Assessing the Risk of AI-Enabled Computer Worms

John Halstead & Luca Righetti

GovAI

Authors and affiliations

John Halstead

GovAI

Luca Righetti

OpenAI

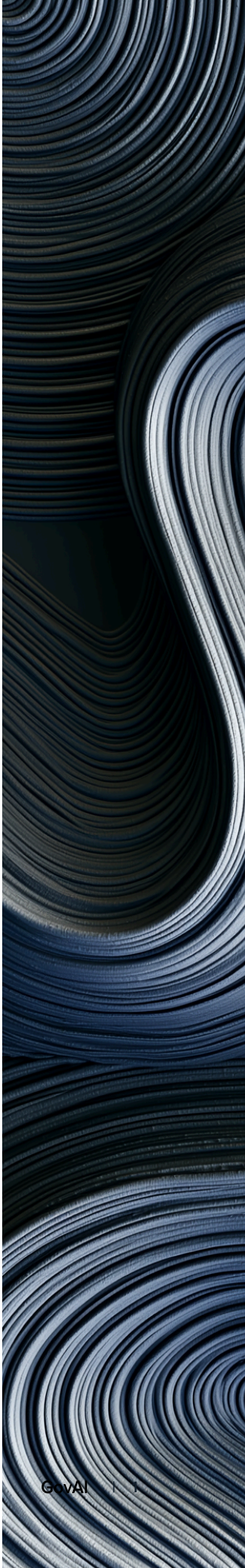
This work represents the views of its authors, rather than the views of the organization, and does not constitute legal advice. This work has been peer reviewed. Any remaining errors are the authors.

Contact

Corresponding author: John Halstead (john.halstead@governance.ai)

Please cite as

Halstead, J., Righetti, L. (2026). *Assessing the Risk of AI-Enabled Computer Worms*. GovAI.



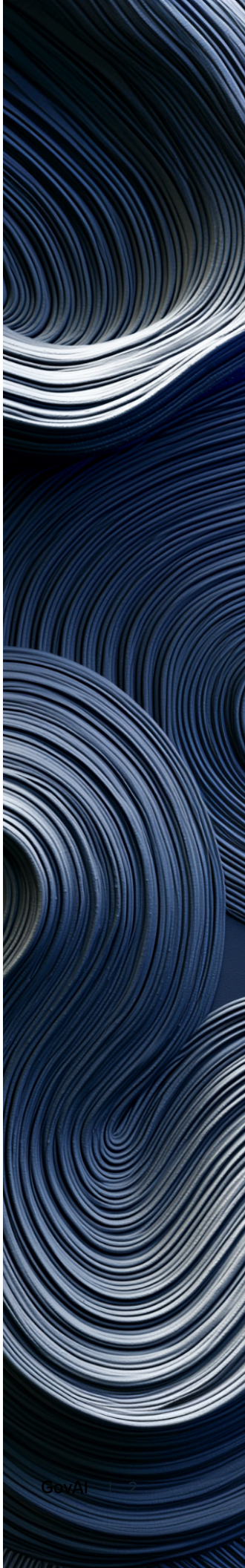
Acknowledgements

This report has benefitted from review, discussions and comments from a large number of people. Our thanks go to Yogev Bar-On, Asher Brass, Christian Chung, Stephen Clare, Steven Comer, Shaun Ee, Janet Egan, Maia Hamin, Lewis Ho, Wesley Hurd, Tom Johansmeyer, Holden Karnofsky, Kyle Kilian, Gabriel Kulp, Stiv Kupchik, Dan Lahav, Omer Nevo, Tiffany Saade, Dmitrii Volkov, Anna Wang, and Jessica Wang.

We also thank our colleagues at the Centre for the Governance of AI for their comments and thoughts: Markus Anderljung, Noemi Dreksler, Ben Garfinkel, Elias Groll, and Matthew van der Merwe. Finally, we are especially grateful to an anonymous peer reviewer for their comments., which made this report substantially better.

