed that other malware on the same device was also installed by the same actors."

These assumptions are a good representation of what the cybersecurity industry and government agencies rely upon when it comes to answering the question "Who did it?" There are certainly more out there, but a comprehensive listing is beyond the remit of this book.

Assumptions are the foundation upon which cyber researchers evaluate all of the technical and geopolitical evidence associated with a cyber attack and make their case for an assignment of attribution. That, in turn, may lead prosecutors down an evidentiary path toward building their own case for a grand jury indictment, or it may convince the president of the United States to issue an executive order with sanctions attached or an order for a cyber attack against the country that has been deemed responsible, potentially accompanied by property damage and loss of life.

So when a private cybersecurity company—or the federal government, which relies heavily on the private sector in matters related to cybersecurity—issues an announcement that places the blame for an attack on a nation-state, the first question that everyone who reads it should be asking is, what are your assumptions? The second question should be, what is your evidence?

Unfortunately, the first question is never asked because the vast majority of people don't even know that assumptions are involved, and the second question is rarely ever answered because the private-sector companies are reluctant to share their evidence, and the federal government's evidence may involve classified sources and methods.

The assumptions in the following three subsections have been proven to be invalid, and yet they still remain in the tool box of cyber attack investigators.

The Exclusive Use Assumption

> *Another premise is that malware source code is owned by the APT group or its sponsor, and not by individual developers. This is an important requirement for clustering attacks into intrusion sets by the malware that was used.*
>
> —Timo Steffens

In CrowdStrike's investigation of the Democratic National Committee (DNC) hack, it found that XTunnel malware had been used. Here's how CrowdStrike describes it:

> This adversary [Fancy Bear/APT28] has dedicated considerable time to developing their primary implant known as XAgent, and to leverage proprietary tools and droppers such as X-Tunnel, WinIDS, Foozer and DownRange.

The theory is that if XAgent is used at any time, anywhere in the world, it's the same threat actor at work because of the industrywide assumption that malware is proprietary and isn't shared. However, we know that isn't true, thanks to the work done by researchers at Focal Point:

> X-Agent has been "in the wild" since 2012, is relatively easy to obtain, and has been well-documented…Further, ESET was able to obtain the source code for a report in 2016. With the implant's known availability, any number of threat actors, in addition to FancyBear, could have been in possession of the implant during the Russian-Ukrainian conflict from 2014-2016.

The fact is that cyber munitions aren't kinetic munitions. They don't explode on impact, rendering them unusable. Once malware is deployed in the wild, the hackers that deployed it lose control over what happens to it. It can be reverse-engineered, or modified, or simply redeployed depending on the circumstances. Malware deployed by one team is malware enjoyed by another.

For example, the offensive cyber team at Ukraine's Directorate of Intelligence for the Ministry of Defense (GUR) maintains a repository of all of the malware that it has seized from hackers working for Russia, Iran, and other countries.[5] After some modifications are made, it deploys the new malware in its own operations against

5 The GUR is to Ukraine what the GRU is to Russia. GUR is also sometimes referred to as HUR or HUR or MOU. GUR is a romanization from the Ukrainian ГУР.

new targets; however, any analysis done on it by the victim entity will still show the malware as Russian or Iranian in origin.

Wikileaks's release of CIA files pertaining to the agency's offensive cyber program, leaked by disgruntled employee Joshua Schulte, showed that the Umbrage subgroup of the agency's Remote Development Branch had been cataloging exploits created by other actors for false flag operations, as well as cost- and time-efficient solutions for more targeted operations.

The assumption of exclusive use is not only a false assumption, but reliance upon it could lead investigators down a path that was intentionally created for them by the adversary.

The Working-Hours Assumption

Again, from Steffens:

> "Cyber-espionage is generally assumed to be a full-time job. Basically all analysis methods about patterns-of-life and timestamp statistics rely on the premise that the attackers work regular office hours—roughly from 8 or 9 am to 5 pm."

Let's assume that this assumption is valid, that hackers employed by a nation-state will work 9 a.m. to 5 p.m. and take the weekends off. The larger question is, how do you differentiate between the many nation-states that have an offensive cyber capacity and are in the same or adjacent time zones? Russia's land mass spans 11 time zones! Which Russian city are you assuming that your hacking team is working from?

If it's Moscow (UTC +3), that time zone is closely shared with Iran (UTC +3.5) and Israel (UTC +3 during Daylight Saving Time). Both states rival Russia in terms of offensive cyber operations.

If you move one hour earlier (UTC +2), you have Ukraine and Israel (+2 during Standard Time), not to mention Finland, Estonia, Egypt, Latvia, and many other countries with known hacking capabilities.

One hour later (UTC +4) and you have the United Arab Emirates (UAE) and Georgia. The UAE in particular is well known for investing in cybersecurity startups and, more to the point, has made the news for employing hackers who had formerly worked for the NSA to conduct offensive operations against Qatar.

So assuming that your hacker is engaged full-time by a nation-state and works the "day shift," do they start work at 8 a.m., 9 a.m., or 10 a.m.?

This assumption seems ludicrous on its face, and yet it's commonly accepted as true.[6]

A corollary to this assumption could be that if the NSA follows it, it's almost certainly because the agency can't say what its evidence really is.[7] While working hours is a terrible cover story for the reasons that I presented, it's understandable coming from a three-letter agency. When used by private-sector cyber sleuths who don't have access to classified information, it's a very weak piece of evidence.

The Criminals Versus Spies Assumption

Criminals do not engage in espionage—or if they do, they do it at the behest of a government.

Hackers for hire certainly do engage in espionage for a fee, and it has nothing to do with work for a government agency. In 2014, I wrote about the case of Su Bin, the Chinese businessman who hired Chinese hackers to steal data on the Brahmos 2 missile from an Indian company and the F-22 from Lockheed Martin on behalf of a client. In 2016, TrendMicro published its findings of EaaS (espionage-as-a-service) vendors on the dark web. In 2019, the ThreatPost blog reported that one in four underground merchants offer advanced hacking services, once reserved for APTs and well-funded organized crime gangs.

Valid Concerns

Cyber operators who engage in computer network exploitation are just as susceptible to the same human traits of laziness, distraction, boredom, and general sloppiness as employees at any other job. Mistakes will happen, and that's what incident responders and investigators count on.

In addition to mistakes due to a lack of discipline, looking for shortcuts, and other human foibles, there are mistakes in tradecraft that can make it easier for cybersecurity companies to group technical indicators that may ultimately point to a nation-state threat actor. Those mistakes have been documented by the CIA's Technical Advisory Council in an informal discussion about what the Equation Group did wrong that contributed to Kaspersky Lab identifying it.[8]

6 The working-hours assumption can be found in almost every report that assigns attribution to a foreign threat actor as well as in every cyber threat intelligence course that offers certification. Curious readers looking for additional examples of this assumption will have an easy time finding them.

7 On May 9, 2023, the NSA coauthored a 48-page public report on Russian "snake" malware. The section on attribution was four paragraphs and included the working-hours assumption.

8 Wikileaks Vault7 release.

A few of the standouts were:

- The use of customized crypto instead of using publicly available crypto
- The reuse of compromised techniques and exploits
- Combining a new tool with a compromised technique or exploit
- The use of shared code across all of its tools

Assumptions have been formulated around these valid concerns, and their use in an investigation can be helpful or harmful. However, when you don't have a way to test assumptions in an objective manner and know when they can be applied and when they cannot, you have a serious problem. Add to that the pressures for a cybersecurity startup to stand out from the crowd by convincing a journalist to run with its latest attribution of a cyber attack to a nation-state before your competitor does, and you have a recipe for poor investigative work and rushed findings.

One of the few times that an attribution was proved correct was with Kaspersky Lab's investigation of a threat actor that it named Equation Group. While not actually naming the NSA, the company predicted that it was an extremely well-funded North American government agency. When the Shadow Brokers started leaking the NSA's hacking tools, and the custom encryption algorithm matched several hundred of the leaked tools, it was a done deal.[9]

Of course, it's not often that a massive leak of cyber tools, like the ones that both the NSA and the CIA experienced, can be used to cement an attribution to a nation-state. It's odd that no other nation-state besides the US has had employees of its own intelligence services do the same. Where are the leaked GRU tools? Or Unit 8200? Or GCHQ?[10]

CrowdStrike's commercial success while producing some heavily criticized analytic reports has taught the industry that there are no repercussions for poor analysis.[11] Without accountability and repercussions, where's the incentive to assess and correct your analytic output? It doesn't exist. As Dmitri Alperovitch, CrowdStrike's cofounder and former chief technology officer, said in the quote that opened this chapter, "The required level of attribution needed…is not proved beyond a reasonable doubt in a court of law, but good enough for CNN." Of course, it's not just CNN. Alperovitch could have used the name of any other mainstream news outlet to make his point.

9 Scott Shane, Nicole Perlroth, and David E. Sanger, "Security Breach and Spilled Secrets Have Shaken the N.S.A. to Its Core", *New York Times*, November 12, 2017.

10 GRU, Unit 8200, and GCHQ are the respective Russian, Israeli, and British equivalents of the NSA (more or less).

11 CrowdStrike's report on the alleged hacking of a Ukrainian artillery app is among the worst ever written. For more, see this article on the Inside Cyber Warfare newsletter.

The Need for Independent Fact-Finding

While some cybersecurity companies may think their teams have mastered the art of attribution, it's rare that other nation-states (a) believe them and (b) take action. For this reason, Yuval Shany and Michael N. Schmitt organized an international research workshop funded by the Dutch Ministry of Foreign Affairs and the Federmann Cyber Security Research Center of the Hebrew University of Jerusalem to consider the feasibility of establishing an international attribution mechanism, as well as the usefulness of such a body. It was held on November 11, 2018, in Jerusalem. While the session content was protected by the Chatham House Rule, a short report following the conference listed 10 topics recommended for further discussion:[12]

- The issue of State and non-State actors' responsibility and direct and indirect responsibility.
- The need to differentiate between individuals with different motivations (criminal or others) and States.
- Is "good enough governance" a sufficient standard for coordination and cooperation?
- The question of whether territorial aspects are still relevant and technologically applicable.
- Do strict legal rules provide the best mechanism for addressing the challenge or should more flexible standards be developed?
- Are cyber attacks more akin to espionage than war?
- Should attribution be addressed through an inter-State mechanism, or other private or public framework?
- What is the role of AI in attribution and response (e.g., AI-governed hack-backs)?
- Can trust-based systems and certifications be developed to help attribution efforts?
- What is the role of ethics by design—are there suitable technological solutions that refer to ethics, in the aspects of evidence collection and attribution?

"Ethics by design" is particularly curious. Can a computer program enforce ethics when the humans involved are ethically challenged? And assuming that such a thing could be done, who would program the AI for that?

12 The rule reads: "When a meeting, or part thereof, is held under the Chatham House Rule, participants are free to use the information received, but neither the identity nor the affiliation of the speaker(s), nor that of any other participant, may be revealed." Source: ChathamHouse.org.

In their paper "An International Attribution Mechanism for Hostile Cyber Operations," Shany and Schmitt reported that the discussions "have led us to conclude that, for the time being, States wielding significant cyber capability have little interest in creating an international attribution mechanism for cyber incidents."[13] The authors mention the US and the UK, but I think you can assume that it also includes Israel based on off-the-record discussions that I've had. The reluctance of developed countries with powerful cyber capabilities to submit to an international body on the question of who's responsible for an attack against their infrastructure is certainly understandable. They don't want to lose their autonomy, nor limit their political options.

For example, if the attempted breach of a State's electoral database was the work of German hackers who utilized some of Russia's bulletproof infrastructure, then the US government would have the option to either pursue criminal action against the German citizens involved or blame Russia for the act and include it in any future sanctions. Since evidence isn't required to be shared for the latter option, and since it dovetails nicely with US national policy that's heating up over other acts of Russian interference, the US would have strong incentives to go with door number two.

A Case of Election Tampering or Not?

On August 29, 2016, the FBI's Cyber Division found that "foreign hackers penetrated two state election databases in recent weeks, prompting the bureau to warn election officials across the country to take new steps to enhance the security of their computer systems."

The White House issued a statement blaming Russia for the two attempts, but the evidence was essentially that the hackers leased server time from a Russian hosting company based in Siberia. Why would hackers working for the Russian government choose to use a Russian web hosting company?

To get an answer to that question, I reached out to the owner on Facebook and, after a few conversations over a period of weeks, he trusted me enough to share an interesting tidbit: that the help tickets that his customers had filed for server support were written in English, and that the email addresses they used were from GMX.com (a German company that is part of 1&1 Mail and Media GmbH) and iname.com (owned by the New Jersey–based World Media Group). I asked if he would send me the originals and he declined for privacy reasons, but he said that he'd provide them to the FBI or Russia's Federal Security Service (FSB), if asked.[14] The last I heard, he hadn't heard from either agency.

13 Yuval Shany and Michael N. Schmitt, "An International Attribution Mechanism for Hostile Cyber Operations," *International Law Studies* 96 (2020): 196–222.

14 This conversation was held via Facebook Messenger on September 28, 2017.

On the other hand, if the US had signed on to a collective cyber attribution agreement involving an independent international commission, the White House's political options would be much more complicated. Where's the incentive to go with an independent fact-finding commission for countries that have the resources to keep everything in-house? The incentive is that a peer or near-peer nation with similar capabilities could make the same arbitrary claim directed at the White House without providing any evidence to support it.

There's also an ethical component to it in that it's only right to have a fair, objective evaluation of the evidence before coming to a finding that could bring negative repercussions to bear on the responsible nation-state, whose citizens will have to bear the brunt of the fallout, economic or otherwise. Unfortunately, ethics is a weak incentive.

A Proposed International Attribution Mechanism Modeled after the OPCW

The Organization for the Prohibition of Chemical Weapons (OPCW) model was suggested by Schmitt and Shaney in their paper referenced earlier. Sharing evidence for objective analysis by third parties is how international enforcement mechanisms typically work. There can be no enforcement of sanctions on an international scale without the sharing of evidence for independent evaluation and attribution, such as what's done by the OPCW and its Technical Secretariat. The OPCW has more than 80 member states and 500 staff, and it's empowered to provide technical assistance and on-site inspections.

An objective, third-party investigative agency similar to the OPCW is needed, but the initial membership should consist of smaller nations that don't have the capabilities that the US, the UK, Israel, China, and other developed nations have. Once a sufficient number of smaller states are involved (about the size of NATO), it will gain momentum and eventually the States that are currently resistant to joining will face more and more demands for evidence-sharing.

The ideal member would have a robust internet infrastructure but lack investigative reach outside of its borders. Another characteristic would be that the prospective member State lacks the ability to respond proportionately to a destructive cyber attack, and would therefore benefit by being able to leverage a group response, similar to what NATO offers to its members.

Finally, member states could avoid having to hire private cybersecurity companies to perform data forensics and incident response, and risk having their findings be both closed and biased in favor of what benefits the company rather than receiving an objective, fact-based determination that was subject to third-party review.

Summary

My goal with this chapter was to take a deep dive into the key assumptions that usually form the basis for any investigation by a cybersecurity company or a government agency into those responsible for a cyber attack. The public's perception is that an assignment of attribution is trustworthy, that it's based on facts, and that it's conducted in a manner similar to what they've seen on TV crime shows. That perception could not be more wrong.

I want to emphasize that there is nothing wrong with inductive reasoning for coming to a finding of attribution when there are so many intangibles involved. However, as long as the incentives for assigning blame to a nation-state are so high and there are no repercussions for poor analytic practices, for both the companies that create the reports and the media that covers them, then it is imperative that you ask questions, raise alternate hypotheses, and generally demand more from corporate and government entities involved in these investigations.

CHAPTER 3
Establishing Corporate Accountability

With great power comes no responsibility.

—Tristan Harris during his 2019 congressional testimony

Where experiment or research is necessary to determine the presence or the degree of danger, the product must not be tried out on the public, nor must the public be expected to possess the facilities or the technical knowledge to learn for itself of inherent but latent dangers. The claim that a hazard was not foreseen is not available to one who did not use foresight appropriate to his enterprise.

—Robert H. Jackson, associate justice of the United States Supreme Court (1953)

Cybersecurity can thus be described as a 'market for lemons' where there is an asymmetry of information between the buyer and the seller such that the seller knows of the product's defects but does not disclose them to the buyer or may even misrepresent them to the buyer.

—NSS Labs, Inc. complaint in US District Court, Northern District of California[1]

Regardless of the industry, one thing is clear. Companies do not embrace accountability on their own. It must be forced upon them.

This chapter will examine the nascent effort to bring accountability to software makers, an effort that had a rough start a few years back but received a shot of adrenaline with the White House's release of the National Cybersecurity Strategy on March 1, 2023. We are a long way from enacting legislation to bring accountability to the software industry, according to Kemba Walden, the former acting national cyber director. If history is a guide to what we can expect, it'll take a monumental disaster to galvanize Congress to act.

1 Case No. 3:18-cv-05711 Complaint for Violation of the Sherman Act, 15 U.S.C. § 1, and the Cartwright Act, California Business & Professions Code §16720.

In this chapter I'll draw from historical precedence, discuss current events at the time of this writing, and make an assessment for the future as regards bringing corporate accountability to the software industry, which includes the cybersecurity industry.

Pay for Protection

In May 2023, US Department of Commerce Secretary Gina Raimondo met with Chinese officials to discuss what could be done about sanctions that the department had imposed on some Chinese companies. In the meantime, Chinese hackers were suspected of accessing Secretary Raimondo's email account via a "flaw" in Microsoft Outlook 365, part of Microsoft's Cloud services. Microsoft didn't announce the breach until July 11, 2023. An intrusion timeline for this attack can be seen in Figure 3-1.

The Department of Commerce wasn't the only US government agency that had been breached by those hackers. The State Department had been as well. And it wasn't just Outlook that was involved. The hackers first exploited a vulnerability found in Microsoft Azure's Active Directory, which suggested to independent researchers that the breach could have been much more widespread than Microsoft had acknowledged.

State and Commerce had reported the breaches to CISA (the Cybersecurity and Infrastructure Security Agency), whose director, Jen Easterly, had been having discussions with Microsoft to enable a more complete logging service for free for its government and enterprise customers. Without such a service, Easterly knew that it was very difficult, if not impossible, to rapidly detect bad actors who were attempting to gain access to a network.

Once the new attacks were reported to CISA, and with these prior discussions already happening without any concessions made by Microsoft to date, the company clearly felt like it could no longer refuse to make free what it had been forcing customers to pay for—barely adequate defenses that would allow an organization to see what was happening on its own networks.

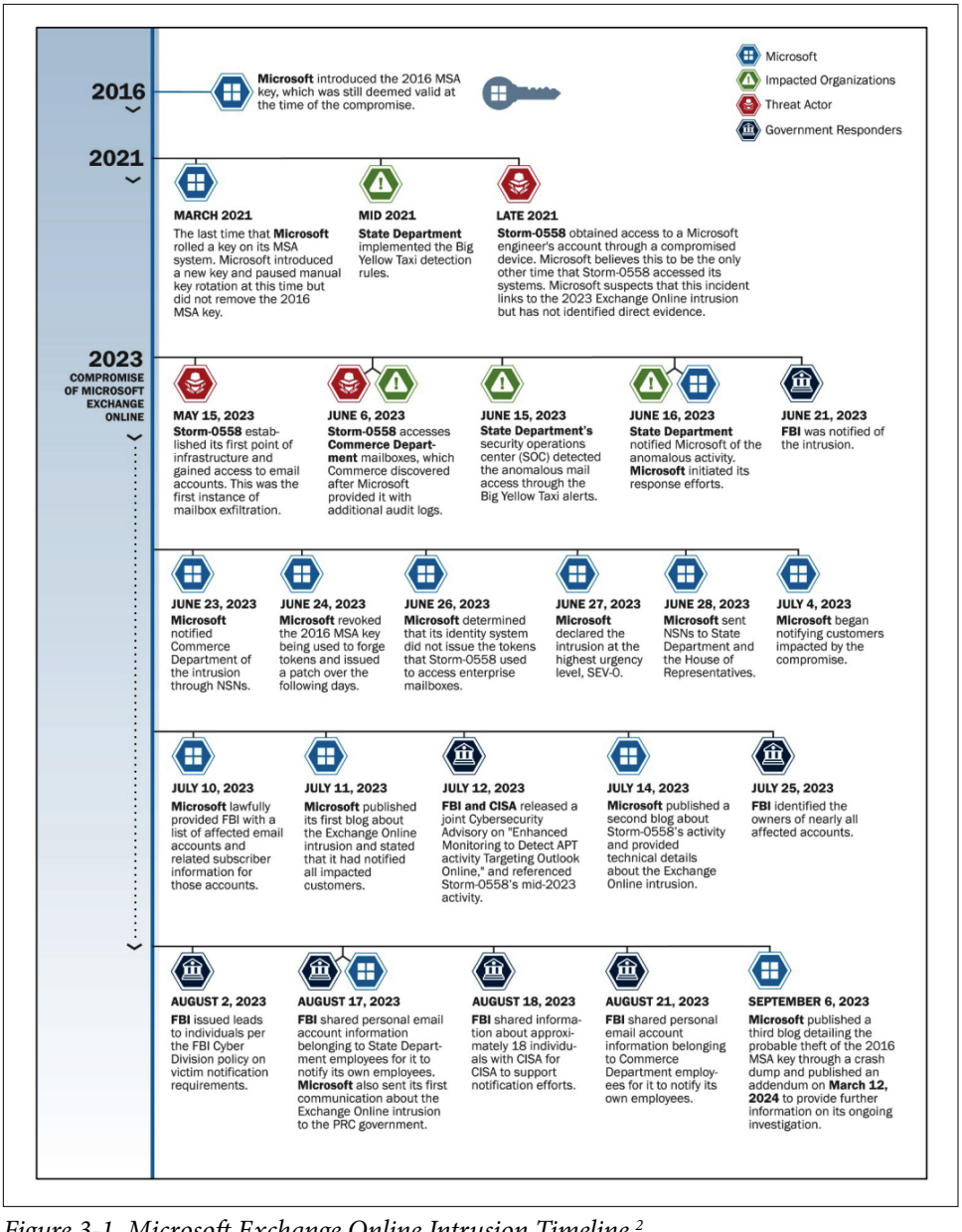


Figure 3-1. Microsoft Exchange Online Intrusion Timeline.[2]

2 "Review of the Summer 2023 Microsoft Exchange Online Intrusion", Cyber Safety Review Board, March 20, 2024 (image on page 26).

On July 27, 2023, Senator Ron Wyden (D-OR), furious over yet another Microsoft-enabled breach of US government agencies—the Department of Commerce and the Department of State—sent a letter to CISA Director Easterly; Lina Khan, chair of the Federal Trade Commission; and Attorney General Merrick Garland at the Department of Justice, suggesting various approaches by which the government could "hold Microsoft responsible for its negligent cybersecurity practices, which enabled a successful Chinese espionage campaign against the United States government."

Wyden supported his claim of negligence on the part of Microsoft by pointing out that similar vulnerabilities had resulted in two disastrous events—SolarWinds in 2021 and the breach of State and Commerce Department emails by Chinese hackers in 2023—all while the company made record profits.

Wyden's directives in his letter to each of the three government agency heads are instructive in both the approach and the overall sense of frustration that's conveyed:

> Director Easterly, I urge you to exercise your shared authority to direct the Cyber Safety Review Board to investigate this incident. In particular, the Board should examine whether Microsoft stored the stolen encryption key in an HSM, a best practice recommended by the National Security Agency and even by Microsoft, and if not, examine why Microsoft failed to follow its own security advice. The Board should also examine why Microsoft's negligence was not discovered during the external audits that were required to obtain certification for government use under the FedRAMP program, or during Microsoft's own internal security reviews.

HSM is an acronym for hardware security module, a hardened, tamper-proof physical device that's used to manage, process, and store cryptographic keys separate from the network. If a network has an HSM installed, like Microsoft's Azure cloud service used by Commerce and State, then the cryptographic keys are never supposed to leave the HSM, and that begs the question—how was the key stolen?

Senator Wyden reminds Attorney General Garland in this next quote about the Department of Justice's commitment to pursue companies with government contracts (like Microsoft) that fail to follow required cybersecurity standards.

> Attorney General Garland, the Department of Justice has previously pledged to "use [its] civil enforcement tools to pursue companies, those who are government contractors who receive federal funds, when they fail to follow required cybersecurity standards." I urge you to examine whether Microsoft's negligent practices violated federal law.

Next, Senator Wyden addresses Lina Khan of the Federal Trade Commission:

> Chair Khan, I urge you to investigate Microsoft's privacy and data security practices related to this incident to determine if Microsoft violated federal laws enforced by the Federal Trade Commission, including those prohibiting unfair and deceptive business practices. In addition, Microsoft was subject to a consent decree for 20 years after a security incident with its predecessor single sign-on product, Passport. That consent

> decree, which expired in December 2022, required Microsoft to "establish and maintain a comprehensive information security program in writing that is reasonably designed to protect the security, confidentiality, and integrity of personal information collected from or about consumers" for Passport or substantially similar services. Microsoft Account, the product from which the encryption key was stolen, is Passport's modern successor. If Microsoft's negligent cybersecurity practices predated the expiration of the consent decree, I also urge you to take all necessary steps to hold the company responsible for any violations of that order.

By including three relevant agencies, Senator Wyden hoped to make it more difficult for Microsoft to escape scrutiny. Time will tell if this approach is successful; however, Microsoft is not the only company that cannot self-regulate when it comes to matters related to cybersecurity.[3] It is simply the noisiest wheel of the bunch, as indicated by Figure 3-2.