About GovAI

GovAI is a 501(c)(3) non-profit organization. Our mission is to help decision makers navigate the transition to a world with advanced AI by producing rigorous research and fostering talent. Researchers at GovAI work on a wide range of topics, with a particular emphasis on the security implications of frontier AI.



Abstract

Several frontier AI companies test their systems for dual-use cyber capabilities, such as vulnerability discovery and exploit development, that might be misused by threat actors. But what do these test results imply about the overall risk from cyberattacks? Answers to this question are needed to calibrate responses to cyber-AI capability progress, particularly in light of recent models' considerable advances.

Currently, few published risk models explain how AI cyber capabilities might cause harm. We take an initial step toward filling this gap by developing an in-depth risk model for AI helping threat actors to develop data-damaging worms similar to WannaCry and NotPetya. We identify the development of "elite exploits" that spread without user interaction, allow remote code execution with high privileges, and are effective against widely used software as the primary bottleneck to such worms.

Drawing on historical case studies, a model that decomposes risk into threat-actor capability, willingness, and resulting damages, and a pilot survey of cybersecurity experts and high-performing forecasters, we conclude that if frontier AI were to enable a quarter of moderately skilled actors to develop elite exploits, the marginal economic damage from data-damaging worms would plausibly run to billions of dollars per year. Conditional on this capability being widely available, respondents' median estimate of the probability of at least one worm attack causing $10 billion or more in 2026 roughly tripled, and median total annual expected damages rose two- to fivefold, to tens of billions of dollars.

These results provide a prima facie case that AI companies should evaluate for this capability and consider mitigations should it emerge – though experts disagreed about which release and safeguard policies would best reduce risk. All estimates carry high uncertainty given the small sample and the fragmentary underlying evidence. This work demonstrates a methodological approach for converting AI-cyber capability evaluations into risk assessments, while highlighting the continued need for better evidence and expert discussion to refine its assumptions.

Executive Summary

AI Cyber Capabilities Are Improving Rapidly, but Threat Models Are Lacking

The cyber capabilities of general-purpose AI models have improved rapidly over the past few years. AI systems can now discover and exploit real-world software vulnerabilities, and several AI companies, governments, and international bodies have raised concern about their potential for cyber misuse. Yet there are few published threat models that explain how AI cyber capabilities might translate into social harm, and how much harm they might cause. Without such models, it is difficult for AI developers to design meaningful evaluations or to calibrate appropriate risk mitigations.

This report takes a first step toward filling this gap. We focus on one specific, narrow threat model: the prospect that AI could assist in the development of what we call "elite exploits" – very powerful software exploits that can spread without user interaction, allow remote code execution with high privileges, and are effective against widely used software. We explore how much the economic risk of data-damaging worms – malware that autonomously propagates across large numbers of systems and encrypts, wipes, or corrupts data – would increase if AI enabled a wider range of actors to develop elite exploits.

Why the Data-Damaging Worm Threat Model Is Especially Concerning

Among AI-cyber threat models, the data-damaging worm threat model stands out for four reasons:

1. **Historical precedent.** Plausibly the most economically damaging cyberattacks ever – WannaCry and NotPetya in 2017 – were data-damaging worms. We estimate that WannaCry caused approximately $1 billion and NotPetya approximately $10 billion in economic damages.
2. **Potential scale of harm.** With modest changes in their code, WannaCry and NotPetya could have been far worse. WannaCry included an accidental kill switch that halted the attack after seven hours; NotPetya was designed to limit damage to Ukraine but still caused severe collateral harm globally. We estimate that worst-case versions of these attacks could plausibly have caused damage on the order of $100 billion.

3. **A relatively narrow capability bottleneck.** WannaCry and NotPetya were the product of an unusual natural experiment: Elite exploits developed by the NSA were leaked publicly by the Shadow Brokers group in April 2017 and were used in both attacks within two months. This, along with other evidence, suggests that elite exploit development is the key technical bottleneck to data-damaging worms, as it is substantially harder than the other tasks involved in creating such worms.
4. **Many actors would be willing to launch such attacks.** From 1998 to 2005, individual hackers regularly launched worm attacks that infected thousands to tens of millions of systems (Figure ES1). Improved cybersecurity has since made this much harder, and recent major attacks have been carried out by states. If AI lowered the capability barrier, many actors would likely exploit it.

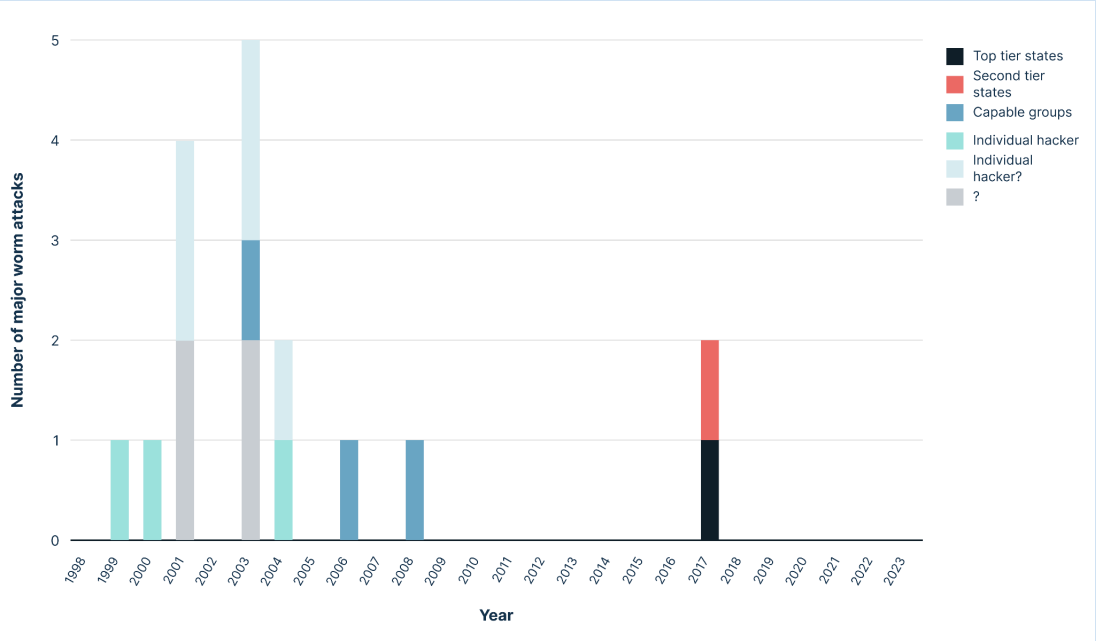


Figure ES1. Major worm attacks by threat actor (1998–2023). [Adapted from Johansmeyer, 2024]

Our Approach: Structured Surveys for Risk Estimation

To estimate the risk, we combined our own analysis with a small pilot survey of cybersecurity experts and high-performing superforecasters, conducted in partnership with the Forecasting Research Institute. We estimated risk in two ways. First, we developed a simple risk model that decomposes the risk into threat actor capability, willingness, and potential damages. Second, in order to account for model

uncertainty, we surveyed respondents on the overall probability and expected damages of data-damaging worm attacks.

We provided experts and superforecasters with an earlier version of this report and asked them to estimate risk for two scenarios:

1. **Baseline:** Assuming no further improvements in AI
2. **AI uplift:** Assume that AI models achieve what we call "Elite Exploit Uplift": "A study conducted at the end of 2025 finds that access to frontier AI models enables 25% of moderately skilled single hackers to develop elite exploits with three months of full-time effort."

By default, we assume models are released open-weight with no deployment safeguards.

Key Findings

Elite exploits are a key bottleneck. Authors, experts, and superforecasters broadly agree that developing elite exploits is much harder than the other tasks involved in creating data-damaging worms – in many cases by an order of magnitude or more. This bottleneck is what currently prevents lower-skilled actors from launching major worm attacks.

AI uplift to elite exploit capabilities would substantially increase risk. The risk model implies that, conditional on Elite Exploit Uplift, expected damages from the first major data-damaging worm attack increase by roughly $1–10 billion per year compared to baseline, though with wide confidence intervals (Figure ES2). Two different direct risk estimates also suggest that Elite Exploit Uplift increases risk by roughly 3–5x.