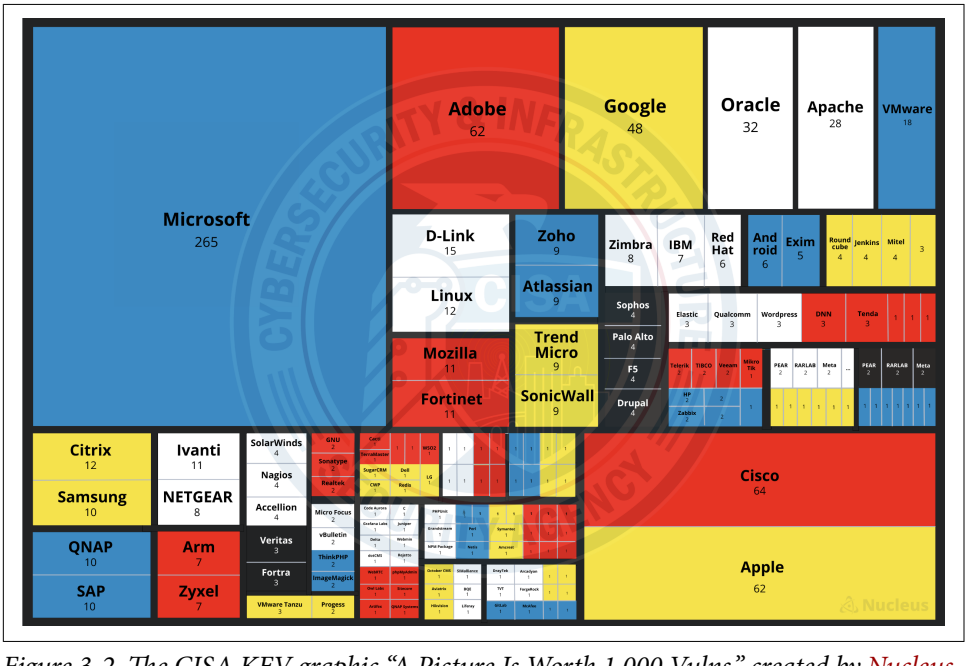


Figure 3-2. The CISA KEV graphic "A Picture Is Worth 1,000 Vulns," created by Nucleus Security.

3 The Department of Homeland Security set up the Cyber Safety Review Board, modeled after the National Transportation Safety Board, in 2022. After an investigation into this Microsoft email breach, it issued a scathing report in March 2024. It has no authority beyond that of an investigatory body. It'll be up to Congress to take action, or not.

Figure 3-2 is a graphical representation of CISA's Known Exploited Vulnerabilities (KEVs) database, currently at 1,000 entries at the time of this writing. It was created by a team at Nucleus Security for exhibition at the Black Hat conference in Las Vegas, Nevada, on August 9–10, 2023. Microsoft dominates the field with 265 KEVs, followed by Cisco with 64, and Apple and Adobe at 62 each.

CISA created the Known Exploited Vulnerabilities database to help organizations prioritize how to address vulnerabilities in the software products that they use. An IT department can be quickly overwhelmed due to the sheer volume of CVEs, and many of them are not important enough to waste time on. For a vulnerability to be included as a KEV, it must meet the following criteria, according to CISA:

- The vulnerability has an assigned Common Vulnerabilities and Exposures (CVE) ID.
- There is reliable evidence that the vulnerability has been actively exploited in the wild.
- There is a clear remediation action for the vulnerability, such as a vendor-provided update.

This ability to prioritize vulnerability patching is the first step toward hardening a company's network.

Before continuing any further, I want to provide some needed historical context in order for you to fully understand the scale of the problem, and what it took for other industries to finally come under the heavy hand of government regulation.

It All Comes Down to Cost Calculation

> *So long as brakes cost more than trainmen, we may expect the present sacrificial method of car-coupling to continue.*
>
> —Ralph Nader's *Unsafe at Any Speed*

There are many historical examples where cost calculations have generated resistance toward regulations of corporate accountability. Next I'll describe three of them, and after that I'll connect them to the current situation for the software industry.

The Railroad

In the previous quote, political activist Ralph Nader was referring to the mode of transportation that the general population primarily relied upon prior to the mass-produced automobile—the train—and the highly dangerous job performed by the trainmen, especially the brakemen, who were responsible for coupling and uncoupling the train's cars using a link-and-pin system.

The brakeman would have to position himself between two moving cars and, using his bare hands and some basic tools, line up the couplers from each car and insert a pin that connects them. Or do the reverse to uncouple the cars. Serious injuries, including crushed hands and limbs, were common, and many brakemen were killed on the job; however, neither the railroad industry nor the government agencies involved saw any need to spend money on replacing the link-and-pin system with something better.[4] Furthermore, the industry's conventional wisdom at the time said that the railroad had an obligation to ensure the security and well-being of its customers. Their employees could take care of themselves.

It took another Ralph Nader–type reformer named Lorenzo Coffin to write about the high rate of accidents and deaths among trainmen until he got the attention of the Iowa legislature in 1890 and the US Congress in 1893. Even after legislation was passed, it took a long time for adoption.[5]

Shipping

On the morning of April 16, 1947, at the port of Texas City, Texas, smoke was sighted coming from Hold 4 of the SS *Grandcamp*, a 10,000-ton cargo ship owned by the Compagnie Générale Transatlantique.

One day earlier it had been loaded with 1,850 tons of ammonium nitrate fertilizer en route to farmers in Europe after the war had ravaged their crop lands. On its way to Texas City from France, the *Grandcamp* had stopped in Belgium and picked up 16 cases of small arms ammunition.

A second vessel, the *High Flyer*, moored next to the *Grandcamp*, was loaded with 1,000 tons of fertilizer in addition to the 2,000 tons of sulfur that it was already carrying.

The *Grandcamp*'s captain, upon notification by his crew that the two fire extinguishers on board were insufficient to put out the fire, made the decision to try to starve the fire by battening down the hatches and covering them with tarps.

When that failed, a fire alarm was sounded and the Volunteer Fire Department responded with its entire contingent of 26 men and 4 trucks. The time was 8:30 a.m.

At 9:12 a.m., it felt like the world came to an end for the citizens of Texas City, Texas. A massive explosion pulverized the SS *Grandcamp* along with its captain and most of its crew. The vessel's 1.5-ton anchor was found two miles away, buried in a hole that was 10 feet deep. Two hundred fifty miles away, a Denver seismologist registered that an earthquake had occurred. The Department of Defense's Strategic Air Command in

4 For more, see *https://oreil.ly/DwW-g*.

5 For more, see *https://oreil.ly/7kWmO*.

Omaha, Nebraska, briefly elevated the nation's defense readiness condition (DEFCON), believing that a nuclear bomb had detonated. Figure 3-3 provides a sense of the scale of the blast.

Figure 3-3. A styrene plant became a roaring inferno on April 16, 1947, in Texas City. Hundreds of other fires started in areas where petroleum was stored.
Source: the Houston Chronicle.[6]

Approximately 600 people were killed and 3,000 injured. One in three homes in Texas City were uninhabitable. Twenty-five thousand people were rendered homeless or jobless. More than 1,000 cars and trucks were destroyed. It remains the deadliest accident in US history.

6 Susan Carroll, "Texas City Disaster Devastated Community, Changed Way Industry Was Regulated", *Houston Chronicle*, updated April 13, 2017.

Thousands of individual lawsuits against the US government were consolidated into a class action that charged negligence in "adopting the fertilizer export program as a whole, in its control of various phases of manufacturing, packaging, labeling and shipping the product, in failing to give notice of its dangerous nature to persons handling it, and in failing to police its loading on shipboard."[7]

The US District Court for the Southern District of Texas ruled in favor of the class by determining that "the Coast Guard and other agencies were negligent in failing to prevent the fire by regulating storage or loading of the fertilizer." The case then went to the Court of Appeals, which had a split decision among its six judges, and was finally taken up by the US Supreme Court in 1953, which found in a split 4-to-3 decision that the US government couldn't be sued due to the age-old legal doctrine of sovereign immunity.

Associate Justice Jackson was one of three Supreme Court judges who dissented to that ruling. Included in his dissent is the quote that I used to open this chapter:

> Where experiment or research is necessary to determine the presence or the degree of danger, the product must not be tried out on the public, *nor must the public be expected to possess the facilities or the technical knowledge to learn for itself of inherent but latent dangers* (my emphasis added).[8]

It took an act of Congress in 1955 to create new legislation as well as to award the victims a total of $16.5 million.

As I'll point out later in this chapter, software companies have been demanding that the public assume the risk for using their insecure products, but the public doesn't "possess the facilities or the technical knowledge" to make an informed decision.

A key takeaway from the Texas City disaster, which also applies to most industrial accidents, is that "ignorance and complacency do not have to be pervasive for catastrophes to occur—these need apply only to crucial junctures in order to initiate a break in the system and set a series of harmful consequences in motion."[9] This observation can be applied to the software industry as well. What is a vulnerability if not a "break in the system"? And it only takes one successful exploit of that vulnerability to create a potentially catastrophic outcome.

7 Dalehite v. United States, 346 U.S. 15 (1953).

8 Ibid.

9 Deanna Meyler et al., "Landscapes of Risk: Texas City and the Petrochemical Industry," *Organization & Environment* 20, no. 2 (June 2007), 204–212.

Automobiles

It took Ralph Nader's excoriation of the automobile industry in his book *Unsafe at Any Speed* (Knightsbridge), along with a National Academy of Sciences report entitled "Accidental Death and Disability: The Neglected Disease of Modern Society," to motivate Congress to draft and pass the National Traffic and Motor Vehicle Safety (NTMVS) Act in 1966.

If you haven't read Nader's book, I encourage you to do so because there are a lot of parallels between the software industry today and the automotive industry back then. For example:

> Highway accidents were estimated to have cost this country in 1964 $8.3 billion in property damage, medical expenses, lost wages, and insurance overhead expenses. Add an additional sum to comprise roughly the indirect costs and the total amounts to over two percent of the gross national product. But these are not the kind of costs which fall on the builders of motor vehicles (excepting a few successful lawsuits for negligent construction of the vehicle) and thus do not pinch the proper foot. Instead, the costs fall to users of vehicles, who are in no position to dictate safe automobile designs.
>
> In fact, the gigantic costs of the highway carnage in this country support a service industry. A vast array of services—medical, police, administrative, legal, insurance, automotive repair, and funeral—stand equipped to handle the direct and indirect consequences of accident injuries. Traffic accidents create economic demands for these services running into billions of dollars. It is the post-accident response that lawyers and physicians and other specialists labor. This is where the remuneration lies and this is where the talent and energies go. Working in the area of prevention of these casualties earns few fees.

The National Safety Council tracks causes of death (see Figure 3-4) and death rates (the number of people who died as a percentage of the total population). What it found (these quotes are from Nader's *Unsafe at Any Speed*):

> A period of rapidly increasing deaths (+26%) and death rates (+9%) occurred between 1961 and 1973. These increases were largely driven by surges in motor-vehicle deaths (+46%) and death rates (+26%).
>
> The longest period of improvement in preventable deaths and rates occurred between 1973 and 1992. During this time, deaths decreased 33% and death rates declined 38%. In 1992, the United States achieved the lowest recorded death rate of 34.0 deaths per 100,000 population. This drop was once again driven by motor-vehicle deaths, which decreased 35%. Throughout the 1960s, there was a comprehensive government response to auto safety issues that can be attributed to this decrease, culminating with Congressional authorization for the federal government to set safety standards for cars in 1966. Within two years, seat belts, padded dashboards, and other safety features became mandatory equipment.

Imagine driving a car that didn't come with airbags, seatbelts, antilock brakes, and other safety measures that we take for granted today.

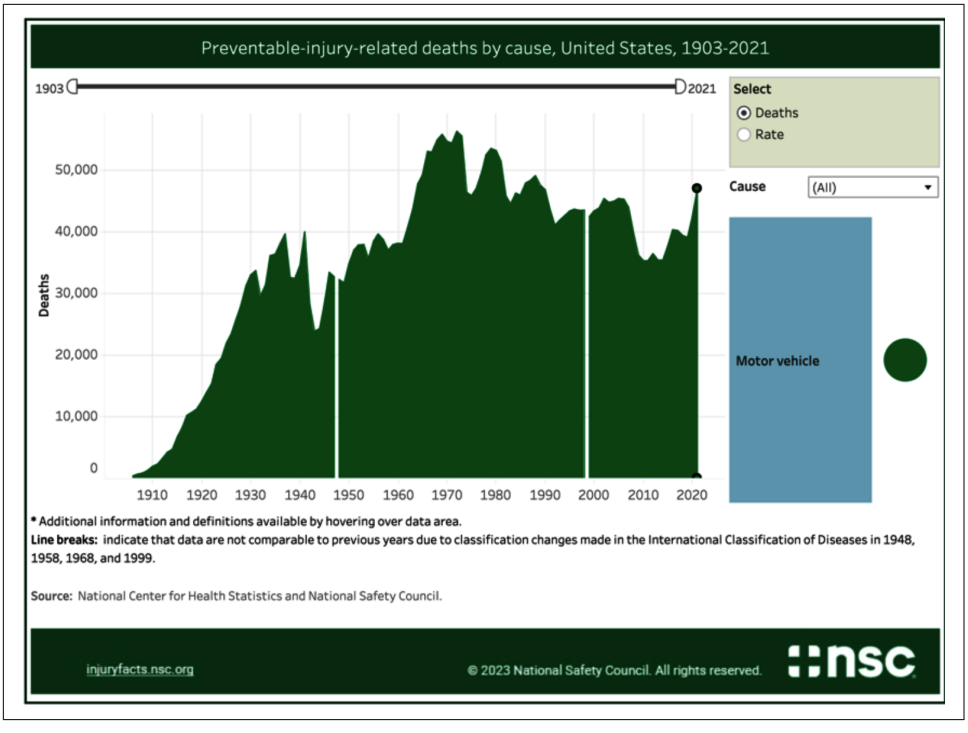


Figure 3-4. Preventable injury-related deaths by cause in the US from 1903 to 2021.

Software

Among the many similarities between the automotive industry and the software industry when it comes to the path toward regulation, safety requirements, and accountability are the ages of the respective industries when the pivotal moment arrived.

The Model T, the first mass-produced automobile, rolled off the assembly line on October 1, 1908. The NTMVS Act was passed on June 24, 1966. The time for legislation to enact safety standards was 58 years.

The first mass-produced personal computer, the Apple I, was released on April 11, 1976. So far, when it comes to legislation, we just have the intent to regulate the software industry in the President's National Cybersecurity Strategy, which came out on March 2, 2023, a total lag to date of 47 years. Acting National Cyber Director Kemba Walden said in an interview at the RSA security conference that same year that she estimated it would be another 10 years before we might see a bill signed into law, which would make it 57 years. The worrisome question that this has generated in my mind is, what would be the equivalent of a Texas City disaster or a year of peak automobile fatalities for the software industry? One thing that we know for sure is that it has to be worse than what we've already experienced.

The FBI's Internet Crime Complaint Center (IC3) issues an annual report that represents the cyber crimes that the agency knows about (see Figure 3-5). Many go undiscovered, and for those that are discovered by the victim company, an untold number go unreported because the company doesn't want the negative publicity that comes from a breach report. In other words, we can safely assume that the IC3 figures are low compared with the actual figures if we had access to them.

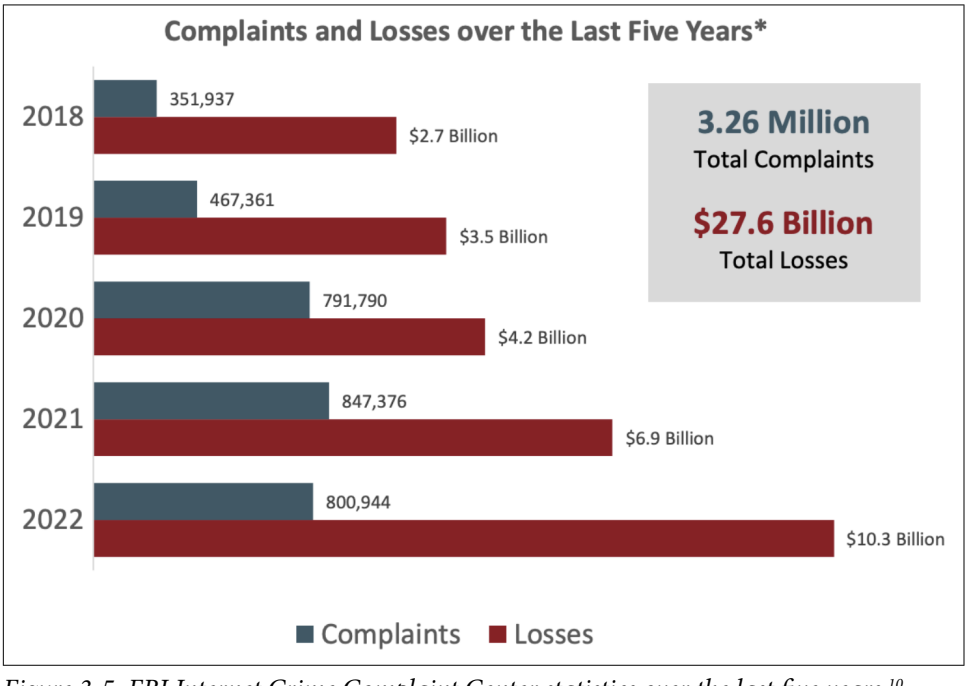


Figure 3-5. FBI Internet Crime Complaint Center statistics over the last five years.[10]

And yet, in spite of the rising losses as depicted in Figure 3-5, we see rising profits for the companies that sell cybersecurity protection. CrowdStrike Holdings, Inc. (CRWD) reported revenue of $677.4 million in the quarter ending April 2023, up from $487.8 million a year earlier, for 38.90% growth. Palo Alto Networks' (PANW) third-quarter earnings topped analyst projections according to *Investor's Business Daily*. Sales, for example, rose 26% to $2.3 billion.

The state of our security is dreadful, and yet the companies responsible for securing our networks, products, and services are making record profits. Like the auto sector before the passage of the NTMVS Act, a lack of regulation has enabled companies to tolerate poor coding practices and focus solely on sales, while the carnage of breaches

10 See the FBI Federal Crime Report 2022.

and ransomware have spawned a separate growth industry of defenders and incident responders.

The Move to Software Regulation

This lack of regulation and a lack of consequences due to "as is" warranties and the limitations of liability found in end user license agreements have brought us to an intolerable condition of vulnerability for all critical infrastructure. Breach after breach, thousands of times over, finally led to the creation of CISA in 2018.

CISA's mission is to lead the national effort to understand, manage, and reduce risk to our cyber and physical infrastructure, and it must do that with the help of the private sector, meaning the cybersecurity industry.

Jen Easterly has been one of the more vocal proponents for a new model of responsibility when it comes to cybersecurity. The CISA website puts it well:

> As a nation, we have allowed a system where the cybersecurity burden is placed disproportionately on the shoulders of consumers and small organizations and away from the producers of the technology and those developing the products that increasingly run our digital lives. Americans need a new model to address the gaps in cybersecurity—a model where consumers can trust the safety and integrity of the technology that they use every day.

In a 2023 speech at Carnegie Mellon University, Director Easterly summed her message up with three key principles:

> First, the burden of safety should never fall solely upon the customer. Technology manufacturers must take ownership of the security outcomes for their customers.
>
> Second, technology manufacturers should embrace radical transparency to disclose and ultimately help us better understand the scope of our consumer safety challenges, as well as a commitment to accountability for the products they bring to market.
>
> Third, the leaders of technology manufacturers should explicitly focus on building safe products, publishing a roadmap that lays out the company's plan for how products will be developed and updated to be both secure-by-design and secure-by-default.

While CISA is providing guidance for what the US government hopes will be voluntary participation by industry, the European Union (EU) is proposing a legislative solution, similar to the way that the EU tackled the privacy issue with its General Data Protection Regulation (GDPR).

The European Union's Cyber Resilience Act posits a twofold problem: "First is the inadequate level of cybersecurity inherent in many products, or inadequate security updates to such products and software…Second is the inability of consumers and businesses to currently determine which products are cybersecure, or to set them up in a way that ensures their cybersecurity is protected."

This concept of holding the manufacturer of software products responsible for the safety and security of what it has built seems like common sense. It applies to every other industry except for the very industry upon which every critical system relies—software.

As Is

If you read the terms of service for any software product, including cybersecurity software, you'll eventually get to the phrase "as is" found in its disclaimer of warranty section. For example, here's an abbreviated version from a Microsoft end user licensing agreement for its Microsoft Defender product (note the capitals are in the original wording):

> "DISCLAIMER OF WARRANTY. THE SOFTWARE IS LICENSED 'AS IS.' YOU BEAR THE RISK OF USING IT."

This is usually followed by an "EXCLUSION OF REMEDIES AND DAMAGES" section, also in all caps, which contains language that attempts to limit liability for any reason to the lowest amount possible—typically the amount paid for the software product or service.

For example, this section comes from CrowdStrike's *terms and conditions*:

> NEITHER PARTY SHALL BE LIABLE TO THE OTHER PARTY IN CONNECTION WITH THIS AGREEMENT OR THE SUBJECT MATTER HEREOF (UNDER ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STATUTE, TORT OR OTHERWISE) FOR ANY LOST PROFITS, REVENUE, OR SAVINGS, LOST BUSINESS OPPORTUNITIES, LOST DATA, OR SPECIAL, INCIDENTAL, CONSEQUENTIAL, OR PUNITIVE DAMAGES.

To be clear, this is not unique to CrowdStrike; every company that makes any kind of software product, app, or service has provisions like this in its end user license agreements. And it has been this way for the lifespan of the industry. Most users have no opportunity to negotiate those terms. In legal parlance, that's known as a contract of adhesion. In other words, if you don't like the terms, don't use the software. The only exception is if the user is a state or federal government agency or a multinational corporation, and that's because of the dollar value of the contract.

The question that Donald MacKenzie asked in his book *Mechanizing Proof* (MIT Press) in 2001 is still relevant today: "How can we know that the computing upon which we depend is dependable?"

Independent Testing

When you're looking to purchase an automobile, there are a number of resources that you can turn to for help in making your selection; the best of those are the ones, like *Consumer Reports,* that do not accept money from automobile manufacturers or any other manufacturer whose products they review as a way to finance their operation. *Consumer Reports* is a nonprofit corporation with no shareholders. It pays for every product that it tests, and its funding comes from subscriptions to its magazine and through grants and consumer donations.

There is no version of *Consumer Reports* when it comes to the cybersecurity industry or the software industry in general, but that's not for lack of trying. There are a very small number of cybersecurity testing companies, but all of them take money from the vendors whose products they're testing. These testing companies use protocols developed by the Anti-Malware Testing Standards Organization (AMTSO), whose stated goal according to its website is one of developing "standards and guidelines for anti-malware testing, providing advice and guidance to the expert testers among our membership and to others starting out in testing."

However, the AMTSO consists entirely of cybersecurity companies, which seemingly renders impossible compliance with the "balance of interests" requirement of the US Standards Development Organization Advancement Act of 2004 (SDOAA)—"that standards development activities are not dominated by any single group of interested persons."[11]

The AMTSO claimed to be a standards development organization in a legal motion to dismiss a lawsuit filed against it by NSS Labs. That motion was responded to by none other than the US Department of Justice in what's known as a "statement of interest," where the US attorney general can take an interest in any case pending in a federal courtroom.

The balancing of interests requirement is not achieved, the statement read, when the standards-setting body is stacked with decision makers sharing their economic interests in restraining competition. In the case of the AMTSO, they have about 60 members, 10 of whom claim to be testing labs.

The other issue is that the cybersecurity companies whose products would be tested are the ones who are creating the testing standards and paying for the tests. The entire structure is as far from independent, objective testing as one could get.

11 "This Act amends the National Cooperative Research and Production Act of 1993 to extend the same antitrust protections to standards development organizations (SDOs) while those organizations are engaged in standards development activity. The Act provides that the antitrust rule of reason applies to SDOs while they are engaged in standards development activities." Per 15 U.S.C. §§ 4301-4306.

The National Cybersecurity Strategy

The US National Cybersecurity Strategy was born out of the pain of countless and relentless cyber attacks that have harmed the nation's national security and its competitiveness in the world. It's an ambitious document that is meant to serve as a framework for future implementation with the goal of making the US a harder target for adversary cyber operations.

The document, whose creation had been overseen by the former National Cyber Director Chris Inglis and whose implementation will fall to the current director, Harry Coker, lays out five key pillars:

Defend critical infrastructure

- Expand the use of minimum cybersecurity requirements in critical sectors to ensure national security and public safety and harmonize regulations to reduce the burden of compliance.
- Enable public-private collaboration at the speed and scale necessary to defend critical infrastructure and essential services.
- Defend and modernize federal networks and update federal incident response policy.

Disrupt and dismantle threat actors

- Strategically employ all tools of national power to disrupt adversaries.
- Engage the private sector in disruption activities through scalable mechanisms.
- Address the ransomware threat through a comprehensive federal approach and in lockstep with our international partners.

Shape market forces to drive security and resilience

- Promote privacy and the security of personal data.
- Shift liability for software products and services to promote secure development practices.
- Ensure that federal grant programs promote investments in new infrastructure that are secure and resilient.

Invest in a resilient future

- Reduce systemic technical vulnerabilities in the foundation of the internet and across the digital ecosystem while making it more resilient against transnational digital repression.
- Prioritize cybersecurity research and development for next-generation technologies such as postquantum encryption, digital identity solutions, and clean energy infrastructure.
- Develop a diverse and robust national cyber workforce.

Forge international partnerships to pursue shared goals

- Leverage international coalitions and partnerships among like-minded nations to counter threats to our digital ecosystem through joint preparedness, response, and cost imposition.
- Increase the capacity of our partners to defend themselves against cyber threats, both in peacetime and in crisis.
- Work with our allies and partners to make secure, reliable, and trustworthy global supply chains for information and communications technology and operational technology products and services.

The most difficult part of these five pillars to implement will be pillar three, which calls for shifting liability to software makers, which, to date, have enjoyed zero liability with the usual exemption of gross negligence. In an interview with Stewart Baker, former NSA general counsel, Inglis was asked about the specifics of making software companies responsible for the performance of their products.

Stage one, said Baker, is voluntary compliance, followed by market forces, and then by legislation.

"Voluntary hasn't worked in 25 years," replied Inglis.

Market forces represent stage two. That would be the customers, for example, only purchasing from companies that will assume responsibility for the security and performance of their software.

One way that this could be done is by the customer demanding changes to the language of the software or cloud provider's terms of service, similar to what the State of New Jersey has done on all of its software-as-a-service and cloud service contracts.

If you want to do business with the State of New Jersey, and a vulnerability in your software product is exploited, leading to a data breach suffered by the State, you're liable for 200% of the fees paid by New Jersey during the past 12 months or a minimum of $1 million, whichever is greater. In addition, you pay the digital forensics and incident response (DFIR) bill, which, at $500 an hour per investigator, with multiple investigators being deployed for the breach of an organization the size of a government agency, means you're looking at another massive amount of money. And that's just for the organization itself. Then the vendor is obligated to pay for a credit monitoring service for the millions of victims, plus establishing a website and an 800 number.

Let's say the result is $1 per victim per month for credit monitoring. The total bill could easily put the vendor out of business.