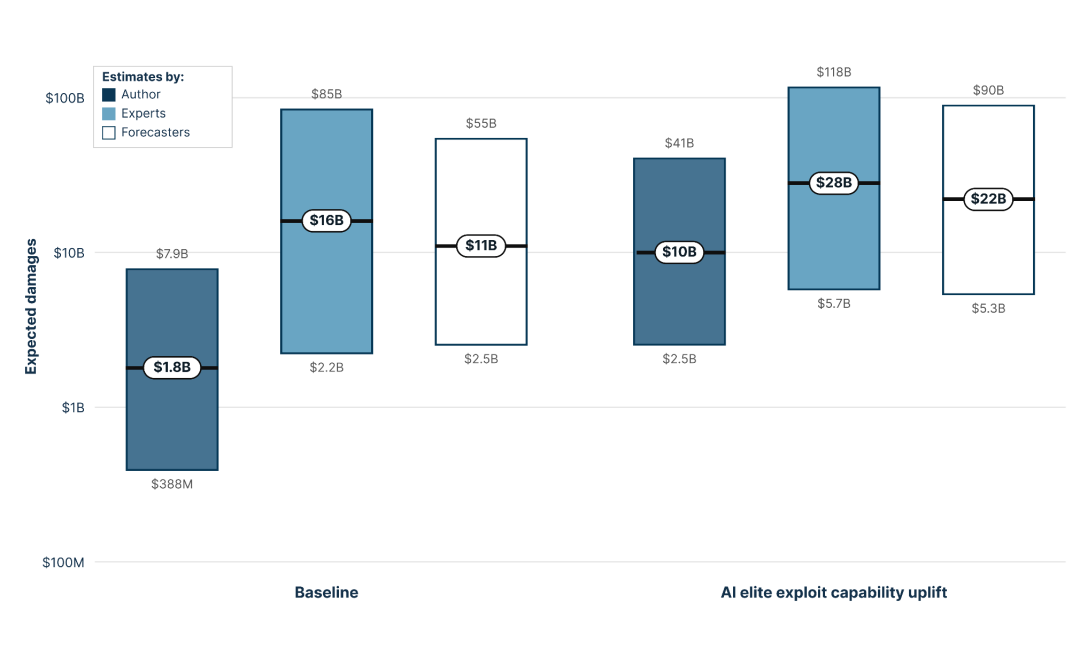


Figure ES2. Expected damages over a year from the first data-damaging worm attack, log-scale y-axis. Bars represent the 90% confidence interval, horizontal lines represent the median.

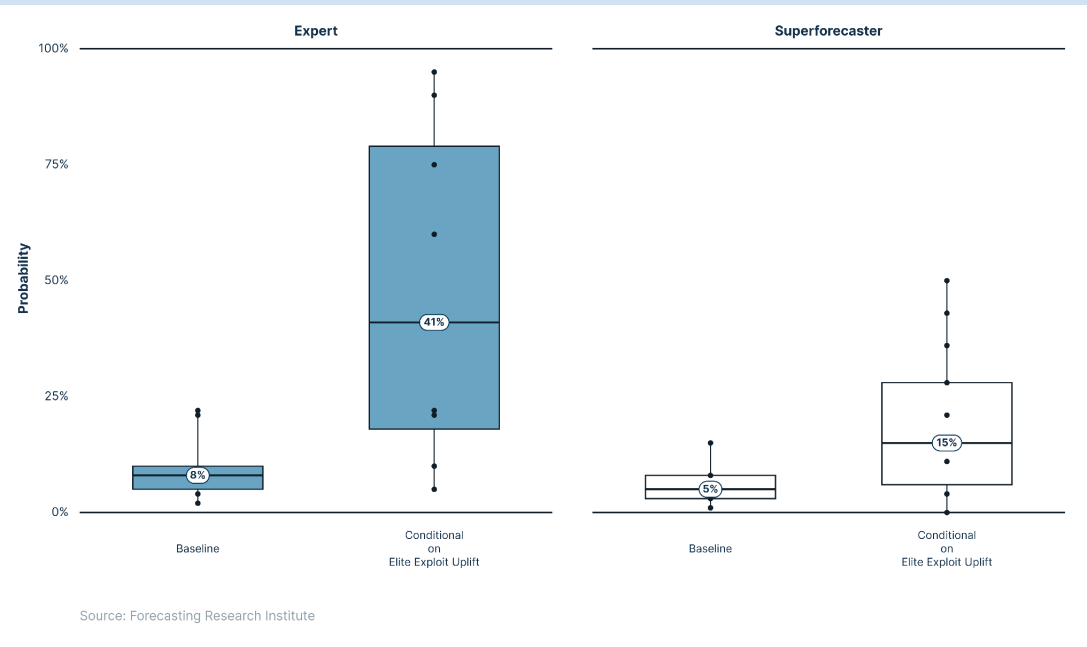


Figure ES3. Probability of at least one data-damaging worm attack causing at least $10 billion in economic damages in 2026, conditional on Elite Exploit Uplift

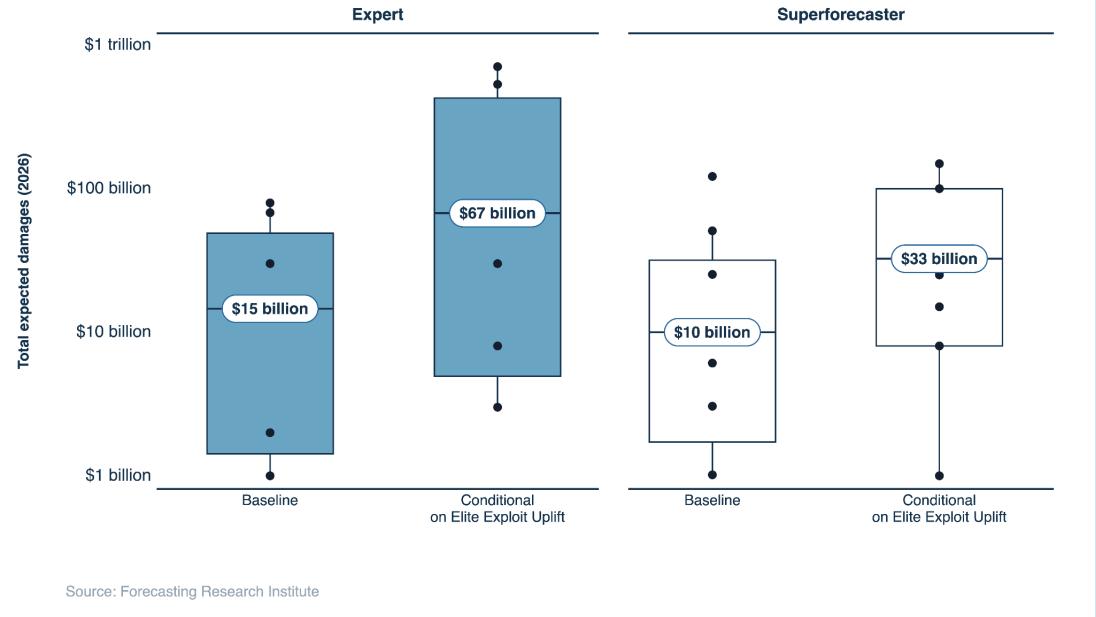


Figure ES4. Total expected damages due to data-damaging worm attacks in 2026

There is broad agreement on risk magnitude, but high disagreement on specific parameters. The authors, median experts, and median superforecasters generally agree on the order of magnitude of the risk. However, there is substantial disagreement – both within and between groups – on the willingness of state actors and the capabilities of mid-tier threat actors. These disagreements point to important cruxes for future research.