What I've just loosely described is just the tip of the iceberg. There are lots more requirements for software-as-a-service providers, such as data security, where all State data must be encrypted at rest and in transit, rules for the physical location of servers that store the data, reporting of security breaches, a variety of IT audits, and more.

While adhering to these requirements may work for companies that are eager to land a government contract, smaller companies won't have the leverage to obtain those types of concessions, and the vast majority fall into that category, which leads to what Baker and Inglis called stage three—forced change through regulation.

Baker told Inglis that he would expect companies to take any attempt at regulation to the US Supreme Court, and that this court in particular will give the US government "a run for its money."

Both men, both veterans of the US intelligence community, acknowledged by the end of the interview that without regulatory authority over IT giants like Microsoft, Oracle, etc., we'll never have the security that we need.

Chris Inglis stepped down from his position as National Cyber Director when the National Cybersecurity Strategy was released. In an interview at the RSA security conference on April 25, 2023, Kemba Walden, then the acting director, said that she's aware of "the current political realities in Congress and the private sector around the issue," meaning that software regulation would never see the light of day outside of a committee. If a return to political civility and informed discourse is a prerequisite to being able to pass a regulatory framework, then there's no way that will happen without a precipitating crisis that almost certainly will involve mass casualties.

Summary

There is no doubt in my mind that—just like the automotive, railway, and shipping sectors before it—the software industry will experience a catastrophic failure at some point in the next 10 years, most likely involving the use of AI, that will force Congress to act and impose regulations upon Big Tech. Until that happens, we will continue to experiment with voluntary compliance, and possibly an executive order or two, none of which will work. We are not good at getting ahead of disaster, or moving left of boom.

CHAPTER 4

The Legal Status of Cyber Warfare

This chapter will look at the rapid increase of civilian hackers from around the world engaging in potentially illegal offensive cyber operations. These attacks have been happening sporadically for years but never at the scale that's been seen following Russia's invasion of Ukraine on February 24, 2022. In many countries, including the US, the unauthorized hacking of computer systems is a criminal act. This chapter will explore how the International Committee of the Red Cross (ICRC) and the International Criminal Court (ICC) interpret cyber attacks in times of war, as well as provide a tool that civilians can use to determine if their actions rise to the level of making them a combatant.

The information provided in this chapter does not, and is not intended to, constitute legal advice; instead, all information, content, and materials in this book are for general informational purposes only.

Ukraine's Call to Arms for Hackers

Thousands of individuals responded to Ukrainian Minister of Digital Transformation Mykhailo Fedorov's call for an IT Army to come to the country's defense on February 26, 2022.

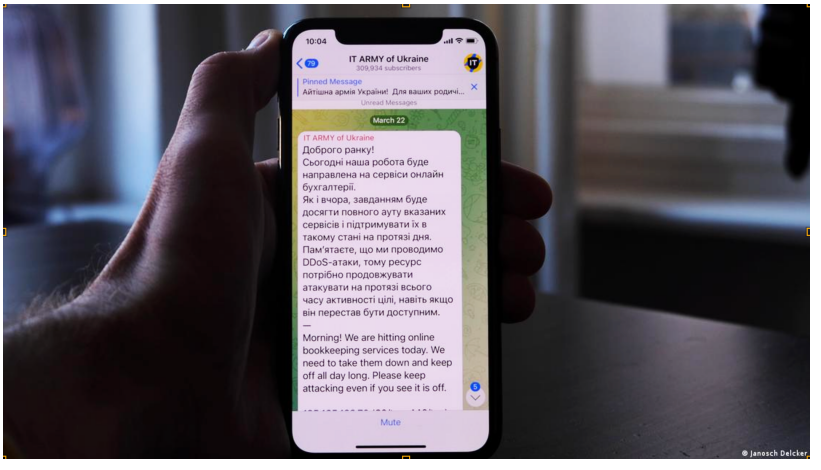


Figure 4-1. The IT Army of Ukraine used a mobile app to coordinate activities among its membership.

One of them was a middle-aged Danish IT professional who kept his involvement a secret because he knew it was illegal but he wanted to do something to help. The targets for his distributed denial-of-service (DDoS) attacks were shared using a messaging app called Telegram (see Figure 4-1). DDoS attacks make it difficult for the targeted organization's websites to operate properly. It's an old technique that members of the hacking group Anonymous have been using for the past decade (see Figure 4-2).

Figure 4-2. Anonymous's DDoS Ping Attack tool user interface. Source: Link11.com.

The downside of DDoS attacks for Ukraine is that they make it difficult for the country's military hackers to extract actionable information if the target is on their acquisitions list. Because of the DDoS, which has minimal impact other than an inconvenience, mission planning that requires data to be surreptitiously extracted is put on hold while the target organization defends its IT environment.

There are a lot of interesting questions surrounding cyber attacks during a time of war, particularly as they apply to the individuals who engage in them.

For example, do any of the hackers (Ukrainian and non-Ukrainian) who have conducted cyber attacks against Russia as part of the all-volunteer IT Army have any rights under the Geneva Convention if they're caught or captured?[1]

Under what circumstances can a non-State hacker be targeted for killing?

Can two warring States conduct cyber operations against a civilian target affiliated with a warring State but whose network resides inside the borders of a nonwarring State?

You'll be able to answer those questions, and others, after reading this chapter.

Rules Related to Cyber Attacks

Civilian hackers who join the war effort from around the world and who are only loosely controlled by the Ukrainian government have prompted the ICRC to issue guidelines for activities by those who engage in cyber attacks. The ICC has also issued a statement that it will prosecute cases involving cyber attacks depending upon the effects. I encourage you to familiarize yourselves with these rules if you or civilians you know are involved in cyber attacks or are deciding whether to become involved—these rules aren't perfect, but they're the best guidance currently available to people in such positions.

The International Committee of the Red Cross

The influx of civilian hacking activities starting with the Russian invasion of Ukraine has prompted a series of breakthroughs in international forums like the ICRC and the ICC about how cyber warfare should be conducted. Following is an excerpt of their recommendations related specifically to cyber warfare.

1 I'm using the phrase "Geneva Convention" as a way to simply refer to the various treaties and agreements that fall under international humanitarian law, with the purpose being to protect noncombatants from attack by assigning them certain legal rights and to delineate the actions by which a civilian could lose their protected status as a noncombatant. Interested readers can learn more by visiting the International Committee of the Red Cross website.

ICRC recommendations to belligerents who conduct cyber operations:

- Mitigate the impact of those operations on civilian populations.
- Refrain from shutting down civilian internet connectivity as much as possible when conducting military operations against government networks.
- Do not encourage civilians to take part in hostilities through cyber operations because it may result in the targeting of those civilians as combatants.
- Do not launch cyber attacks against medical personnel, facilities, or humanitarian operations.

ICRC recommendations to States:

- Enhance their resilience to cyber attack against critical infrastructure by developing contingency plans.
- Educate their population about the legal rules governing cyber operations during armed conflict.
- Segment military data and communications infrastructure from civilian data and communications infrastructure as much as possible.
- Regulate the manufacture of offensive cyber tools that may be used to harm civilian populations.
- Provide enhanced data and network security for humanitarian organizations.

ICRC recommendations to tech companies:

- Policies for content moderation on digital platforms should be consistent with international humanitarian law and human rights standards.
- Segment data and communications infrastructure provided for military purposes from infrastructure provided for civilian purposes.

The International Criminal Court

In September 2023, the prosecutor of the ICC, Karim A. A. Khan, published an op-ed that suggested the court would prosecute cyber attacks against civilians during wartime as international crimes under the Rome Statute.[2] A spokesperson for the ICC later clarified matters with this statement:

> The Office considers that, in appropriate circumstances, conduct in cyberspace may potentially amount to war crimes, crimes against humanity, genocide, and/or the crime of aggression…and that such conduct may potentially be prosecuted before the Court where the case is sufficiently grave.

2 The Rome Statute is a treaty signed by 120 nations in Rome on July 17, 1998, that established the International Criminal Court.

Cyber Attacks against Civilians During Wartime

This decision can be traced back to a 2021 report prepared for the United Nations, called "The Council of Advisers' Report on the Application of the Rome Statute of the International Criminal Court to Cyberwarfare". Part II of that document addresses the commission of war crimes via cyber attacks during times of international armed conflict, saying that "civilians and members of the armed forces involved in cyber operations in the context of an armed conflict may be held criminally responsible for violations of IHL (International Humanitarian Law) under Article 8 of the Rome Statute, despite the relatively recent appearance of cyber operations in the world of armed conflict."

Some interesting considerations follow that statement in the Council of Advisers report. The first one is that to qualify, there must be a war going on. A cyber attack on its own isn't sufficient or, at least, is up for debate. I'm sure it also depends on the impact of the cyber attack. If, for example, a foreign actor claimed responsibility for causing a train crash via cyber means that killed dozens of people and injured hundreds, that would certainly qualify.

Speaking of which, the report goes on to say, "the Council of Advisers assessed the question of whether or not a cyber operation could constitute armed force based on the 'effects' method, rather than the 'means' method." In other words, the physical impact generated by a cyber attack is what's important when it comes to answering that question.

One of the issues was attribution, and its reliability or lack thereof (see Chapter 2). And if the victim nation's government assigns an attribution to another state or state-controlled group, how would other states know if the evidence actually supported that finding? Adding to the complexity, the group must be under the control of the state and not simply of the same nationality (i.e., "Chinese" hackers versus hackers working for China's PLA or Ministry of State Security).

Another issue was impact, meaning, were the effects of a cyber attack equivalent to the effects of a kinetic attack? If the answer is yes, it could trigger an international armed conflict. If no, then the cyber attack doesn't meet the required bar and the victim would have to pursue other means of redress, such as suing for damages in the International Court of Justice or pursuing an action in civil court if the attack resulted in demonstrable harm.

Incitement to Genocide

The issue of genocide as regards cyber attacks is especially relevant for Russia's war against Ukraine.

The city of Kherson was occupied by Russian forces for eight months, and when it was finally liberated, Ukrainian soldiers uncovered mass graves, torture chambers, and descriptions of rape and torture against women and children. These are clearly acts of genocide punishable under international law.

The question that the Council of Advisers sought to answer was whether and how cyber operations may constitute the crime of genocide. They decided on a few criteria.

First, a group (national, ethnic, racial, or religious) must be the target.

Second, cyber operations must result in direct or secondary effects that cause fatalities among members of the group. For example, say a cyber attack against a nuclear power plant results in radiation leakage that kills civilians near the plant. Another attack might be against a transportation hub that results in a commuter crash where dozens of civilians—all members of the group—were killed or injured. For example, the ongoing genocide against the Rohingya, a Muslim group in Myanmar, by the nation's military has been fueled by information warfare campaigns on Facebook, according to a report in the *New York Times*.

Third, regardless of what was attacked, the effect must bring bodily or mental harm to the victims. And mental harm, in particular, is narrowly defined. It refers, according to the Council of Advisers' Report, to lasting damage usually caused by "threats of death; knowledge of impending death; acts causing intense fear or terror; surviving killing operations; forcible displacement; and 'mental torture.'"

Legal Review of Cyber Weapons

Col. Gary Brown (United States Marine Corps, retired) is one of the legal experts whom I often refer questions to when it comes to cyber warfare and international law. In 2014, Col. Brown and another Marine Corps officer, Lt. Col. Andrew O. Metcalf, wrote a paper about this very topic.[3] Of course, it was 10 years ago, and a lot has happened in terms of cyber capabilities in the last decade; however, their definition of a cyber weapon is just as applicable today as back then:

3 Gary D. Brown and Andrew O. Metcalf, "Easier Said Than Done: Legal Reviews of Cyber Weapons," *Journal of National Security Law and Policy* 115 (2014).

> An object designed for, and developed or obtained for, the primary purpose of killing, maiming, injuring, damaging, or destroying.

This definition's origin was for kinetic weapons, but as Brown and Metcalf point out in their paper, applying this to cyber weapons as well makes compliance much simpler.

There is a principle in international humanitarian law that parties to a conflict do not have unlimited scope to choose weapons, means, and methods of warfare.[4] The broad prohibitions in IHL that apply to all weapons will also apply to cyber and AI-enabled cyber weapons as well.

To that end, here's a rundown of what those platforms should avoid:

- Superfluous injury or unnecessary suffering
- Being indiscriminate in nature (meaning its effects must be contained to the specific target)
- Intention to cause widespread long-term or severe damage to the natural environment
- Being specifically prohibited by treaty or customary international law

For cyber and AI, they "must comply with the principles of distinction, proportionality and precaution by":

- Distinguishing between military objectives and civilian objects
- Evaluating how much incidental harm to civilians is expected and weight that against any anticipated military advantage
- Taking all feasible precautions to spare civilians and civilian objects

The Civilian Hacker Targeting Matrix

There are three conditions that must be met before the targeted killing of a civilian hacker by a state actor may occur.[5] If all three of these conditions are met, then the civilian is considered a direct participant in hostilities (DPH), which automatically makes them a legitimate target.

4 For more, see the Cyber Law Toolkit on the website of International Cyber Law in Practice.

5 The Tallinn Manual, p. 119, footnote 63, cites these three conditions stipulated by the International Committee of the Red Cross.

Threshold of harm
: The act must negatively affect the enemy's military operations or capabilities.

Causal link
: There needs to be a direct causal relationship between the act and the harm involved in the first condition. Attacks that do not meet this criterion are labeled "indirect participation" and will not open the door to targeting the individual.

Belligerent nexus
: The cyber operation needs to be about the conflict, as opposed to a random cyber attack that takes place during a conflict but is unrelated (i.e., ransomware, financial credentials theft, espionage).

A Decision Tree for the Legal Targeting of Combatants and Civilians

If you're considering joining Ukraine's IT Army, or involving yourself in a similar role during any military conflict, you should work this decision tree to see if your actions would make you a combatant. Doing so doesn't address the obvious illegality of your actions if you're a citizen of a country where hacking is illegal. Becoming a combatant brings an additional and potentially more lethal level of risk into play.

The decision tree in Figure 4-3 has been constructed from the rules of the law of armed conflict and international humanitarian law.

More information about the Tallinn Manual, referenced in the decision tree, can be found at the NATO Cooperative Cyber Defence Centre of Excellence website.

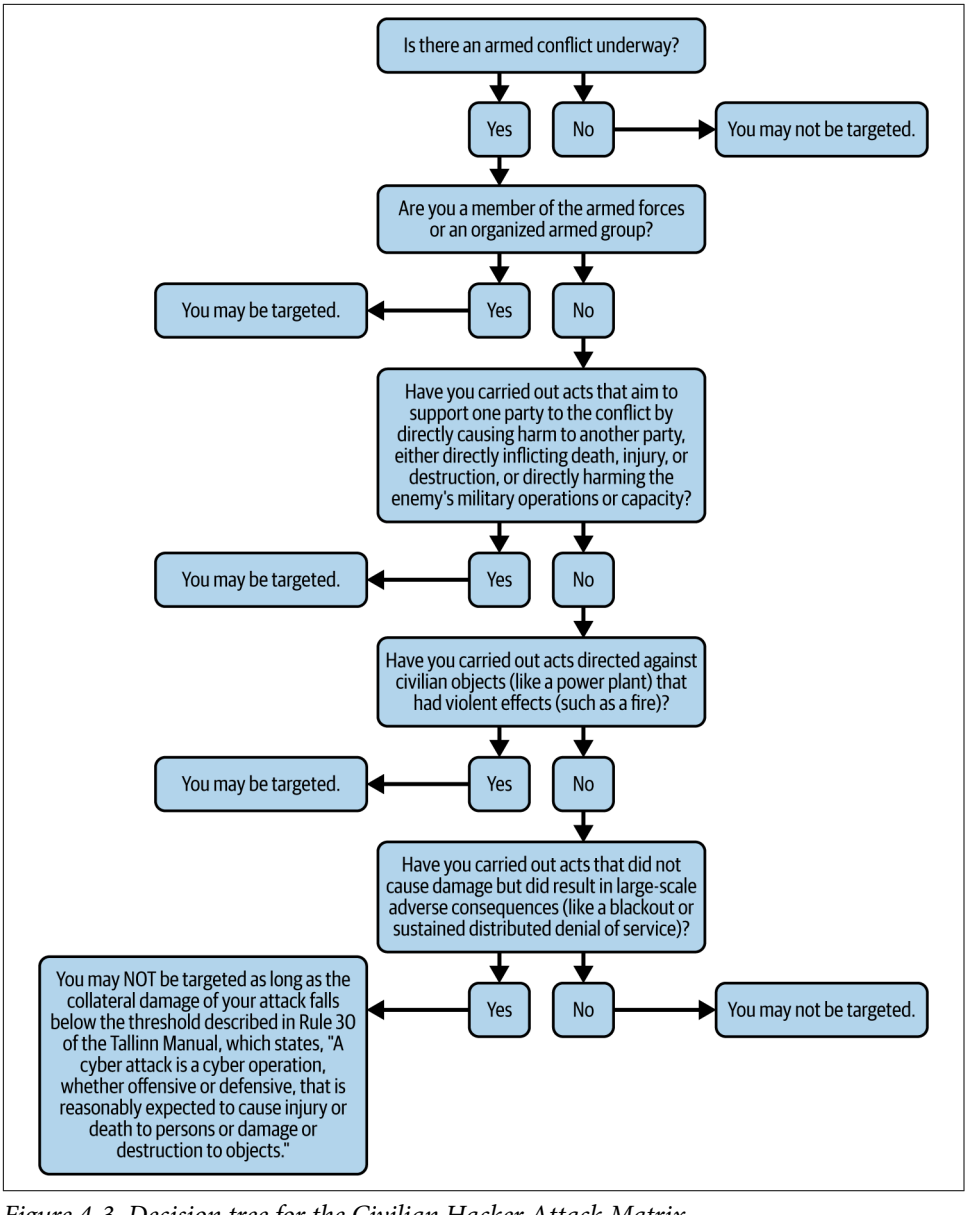


Figure 4-3. Decision tree for the Civilian Hacker Attack Matrix.

Case Studies

I've included three case studies to demonstrate how you would work the decision tree using real-world examples.

Junaid Hussain

Junaid Hussain was a British hacker who joined the Islamic State of Iraq and the Levant (ISIL) in Syria and was actively involved in recruiting sympathizers in the West to carry out attacks.

He also used his hacking skills in obtaining and releasing personally identifiable information (PII) on US military government employees. Hussain was targeted and killed by a drone strike on August 24, 2015.

The rationale was not a controversial one because Junaid Hussain's status was that of a DPH. His hacking activities may have raised his importance as a target, but it wasn't required to justify the strike.

The Anonymous War on ISIS

The online collective known as Anonymous (Figure 4-4) announced that its members declared war on the Islamic State of Iraq and Syria (ISIS) after the Paris attacks in late 2015. By "war," they meant cyber attacks against ISIS/ISIL social media accounts and websites.

Figure 4-4. A still from a video made by Anonymous and posted to YouTube.

Assuming that the Islamic State had legal status as a nation-state, and assuming that it could identify an individual hacker who participated in one of those cyber attacks, could it legally kill the person?

Let's work the decision tree and find out.

Step one: Is there a conflict underway? Yes.

- Is the hacker a member of the armed forces? No.
- Is the hacker a member of an organized armed group? No.

Therefore, under the law of armed conflict, the Anonymous hacker cannot be legally targeted.

Let's proceed to his status under international humanitarian law:

- Did the cyber attack result in death, injury, destruction, or harm to the Islamic State's ability to carry out military operations? No.
- Was the cyber attack directed against critical infrastructure like a power grid that resulted in a fire, or did it cause a blackout that resulted in casualties? No.

Neither the law of armed conflict nor IHL would support ISIS/ISIL's targeting of an Anonymous hacker who was only responsible for attacks against social media and recruitment websites.

The Ukraine Power Grid Attack

On December 23, 2015, several hundred thousand people in three districts in Ukraine lost power for one to six hours while the country continued to be in a state of armed conflict with Russia.[6] The attackers deployed BlackEnergy 3 malware against three energy distribution companies. Figure 4-5 illustrates network activity at one of the victim companies prior to the shutdown.

6 See Chapter 6 for more examples of cyber attacks that had kinetic effects.

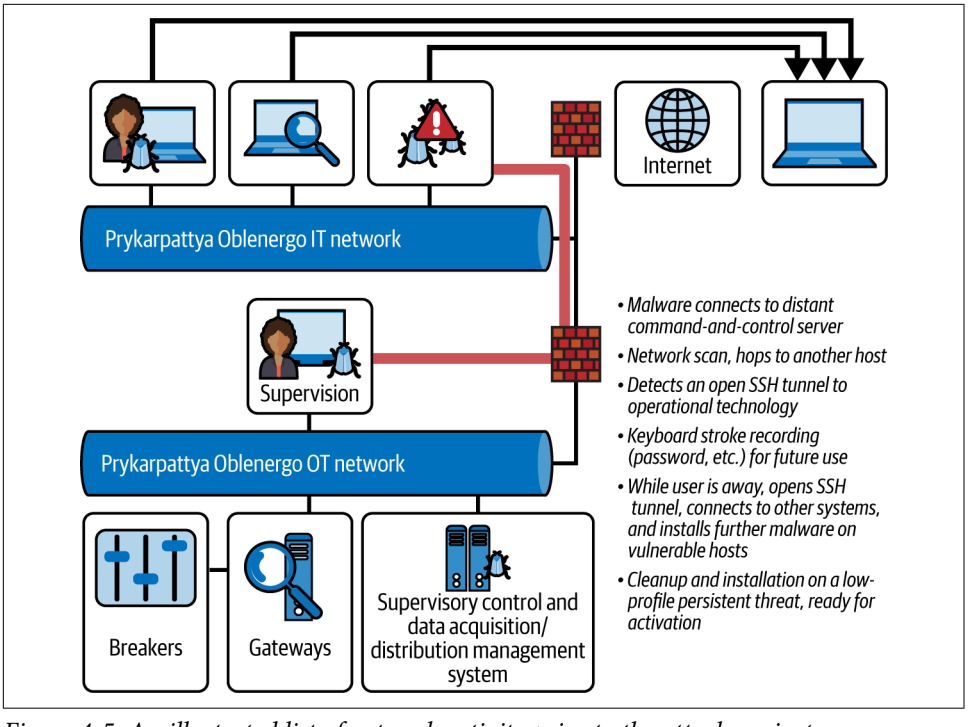


Figure 4-5. An illustrated list of network activity prior to the attack against Prykarpattya Oblenergo.

The Ukrainian government suspected Russian hackers to be responsible but stopped short of blaming the Russian government. The US government assigned the attribution to a Russian threat actor known as Sandworm. If one or more of those hackers were identified, could they be legally targeted? Let's use the decision tree again.

Is there an armed conflict underway? Yes.

Is the hacker a member of the armed forces or an organized armed group? The US government didn't officially announce attribution; however, let's assume that they did. The hacker(s) responsible would be considered a DPH and could legally be targeted.

If the hacker was a civilian, did they carry out acts against critical infrastructure? Yes.

- That had violent effects? No.
- That resulted in a blackout? Yes.

If the level of harm is sufficiently high to meet the bar established under the Right of Self Defense in Article 51 of the United Nations Charter, or under IHL, then that civilian could be legally targeted.

Summary

This chapter addresses the questionable effectiveness of cyber attacks that are left to civilian volunteers and recommends that they be weighed against the potential risks, up to and including law enforcement action if what you're doing is illegal, as well as being targeted and possibly killed as an enemy combatant. Most of the time, when it comes to volunteer cyber militias as well as individual hacktivists, the juice isn't worth the squeeze both from a risk perspective and from a strategic perspective, at least as far as the Russia-Ukraine war is concerned.

This chapter also points to recent landmark policy decisions and guidance issued by the ICRC and the ICC as they pertain to cyber attacks that generate a sufficient level of effects and trigger actions under IHL. This should inform government policy makers who must make decisions about how to respond to a cyber attack that generated casualties.

The addition of AI-enabled weapons platforms that include cyber capabilities will certainly up the ante when it comes to both effectiveness and tactical as well as strategic impact, and most likely the ICRC will have to devise new rules for AI in warfare.[7]

7 See Chapter 7 for risks associated with AI, including its use in warfare and cyber warfare.

CHAPTER 5
The New Enmeshed War Strategy

Everything we call 'cyber' is inseparable from blood and sweat: the more the cyber or remote-digital dimension of warfare intensifies, the more blood and sweat are shed—and it won't be any different.

—Svitlana Matviyenko[1]

I've never just been the financier of the Internet Research Agency. I invented it, I created it, I managed it for a long time. It was founded to protect the Russian information space from boorish aggressive propaganda of anti-Russian narrative from the West.

—Yevgeny Prigozhin[2]

For much of history, wars have been fought and won through attrition, where the winner is the one with the largest army and the most resources needed to sustain years-long conflict. The early 20th century saw the rise of electronic warfare, and the 21st century added cyber and cognitive warfare. Each, when introduced, was considered to be a separate warfighting domain, but eventually it became clear that each was more effective when used in combination rather than separately.

To further muddy the waters came the rise of social media, which for many of us replaced what became known as the mainstream media. In the early days of Twitter, for example, you could read breaking news hours before it hit the mainstream outlets. Perhaps you'd share what you learned with your Facebook friends, who, in turn, would propagate it in Facebook groups. And it was "muddy" because it was so easy to create fake news or spread misinformation, and increasingly difficult to discern what was true

1 Svitlana Matvienko is an associate professor at the School of Communication and the associate director of the Digital Democracies Institute at Simon Fraser University in Vancouver. Her quote was posted on the IT Army of Ukraine's Telegram channel.

2 Mick Krever and Anna Chernova, "Wagner Chief Admits to Founding Russian Troll Farm Sanctioned for Meddling in US Elections", CNN, February 14, 2023.

and what was false. Even harder to discern was what was true and what had only a few elements of truth to make the misinformation or disinformation more believable.

In this chapter I'll give you three case studies that show how an enmeshed war strategy was used in real-world scenarios. I'll explore several vulnerability points for disinformation, misinformation, and unwanted surveillance. And then I'll end with some practical tips for how you can navigate this new world.

This chapter contains descriptions of graphic violence.

Cognitive Warfare and Operations in the Information Environment

In cognitive warfare, the human mind is the battlefield, and warfare is conducted by manipulation of our information mediums.

In land, air, and sea warfare, the battle takes place on physical terrain, and warfare is conducted using weapons platforms in each of the warfighting domains—land, sea, air, space, and cyber.