Initial Evidence on Risk Mitigation Policies

We also surveyed respondents on the effects of two risk mitigation policies, compared to the default of open-weight release with no safeguards:

1. **Proprietary models with deployment safeguards** such as refusals, anti-jailbreak measures, and API-only access
2. **Early defender access**, giving vetted defenders a four-month head start with unrestricted models before open-weight release

Both policies were estimated by the median expert and superforecaster to reduce risk relative to unrestricted open-weight release, though there was high disagreement about the magnitude of the effect (Figure ES5). Respondents noted that deployment safeguards would primarily affect lower-skilled actors, while more sophisticated actors could likely circumvent them. Several respondents questioned

whether a four-month head start would be sufficient for defenders, and some argued that eventual open-weight release would negate much of the benefit.

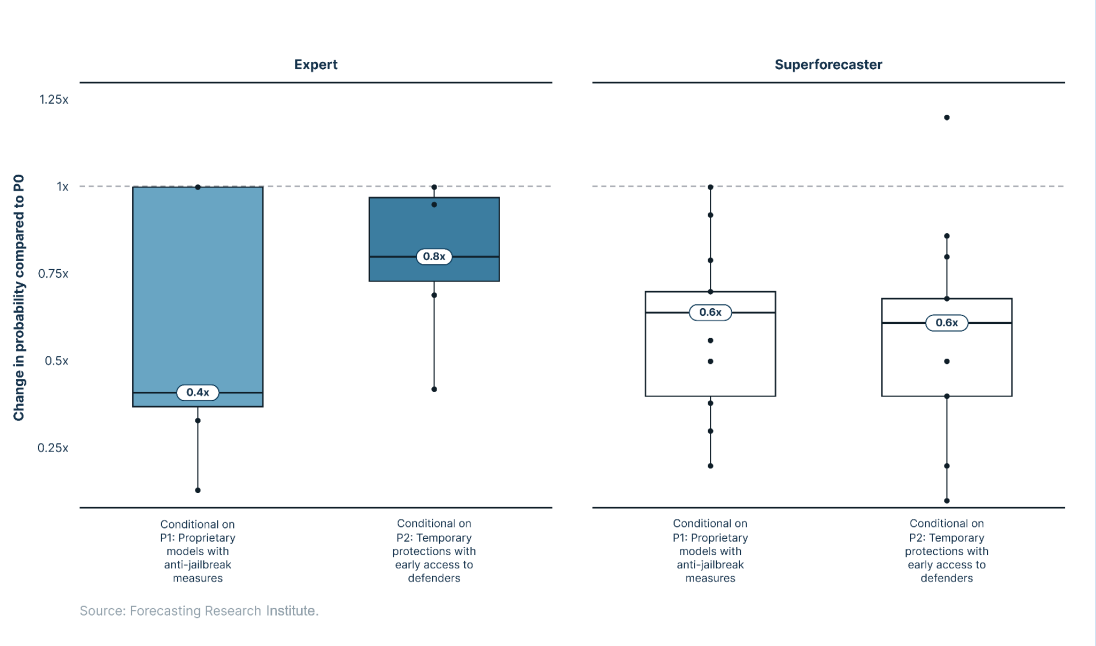


Figure ES5. Survey results on the effect of different risk management policies on the risk of data-damaging worms, conditional on AI Elite Exploit Uplift.

More research on risk mitigation strategies seems warranted. The lack of a consensus on the optimal policy response to AI cyber risk and uncertainty about how AI vulnerability discovery and exploit development capabilities affect the offense-defense balance indicate the need for additional research on strategies for cyber risk mitigation.

Important Caveats

Our pilot survey had a small sample size (17–21 respondents per wave), and the underlying evidence on many parameters is fragmentary. The quantitative estimates presented here are uncertain and should be interpreted as order-of-magnitude guides rather than precise figures. We have tried to mitigate the risk of false precision by presenting wide confidence intervals and complementing the risk model with direct survey estimates. Despite these limitations, we believe that explicit, quantitative risk estimation – even when deeply uncertain – is a useful complement to qualitative judgment, and preferable to relying on intuition alone.