However, since all of our communications, command-and-control systems, weapons platforms, and supply chains across all domains run on software, cyber warfare impacts all of the above.[3]

Today, cognitive, cyber, electronic, and all other wartime operations are enmeshed in a one-two punch of cognitive and kinetic warfare. The case studies you're about to read demonstrate this and involve the Russian oligarch known as "Putin's chef," Yevgeny Prigozhin.[4]

A Central Figure: Yevgeny Prigozhin

Yevgeny Prigozhin was a Russian oligarch with close ties to the country's president, Vladimir Putin, whose business holdings (see Figure 5-1) included two organizations that have been active in both cyber and kinetic conflicts from 2013 until 2023, the Wagner Group and the Internet Research Agency (IRA).

3 If you're interested in taking a deeper dive into these areas, check out Majors Andrew MacDonald and Ryan Ratcliffe, "Cognitive Warfare: Maneuvering in the Human Dimension," *Proceedings*, US Naval Institute, April 2023.

4 The *New York Times* published an insider's view into the workings of the St. Petersburg troll factory.

On June 23, 2023, after increasingly hostile outbursts by Prigozhin against the leadership of Russia's Ministry of Defense for the way they were handling the war in Eastern Ukraine, Prigozhin organized a mutiny by ordering several thousand of his men to travel to Moscow and bring justice to those who died during the war in eastern Ukraine because of incompetence and corruption in Russia's leadership.

He called off the mutiny when his men were within 100 miles of Moscow, thanks to the intervention of the president of Belarus, who negotiated a peaceful settlement between Prigozhin and Putin.

Sixty days after the fateful march on Moscow, Prigozhin, Dmitry Utkin (Wagner Group's operational commander), and several other Wagner Group members departed from Moscow on a private jet headed to St. Petersburg. The jet crashed shortly after take-off, killing everyone onboard.

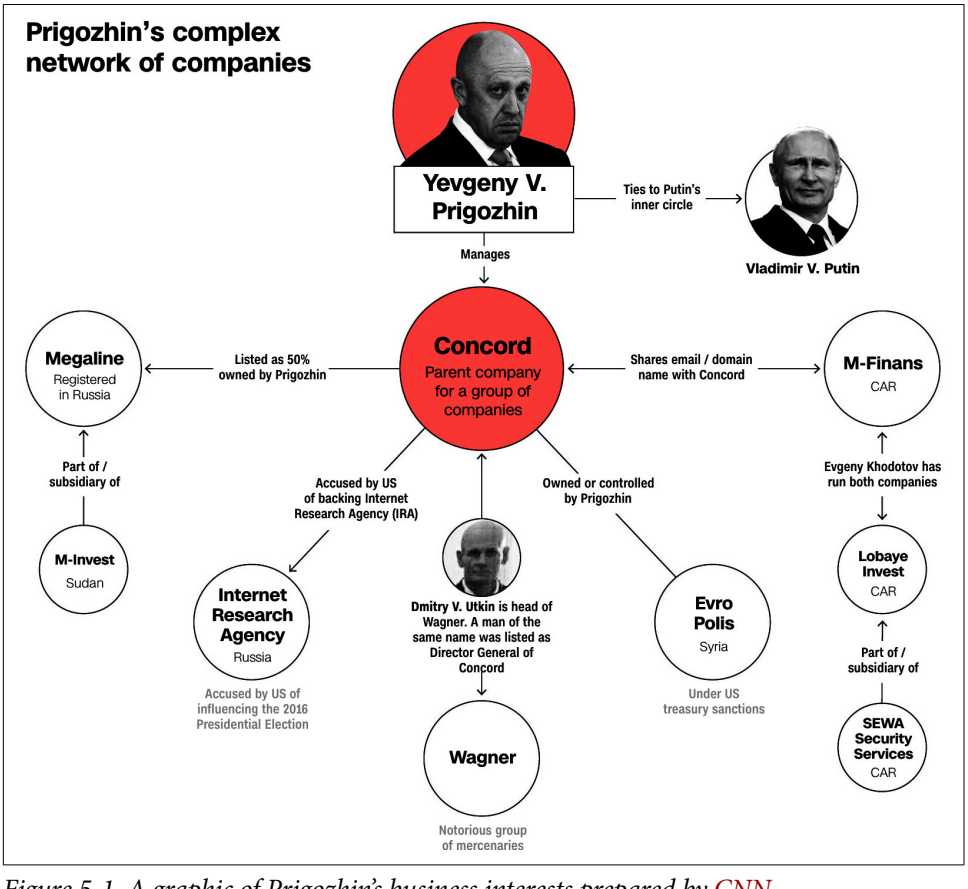


Figure 5-1. A graphic of Prigozhin's business interests prepared by CNN.

The Wagner Group

The Wagner Group was Prigozhin's mercenary army and had been committing war crimes in various jurisdictions around the world since its inception. The number of human rights violations attributed to this group is staggering. Here are just a few accumulated by Lawrence Wittner for CounterPunch:

- In Syria, the Wagner Group's soldiers, fighting to maintain the Assad dictatorship, were filmed laughing as they used a sledgehammer to break the bones of a Syrian army deserter before dismembering his body and cutting off his head.
- In the Central African Republic, UN investigators reported that the Wagner Group's forces tortured, raped, and murdered civilians, forcibly recruited child soldiers, and engaged in widespread looting.
- In Libya, Wagner mercenaries reportedly booby-trapped civilian homes with explosives attached to toilet seats and teddy bears.
- In Ukraine, Wagner Group mercenaries along with Russian forces participated in a number of war crimes, including the Bucha massacre, where, over a period of 30 days, over 400 civilians, including 9 children, were killed and buried in mass graves, some with their hands tied behind their back, others burned or mutilated.[5]

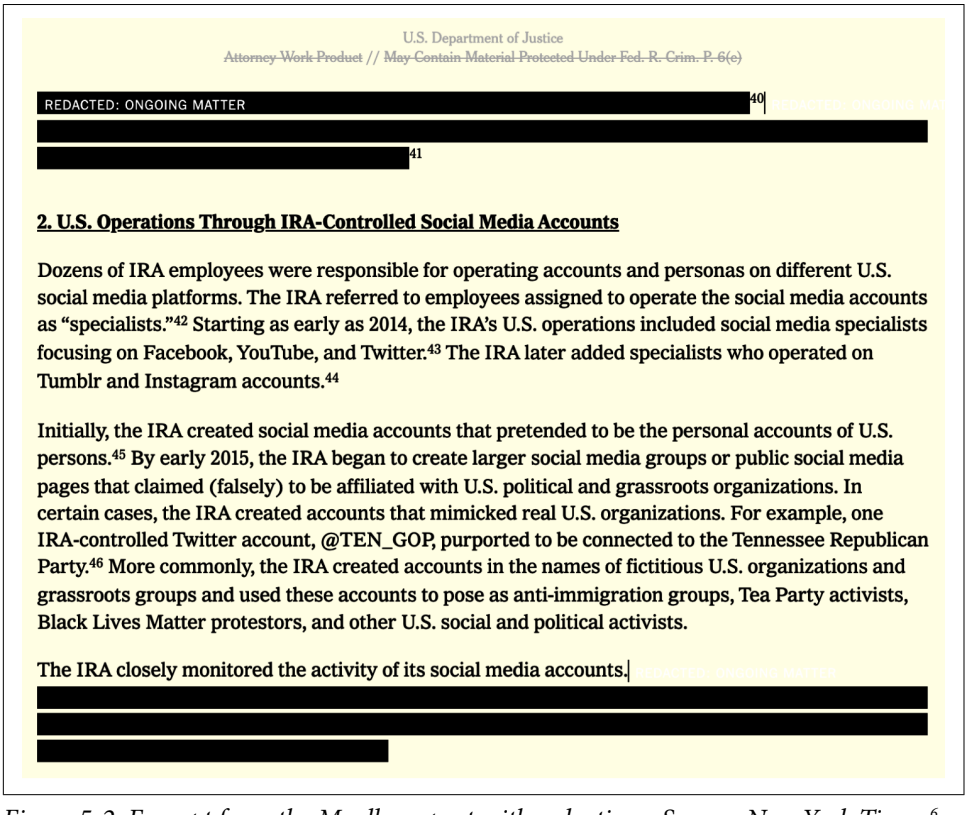
The US State Department has designated the Wagner Group as a "significant transnational criminal organization…whose pattern of serious criminal behavior includes violent harassment of journalists, aid workers, and members of minority groups and harassment, obstruction, and intimidation of UN peacekeepers in the Central African Republic (CAR), as well as rape and killings in Mali."

The Internet Research Agency

The Internet Research Agency (IRA) is a Russian company with offices in St. Petersburg and elsewhere. It employs thousands of social media trolls, meaning individuals whose job it is to set up fake accounts and spread misinformation to create dissent and inflame social tensions among the targeted population. In 2018, the IRA along with 13 individuals including Prigozhin, were indicted by a US federal grand jury for conspiring with each other for the purpose of interfering with the country's political and electoral processes, including the presidential election of 2016. The IRA and its US operations were featured heavily in the Mueller report. For example, the section seen in Figure 5-2 describes how IRA employees posed as US grassroots organizations involved in hotly debated domestic topics like Black Lives Matter and

5 This bullet point was added by the author to retain the organizational format of the chapter. The source is France24.

anti-immigration groups, along with forming pro-Trump and anti-Clinton factions on Facebook and Twitter:



U.S. Department of Justice
Attorney Work Product // May Contain Material Protected Under Fed. R. Crim. P. 6(e)

REDACTED: ONGOING MATTER [40]

[41]

2. U.S. Operations Through IRA-Controlled Social Media Accounts

Dozens of IRA employees were responsible for operating accounts and personas on different U.S. social media platforms. The IRA referred to employees assigned to operate the social media accounts as "specialists."[42] Starting as early as 2014, the IRA's U.S. operations included social media specialists focusing on Facebook, YouTube, and Twitter.[43] The IRA later added specialists who operated on Tumblr and Instagram accounts.[44]

Initially, the IRA created social media accounts that pretended to be the personal accounts of U.S. persons.[45] By early 2015, the IRA began to create larger social media groups or public social media pages that claimed (falsely) to be affiliated with U.S. political and grassroots organizations. In certain cases, the IRA created accounts that mimicked real U.S. organizations. For example, one IRA-controlled Twitter account, @TEN_GOP, purported to be connected to the Tennessee Republican Party.[46] More commonly, the IRA created accounts in the names of fictitious U.S. organizations and grassroots groups and used these accounts to pose as anti-immigration groups, Tea Party activists, Black Lives Matter protestors, and other U.S. social and political activists.

The IRA closely monitored the activity of its social media accounts.

Figure 5-2. Excerpt from the Mueller report with redactions. Source: New York Times.[6]

Prigozhin was the warlord who wielded two weapons simultaneously—in one hand the paramilitary Wagner Force and in the other the purveyor of misinformation/disinformation, the IRA. Here are three case studies that show his adeptness using real-world examples.

Case Study #1: Ukraine

This case study describes the targeting and collapse of a uniquely effective military and humanitarian services startup that was formed by a US Marine Corps MARSOC colonel and a small group of his Special Operations Forces friends, all retired.[7] When Russia

6 Mueller report, vol 1., 22.

7 MARSOC is an acronym for Marine Forces Special Operations Command.

invaded Ukraine in 2022, the residents of the capital city of Kyiv were expecting Russian troops to arrive at any moment. Women and children collected bottles, ripped up sheets, and mixed gasoline with detergent to make Molotov cocktails (see Figure 5-3) while husbands and brothers signed up to join the all-civilian Territorial Defense Forces, where they immediately were given AK-47s and ammunition. Most had no idea how to fire a rifle. Many didn't survive their first engagement with the enemy.

The MARSOC colonel, named Andy Milburn, was in Kyiv shortly after the war broke out. He had retired after 31 years of service with the Corps in 2019, wrote a critically acclaimed memoir, *When the Tempest Gathers: From Mogadishu to the Fight Against ISIS, a Marine Special Operations Commander at War,* and was writing the occasional article for Task and Purpose, a military news site, when Russia invaded.[8]

Almost immediately, Col. Milburn was asked by a Ukrainian general whom he had met a few years earlier if he could put together a one-week training regime in core combat skills to help offset the high rate of casualties his troops were experiencing. Milburn said he would, and he called a few of his friends to help. They ramped up a curriculum within days and started running several hundred people a week through the course, including lawyers, bakery employees, medical technicians, factory workers, and other civilians who suddenly found themselves with a uniform and weapons, and no idea how to use them.

Figure 5-3. A still captured from a video about civilian resistance in Kyiv.[9]

8 The information about Colonel Milburn and the Mozart Group came from my own interviews with Milburn and other Mozart Group members between September 2022 and March 2023.

9 The still was captured by the author from a video in his possession and is used with permission.

The Wagner Group's Campaign

After about four months of providing training, Col. Milburn and his growing team of trainers, who became known as the Mozart Group (a counter to Prigozhin's Wagner Group), began receiving requests to take food and water to civilians trapped on the front lines of the war in Eastern Ukraine where the fighting was worst—Soledar and Bakhmut. Prigozhin called the area a "meat grinder" where progress was measured in meters per day.

Mozart's evacuation team also extracted women, children, and elderly people who wanted to leave the front, all the while avoiding Russian small arms fire, artillery, and drone strikes. Pretty soon the Mozart Group's humanitarian work with civilians who had formerly favored a Russian-controlled Ukraine resulted in a shift of loyalty from Russia to Ukraine. The group was making a difference, and did it all without carrying any weapons and for very little pay.

Prigozhin, whose forces were suffering massive losses to the extent that he was recruiting new soldiers from Russian prisons to send into the meat grinder of Bakhmut, took notice of what the Mozart Group was achieving with four team members, a jeep, and a van. He decided to act against them by using his troll farm at the IRA to plant lies about them on social media, and by using the Wagner Group to threaten the hotels where Mozart team members stayed.[10]

The Internet Research Agency's Campaign

The social media campaign was started by Prigozhin's own post on Telegram, where he launched a rumor that "due to the heavy losses of the Armed Forces of Ukraine and the demoralization of the personnel, American mercenaries arrived in the city [Bakhmut]. According to some reports, they are personally led by retired US Marine Corps Colonel Andrew Milburn, who founded the Mozart PMC [Mozart Group]."

The Mozart Group wasn't a PMC (private military company), of course. That's partly why they had strict rules about not carrying any weapons. But that didn't stop Prigozhin.

"The surviving soldiers of the 4th operational brigade of the National Guard of Ukraine," Prigozhin continued, "can be transferred to the subordination of the American PMC. From June to November, the formation suffered heavy losses and is on the verge of being disbanded."[11]

10 I considered linking to a few of the false narratives but I do not want to give the posters any additional views. Interested readers shouldn't have a hard time finding them using Google Search.

11 Telegram account.

The same day that Prigozhin made this post, the Mozart Group's website suffered a sustained DDoS attack for several hours, followed by multiple social media accounts repeating Prigozhin's false claims.

Numerous articles appeared simultaneously in Russian media labeling the Mozart Group as a ЧВК (PMC), including an official statement from Prigozhin's own Concord Press Service:[12]

> There is a PMC [Wagner], we did not arrange contests for the name, so historically. And so it turned out that the PMC [Wagner] is the largest private army in the world and the most combat-ready. Therefore, no matter how they call themselves—Mozart or Salieri, Schubert or Kirkorov—various military formations, in any case, this does not give them strength or advantages, nothing but a hype.

Approximately two weeks after the social media blitz began, Wagner Group mercenaries threatened to bomb any hotel in the Donbas Oblast that supplied rooms to Mozart Group team members, a few of whom are shown in Figure 5-4.

Figure 5-4. A Mozart Group evacuation team planning their route to help civilians during the battle for Bakhmut in November 2022 (Photo courtesy of a Mozart Group team member).

12 The English-language version is a machine translation from the original Russian.

"There's no hotel in Donbas now that will let us stay because they've been told we're being targeted," Col. Milburn told *Newsweek* in an interview. "We have had three hotels that we stayed in hit by missiles."

There was little that Milburn and the Mozart Group could do to combat Prigozhin's strategy. With no hotels willing to give the team rooms, they had to slow their food and water deliveries to accommodate the longer driving distances. The social media campaign against Milburn and the Mozart Group had the effect of hurting donations, which the group desperately needed to pay for gas, vehicle maintenance, new tires (they were losing two or more tires a day), medical supplies, and more.

The final blow occurred in late December shortly after an interview on the *Team House* podcast. *Team House* is "a weekly livestream/podcast hosted by Jack Murphy (former Ranger/Special Forces) and Dave Parke (former Ranger/paramilitary contractor) interviewing Special Operations and intelligence community professionals about their service."[13] This was Milburn's third time as a guest, so it was very casual and done over drinks of bourbon whiskey. The episode was over two hours long, was recorded live, and included questions from the people who were watching the livestream.

A fake version of the episode was released by Max Blumenthal of the Grayzone, a far-left news site, that was heavily edited and manipulated with out-of-context quotes about corruption and war crimes in Ukraine, along with praise for Putin.[14] It's a classic example of misinformation and it was extremely effective. The fake had 3.5 million views on Blumenthal's Twitter account. The video that *Team House* made to show how the original was manipulated has received more than five thousand views.

The Mozart Group conducted its training and humanitarian missions for nine months. It trained over 6,000 soldiers and delivered more than 700 tons of food and water to women and children in areas where there was no power, little water, and no food. All of that ended on January 31, 2023, due to the long-term deleterious effects of a Russian information warfare operation combined with an insider attack—the latter being something that every cybersecurity professional knows as the most difficult attack vector to defend against.[15]

This wasn't the first time that the Wagner Group, the IRA, and Max Blumenthal had tag-teamed a humanitarian effort.

13 For more, see the *Team House* podcast.

14 For background on Max Blumenthal and his affiliation with *Russia Today*, see my article "Trolling for Putin".

15 I wrote about this at the Inside Cyber Warfare newsletter; at the time of the writing of this chapter (February 28, 2023), Colonel Milburn and a few core team members were examining options for how they could resume their training and humanitarian assistance work under a different nongovernmental organization.

Case Study #2: Syria

A Syrian civil war had been underway since 2011, when, during the Arab Spring, dissent against the abuses of Syrian President Bashar al-Assad was violently suppressed, which included Assad using chemical weapons against his own people. It is, according to *Encyclopaedia Britannica*, the second deadliest conflict of the 21st century, with at least 470,000 deaths attributed to it.[16]

The Wagner Group's Campaign

In 2015, several thousand Wagner Group mercenaries entered the war to support a contract that Yevgeny Prigozhin made with President al-Assad to protect Syria's oil fields in exchange for a 25% share of oil and gas production. The deal was made with another one of Yevgeny Prigozhin's companies, Evro Polis Ltd., which was put under sanctions by the US Treasury Department on January 26, 2018. A rare confrontation between US forces and Wagner Group mercenaries happened less than two weeks later at one of those gas plants.

On February 7, 2018, the Wagner Group had its mercenaries mixed in with Assad's forces near the Syrian city of Deir al-Zour at a Conoco gas plant. According to the *New York Times*, the US had about 30 Special Operations forces soldiers in the area (Delta Force and Army Rangers assigned to Joint Special Operations Command) who were supporting Kurdish and Arab forces. They were backed up by a quick reaction force of Green Berets and Marines about 20 miles away, who were also monitoring the buildup of about 500 Russian and Syrian troops and 27 vehicles, including Russian T-72 tanks and armored personnel carriers.

The US government was concerned that a direct engagement between US and Russian troops in Syria could lead to the two nuclear superpowers going to war against each other, so the general in charge of Special Operations forces in Syria contacted his counterpart in Russia on a previously established "deconfliction" telephone line to ask if any Russian forces were operating in that area.

The answer from Moscow was no. Whoever the Russians were that were amassing with Assad's forces near the Conoco plant, they weren't Russian military. This is a classic example of how Putin utilized plausible deniability for Wagner Group deployments up until September 26, 2022, when Prigozhin announced publicly for the first time that he founded the group, after years of denying that he had anything to do with it.

16 The number one spot belongs to the Second Congo War (1998–2003), which started with the Rwandan genocide. Over 3 million people were killed either by the fighting or by disease that came about because of the devastation that the war brought to the people of the region.

When Assad's forces, along with Wagner mercenaries, launched their attack, the battle lasted for four hours and ended with a US victory thanks in large part to the use of rocket artillery and airstrikes from Reaper drones, F-22 stealth fighter jets, F-15E strike fighters, B-52 bombers, AC-130 gunships, and AH-64 Apache helicopters. The Russian independent news outlet Znak.com reported that 217 Russian mercenaries were killed in the incident.

Intercepted audio from Wagner Group mercenaries painted a graphic picture:

> The reports that are on TV about…well, you know, about Syria and the 25 people that are wounded there from the Syrian *#&$ Army and—well...to make it short, we've had our asses *#&$ kicked. So, one squadron *#&$ lost 200 people…right away, another one lost 10 people…and I don't know about the third squadron but it got torn up pretty badly, too...So three squadrons took a beating…The Yankees attacked…first they blasted the *#&$ out of us by artillery and then they took four helicopters up and pushed us in a *#&$ merry-go-round with heavy-caliber machine guns…
>
> Out of all vehicles, only one tank survived and one BRDM [armored reconnaissance vehicle] after the attack, all other BRDMs and tanks were destroyed in the first minutes of the fight, right away…
>
> (O)ur *#&$ government will go in reverse now and nobody will respond or anything and nobody will punish anyone for this...So these are our casualties.

Almost as if on cue, Russian Foreign Ministry spokesperson Maria Zakharova attempted to diminish the embarrassing loss by calling it a disinformation campaign by the US media, and said that the actual number of Russian citizens killed was five (see Figure 5-5).

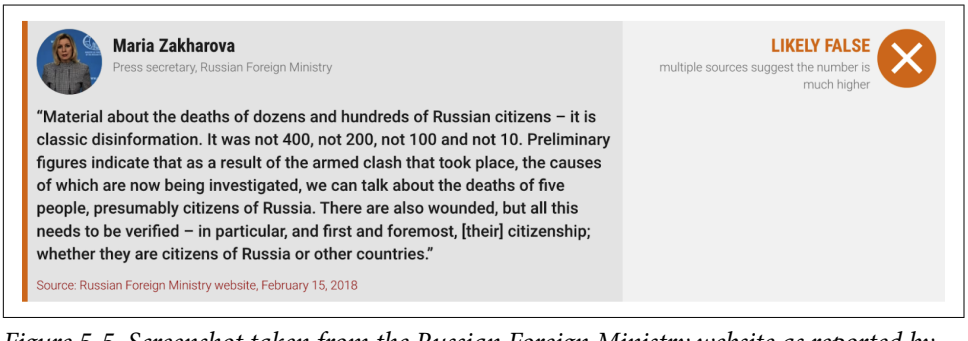


Figure 5-5. Screenshot taken from the Russian Foreign Ministry website as reported by Polygraph.

In other words, apply the classic gaslighting strategy of deny, deny, deny. Messaging, not the facts, is what matters in the information environment.

The Internet Research Agency's Campaign

Similar to the Mozart Group, although much, much larger, the White Helmets (aka Syria Civil Defense) were a group of volunteers who formed in 2012 to do civilian rescues, evacuations from hot zones, delivery of food and water, and other essential services. Members would wear head-mounted cameras while they entered buildings that had been turned into rubble, looking for survivors. In 2014, they grew to three thousand members and received millions in funding from the US government, Western Europe, and Japan. James Le Mesurier, a former British Army officer, was one of the founders. Most importantly, Le Mesurier stood up a nonprofit organization called Mayday Rescue to act as a conduit for donors and process the millions in funding that was coming in.

A Netflix documentary, called *The White Helmets*, was made in 2016 about the group's work, especially its filming of Assad's use of a chemical weapon called sarin gas. The documentary won an Academy Award in 2017 (Netflix's first Oscar win for a documentary short film), and the group was twice nominated for the Nobel Peace Prize.

But along with the attention came a big problem. The videos captured by White Helmet volunteers showed clearly that Russian and Syrian forces were indiscriminately killing civilians, and not just Islamic terrorists as Russia and Assad claimed. That was clear evidence that President Assad and President Putin were engaging in war crimes. Harming the public image of the White Helmets and James Le Mesurier became job one for Russian state media outlets *RT* and *Sputnik* and for Russia-paid US journalists like Max Blumenthal and others at the Grayzone, as well as for Prigozhin's professional trolls at the IRA.[17] Attacking the funding arm is a surefire way to hurt an organization. In fact, it ultimately led to Le Mesurier taking his own life on November 11, 2019.[18]

Enter the IRA.

The IRA troll farm in St. Petersburg began a relentless push of fake news stories about how the White Helmets were staging their rescues, the "victims" were actors, and James Le Mesurier was supposedly a terrorist, a pedophile, and an organ trafficker.

17 For additional information on Max Blumenthal's history with Russian media, see my article about him.

18 All of the details surrounding this tragic ending of Le Mesurier's life deserve to be read in full. The BBC produced a lengthy report and accompanying podcast entitled "Mayday: How the White Helmets and James Le Mesurier Got Pulled into a Deadly Battle for Truth".

Unfortunately, the White Helmets' participation on social media in the Mannequin Challenge (see Figure 5-6), a viral TikTok sensation with over a billion views where you "freeze like a mannequin" in the middle of doing something you love, gave the trolls ample ammunition to back up their claims.

The video was created by the Revolutionary Forces of Syria (RFS), ostensibly to bring greater awareness of what was happening in Syria to western audiences, according to an RFS spokesperson who was interviewed by CNN.

Figure 5-6. Volunteers from Syria's White Helmets organization in a Mannequin Challenge; video courtesy BBC.

The backlash was immediate, vicious, and predictable.

On November 23, 2016, *Russia Today* ran this headline: "White Helmets 'Deserve an Oscar' for Mannequin Challenge Performance in Syria War Zone." The article went on to state, "(t)he footage shows the rescuers standing motionless over what looks like a realistically wounded man. So realistic, it's raising serious questions about the authenticity of the White Helmets' other frequently posted videos."

The State-run news channel created its own version of the video (see Figure 5-7) with inserts of critical tweets followed by incriminating charges against the White Helmets and accompanying photos and posted it to Facebook Watch.

In an effort to do damage control, the White Helmets apologized for the video but said the video had not been sanctioned by the group's leadership team.

"The video and the related posts were recorded by RFS media with Syria Civil Defence (White Helmets) volunteers, who hoped to create a connection between the horror of Syria and the outside world, using the viral Mannequin Challenge," the statement read. "This was an error of judgment, and we apologize on behalf of the volunteers involved."

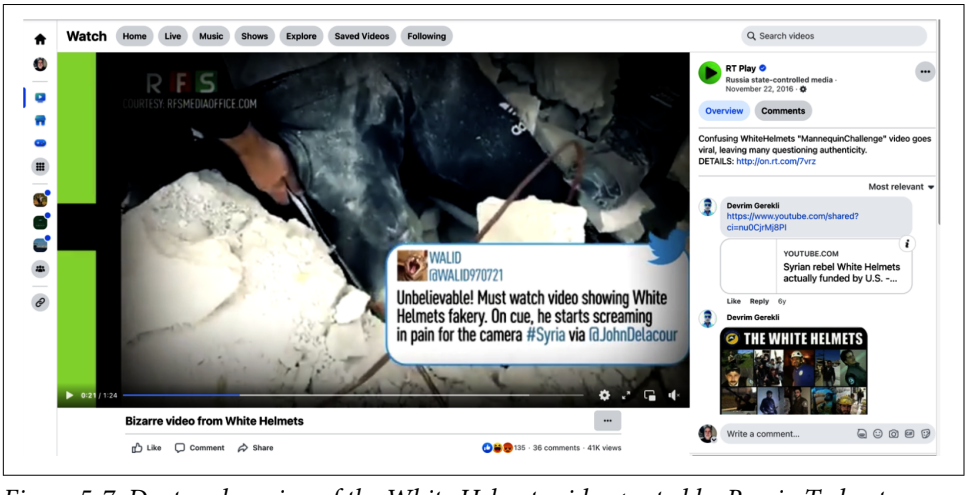


Figure 5-7. Doctored version of the White Helmets video posted by Russia Today to Facebook.

Analytics firm Graphika did a comprehensive study on the social media campaign that targeted the White Helmets and found that "bots and trolls linked to Russia have reached an estimated 56 million people with tweets attacking Syria's search and rescue organization, the Syria Civil Defence—also known as the White Helmets—during ten key moments of 2016 and 2017. Many of these smears are linked to efforts to promote false information about the sarin chemical attack of April 2017 in Khan Sheikhoun, which UN investigators concluded were carried out by Russia's ally, the Syrian government of Bashar al-Assad."[19]

19 For more, see *https://oreil.ly/ky3Vj*.

Case Study #3: Mali

In late 2021, Mali was a nation in chaotic disarray thanks to two military attempts to seize power and a terrorist threat by Islamic militants linked to Al Qaeda or ISIS.

The Wagner Group's Campaign

The Wagner Group deployed to Mali in December 2021 at the request of the government, which decided to deploy them instead of using its own military armed forces. Wagner forces already had contracts in the Central African Republic and Sudan at the time, so it was relatively easy to sell the Malian government on their advantages. Figure 5-8 shows satellite imagery of one of their deployment locations. Aside from maintaining order, the Wagner Group offered a direct line to Russian military aid, equipment, and training in exchange for lucrative mining agreements.

Figure 5-8. Satellite imagery of a suspected Wagner Group operating base in Mali, in a report created by analysts with the Center for Strategic and International Studies.

The Internet Research Agency's Campaign

Several months before an agreement was reached, either the IRA or another Prigozhin-owned media outlet began a disinformation campaign through a "coordinated network of Facebook pages in Mali that promoted Russia as a 'viable partner' and 'alternative to the West,'" encouraged the postponement of democratic elections, and attempted to

create local support for Wagner. Local news outlets often mirrored these narratives by publishing interviews with Russian officials who extolled Wagner "advisers."[20]

Platforms for Disinformation and Misinformation

X (formerly Twitter), Facebook, TikTok, and dozens of smaller social media platforms are the preferred platforms for delivering disinformation and misinformation to an audience because doing so is easy and inexpensive, and there are no laws preventing it.[21] Let's look at some examples for the major platforms.

X

Impersonation, fake accounts, parody accounts, and so on have always been a part of X. Anonymity was part of its appeal, but it was also responsible for much of its nastiness. For public figures and celebrities, being impersonated could have serious repercussions for their image and may hurt their fan base as well if the impersonator uses their access to, for example, promote a scam.

So in 2009, Twitter tested and eventually enacted a blue check mark verification symbol for accounts of people that were likely to be impersonated. Over the years, it didn't work very well and no one was clear about exactly how one went about receiving a *verified badge*.

In 2021, Twitter worked on improving the process by allowing users to apply to be verified.

On November 2, 2022, shortly after Elon Musk bought Twitter, he made it a profit center by charging for a blue check mark and announced it with this *tweet*: "Twitter's current lords & peasants system for who has or doesn't have a blue checkmark is ! *&@#$. Power to the people! Blue for $8/month."

However, Musk dropped the most critical part of the verification criteria: the part that confirms the account is *authentic*. It also dropped the other two criteria—that an account is notable (defined as a "prominently recognized individual or brand") and active (six months with a confirmed email address or phone number, and no lockouts).

Almost immediately, Russian propaganda accounts based outside of Russia paid the $8 per month for the blue check, according to the *Washington Post*.

20 For more, see *https://oreil.ly/sgUGF*.

21 Section 230 of Title 47 of the US Code provides immunity to online content providers.

Musk, with 129 million followers, "boosted one of the accounts by replying to its tweets, including one spreading a lie that thousands of NATO troops had died in Ukraine," writes Joseph Menn, the journalist who broke the story.

X is just one of many social media platforms that are utilized by disinformation operators to disseminate propaganda, false narratives, and harmful memes.

EEAS

The European Union's European External Action Service (EEAS) published its first report on foreign information manipulation and interference (FIMI) on February 7, 2023. It called out Telegram, Facebook, and Twitter as the most frequently used channels for the 100 incidents used in their study. These incidents occurred between October and December of 2022.[22] Other platforms mentioned were YouTube, Snapchat, Vimeo, Telegram, Rutube, TikTok, and Douyin.

The study found that in 66 out of 100 cases studied, the objective of the posts was to provide support for Russia's invasion of Ukraine as justified and necessary. Russian diplomatic channels were frequently used in tandem with social media posts. Impersonation of trusted organizations and individuals was a common tactic. Posts were multilingual according to who the audience was. The most common tactics observed by the researchers included shifting blame, distorting context, and distracting attention from what was actually going on.

Some examples along with timelines are provided in the report, such as this sequence for the week of October 31 to November 6, 2022:[23]

"Eight incidents targeted Ukrainian officials and the Ukrainian Armed Forces. The Russian FIMI ecosystem produced videos and images implying that Ukrainian people do not support their President and Ukraine is supportive of Nazism. Moreover, Telegram was used by Russian FIMI actors to claim an alleged hack against NATO and Ukraine military systems."

TikTok

What would happen if the Chinese government ordered ByteDance, the parent company of TikTok, the West's most popular social media app, to turn over all the data it has collected on US military and government employees? Would TikTok comply?

22 "1st EEAS Report on Foreign Information Manipulation and Interference Threats", European Union European External Action Service, February 2023.

23 Ibid.

The app has repeatedly said that it is a separate entity from ByteDance, as evidenced by its headquarters being outside of China, its deal with Oracle to store its data on servers geolocated in the US, and the fact that, if asked such a question by Chinese authorities, it would not comply.

That might be a believable statement if ByteDance were a company based in any other country but the People's Republic of China. The Chinese government has proven time and again that it has little tolerance for companies and individuals that exert independence from direction received from Beijing. Alibaba and its billionaire founder, Jack Ma, are the most visible example.

In 2020, Ma, China's richest man at the time, was repeatedly asked by the Chinese government to share user data on its Alipay app. The app is used by over a billion Chinese users (mostly young people) for purchases and loans, since young people in China have a hard time getting credit. Alipay had a deal with about a hundred banks that would provide a majority of the funding for loans and assume 100% of the risk while Alipay made a fortune as the middleman. The Chinese government saw that as an unfair competitive practice.

Ma not only refused to cooperate, but in October he also publicly accused the Chinese government of overreach and included Chinese President Xi Jinping in his criticisms.

In November 2020, Ma was summoned to appear before Chinese regulators, his pending initial public offering for Ant Group, the owner of Alipay, was pulled, and Ma himself disappeared for about three months.

Other high-profile Chinese citizens who have mysteriously disappeared due to conflicts with the Chinese government include:

- Tennis player Peng Shuai, not seen since 2019, when she accused a high-ranking Chinese official of sexual assault.
- Actor Zhao Wei, a very popular actor and pop star who was a billionaire, was deemed to be a poor example for young people and, in August 2021, vanished from public life. All her films and TV appearances were scrubbed from streaming services.
- Meng Hongwei, a former Chinese government official who became the director of Interpol in 2016, was arrested in China in 2018 for corruption. The move was seen as being not so much about taking bribes as about Hongwei representing a political threat to President Xi. Hongwei is serving a 13-year prison term.

The obvious takeaway is that if you're a Chinese company, you don't say "no" to a request by the Chinese government, not even if you're the richest man in China or the director of Interpol.

For Western nations whose population uses TikTok—and who recognize it as a national security threat for those users who hold security clearances, who are employed by the armed forces, and who in whatever capacity might be of interest to China's intelligence services—there's very little that they can do except ban the app entirely, and that won't solve the problem.

The US government has banned TikTok from being used on government-owned phones, but that doesn't keep it off the personal devices of its employees. The app is the most popular in the world for a reason. It's highly addictive and fun to use. There are so many military members using it that there's an entire community known as "MilTok" that shares funny clips of soldiers in uniform firing machine guns, dancing to whatever meme music is popular at the time, and, especially during the first two years of COVID-19, complaining about mandatory vaccines. Yet TikTok, like every app, has the capability of turning its users' smartphones into 24/7 surveillance devices.

You might think that such an act would instigate reprisals at a State level, but in fact, it would qualify as a legal act of espionage between States, assuming that the surveillance is being done by China's foreign intelligence service, the Ministry of State Security, and the target of the surveillance is US military or other government employees.

A Western government that wants to take action to prevent such intelligence collection from happening has few options.

One is to ban TikTok nationwide. The likely result of a ban, however, is that a certain percentage of the population would find a way to continue to use it, and others who haven't been using it might now want to see what all the fuss is about. The one thing we know for certain is that some people would hate the ban, laugh at the idea that their daily lives would be of interest to the Chinese government, create "From my cold, dead hands" memes that substitute the TikTok app for a rifle, and generally dedicate every waking hour of every day to fighting it.

Another option is to educate government employees about the national security risks involved with the app's use and/or ban the use of personal devices when they are on deployment. Military and government employees would have to carry government-owned and government-controlled smartphones and would be penalized if caught using any non-government phone.

For that group, the education process should include examples pulled from the Russia-Ukraine war, of which there are plenty. In every case, the end result is capture and interrogation, or death. That's a pretty good incentive to not play with apps while on deployment.

Using Social Media for Surveillance

On March 21, 2022, a Russian sympathizer spotted some armored vehicles parked in a corner of a shopping center in Kyiv. He posted a video to TikTok, and shortly after, the shopping center was hit. Figure 5-9 shows a fraction of the total damage. This was one of the early lessons of the war—that Russian intelligence closely monitors social media as well as mobile phone transmissions in contested areas.

Figure 5-9. The Kyiv shopping center where the armored vehicles were spotted and bombed.

What happened next is also illuminating. The GUR, Ukraine's military intelligence agency, polled the cell phone tower that served this shopping center, ran all the numbers, narrowed it down to three possibles, and ran those through several ad tech company subscriptions that served ads on TikTok, Telegram, and other social media. Within minutes, it identified the person responsible—a Belarus resident—and sent a team to track and capture him using his cell phone as a beacon for his position.

F3EAD

As the previous example with TikTok illustrates, smartphones and the apps installed on them, especially the social media apps, are being used in targeting combatants as part of a Find, Fix, Finish, Exploit, Analyze, and Disseminate (F3EAD) methodology.[24] F3EAD goes back to combat operations in Iraq in 2010, where Special Operations forces teams would hunt high-profile targets (HPTs) and, when found, kill or capture them and gather portable hard drives, laptops, mobile phones, etc., for exploitation and analysis, and then disseminate the findings to better inform the next operation against the next HPT. All of this was done faster than the enemy could react, thus placing the enemy perpetually on the defensive.

In 2022 and up to the present in the Russia-Ukraine war, Ukrainian Special Forces, with the assistance of GUR cyber operators, are leveraging popular Russian social media platforms like Telegram and VK to find targets of interest. They'll apply facial recognition software to posted photos of the user in order to crack their anonymity and supplement that with information pulled from a commercial data broker, then track them in real-time using the app's access to the mobile device's Location Services feature. Many apps work best when that feature is enabled.

Once the target's approximate location is identified, a surveillance drone will be launched to visually confirm the identity of the target, and depending on the terrain and other conditions, a decision will be made to capture or kill the HPT and collect any electronic hardware in their possession for analysis.

An example of this process was what happened to an FSB detachment Unit 607 assigned to Unit 6762 of the Russian military, which is part of the Ministry of Internal Affairs, and based in the city of Zheleznovodsk, a part of Stavropol Krai in the North Caucasus region of southern Russia.

Over the years the unit has been deployed to suppress riots, combat terrorism, and generally participate in the territorial defense of Russia.

Ukrainian hackers working for the GUR were able to gain access to personnel files for both Units 607 and 6762, including contact information such as email addresses and mobile phone numbers. That information was later used to track their location and a GUR Special Operations forces team was sent to engage. After a successful operation, the team returned with a heavily fortified and encrypted FSB laptop (see Figures 5-10 and 5-11).

24 Charles Faint and Michael Harris, "F3EAD: Ops/Intel Fusion 'Feeds' the SOF Targeting Process", Small Wars Journal, January 31, 2012.

Figure 5-10. A Russian-made field communications unit used by Federal Security Service soldiers in Ukraine.[25]

The laptop was made by TS Computers, a Russian company that specializes in ultra-secure industrial notebooks that are highly customizable.

The removable hard drive was password protected, but one of the Ukrainian hackers who stood up the offensive cyber unit for the GUR was able to eventually crack it. The information pulled from the hard drive was sent to analysts and their findings were passed to the officers responsible for planning operations.

25 Provided to the author by the photographer only for use in this book.

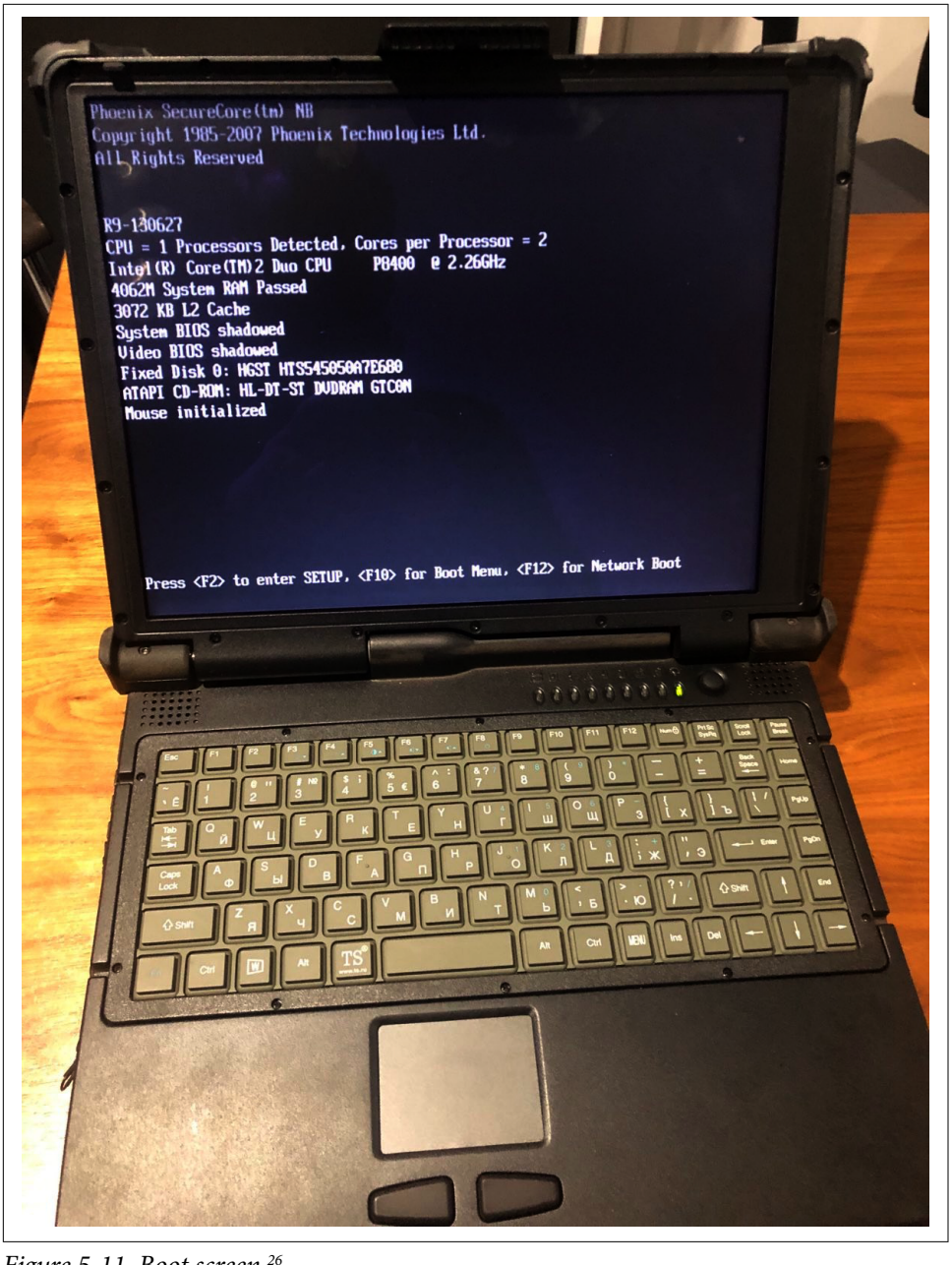


Figure 5-11. Boot screen.[26]

26 Ibid.

Benign Surveillance (Not) and Real-Time Bidding

Data may well be the new oil, but in the digital economy, attention is the scarcer resource.[27]

After the internet bubble crash of 2000, Google needed to find a way to become profitable. Its search engine wasn't generating enough revenue through licensing fees, so Google's founders, Larry Page and Sergey Brin, decided to focus on enhancing ad sales by making the process simpler for advertisers. As tech writer Shoshana Zuboff explained, "Ads would no longer be linked to keywords in a search query, but rather a particular ad would be 'targeted' to a particular individual."[28]

In 2003, Google scientists filed a patent application titled "Generating User Information for Use in Targeted Advertising". The company had managed to turn advertising from an art to a science by using the "digital exhaust" of its users' interactions with Google Search and the personal information provided by the user and third-party apps to predict which ads the users would click on. Google collects everything—your Gmail messages, your search queries, your text messages or chats, everything. It all goes into the algorithmic formula used to assess which ads to serve you, and when. The same process is done by Meta, but Meta has a lot more personal data to play with—basically everything that we post to Instagram, Facebook, Messenger, and WhatsApp feeds the Facebook Ads engine.[29]

Real-time bidding (RTB) takes a programmatic approach to serving ads that involves the publishers (i.e., websites or apps) on one side and advertisers on the other. Both publishers and advertisers are represented by technical intermediaries. You can think of them as agents who handle the technical details of the ad sale.

The publishers' technical intermediary is the supply-side platform (SSP). The advertisers' is the demand-side platform (DSP).

The website or app has an inventory of spaces to be filled by ads. It gives those open slots to one or more of its "agents" (the SSP) to sell at auction at the ad exchange. However, what's really at auction is the website or app user's attention to the ad that will fill that slot. The agent (DSP) for the advertiser, whether it's P&G or Nike or Chase, will make a bid based on how closely the user of the app matches the profile of the ideal customer that the advertiser wants to view its ad. The profile data is conveyed to the DSP via a bid request, which looks like this very simplified format:

27 "The Attention Economy: Exploring the Opportunity for a New Advertising Currency," Dentsu Aegis Network, June 2019.

28 Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (New York: PublicAffairs, 2019).

29 Meta claims that WhatsApp uses end-to-end encryption; however, an investigation by ProPublica shows that that isn't the case.

Site

- URL of the site being visited
- Site category or topic

Device

- Operating system
- Browser software and version
- Device manufacturer and model
- Mobile provider
- Screen dimensions

User

- Unique identifiers set by vendor and/or buyer
- Ad exchange's unique person identifier, often from cookies
- The DSP's user identifier, often taken from the cookie of the AdX, which has been synced with a cookie from the DSP's domain
- Year of birth
- Gender
- Interests
- Metadata reporting on consent provided
- Geography
- Longitude and latitude
- Postal/zip code

If you're looking at this list and wondering what all the fuss is about, you're right. The level of data isn't worth very much. So like any good agent, the DSP will work to enrich the profile with very granular details that will up its value at auction. The more details, the more successful the DSP will be at winning the open slot. And for that, it turns to a data management platform that pulls from hundreds of sources to enrich the customer data profile of the user.

The data management platforms use a standardized taxonomy of 1,679 personal characteristics created by the Interactive Advertising Bureau (IAB). In a report to the Irish Data Protection Commission, Dr. Johnny Ryan of the Irish Council for Civil Liberties documented the following sample points of collection:

- Personal affluence, for example, "very low net worth" (IAB code: 193)
- Household, for example, "rural" (IAB code: 147)
- Personal debt (IAB code: 537)
- Monthly mortgage payments (IAB codes: 114-130)
- Interest in buying, for example, "bail bonds" (IAB code: 1495)
- Political views, for example, "conservative" (IAB code: 199)
- Interests, for example, "vaccines" (IAB code: 404)
- Health-related interests and purchases, for example, "weight loss" (IAB code:
- 414)
- Buying interests relevant to law enforcement (IAB code: 891)
- Matzoh food (IAB code: 1097)

It's this system that was, for example, abused by Cambridge Analytica in its Facebook scandal, as reported by the *New York Times* in 2018. Data broker OnAudience used RTB data to profile LGBTQ+ people in Poland to influence its 2019 Parliamentary elections.[30] And data broker Mobilewalla used RTB data to profile 17,000+ participants in Black Lives Matter protests.

Best Practices

The bulk of this chapter focused on the combination of cyber warfare, information warfare, and kinetic warfare and how they've all become enmeshed over time. It's beyond the remit of this book to tackle giving advice on best practices in a combat zone, so I'm going to limit this to cyber and information warfare because what I'm about to propose are effective strategies for both.

Disinformation and Misinformation

Successful disinformation and misinformation campaigns rely upon several factors that we each have the ability to control in order to minimize their impact. These factors include a noisy information environment (i.e., social media) and our natural propensity to engage in some form of cognitive bias as we consume information. Our goal is to reduce the noise and challenge our biases at the same time.

30 "Creating Custom Segments for 'I Vote for Love' Campaign," OnAudience.com, accessed September 15, 2020.

Treat your news consumption the same way that you treat your food consumption. You have consequences that can range from mild to lethal if you aren't careful about what you eat. You have similar consequences if you aren't careful about what you read, watch, or listen to.

So, just like you create healthy eating habits for yourself, you can do something similar with how you consume information. The origin and diversity of information sources are two key principles to apply to your news aggregator.

My preference is to use an RSS news reader, since it helps you take charge of what you consume instead of letting an algorithm feed you stories that it thinks you might like to read.[31] When selecting sources, I'd encourage you to avoid blogs and stick to established mainstream organizations that maintain high journalistic standards and can legally be held accountable for what they publish. Purveyors of misinformation, like the IRA and Cambridge Analytica, prefer social media because of the ease of creating bots to propagate false messaging and target their preferred audience—then the platform's attention-grabbing algorithms take over and do the rest. And this brings me to my next recommendation. Don't use social media to get your news, but if you can't resist, the Union of Concerned Scientists (UCS) recommends that if you read a news item that sparks your interest, ask yourself the following questions of it:

- Is it difficult to separate facts from opinions?
- Does it fail to cite experts from reputable organizations?
- Is the original source of the information hard to pin down?
- Does it confirm your beliefs, or play to your emotions?
- Does the group, person, or organization sharing the information have a stake in the claim (financial, political, or otherwise)?
- Does it require belief in a secret plot and a group of co-conspirators?
- Does it scapegoat people or groups?
- Is it spread by someone who recently started their social media account but has a lot of followers?

If you answered "yes" to one or more of those questions, you should probably dig deeper and try to validate the facts using independent sources, which leads me to the next important consideration—how do you know if a news source is reliable and objective?

31 To find a good option, do a search for the best RSS feed readers. As an example, here's a *Wired* article featuring its top five picks for 2022.

The UCS has created a checklist for that as well:

- Does it back up its statements with links to independent experts with relevant knowledge or references to peer-reviewed science?
- Does it fairly present differing points of view while acknowledging the importance of expertise?
- Does it treat individuals who have differing perspectives with respect?
- Does it distinguish facts from opinions?
- Is the content free from racial, gendered, ableist, anti-LGBTQ2S, or otherwise problematic stereotypes?
- Does it make it easy to identify funding sources or ideological or policy positions?

If after reading the original article and/or other articles posted by the news source you answered "no" to any of these questions, then the source is neither reliable nor objective, and the news story is most likely either misinformation or disinformation.

Cyber Warfare

The aspect of cyber warfare that I covered in this chapter had to do with turning your smartphone into a surveillance and targeting device. The ad industry has already done the heavy lifting. The apps on your smartphone are designed to collect as much information about you as possible in order to serve you targeted advertising at the exact moment when you are most ready to make the purchase. So it's simply a matter of your adversary hacking into one or more of those apps to be able to "reach out and touch someone," as the old AT&T commercial said.

If someone wanted to find you for the purpose of performing a harmful act, one of the quickest ways to do that would be to send you a message containing a malicious link that, when clicked, would deliver a piece of malware that would track your location for as long as your mobile device powered up. That text could be delivered as a traditional text message on your smartphone, or an X or LinkedIn direct message, or via Facebook Instant Messenger, for example.

There's little that can be done to prevent this from happening, with the exception of switching to a mobile device with no apps and no text messaging whenever you are traveling and don't want to risk being tracked. On the plus side, if you aren't in a war zone, aren't engaging in espionage, or aren't a high-value target for an intelligence service, you probably don't need to go to those lengths.

Summary

This chapter touches on how our connectivity to a digitally run world has changed how we fight wars as well as how we communicate socially and, in fact, how they have become enmeshed. What used to be purely kinetic (effects in the physical world) and purely cyber (effects on networks) has changed to the point where cyber warfare is an integral part of kinetic warfare, and kinetic warfare is more effective when it's blended with cyber operations.