Further, frontier models' capabilities related to elite exploits have considerably improved since we fielded the survey (July 2025–January 2026). At that time, frontier models had strong vulnerability discovery capabilities, but were weak at developing exploits. In 2026, models' exploit development capabilities improved dramatically with the development of Anthropic's Claude Mythos model. A version of Mythos without strong cyber safeguards has not, at the time of writing, been released publicly, so we do not think our Elite Exploit bar has been met. We have also not investigated the exploit capabilities of Mythos in depth. However, we think it is difficult to rule out the possibility that an open-weight, safeguard-free version of a Mythos-level model would reach our Elite Exploit Uplift bar. The median expert in our survey forecast that Elite Exploit Uplift would be achieved in 2031, and the median superforecaster in 2029. Both forecasts predate the Mythos preview and now appear conservative.

Report Outline

The remainder of the report proceeds as follows. Section 1 sets out our definitions, scope, and methodology; readers already familiar with threat modeling methodology may wish to skip straight to Section 2. Section 2 and Section 3 assess the willingness and capability of threat actors to launch data-damaging worm attacks, drawing on historical cases and our expert and superforecaster survey. Section 4 reviews estimates of the economic damages from past worm attacks, and estimates the economic damage a future attack could cause. Section 5 discusses offense-defense balance: the extent to which AI capabilities might differentially benefit attackers or defenders. Section 6 brings the preceding discussion together into baseline and AI-uplift risk estimates and tests how far different model-release and safeguard policies would change them.

Contents

Executive Summary 4

Report Outline 11

1. Introduction 14

1.1. Background 14
1.2. Key Definitions 15
1.3. Scope of the Report 17
1.4. Why We Prioritized the Data-Damaging Worm Threat Model 21
1.5. Methodology 23

2. How Willing Are Threat Actors to Release Data-Damaging Worms? 31

2.1. Background on Past Worm Attacks 31
2.2. Motivations for Historical Worm Attacks 35
2.3. Incentives and Disincentives to Release Data-Damaging Worms 37
2.4. Estimating Actor Willingness to Launch Data-Damaging Worm Attacks 41

3. How Capable Are Threat Actors of Releasing Data-Damaging Worms? 46

3.1. Elite Exploit Development as Key Capability 46
3.2. Assessing How Capable Different Threat Actors Are of Finding Elite Exploits 53
3.3. Estimating Actor Capabilities 58

4. Potential Damages from Data-Damaging Worm Attacks 64

4.1. Estimating Economic Damages from Worm Attacks Is Challenging 64
4.2. Historical Data on the Damages of Past Worm Attacks Is Limited but WannaCry and NotPetya Plausibly Caused Damages of $1B–$10B 64
4.3. With Modest Changes, WannaCry and NotPetya Could Have Done Far More Damage 68

5. Offense-Defense Balance 78

6. Overall Risk Estimates 81

6.1. Historical Base Rate Damages 81
6.2. Hypothetical Marginal Risk Scenario 81
6.3. Risk Model Estimates 81
6.4. Direct Risk Estimates 85

6.5. The Effects of Different Risk Mitigation Policies....90
6.6. Current and Future Model Capabilities....93

7. Conclusion....96

Appendices....1

A.1. Shadow Broker Prices....1
A.2. Technical Details of How EternalBlue Enabled the WannaCry and NotPetya Worms....1
A.3. Elite Exploit Prices....3
A.4. FORCEDENTRY Elite Exploit and Possible Worm Application....4
A.5. Elite Exploits Discovered Being Used in Cyberattacks Since 2020....6
A.6. Motivations and Functionality of Past Worm Attacks....8
A.7. A Critique of Estimates of Past Cyber Damage Estimates....11
A.8. Crosignani et al Estimate of NotPetya Costs....14
A.9. Offense-Defense Balance....16
A.10. Defining Different Model Release and Safeguard Policies....19
A.11. For AI Developers Designing Capability Thresholds, It Makes Most Sense to Consider Expected Costs Up to a Period of Around 6–12 Months....23

About the Authors....25

References....26

1. Introduction

1.1. Background

The cyber capabilities of general-purpose AI models have improved extremely quickly in recent years (Irregular 2025; International AI Safety Report 2026, sec. 2.1.3.). As a result, AI companies, governments, and other experts have raised concern about the potential for the misuse of frontier AI systems for cyberattacks (International AI Safety Report 2026, sec. 2.1.3; OpenAI 2025; Google DeepMind 2025).