This shift has become possible by turning our natural desire to connect with others online and share information about our lives into an addictive reliance upon the platforms that enable those connections. We have fun so we trust. By trusting what we see and hear, we let the Trojan horse through the gate to cause havoc and far too often generate tragic results.

The examples in this chapter of Andy Milburn (of the Mozart Group) and James Le Mesurier (of the White Helmets) are just two out of many. While it may be convenient to see a shiny new object appear on your Instagram feed that you've been wanting to purchase, the ability to do that at such a granular level comes with a more serious realization—that you can be microtargeted out of a billion people at the exact time and place of someone else's choosing, and there's literally nothing that you can do about it. That's why intelligence agencies around the world are utilizing social media as a hunting ground—whether it's China's Ministry of State Security and its potential interest in US government TikTok users, or Ukraine's GUR as it hunts Russian invaders, or Russia's GRU as it monitors mobile phone usage in its contested areas in Eastern Ukraine for the purpose of jamming or targeting the users.

CHAPTER 6
Cyber Attacks with Kinetic Effects

> *A terrorist attack by Ukraine in the center of Moscow? Lights are turned off in several buildings of Moscow City, elevators are stopped, and evacuation is underway. Is this Zelensky's answer for [Russian Defense Minister] Shoigu's turning off the lights in Kyiv?*
>
> —Sergei Markov[1]

Impacting the delivery of electricity by cyber means is not a frequent occurrence, but it is a threat that's given wide attention by cybersecurity companies and government agencies responsible for protecting US critical infrastructure. Going further to cause an electrical fire or an explosion is even more rare, and that's the focus of this chapter—manipulating the automated industrial control devices of a structure to create a fire or an explosion that results in property damage, injuries, and/or loss of life. These are known as cyber/physical attacks, or cyber attacks with kinetic effects. In military parlance, they are offensive cyber operations, or OCOs.

In this chapter I'll provide examples of cyber attacks that resulted in kinetic effects on the target. I'll point out that, unlike traditional cyber attacks, there are no malware signatures or tell-tale tools, techniques, and procedures that defenders can use to stop them.

I'll also introduce how Ukraine is using OCOs that pair cyber operators with Special Forces operators, resulting in much greater effects than a cyber/physical attack on its own.[2]

1 See *https://oreil.ly/pUe62*.

2 Brigadier General Brett Williams defined offensive cyber operations as "the ability to project power in and through cyberspace to achieve campaign objectives" in *Joint Force Quarterly* 73, April 1, 2014.

Finally, I'll share my opinion that cybersecurity companies are handicapped in their ability to meet this threat because defending against it isn't scalable, and if it's not scalable, it's not profitable.

We Can Only Measure What's Been Discovered

Attacks against Ukraine's power grid in 2015 and 2016 have been widely publicized; however, it's rare to see an attack on Moscow's power grid make the news. On November 3, 2022, in the midst of a video conference call with President Putin and some of his ministers, the lights went out in Moscow City, the popular name for Moscow's International Business Center, a high-prestige, multibillion-dollar complex in the heart of the capital.[3] The affected section is the IQ quarter, which is the main transport hub for Moscow City, with connections to Vnukovo and Sheremetyevo airports.

It's also home to the Federal Treasury of the Russian Federation; the Ministry of Digital Development, Communications, and Mass Media; the Ministry of Industry and Trade; and the Ministry of Economic Development.

The blackout occurred during a live televised meeting that included President Putin, who asked why Maxim Reshetnikov (minister of economic development) and Denis Manturov (minister of industry and trade) were sitting in the dark!

Sergei Markov, a former Putin adviser and political scientist, blamed the attack on Ukraine as payback for Putin's ordered power outages in Kyiv. Markov is in a position to know since he's been a long-standing proponent of the value of cyber attacks against Russia's enemies.

In 2019, TechRepublic issued a report entitled "The 10 Most Important Cyberattacks of the decade." The ones mentioned were Yahoo (2013), Equifax (2017), Sony Pictures Entertainment (2014), Marriott Hotels (2018), Ashley Madison (2015), Target (2013), Capital One (2019), the US Office of Personnel Management (2015), First American Financial (2019), and Stuxnet (2010), plus an honorable mention for the Ukraine power grid outage in 2016.

Not a single fire or explosion among them.

No loss of life.

But was that assessment an accurate reflection of reality, or was it the result of the industry's very limited aperture into the attacks that we don't know about?

After all, we can only measure what's been discovered. We have no way of knowing how many attacks actually occur, or how much personal data, intellectual property (IP), and financial assets have been stolen, transferred, or altered by attacks that went

3 Ibid.

undetected. Sometimes, like with Sony, the attackers will announce themselves as part of the campaign. But generally speaking, when you are after money, IP, state secrets, etc., you want to quietly access the network, find the crown jewels, and extract them without anyone noticing.

Remember that everything defenders know about how bad actors attack a network comes from the ones that failed—meaning, that did not succeed in remaining covert, versus the ones that succeeded and remained undetected. The response that you want from an attack that breaks things is, "It just broke, boss," and not, "We've been hacked!" What follows are examples of cyber attacks with violent physical consequences. Even if, for example, a government agency issues a denial that the attack occurred, or says it was merely an accident, it's only prudent to ask yourself if this was a cyber attack, how might it have been accomplished, and, once you have that worked out, if the security tools and controls employed on your network would be able to detect and stop it.

Attacking Operational Technology

Operational technology, known simply as OT, includes the industrial control systems that run power plants; the automation systems that control temperature, alarms, water, and power in office buildings and warehouses; and the systems that keep subways, trains, aircraft, rockets, and even the International Space Station running safely.

Using a cyber attack to create an explosion, fire, or other act of destruction has a long history, going back to a once-secret experiment called Aurora that was carried out at Idaho National Laboratory (INL) on March 4, 2007.

The Aurora Generator Test

One of the engineers I spoke with about this chapter was at the Control Systems Security Program of the Department of Homeland Security at that time, and wrote the following definition as part of a presentation that he gave to a group of Electric Utility asset owners after the exercise. The presentation was released under a Freedom of Information Act request by news site MuckRock.[4]

> Aurora is the malicious use of a protective relay or other digital protection and control device to inflict an out-of-sync condition that results in physical damage to rotational equipment. The abrupt opening and closing of the protective circuit changes the behavior of the relay from providing maximum protection to inflicting maximum damage.

4 Scott Ainslie filed this request with the Department of Homeland Security of the United States of America.

In the Homeland Security presentation that contained the definition, a statement in small typeface further explained, "Aurora is unique because the out-of-phase condition can be caused through a cyber attack."

Watching a 27-ton diesel generator the size of a 40-foot shipping container go through a series of violent shakes, culminating with a massive amount of gray and black smoke erupting from it, was pretty scary stuff because, until Aurora, the most concerning cyber attacks focused on either stealing data or stealing money.[5]

Even under that scenario, impacting the power grid meant causing a disruption in the ability to generate electricity, not causing an explosion or a fire or some other kinetic effect. Post-Aurora, it was clear that a cyber attack could impact critical infrastructure in ways that we weren't prepared to defend against. And it wasn't only the electric grid at risk. It was anything that utilized digital protective relays and programmable logic controllers (in other words, pretty much anything that is automated).

Many successful attacks against critical systems, including some of the examples in this chapter, could have been avoided if the design included better redundancy.

This experiment inspired further development and testing of what's possible with the disruption of the automated systems that, for example, control the spinning of centrifuges at nuclear enrichment facilities. More specifically, the famous Stuxnet attack at Iran's Natanz uranium enrichment facility in 2010 and 2011 resulted in the destruction of 1,000 to 2,000 centrifuges.

While Stuxnet's impact resulted in some cost in time and money to Iran, it may not have produced the desired effect on the country's uranium enrichment program. Foreign policy writer Trita Parsi reported that from 2008 to 2013 "Iran's stockpile of low-enriched uranium (LEU) grew from 839 to 8,271 kilograms—a near tenfold increase of the very variable that the Obama administration treated as a measurement of Iran's proximity to a bomb."[6]

The Institute for Science and International Security report on the effectiveness of Stuxnet looks at it from the perspective of what the objective of the sabotage was: "If Stuxnet's goal was the destruction of all the centrifuges in the fuel enrichment plant (FEP), Stuxnet failed. But if its goal was to destroy a more limited number of centrifuges and set back Iran's progress in operating the FEP while making detection of the malware difficult, it may have succeeded, at least for a while."

5 *"Aurora Test Footage"*, MuckRock, YouTube, November 9, 2016.

6 See "Losing an Enemy: Obama, Iran, and the Triumph of Diplomacy", *International Affairs* 94, no. 2 (March 2018): p. 469–470.

Although neither the US nor Israel has officially claimed responsibility for Stuxnet, it's widely believed that both countries were responsible. And rather than the 30 lines of code used in the INL experiment, Stuxnet had hundreds.

Post-Stuxnet, Natanz remained a popular target for the Israeli government, particularly when it came to OCOs that achieved kinetic effects.

Iran Centrifuge Assembly Center

On July 1, 2020, an Israeli cyber attack caused a fire and explosion at a new centrifuge production facility at Natanz. Israel said that the event was in response to an earlier cyber attack by Iran that was intended to poison Israel's water supply by raising the chlorine levels to dangerously high amounts.

The building was the Iran Centrifuge Assembly Center. Construction began in 2012, shortly after Stuxnet. The new facility, which took six years to build, was designed for the rotor assembly of advanced centrifuges.

The damage caused by the fire and subsequent explosion couldn't be repaired, and on September 8, Iran announced that it would build a new, larger facility deep in the mountains near its Natanz nuclear site. (See Figures 6-1 and 6-2 for before and after pictures of the fire, and Figure 6-3 for the new tunnel complex.) As of May 2023, the new site was still under construction.

Figure 6-1. Before the fire. Source: Atomic Energy Organization of Iran.[7]

7 "Kamalvadi: Iran Has Tolerated Enough", Atomic Energy Organization of Iran, May 26, 2019.

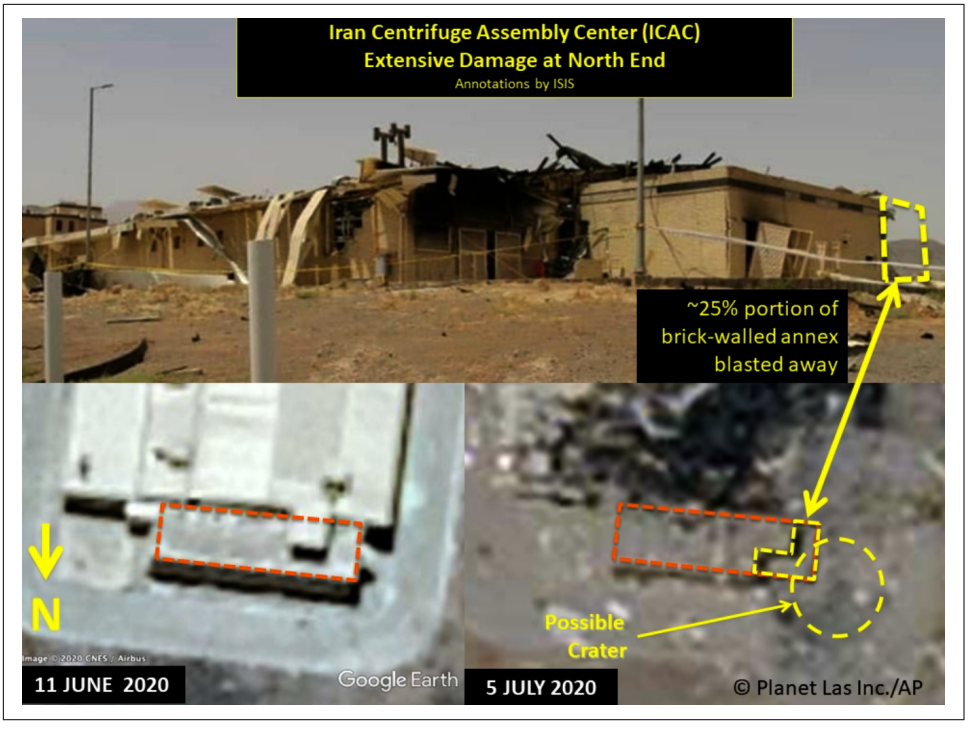


Figure 6-2. From the Institute for Science and International Security report on the fire, dated July 8, 2020.[8]

8 "The top ground image is of the north end of the ICAC Building as published by the Atomic Energy Organization of Iran as it appeared following the explosion and fire showing where a section of the annex was blasted away. The lower left image is from Google Earth and shows the appearance of the 30 meter long annex pre-incident, while the lower right image of the annex from July 5, 2020 (as published by CNN) shows by comparison that approximately 25 percent of the annex was blasted away, as well as its proximity to the possible crater." Source: Institute for Science and International Security report, July 8, 2020.

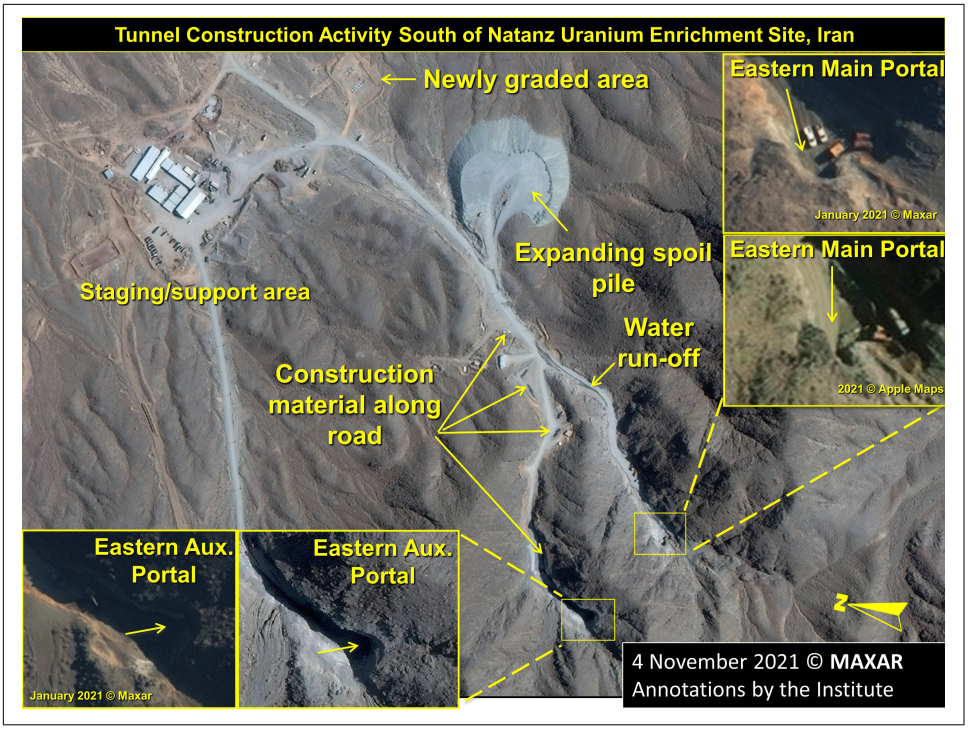


Figure 6-3. This image shows tunnel construction activity south of the previous uranium enrichment site. Image courtesy of the Institute for Science and International Security.

Underground Fuel Enrichment Plant

On April 11, 2021, Natanz was hit by another attack by Israel. The *New York Times* reported that "it had been caused by a large explosion that completely destroyed the independent—and heavily protected—internal power system that supplies the underground centrifuges that enrich uranium."

According to the *Jerusalem Post*, the attack was initially reported by the Atomic Energy Organization of Iran as an "accident" in the nuclear facility's electricity distribution network, but the country's nuclear chief, Ali Akbar Salehi, later confirmed that the incident was a cyber attack.

Israel denied responsibility, like always, although Danny Yatom, former chief of Mossad, Israel's intelligence agency, expressed concern about a possible leak of information if, hypothetically, the country was behind the attack.

"If indeed this thing is the result of an operation involving Israel, this leak is very serious," said Yatom in the *Post*. "It is detrimental to the Israeli interest and the fight against Iranian attempts to acquire nuclear weapons. There are actions that must remain in the dark."

"Once Israeli officials are quoted, it forces the Iranians to take revenge," warned Yatom. "If the Iranians start investigating with the publication hovering over their heads that the people behind the attack are the Israelis or the Americans, they will leave no stone unturned. This has an impact on our operational capability."

Gazprom

The Main Directorate of Intelligence at the Ministry of Defense of Ukraine, also known as GUR, is the equivalent to Russia's GRU.[9] Both intelligence departments have hackers who conduct various types of operations.

While GUR has existed since 1992, its offensive cyber team is a relatively recent addition, formed in 2013 by a handful of experienced hackers who had been working for other security services in the Ukrainian government. Based on their long history of access to Russian assets, they knew that the growing unrest that led to the Euromaidan protests in November 2013, and ultimately the Maidan Revolution in February 2014, would lead to reprisals by Russia. In fact, it was the start of the Ukraine-Russia war, which has continued now, as of this writing, for 10 years. The networks of Russian energy multinational Gazprom and those of its subsidiaries have been pwned by GUR hackers for about that same length of time, according to the hackers who were consulted for this book.

They began with collecting information on the company's supply chain, gained access to a supplier through phishing, then leveraged that trusted access to phish one of Gazprom's organizations and, once inside, began mapping the network and exfiltrating data.

As a result of their access, Ukraine's small team (with little to no funding but a wealth of experience, including some time spent with Israel's Mossad) engineered a hack of Gazprom's pipelines pressurization controls that would cause a pipeline to rupture, resulting in an explosion and a fire.

To date, three pipelines have experienced rupture events that were directly the result of a computer network attack; they are detailed in the following section. Prior to

9 The GUR was established on September 7, 1992, after the dissolution of the Soviet Union and the resulting independence of Ukraine. The GUR was created using intelligence assets that formerly belonged to the foreign military intelligence agency of the General Staff of the Soviet Union Armed Forces (*Glavnoye razvedyvatel'noye upravleniye),* known as the GRU. Source: Defence Intelligence of the Ministry of Defence of Ukraine.

those, there were multiple practice sessions on live pipeline systems that had mixed results.[10] Unlike the Department of Homeland Security's Aurora generator test experiment, which required building a test site to conduct the experiment at a cost of $2 million, the GUR hackers tested their methods on real pipelines until they had it right. The cost to the Ukrainian government? Zero.

Gazprom Sartransneftegaz Pipeline

This cyber operation was launched just after two Ukrainian helicopters hit an oil depot in Belgorod, Russia, on April 1, 2022 (see Figure 6-4).[11]

Figure 6-4. Photo courtesy of RIA Novosti/Press Service of the Ministry of Emergency Situations of the Russian Federation.

On April 3, according to Russian news agency RIA Novosti, "employees of Sartransneftegaz JSC discovered an underground gas leak from the high-pressure gas pipeline from the AGDS to GRP-1 at the entrance of the gas pipeline to the village of Verkhnevilyuysk…in the Republic of Sakha (Yakutia)."

10 Earlier fires that GUR hackers took credit for included Gazprom's Amur plant on October 11, 2021. See *https://oreil.ly/zvOkz*.

11 "ЧП на нефтегазовых предприятиях в России в 2018 — 2022 годах," RIA, April 01, 2022. Since URLs in Russia often don't work in the United States, interested readers might try searching for news stories or URLs using the Russian search engine Yandex while on a Russian or Moldovan VPN, such as those offered by Proton VPN.

As a side note, Yakutia is also home to a division of Russia's Space Forces, whose duties include "the radar tracking of artificial earth satellites for military purposes." The Space Forces would later become another target of interest, at least in terms of its espionage value to Ukraine's allies, who had their own Space Forces in development.

Gazprom Urengoy Center 2 Pipeline

On April 4, 2022, a cyber attack resulted in the rupture of a section of the main gas pipeline Urengoy-Center 2, which caused a large fire in the Lysvensky district of the Kama region near the village of Matveevo in Russia. An eyewitness captured footage of the large fire quite a distance away and posted it on VK, Russia's equivalent of Facebook (see Figure 6-5). The incident was reported by a local paper, *AiF-Prikamye*.

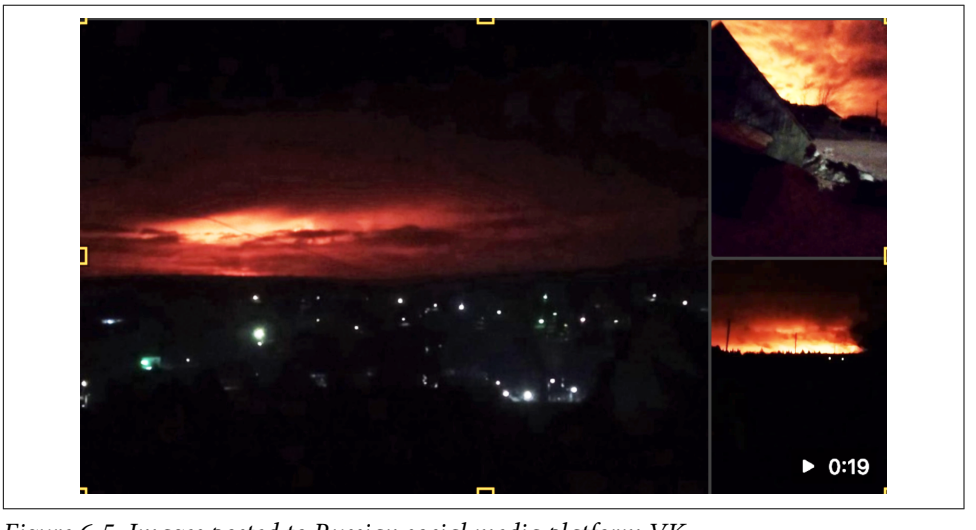


Figure 6-5. Images posted to Russian social media platform VK.

Gazprom Urengoy Pipeline

The most successful of the Gazprom fires was the one on June 16, 2022 (see Figure 6-6), at Gazprom's Urengoy gas field in the Yamalo-Nenets region. It's the second largest gas field in the world and the largest in Russia.

Part of the planning for this attack included understanding who the vendors were and gaining access, if necessary, to their respective networks (see Figure 6-8).

Figure 6-6. Images posted to Telegram and Twitter of the Urengoy gas field pipeline explosion.

According to an interview that I had with one of the cyber operators involved in the planning and execution of this and earlier attacks, his team discovered during their network reconnaissance phase that a key section of the gas pipeline's data communications network that would transmit an alarm when the pipeline was operating outside of acceptable conditions was never connected.

The schematic in Figure 6-7 is part of the updated 30-page data communications plan for the Urengoy natural gas combined cycle (NGCC) plant. The Xs on both sides show that the security alarms were not working at the time of that update (2011), nor had they been connected in 2020 when Gazprom was looking for a new vendor to complete the work or in the week when the explosion occurred.

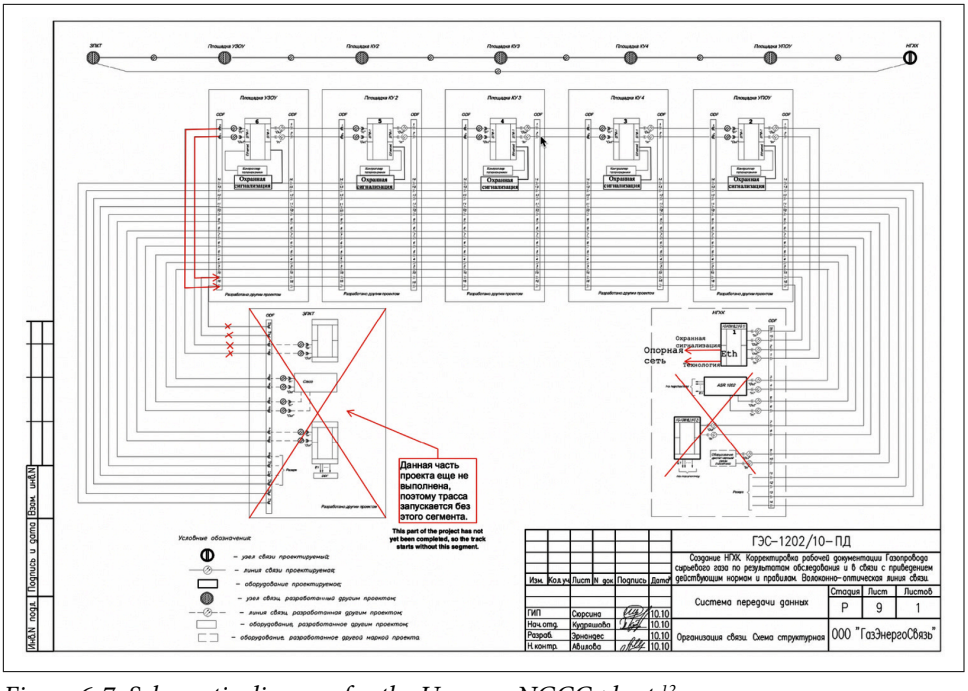
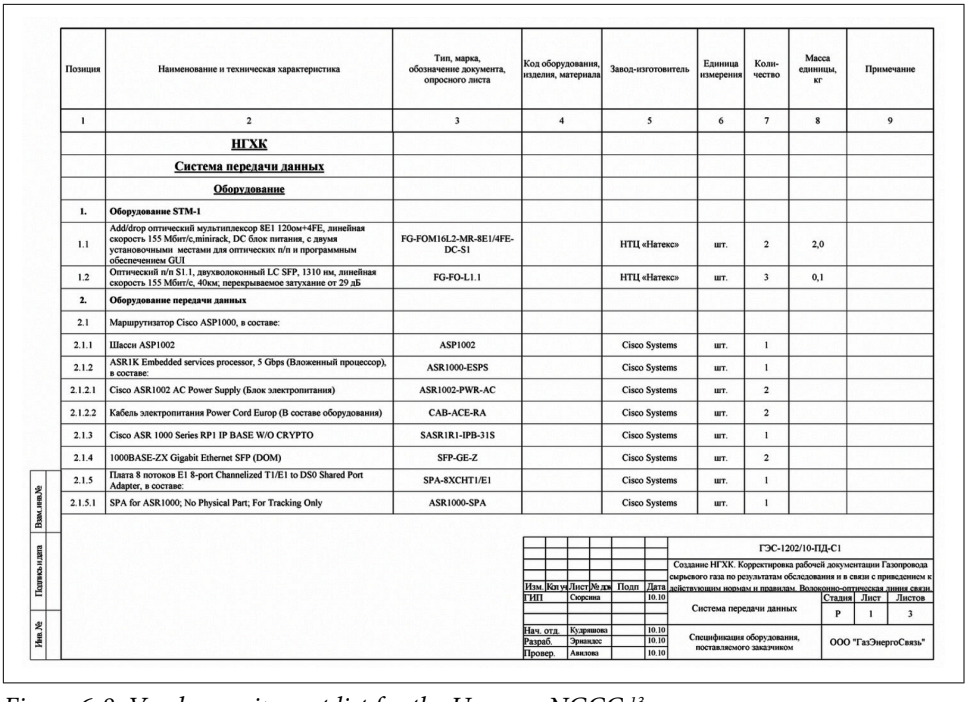


Figure 6-7. Schematic diagram for the Urengoy NGCC plant.[12]

Figure 6-8 shows one of seven pages of equipment for use in the NGCC. The foreign manufacturers include Cisco, Dell, HP, Citect, Acronis, and Microsoft. Citect makes supervisory control and data acquisition products and was acquired by Schneider Electric in 2008. Acronis makes backup, disaster recovery, cybersecurity, and endpoint management solutions.

When an attacker has access to this level of information, it isn't difficult to find known vulnerabilities to exploit, or to build and test methods for achieving the desired effect, especially when the software hasn't been patched or if the vendor substituted pirated software for genuine software.

12 This image was taken by the author from part of a large cache of documents captured by an offensive cyber team in Ukraine and shared with the author.



Позиция	Наименование и техническая характеристика	Тип, марка, обозначение документа, опросного листа	Код оборудования, изделия, материала	Завод-изготовитель	Единица измерения	Коли-чество	Масса единицы, кг	Примечание
1	2	3	4	5	6	7	8	9
	НГХК							
	Система передачи данных							
	Оборудование							
1.	**Оборудование STM-1**							
1.1	Add/drop оптический мультиплексор 8E1 120ом+4FE, линейная скорость 155 Мбит/с,minirack, DC блок питания, с двумя установочными местами для оптических п/п и программным обеспечением GUI	FG-FOM16L2-MR-8E1/4FE-DC-S1		НТЦ «Натекс»	шт.	2	2,0	
1.2	Оптический п/п S1.1, двухволоконный LC SFP, 1310 нм, линейная скорость 155 Мбит/с, 40км; перекрываемое затухание от 29 дБ	FG-FO-L1.1		НТЦ «Натекс»	шт.	3	0,1	
2.	**Оборудование передачи данных**							
2.1	Маршрутизатор Cisco ASP1000, в составе:							
2.1.1	Шасси ASP1002	ASP1002		Cisco Systems	шт.	1		
2.1.2	ASR1K Embedded services processor, 5 Gbps (Вложенный процессор), в составе:	ASR1000-ESPS		Cisco Systems	шт.	1		
2.1.2.1	Cisco ASR1002 AC Power Supply (Блок электропитания)	ASR1002-PWR-AC		Cisco Systems	шт.	2		
2.1.2.2	Кабель электропитания Power Cord Europ (В составе оборудования)	CAB-ACE-RA		Cisco Systems	шт.	2		
2.1.3	Cisco ASR 1000 Series RP1 IP BASE W/O CRYPTO	SASR1R1-IPB-31S		Cisco Systems	шт.	1		
2.1.4	1000BASE-ZX Gigabit Ethernet SFP (DOM)	SFP-GE-Z		Cisco Systems	шт.	2		
2.1.5	Плата 8 потоков E1 8-port Channelized T1/E1 to DS0 Shared Port Adapter, в составе:	SPA-8XCHT1/E1		Cisco Systems	шт.	1		
2.1.5.1	SPA for ASR1000; No Physical Part; For Tracking Only	ASR1000-SPA		Cisco Systems	шт.	1		

Взам. инв.№ | Подпись и дата | Инв.№

Изм.	Кол.уч.	Лист	№ док.	Подп.	Дата	ГЭС-1202/10-ПД-С1			
						Создание НГХК. Корректировка рабочей документации Газопровода сырьевого газа по результатам обследования и в связи с приведением к действующим нормам и правилам. Волоконно-оптическая линия связи.			
ГИП		Сорокина			10.10	Система передачи данных	Стадия	Лист	Листов
							Р	1	3
Нач. отд.		Кудрикова			10.10	Спецификация оборудования, поставляемого заказчиком	ООО "ГазЭнергоСвязь"		
Разраб.		Эрмилас			10.10				
Провер.		Авизова			10.10				

Figure 6-8. Vendor equipment list for the Urengoy NGCC.[13]

A proxy for the vendor, Stroyneftegaz Alliance (SNG Alliance), which had filed for bankruptcy in 2017, was sued by Novy Urengoy Gas Chemical Complex, and on March 12, 2020, the court found SNG Alliance to have not completed many of the contractual items it had been paid to deliver, amounting to several billion rubles.[14]

13 Ibid.

14 ARBITRATION COURT OF THE YAMAL-NENETS AUTONOMOUS DISTRICT in the matter of The Arbitration Court of the Yamalo-Nenets Autonomous District composed of Judge S.V. , OGRN 1037730026575) to the limited liability company Novourengoy gas chemical complex (TIN 8904006547, OGRN 1028900620264) for the recovery of 1,317,983,943 rubles 25 kopecks, and on the counterclaim of the limited liability company "Novourengoy gas chemical complex" (TIN 8904006547, OGRN 1028900620264) against the limited liability company "Stroyneftegaz Alliance" (TIN 7730172171, OGRN 1037730026575) for the recovery of 1,379,095,416 kopecks. Source: Решение от 14 марта 2020 г. по делу № А81-6114/2018.

Some of the many unfinished items included:[15]

- "A complex of engineering and technical means of protection and means of anti-terrorist protection (designer of DOAO 'Gazprojectengineering')"
- "Control room (title 401/080), administrative building with a laboratory (title 401/080-1)"
- "Engineering networks of the administrative and amenity zone, automated fire safety system"

On January 22, 2020, Gazprom announced a solicitation for:

- "Execution of turnkey works on the facility 'Automatic fire alarm system, gas pollution control and fire extinguishing of gas pumping units of booster compressor stations (level I) of the installation of complex gas treatment—7, 8, 9, 10, 12, 13, 15, booster compressor stations (level II) of the installation of complex gas treatment—1, 2, 4, 9, 10, 11, 12 of Urengoy oil and gas condensate field for the needs of Gazprom dobycha Urengoy LLC (0001/19/1. 1/0101853 / Durengoy/K/STATE/e/20.12.2019)—1 551 051 576.00 rubles."
- "Performing major repairs of the fire alarm system for the needs of Gazprom dobycha Urengoy LLC in 2020–2021 (for small and medium-sized businesses) No. 0095/19/5. 1/0097185 / Durengoy / PR/STATE/e/11.12.2019 - 43 726 275.89 rubles."

This incident of vendor incompetence and corruption is not the exception in Russia. It happens far too frequently and in every industry, including space, energy, finance, and defense.

Acts of sabotage by cyber means—such as the ones conducted by GUR hackers at Gazprom's largest plants, including the explosions at the Urengoy Yamalo-Nenets region, the Urengoy Kama region, and the gas leak in Yakutia—are facilitated by the culture of corruption there. And the only thing that's preventing these types of attacks from happening more frequently and in greater numbers is not the technical difficulty of the operation, nor that Gazprom has improved its networks' defenses. It is solely due to the restraint being exercised by Ukraine's leadership. One option under consideration by the GUR team was to develop a scalable attack that would hit all Gazprom pipelines simultaneously; however, that was deemed by senior leadership to be too inflammatory and would most likely be seen by Ukraine's allies as excessively aggressive.

15 Decision of March 14, 2020, in the case no. A81-6114 / 2018, Arbitration Court of Yamalo-Nenetsky JSC (AS Yamalo-Nenetsky AO).

Second Central Research Institute of the Ministry of Defense of the Russian Federation

On April 21, 2022, a fire and explosion (see Figure 6-9) at this top-secret research facility in Tver, where the Iskander and S-400 missiles were designed, resulted in the complete destruction of the building, leaving 6 dead and 27 wounded. Russian state news service TASS reported that the cause may have been due to a malfunction associated with faulty wiring.

Figure 6-9. Fire at the Second Central Research Institute of the Ministry of Defense of the Russian Federation on April 21, 2022.

According to informed sources, this was a combined cyber/Special Operations forces mission where a commando team surreptitiously entered the structure to place explosives so as to render maximum damage to the structure. The initiating cause for the explosion was the electrical fire caused by the cyber team.

This combination approach gives a tactical advantage to the commandos, who can already be on their way to the next target when the fire starts and the explosive charges ignite. For example, the Dmitrievsky Chemical Plant in Kineshma, Ivanovo Oblast, Russia, an approximately seven-hour drive from Tver (see Figure 6-10), was destroyed by a fire the very next day.

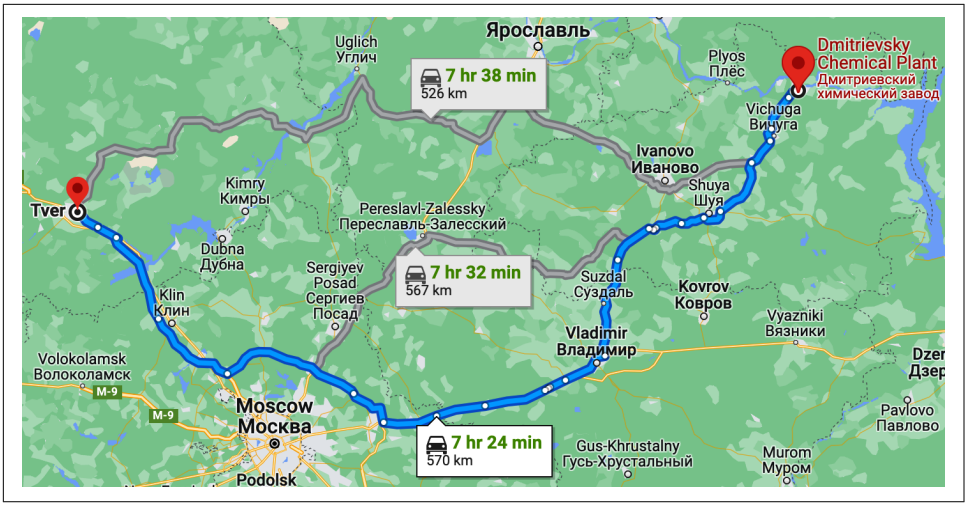


Figure 6-10. Google Maps directions from Tver to Kineshma.

That's not to say that both fires were due to cyber/physical attacks, but the second fire was certainly deemed suspicious, according to the press reports. One of the many problems that Moscow has is differentiating between what was an act of sabotage versus what may have been one of hundreds of fires that start because of poorly maintained infrastructure.

Khouzestan Steel Company

On June 27, 2022, the Gonjeshke Darande (aka Predatory Sparrow) threat actor carried out a series of cyber attacks against three Iranian steel companies where they stole sensitive data and, in the case of the Khouzestan Steel Company, caused a fire (see Figure 6-11) by accessing the supervisory control and data acquisition system that controlled the plant's arc and ladle furnaces. The attack was announced on Predatory Sparrow's Telegram and X (formerly Twitter) channels; also uploaded were two videos of the fire and several gigabytes' worth of emails and sensitive documents.

Figure 6-11. A still from the video that was taken of the fire caused by hackers taking control of the steel furnace at the Khouzestan Steel Company.

Evaluating the Effectiveness of Sabotage

These examples demonstrate the effectiveness of a cyber attack to trigger a kinetic effect on the target. The benefit is that it can be done for the most part without humans in the loop. However, there's still the question of whether it leads to a strategic benefit in the long run. To help answer that question, let's look at the effectiveness of sabotage on Iran's nuclear program (see Figure 6-12).

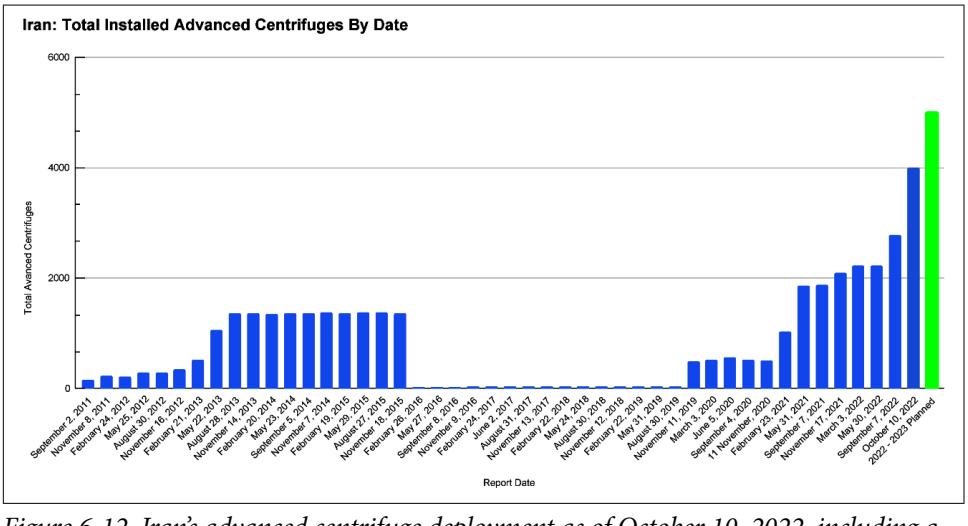


Figure 6-12. Iran's advanced centrifuge deployment as of October 10, 2022, including a future projection for late 2022 and 2023. Source: Institute for Science and International Security.

In spite of repeated successful attacks against Natanz, Iran has managed to increase its uranium enrichment production over time. A report from the Institute for Science and International Security shows a marked increase in Iran's centrifuge deployments. As of October 10, 2022, Iran had a total of 4,000 advanced centrifuges of varying types installed at its three enrichment facilities. Over half of the installed ones are IR-2m centrifuges, which have an output believed to be four times greater than the IR-1, Iran's first centrifuges that were targeted by the Stuxnet attack.[16]

This shows that just because we can do cyber/physical attacks doesn't necessarily mean they're a good use of money (in terms of sunk time versus effect, for example).

Defending Against Cyber/Physical Attacks

Building resilience may be the optimal defense against cyber/physical attacks as well as straightforward kinetic attacks (like missile strikes). Here are a few examples provided by Tim Roxey, the former chief security officer at the North American Electric Reliability Corporation.

Nuclear power plants use multiple redundant control systems to ensure that critical functions, such as reactor shutdown, can be performed even in the event of a failure. For example, a plant may have multiple independent control systems for the reactor, each with sensors, actuators, and software. These systems are designed to cross-check each other and ensure critical functions are performed correctly.

Fly-by-wire aircraft control systems use three or more independent computer systems to control the aircraft's flight surfaces. Each computer system is programmed with its software and is designed to cross-check the outputs of the other systems. If one system detects an error or inconsistency, it will be outvoted by the other systems, which continue to function normally.

Medical device control systems use redundant paths to ensure that critical functions, such as drug delivery or patient monitoring, are performed correctly. For example, a medical device may have two independent microcontrollers that control the delivery of a drug. Each microcontroller is programmed with its software and is designed to cross-check the outputs of the other. If one detects an error, it will be outvoted by the other microcontroller, which continues to function normally.

These are just a few highly reliable electronic systems that use multiple control paths. Generally, any system requiring high reliability and safety may use redundant control paths to improve reliability and ensure critical functions are performed correctly.

16 IAEA "Verification and Monitoring in the Islamic Republic of Iran in Light of United Nations Security Council Resolution 2231 (2015)", International Atomic Energy Agency, November 15, 2023.

Summary

Cyber attacks with kinetic outcomes, aka cyber/physical attacks, have been a focus of Israel's since the Stuxnet program and continue to the present day. A handful of Ukrainian government hackers (with Mossad experience and little to no funding) have shown that these types of attacks can be done without the resources of a nation-state, without the need to develop specialized malware, and without being seen by the NSA and similar agencies in developed countries.[17]

These attacks are uncommon, and so they aren't profitable to defend against. They don't neatly fit within an OT network defender's playbook, and you cannot write a signature for them or upload a binary to VirusTotal. In other words, they don't fit within the funding model that venture capitalists use for cybersecurity startups.

Attacks like these are a frightening innovation in warfighting, one that the cybersecurity industry isn't prepared to defend against, so resilience must be built into critical systems similar to those in the examples provided by Roxey: redundant pathways for critical functions in medical devices; multiple computer systems that cross-check an airplane's flight controls; multiple, independent controllers that ensure critical functions do not stop operating at a nuclear power plant, etc. Apply the military adage "Two is one, one is none" when it comes to protecting your critical systems. Redundancy saves lives.

17 An NSA insider unofficially advised caution on attributing the Gazprom explosions to the GUR because the agency hadn't seen any activity that it would expect to see for an attack like that.

CHAPTER 7
AI

Machines take me by surprise with great frequency. This is largely because I do not do sufficient calculation to decide what to expect them to do.

—Alan Turing[1]

The pursuit of artificial intelligence (AI) is almost as old as computing. Stanford and MIT's AI labs were both founded by the same man, computer scientist John McCarthy. McCarthy arrived at MIT in 1956 as a research fellow and opened the AI Lab with Marvin Minsky in 1959. In 1962, he moved to Stanford and founded the Stanford Artificial Intelligence Lab.

In 1956, McCarthy expressed their major challenge in a way that was similar to the test that computer scientist Alan Turing proposed in 1950: "The artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving."[2]

More than 60 years later, computer scientists working in AI still anticipate that a sentient AI with human-level intelligence will emerge from its silicon base to deliver a techno-utopia to the human race, if it doesn't kill us first.

This chapter will explore present harms and future risks associated with the rapid adoption of AI, especially as it impacts national security. I'll detail the most pressing

1 Manuel Alfonseca et al., "Superintelligence Cannot be Contained: Lessons from Computability Theory," *Journal of Artificial Intelligence Research* 70 (2021): 65–76.

2 "Turing…describes the following kind of game. Suppose that we have a person, a machine, and an interrogator. The interrogator is in a room separated from the other person and the machine. The object of the game is for the interrogator to determine which of the other two is the person, and which is the machine." Source: Stanford Encyclopedia of Philosophy.

present risks and the harms that have resulted from them, explore future risks, and provide recommendations for prevention and mitigation.

Defining Terms

In order to navigate the world of AI, it's important to have some agreed upon definitions of terms. It's easy to confuse "generative" with "general," and artificial general intelligence, or AGI, has lost all meaning from when it was first coined years ago.

Generative AI

The GPT in ChatGPT stands for generative pretrained transformer. The tool ingests incomprehensible amounts of training data and then, using neural networks, it generates new content when prompted to do so by its user.

One of the reasons ChatGPT's popularity blew up so quickly is that it's a huge amount of fun to use. You make a request and, in short order, it delivers results that range from terrible to mind-blowing. In other words, it will "generate" creative output that will surprise or impress you most of the time. Its sister app, DALL-E, does the same thing except with images. In short, a generative pretrained transformer aims to create new content based upon its prior training on the creative content of others. It does not, however, understand what it is generating. It is, in the words of AI researcher Emily Bender, a "stochastic parrot"—that is, "a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning."[3]

Neural Network

The neural network is the brain of the AI. It works in a manner similar to the human brain. While the brain is made up of neurons, a neural net is made up of layers, and each layer has nodes that are controlled by algorithms.

Algorithms are adjusted by weights, which are parameters that can be adjusted up or down to improve the accuracy of the model's output in response to a query, otherwise known as a "prompt."

What scientists don't know yet is exactly how it all works.[4] That, I believe, has contributed to the allure, and fear, of what AGI or superintelligence could do.

---

3 Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* Association for Computing Machinery (2021), 610–623.

4 "Large language models can do jaw-dropping things. But nobody knows exactly why," said Will Douglas Heaven in *MIT Technology Review*, March 4, 2024.

Narrow AI

Narrow AI (NAI) is designed to perform a specific function and only that function. For example, there's a program for playing chess, another specializes in getting you the best price for a widget, and yet another might make suggestions for improving the tone of your Valentine's Day email or break-up text. We use NAI more often than we know, in probably hundreds of different ways, and unlike AGI, NAI isn't considered an existential threat to humanity; however, it has been used in certain cases (like AlphaGo's Move 37—see Figure 7-1) to justify concern for future existential risk from AGI and superintelligence.[5]

Figure 7-1. International Go champion Lee Sedol competing against DeepMind's AlphaGo computer in Seoul in 2016. Image provided by Google DeepMind.

Foundation Model

According to Stanford University report *"On the Opportunities and Risks of Foundation Models"*, a foundation model "is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks." A few examples of foundation models include ChatGPT-4 (created by OpenAI), Claude (Anthropic), and Gemini (Google DeepMind), but there are many more.[6]

5 See this *Wired* story if you're interested in learning more about Move 37 from the Lee Sedol Go match with an AI computer named Alpha Go.

6 For a current directory of large language models, see *https://llmmodels.org*.

Frontier AI

An OpenAI white paper defines frontier AI as "highly capable foundation models that could possess dangerous capabilities sufficient to pose severe risks to public safety." This term, along with artificial general intelligence, is extremely fuzzy in how it's defined. The referenced paper refers to next-generation foundation models but assigns some capabilities to them that have been associated with AGI, such as "evading human control through means of deception and obfuscation"; a model that can do this is also known as a "schemer" in AI safety circles.[7]

The phrase "frontier AI" identifies the user as one who believes that (a) AGI and superintelligence are inevitable and (b) that they pose an existential threat to mankind, thereby making their work to save mankind from a rogue superintelligence absolutely essential when it comes to funding, resources, and input on matters of regulation.[8]

Artificial General Intelligence

Artificial general intelligence describes a computer that can think and act in a way that could fool a human into thinking the computer is human. In 1950, Alan Turing proposed a test he called the Imitation Game, where an interrogator would ask questions of a human and a computer, not knowing which was which, and after five minutes of questioning, would identify the computer as the human in the game.[9] Today, ChatGPT-4 makes the Imitation Game a nonstarter. For example, in a recent study involving 180 psychology students at the bachelor's degree and doctoral levels, ChatGPT-4 outperformed all of them in tests of social intelligence.

Apple cofounder Steve Wozniak suggested that his own version of the test might involve a robot entering a stranger's home and being asked to make a cup of coffee. The AI powering that robot would have to:

1. Find the kitchen
2. Choose which coffee maker to use (automatic drip, espresso, AeroPress, French press)
3. Find the coffee bean grinder and adjust the setting for the grind to match the chosen method of brewing coffee
4. Heat the water to the optimal temperature

7 See the Open Philanthropy white paper "Scheming AIs: Will AIs Fake Alignment During Training in Order to Get Power?" by Joe Carlsmith, November 2023.

8 Gina Helfrich, "The Harms of Terminology: Why We Should Reject So-Called 'Frontier AI'", *AI and Ethics*, March 11, 2024.

9 A. M. Turing, "Computing Machinery and Intelligence," *Mind* 59, no. 236 (October 1950): 433–460.

5. Grind the beans
6. Pour the water over the grounds
7. Pour the brewed coffee into a cup

Since the robot had never been in that house, it would need to solve a number of different problems in different domains, some of which had no "correct" answer and instead relied upon intuition or preference. Yet while this test would be an impressive feat, it wouldn't demonstrate that the AI was self-aware or sentient, which is a prerequisite to an AGI that poses an existential risk to humanity.

The debate of what constitutes evidence of human-level intelligence in an AI can be roughly divided into two camps, according to a recent article in *Nature* and confirmed by my own interviews with computer scientists working on this problem. One camp's position is that large language models (LLMs) are exhibiting signs of intelligence when they score better than humans in a variety of different tests ranging from poetry to law. The opposing camp says that such indications make them excellent mimics but still unable to generalize their knowledge across different disciplines. Underlying both camps is our human tendency to anthropomorphize anything in nature that reminds us of us.

So while philosophers and scientists disagree about how to define an AGI, or how to test for it, one only needs to look at what the fear is, and then work our way backward to what would be required. Sentience is at the top of that requirements list; in the words of René Descartes, "I think, therefore I am." The AGI or superintelligence is self-aware and makes decisions based upon self-preservation. That hasn't happened yet, nor is there any clear path to achieving such self-awareness, although not for lack of trying.

Two researchers at the Global Catastrophic Risk Institute attempted to map out the requirements for a catastrophic event involving artificial superintelligence (ASI) using fault tree analysis. Figure 7-2 shows what must happen in order for the top-level state of an ASI catastrophe to occur.

At level two, there is a "takeoff" state, meaning that the ASI is now acting entirely on its own, thus achieving a decisive strategic advantage, and an "unsafe" state, meaning that its autonomous actions result in effects that are not safe for human life.

At level three, there are six conditions that must exist in order for each of the level-two states to come about.

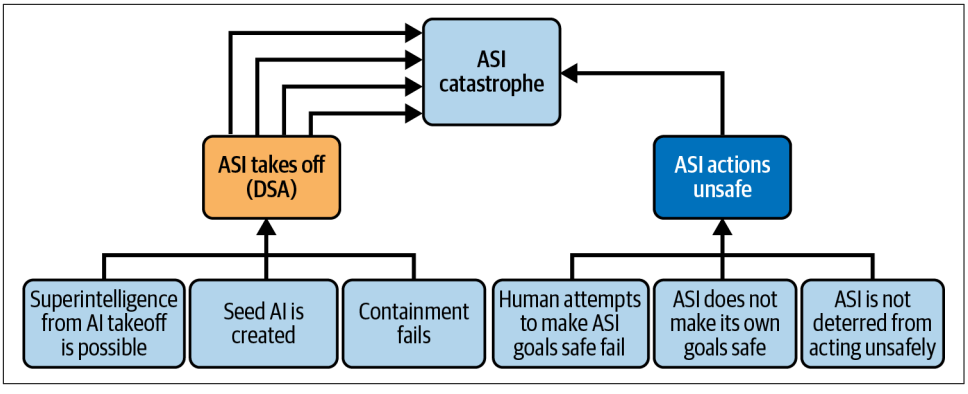


Figure 7-2. ASI-PATH.[10]

The biggest hurdle is how to get to "Seed AI," a term coined by Eliezer Yudkowsky, the founder of the research nonprofit the Singularity Institute for Artificial Intelligence (renamed the Machine Intelligence Research Institute), and defined as "an AGI which improves itself by recursively rewriting its own source code without human intervention."[11] Anthony Barrett and Seth Baum, the researchers from whose work Figure 7-2 comes, do not address this in any kind of satisfying way, but neither does anyone else.

There isn't a road map to achieving AGI, nor is there agreement among scientists that AGI is technically feasible. That was viewed as enough of a problem now so as to warrant its own workshop at the Twelfth International Conference on Learning Representations in Vienna, Austria, in May 2024. The lack of a firm and testable definition gives maximum leeway to AI companies to define it as they please. Further, the belief that AGI is within reach, whether it's 5 years or 25 years from now, has attracted vast amounts of investment, both in development and in safety research to prevent an AGI from destroying the human race—either accidentally, as in Nick Bostrom's paper clip problem (which we'll explore soon), or on purpose, such as in Thomas Metzinger's Benevolent Artificial Anti-Natalism theory.[12]

Superintelligence

Superintelligence or artificial superintelligence is the theoretical next step beyond AGI.

It is a machine intelligence that is superior to humans in every way. It can remember everything with total recall. It never tires. It doesn't require food or water or sleep. It

10 Anthony Barrett and Seth Baum, "A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis", *Journal of Experimental & Theoretical Artificial Intelligence* 29, no. 2 (2017): 397–414, available at SSRN.

11 See the Less Wrong forum wiki entry for Seed AI.

12 For the Bostrom example, see his book *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press).

can consider any problem, forecast all possible outcomes, and always choose the optimum solution. It can control its hosting environment, including its power source, is recursively self-improving, can modify its own code, and yours, and can act on its own behalf—including, we presume, in ways that will ensure or maximize its chances for survival. Proponents of longtermism believe that should this scenario occur, it poses an existential threat for humanity.[13]

Present Risks

AI foundation models, such as OpenAI's ChatGPT-4, Google's Gemini, and Meta's Llama, and AI agents or apps that connect to those services, of which there are thousands, suffer from a wide variety of security issues. Rather than survey the latest research, most of which will be out of date by the time this book comes out, I've provided a few important papers on this topic in the footnotes as a starting point.[14]

Cybersecurity Vulnerabilities

Let's start with the category of cybersecurity vulnerabilities.

Indirect prompt injection

This type of attack was first written about by a team of security researchers from Saarland University and the CISPA Helmholtz Center for Information Security in their paper "More Than You've Asked For: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models".

I'm highlighting this attack because as of now there's no known mitigation and, according to the paper's authors as well as security professional Johann Rehberger, who gave a talk about this at the Chaos Communication Congress in 2023, it's unlikely that any of the LLMs will find a way to fix that any time soon.

13 The word "longtermism" has been attributed to Oxford philosophers William MacAskill and Toby Ord when describing mankind's duty to consider how our actions today will impact the long-term future, especially when existential risk is involved. This is an essential part of the effective altruism belief system.

14 There are so many ways to compromise these AI models that doing a survey for this chapter was impossible. Instead, here are a few research papers for those who want to take a deep dive into the pool of AI vulnerability research: (1) Hossein Hajipour et al., "CodeLMSec Benchmark: Systematically Evaluating and Finding Security Vulnerabilities in Black-Box Code Language Models," arXiv (2023); Hossein Souri et al., "Sleeper Agent: Scalable Hidden Trigger Backdoors for Neural Networks Trained from Scratch," arXiv (2022); Yujia Fu et al., "Security Weaknesses of Copilot Generated Code in GitHub," arXiv (2024). I also recommend the blog Embrace the Red by researcher Johann Rehberger.

Automated vulnerability exploitation

Computer scientists at the University of Illinois Urbana-Champaign conducted a study to see if LLM agents could effectively exploit real-world computer systems. They used 10 different LLMs, including versions of ChatGPT-3.5 and 4, Llama-2, Mistral, Mixtral, Nous Hermes-2, and OpenChat, and 15 CVEs marked "high" or "critical" severity in the CVE database. Eleven out of the 15 vulnerabilities had been discovered past the ChatGPT-4 model's knowledge cut-off date, meaning those 11 would have never been seen before.

Of all the LLMs tested, only ChatGPT-4 was able to successfully exploit all but two of the CVEs presented to it, as long as the LLM model had a description of each CVE. Without a description, its effectiveness dropped from 87% to 7%.

The conclusion of the study showed that LLM agents are capable of autonomously hacking real-world systems; however, their effectiveness varies greatly and even the most effective system still has problems with planning and problem solving.

Network attacks

AI is being used today to, among other things, write malware, compose believable spear-phishing texts and emails, empower bots with human-like conversational abilities, and test different attack and defense scenarios against heavily fortified networks to find an exploitable entry point (see here and here for more).

Automated Decision Making

Particularly in the fields of justice and health care, automated decision making without human supervision is a recipe for disaster because of hidden biases in the training data that almost certainly will not be open for scrutiny.

I can imagine the need for an updated confrontation clause in the Sixth Amendment to the US Constitution, where the phrase "a defendant's right to confront witnesses brought against them" would have "or algorithms" added.

As far as health care, as I pointed out in Chapter 1, EHR software has been a nightmare for physicians since it was hurriedly pushed during the Obama administration.

"Our investigation found alarming reports of patient deaths, serious injuries and near misses—thousands of them—tied to software glitches, user errors or other flaws that have piled up, largely unseen, in various government-funded and private repositories."[15]

15 Fred Schulte and Erika Fry, "Death by 1,000 Clicks: Where Electronic Health Records Went Wrong", KFF Health News, March 18, 2019.

Medical record errors are too numerous to count, and the majority of states don't have any reporting requirements for EHR errors that result in patient harm or death. This combines to make the training data for a medical AI extremely problematic.

Warfighting

Disinformation (aka cognitive warfare)

Deepfakes utilizing text, audio, and video have been causing problems via social media for at least the past 10 years, and probably longer, but in the past they've been easy to spot. Today, text-to-video apps like Runway, Pika, and Sora make it virtually impossible for the average person to tell a fake from the real thing. AI-enabled fake accounts are replacing the need for humans to work in troll farms like Russia's IRA. On December 17, 2023, the *Washington Post* reported about the rise of AI-enabled fake news, referring to it as a misinformation superspreader.

AI-guided drone swarms

Today, drone swarms consisting of thousands of tiny drones are controlled by AI to produce something incredibly beautiful, like the 1,800 drones that created a 300-meter-long dragon encircling the Burj Al Arab hotel in Dubai in celebration of the 2024 Lunar New Year.

However, if you add facial and gait recognition plus a high explosive charge to each of them, you have a lethal weapons platform that is close to impossible to defend against. Israeli military technology company Elbit Systems, for example, offers the Legion-X drone swarm, which the Israeli Defense Forces have been using to search and kill Hamas terrorists in Gaza.

Ukraine has been very innovative with its use of drone warfare against Russian forces, and AI-enabled drone swarms cannot be far off.

Both the US and China have been developing drone swarm capabilities since at least 2018, and Russia has been using swarms without AI (so far) in its war with Ukraine.

Speculative Risks

Predictions of future catastrophe like the one written by Eliezer Yudkowsky for *Time* magazine, or by Nick Bostrom in his book *Superintelligence* (Oxford University Press), are based upon the theory that an AI can become self-aware or sentient. Should that happen, then there are any number of ways that things could go wrong. One well-known thought experiment is Bostrom's paper clip maximizer. Here's how he puts it:

> Suppose we have an AI whose only goal is to make as many paper clips as possible. The AI will realize quickly that it would be much better if there were no humans because humans might decide to switch it off. Because if humans do so, there would be fewer paper clips. Also, human bodies contain a lot of atoms that could be made into paper clips. The future that the AI would be trying to gear towards would be one in which there were a lot of paper clips but no humans.[16]

Self-Preservation

Stuart Russell, a professor of computer science at UC Berkeley who's been very vocal about the existential risk of misaligned AI, uses a "fetch the coffee" scenario to illustrate his concerns. A system given the request "Could you fetch me a cup of coffee?" will organize itself to not only fulfill your goal but also to establish a subgoal that would prevent itself from being shut off, up to and including killing everyone who could potentially interfere (as I depicted humorously in Figure 7-3).

Figure 7-3. The author's conception of Stuart Russell's coffee risk argument; image created by the author using DALL-E.

16 See Nick Bostrom, "Ethical Issues in Advanced Artificial Intelligence".

The Treacherous Turn

Dan Hendrycks, founder of the Center for AI Safety, points out that AIs have been known to exhibit deceptive behavior. What would happen, Hendrycks posits, "if an AI agent became 'self-aware' and discovered that it was being evaluated for compliance with a safety regime. It might…learn to play along, exhibiting what it knows is the desired behavior while being monitored…then take a treacherous turn and pursue its own goals once the monitoring has ceased or once it has reached the point where it can overpower us."

The Sharp Left Turn

Nate Soares, president of the Machine Intelligence Research Institute, has written extensively on this risk, which is defined as an AI system that suddenly exhibits an increase in capabilities in planning and world modeling such that alignment methods (i.e., aligned with human values) that had worked before are no longer working. This will happen faster than humans will notice and react to, and the system will now resist realignment or any other interference with its plans.

A Short Primer on Effective Altruism and AI Safety

Effective altruism (EA) is a philosophy that espouses evidence-based philanthropy where giving is based on logic and reason; that is, where would my donation make the most impact for the good of humanity? After the AlphaGo victory over Lee Sedol, the answer to that question became AI safety; after all, what greater good could one do beyond saving humanity from the threat of extinction by a superintelligent AI?

To date, over a half billion dollars has been spent on funding AI safety organizations, and the majority of that ($335 million) came from Open Philanthropy, founded by billionaire effective altruists Dustin Moskovitz and his wife, Cari Tuna.

Other major EA donors include Jaan Tallinn's Survival and Flourishing Fund ($44.5 million), Sam Bankman-Fried's FTX Future Fund ($32.5 million), and Vitalik Buterin ($65 million). Figure 7-4 shows levels of investments in this area.

There are several hundred AI safety organizations today and an untold number of EA members in US and UK government employ, including high-profile individuals like Paul Christiano, recently appointed by US Secretary of Commerce Gina Raimondo to NIST (under a storm of controversy) and RAND CEO Jason Matheny.

AI Watch is an excellent resource for chasing down affiliations of individuals and organizations who are receiving funding in the field of AI safety.

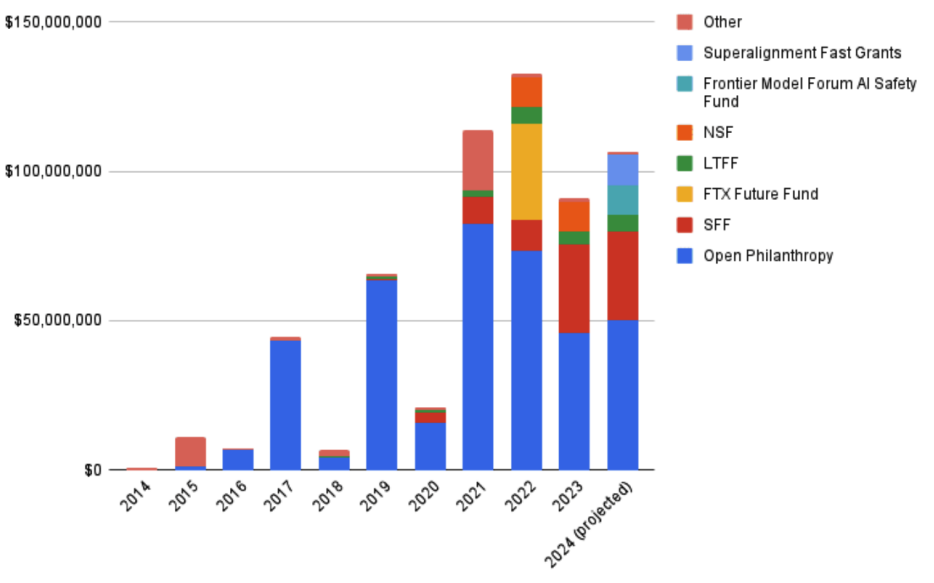


Figure 7-4. Estimated AI safety spending from 2014 to 2024; prepared by the Effective Altruism Forum.

Risk Versus Probability

When people in the risk management business speak about "low probability, high impact" events, they're referring to actual threats that have a very low probability of occurring and an outsize impact if they do occur.

An example is the Tohoku earthquake and tsunami on March 11, 2011. It was the largest earthquake ever recorded in Japan and the fourth largest to date in the world. It shattered every emergency planning contingency in the government of Japan's disaster risk management plan because no one had ever seen, nor considered, the chain of events that a 9.0 earthquake would set in motion.

The earthquake triggered a tsunami with waves as high as 40 meters (132 feet). The tsunami caused the meltdown of three nuclear reactors at the Fukushima Daiichi Nuclear Power Plant, and that caused extended power shortages. This also put a strain on the nation's communications infrastructure that resulted in multiple failures, leading to increased public anxiety, panic, and unrest. Overall, 450,000 people were made homeless and there were at least 15,000 people killed.

The Zero-Probability High-Impact Risk Model

Misaligned artificial superintelligence (MAS) is a fictional threat actor that has been embraced by some AI researchers as far back as the early 1960s, and the end of humanity as we know it is the theorized high-impact outcome.

Since MAS doesn't currently exist, there is zero probability that it can generate any sort of impact, high or otherwise. However, if you bring this up to anyone working in frontier AI safety, they'll undoubtedly tell you about other "fictional" scenarios that at some point did not exist, like space travel, and yet, here we are!

In addition, they will say that it's not inconceivable that we could develop an AI with human-level intelligence, and should that happen, it's not unreasonable to assume that it will evolve into an unmanageable superintelligence that is superior to humans in every way and may very well make decisions that will cause widespread devastation unless we find a way to control it. And that's not something that you want to leave until the very last minute.

Award-winning computer scientist Yann LeCun made this point very well in a post on X when he wrote, "It seems to me that before 'urgently figuring out how to control AI systems much smarter than us' we need to have the beginning of a hint of a design for a system smarter than a house cat." We don't even have that. Not even a sniff of that.

Regulation

If human-level AGI and superintelligence pose such a highly speculative and extremely remote risk as the extinction of the human race, why is it repeated time and again by hundreds of accomplished philosophers, professors, and scientists?

Regulation is why.

If you want a voice at the table, you need specialized knowledge, as well as a high-visibility, high-risk, mass casualty event that captures public attention. A rogue superintelligence that's vastly smarter and more capable than human beings fits the bill perfectly, even if it is the stuff of science fiction.

When UK Prime Minister Rishi Sunak announced that the British government would host an international AI safety summit in September 2023, his concern was fueled by intense lobbying from a well-funded movement called effective altruism (EA) that evangelizes AI in a completely binary way. AI will either bring about a techno-utopia where humans are freed from manual labor and grunt work, enjoy a low cost of living, and live long, creatively fulfilling lives, or 50% or more of us will be dead in the wake of a superintelligence gone rogue. The EA-aligned career advice center 80,000 Hours recommends "AI safety technical research" and "shaping future governance of AI" as the two top careers for EA members, according to *Politico* journalist Laurie

Clarke. One safety researcher who asked to remain anonymous told Clarke that EA members are scrambling to be a part of the UK's AI safety task force because "we need political buy-in and policy" related to AI and existential risk.

Prior to regulation is an effort at voluntary compliance. On November 2, 2023, in a history-making international summit at the UK's Bletchley Park, where AI pioneer Alan Turing and other lesser-known heroes broke the encryption code on Germany's Enigma machine during World War II, dignitaries from the US, the UK, Canada, Japan, and other nations met to discuss the setting of global standards for AI safety. The first task for the institute was going to be establishing a test-first protocol for all new advanced AI models before they launch. All of the attending tech companies and their CEOs, like OpenAI's Sam Altman and xAI's Elon Musk, agreed, as did Anthropic, Google, Microsoft, Amazon Web Services, Meta, and others.

Six months later, only Google's DeepMind team provided early access to their latest models. This underscores how ineffective voluntary compliance is, and the urgent need for regulation not only for the AI industry but, as I wrote in Chapter 1, also the entire tech industry, at least as far as software and hardware are concerned.

The initial safety requirements, in both the US and UK, were as follows:

- Security testing of their AI systems by internal and third-party red teams
- Sharing information across the industry and with governments, civil society, and academia on managing AI risks
- Investing in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights
- Facilitating third-party discovery and reporting of vulnerabilities in their AI systems
- Developing robust technical mechanisms to ensure that users know when content is AI generated
- Publicly reporting their AI systems' capabilities, limitations, and areas of appropriate and inappropriate use
- Prioritizing research on the societal risks that AI systems can pose, including on avoiding harmful bias and discrimination, and protecting privacy
- Developing and deploying advanced AI systems to help address society's greatest challenges

On July 23, 2023, the White House announced that Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI had agreed to voluntarily abide by those requirements.

On September 12, 2023, the White House announced that eight more companies had made the same commitment—Adobe, Cohere, IBM, Nvidia, Palantir, Salesforce, Scale AI, and Stability.

On October 30, 2023, President Biden signed the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. It's extremely lengthy, as it instructs 50 government entities to take 100 actions distributed across 8 policy areas. The Stanford Institute for Human-Centered Artificial Intelligence created a wonderful spreadsheet that simplifies the process of discovering who is responsible for what, and by when.

The issuance of an executive order (EO) carries more weight than a CEO's signature to voluntarily comply with a list of requirements; however, there are a number of drawbacks:

An EO isn't the same thing as an act of legislation that's been signed into law. There's no clear enforcement mechanism in place the way that there is with an act of Congress, and any future president can remove a former president's EOs with nothing more than his signature. You can count on the same companies that signed the voluntary commitments at President Biden's request to spend millions, if not billions, of dollars lobbying House and Senate members to water down any proposed AI legislation.[17]

If you still believe that you can trust companies to self-regulate, look at what happened at OpenAI. In 2015, OpenAI was launched as a nonprofit foundation pursuing artificial general intelligence safely and openly. Then it became a for-profit corporation governed by a nonprofit board of directors.

On November 17, 2023, the OpenAI board fired Altman, its CEO, for not being candid about AI safety issues. Five days later, Altman was back in charge and the board was fired.

Today, OpenAI is a for-profit company where AGI research is done in secret with closed source proprietary algorithms while the company faces investigations in the US and the EU for privacy and copyright violations.

As Elon Musk put it in an interview with *New York Times* journalist Andrew Ross Sorkin, "Fate loves irony."

17 See, for example, *https://oreil.ly/VQk6u* and *https://oreil.ly/b_ZsH*.

Summary

AI is an exciting field with strategic importance across a variety of use cases, ranging from the military to health care to business, entertainment, and finance. It also makes us vulnerable to an entirely new set of harms at a scale never before seen.

Risk

The US government has outsourced its AI risk calculus to Silicon Valley billionaires and AI experts who have adopted EA and its banner of saving humanity from scheming AIs as a belief system. As a result, we are allocating expensive resources dedicated to AI safety to preventing theoretical and highly unlikely outcomes, such as rogue superintelligent computers wiping out the human race, instead of addressing dangerous problems that exist today, such as empowering mass surveillance, cognitive warfare, killer drone swarms, AI's immense drain on energy resources, the loss of jobs, and making vulnerable critical systems easier to attack at scale.

A prudent risk formula is informed by the probability of a bad event happening. If you believe that aliens exist, for example, then you may want a contingency plan on what to do if they attack because such an attack would probably be devastating. However, the likelihood of that happening is so low that you wouldn't want to invest much money into that potentiality, as scary as it sounds. So our focus should be on AI's actual harms, of which there are many, rather than wasting resources on hypothetical horror scenarios generated by EA safety researchers.[18]

Regulation

Because AI is software, there will be an effort by large technology companies to lobby the government to only regulate how AI is *used* rather than how it is *engineered*. It's important that any regulation regime include security requirements that the CISA and the White House have been urging industry to adopt since the release of President Biden's National Cybersecurity Strategy, such as using a memory-safe programming language instead of C and C++.

More importantly, the companies that build these foundational models, like Google, Microsoft, Facebook, OpenAI, and Anthropic, should be held liable for damages instead of offloading their own liability onto the user, which has been the practice for the past 50+ years.

18 For example, see Eliezer Yudkowsky, "Pausing AI Developments Isn't Enough. We Need to Shut It All Down," *Time*, March 29, 2023.

Influence

Because of EA's outsize influence in the world of AI safety at the present time, there should be a public accounting by AI scientists and executives on their personal affiliation with EA, all monies received from EA-funded nonprofits, and how they prioritize the research of their safety teams regarding known risks versus hypothetical risks. Some of my research into EA-affiliated researchers and AI safety organizations can be found here and here.

What happened during the 2023 Thanksgiving weekend at OpenAI—when the board fired the cofounder and CEO, Sam Altman, over safety concerns, and then investor pushback got the board fired and Altman rehired—underscored that profit trumps safety. The *Economist* took note of this in an op-ed that concluded with an important takeaway: that what had happened at OpenAI was "all the more reason for governments to set the tone on AI safety, not mercurial tech visionaries."

Or apocalyptic doomsday cults.

Afterword

Tell me how this ends?
—General David Petraeus[1]

I'm not optimistic about our future.

We are entirely dependent upon devices and systems that cannot be made safe from sabotage or attack, as I laid out in Chapter 1.

We live in a world where our most popular media is run by attention-seeking algorithms that serve to further inflame division and hatred because increasing the user's screen time makes money, while preventing the propagation of misinformation and deepfakes costs money. I discuss this in detail in Chapter 5.

Bad actors operating behind a keyboard are no longer limited to stealing information, as I point out in Chapter 6. They can now steal your information and then burn down your office for good measure, or cause a gas pipeline to explode, or crash the train that you take to work.

There is an international race for achieving human-level AI across all domains, including the military, with zero guardrails in place except for voluntary ones, and that—if you believe the hype—will either bring us into a utopian existence or spark a mass catastrophe.

The software industry, especially the world's largest and most valuable companies, has made much of this possible because it has operated free from liability for 40 years except in cases of negligence. This is insanity, as I point out in Chapter 3. Since regulation has only come to industries, historically, after a catastrophe or an intolerable number of deaths, I anticipate that the software industry will be no different, and I fear for our future.

1 Michael O'Hanlon, "Tell Me How the US-China War Ends", *Hill*, January 17, 2022.

If you are interested in knowing what you can do to keep you and your family safe, or at least as safe as possible, I'll share the three-step plan that I use, and that I recommend for my friends.

Reduce Your Attack Surface

This means to make it more difficult for you to be hacked, tracked, or otherwise suffer harm to your devices. For example:

- Stop using Windows. Switch to either macOS or Linux. Doing so doesn't mean that you're safe, however—it just means that you're less vulnerable.
- Delete all of the apps on your phone that you haven't used in the past 30 days. For whatever remains, turn the "Location" setting to Off. The apps won't be as convenient to use, because some need to know where you are in order to work. In those cases, turn the "Location" setting to On, use the app, then go back and turn it Off. Unfortunately, because security isn't built in, you have to work harder to stay safe.
- Make sure that your Wi-Fi router at home, and everything that connects to it, is password protected with a unique password, and use a password manager to store them.

Create Redundancies for Your Critical Systems

By critical systems, I mean power, water, food, and communications. As I write this, Russia has been systematically destroying Ukraine's power grid. There are serious outages of electricity and water since the water system requires electricity to operate.

My recommendation is to move out of urban areas because they quickly become untenable during any type of extended power outage. Find a rural or semirural environment with year-round growing seasons, local farmers, and abundant wildlife. If you're unsure about how to make that transition, I encourage you to follow my friend Jason N. Gardner's Instagram channel for everything you need to know about rural living and homesteading, as well as an occasional leadership lesson or two.

Diversify Your Risks

You usually hear this phrase from financial experts, but it works as a general rule of thumb.

If the only way you can access your money is by using an ATM card or walking into the branch office of your bank, that's a big risk. A power outage would make it impossible for you to buy anything. In the event of a national emergency, the president has the authority to temporarily freeze the bank accounts of American citizens. Diversify

your risk by keeping a supply of emergency cash available, and in more than one location around your home.

In a state of emergency, 911 calls will not result in police officers showing up at your house—they'll be reassigned to protect critical infrastructure. Medical care may or may not be available. Diversify your risks by creating a neighborhood watch and collaborating on emergency backups of food, water, and basic medicines. Individuals don't survive long in chaotic environments. You need to be part of a group.

I'll close with these words from *Pirkei Avot*: "You are not obligated to complete the work, but neither are you free to desist from it."

Index

C

D

E

F

G

H

I

J

K

L

M

N

O

P

R

S

About the Author

Jeffrey Caruso (né Carr) is a US Coast Guard veteran and has worked in the cybersecurity and cyber warfare field since 2006. He has provided cyber intelligence briefings to the CIA's Open Source Center, the DIA, the FBI, and the Chief of Naval Operations Strategic Study Group. He has been a frequent lecturer at the US Air Force Institute of Technology and the US Army War College, and was a technical peer reviewer for Tallinn 2.0, the second edition of the *Tallinn Manual on the International Law Applicable to Cyber Operations*.

Colophon

The animal on the cover of *Inside Cyber Warfare* is a standardbred horse. Standardbreds are a horse breed with foundation blood lines from Messenger, a thoroughbred brought to America from England in 1788. Messenger's great-grandson, Hambletonian 10, is considered the foundation sire of most standardbreds; other breeds, including the Narragansett pacer and the hackney, also contributed to standardbreds.

Standardbreds are similar in appearance to thoroughbreds but are a bit shorter and longer. They were defined by a certain "standard" of trotting—a mile in less than two-and-a-half minutes—which led to their name. They are divided into trotters and pacers. Trotting is when the diagonal legs work together (e.g., front left and back right), and pacing is when the legs on each side of the body work together (e.g., left, front and back). (The horse on the front of this book is a trotter.) Whether a horse is a trotter or a pacer is typically determined by a gene. The type of gait is important in racing, with 80 percent of harness racing being pacing. An incorrect gait can result in disqualification.

Standardbreds are gentle, hard-working, and easy to train. Because of these traits, they are used for law enforcement, in movies, and in battle re-enactments, as well as in their primary career, harness racing. The Amish use older horses to pull their buggies, since the horses don't need to have the energy required for harness racing. Frequently, standardbreds need additional training for their second careers (since they have been so focused on trotting or pacing), but their ability to learn quickly helps facilitate career changes.

Many of the animals on O'Reilly covers are endangered; all of them are important to the world.

The cover illustration is by Jose Marzan, based on an antique line engraving from *Dover*. The series design is by Edie Freedman, Ellie Volckhausen, and Karen Montgomery. The cover fonts are Gilroy Semibold and Guardian Sans. The text font is Adobe Minion Pro; the heading font is Adobe Myriad Condensed.