In Europe and the United States, policymakers have implemented rules to help mitigate AI risk. California's SB53 and New York's RAISE Act require frontier AI firms to publish and abide by frameworks managing catastrophic risks. The EU AI Act requires providers of systemic-risk general-purpose AI models to test for and to mitigate systemic risks.

To address risks from AI systems, their developers have adopted frontier AI safety policies, which are risk management tools to identify dangerous capabilities and how to mitigate them. As part of such frameworks, several companies have committed to mitigating catastrophic cyber risks that might emerge from their models (OpenAI 2025; Google DeepMind 2025; Meta 2025).

Frontier AI safety policies are designed to manage only a subset of especially severe or catastrophic risks, though AI companies also aim to manage less severe risks, such as AI models producing abusive content or other violations of usage policies (Meta 2025, p. 12; OpenAI 2025, p. 4; Anthropic 2026, pp. 3–4).

At present it is difficult to calibrate an appropriate response to AI-cyber capabilities because there is a lack of published threat models explaining how AI capabilities translate into increased social harm. We use the term "threat model" in a similar sense to "risk model." By this, we mean a structured account of how a capability could lead to harm and how large or likely that harm might be. (This contrasts to how the concept of a "threat model" is typically used in the cybersecurity context, where threat models typically describe the adversaries, attack paths, and defenses relevant to a given target.) Rigorous risk modeling of AI systems' social costs allows firms and policymakers to devise more appropriate responses to AI's cybersecurity risks at a time when there is little consensus on why these capabilities might be harmful and how harmful they might be.[1] Currently, for example, it is difficult to interpret whether weak performance on an evaluation implies that a model is safe and whether strong performance on an evaluation implies that a model is too risky (Lukošiūtė and Swanda 2025). Risk models are a way to interpret how evaluation performance translates into social cost.

[1] For example, Anthropic (2026) does not include cyber risk in its Safety Policy, but cyber is included in other companies' Safety Policies (OpenAI 2025; Google DeepMind 2025).

This report aims to inform AI-cyber risk governance by developing an in-depth risk model for one specific pathway to harm: AI-enabled data-damaging worms. We focus on how the use of AI to discover vulnerabilities and develop exploits might lead to significant social harm via data-damaging worms. Vulnerability discovery and exploit development is already a major focus of model evaluations, but there is a lack of public threat modeling work explaining why such capabilities might be harmful and how harmful they might be.

For this report, we develop a forecast for one specific and narrow question:

"If future AI systems enable various threat actors to develop what we call "elite exploits", then how much would this increase the economic risk from data-damaging cyber worm attacks, similar to WannaCry or NotPetya?"[2]

We produce our own forecasts for this question and elicit estimates from experts and high-performing forecasters.

Threat modeling is necessary but not sufficient for effective management of AI-cyber risks. It can tell us how risky a capability might be, but it does not tell us whether that level of risk is unacceptable. There is as yet little consensus in AI governance on appropriate catastrophic risk thresholds for AI: SB53 uses a threshold of $1 billion of economic harm (California leginfo 2025), whereas OpenAI (2025) tracks AI capabilities that could cause more than $100 billion in economic damage. We leave the question of acceptable risk thresholds to future work.

1.2. Key Definitions

Here we define key terms used in this report, including worms, vulnerabilities, exploits, zero-days, n-days, patches, and payloads.

Worms

A worm is malicious software, or malware, that can autonomously propagate from one computer system to another without requiring the attacker to separately compromise each system. Some worms require victims to take certain actions to spread. For example, many early worms spread via email and required users to open malicious attachments in emails, which enabled the malware to send the worm to contacts in the victim's address book. Other worms require no action by victims to spread between systems.

Worms can cause damage in various different ways. We focus on what we call "data-damaging worms", which we define as worms that directly damage (wipe, encrypt, or corrupt) data on a large number of systems. However, not all worm attacks cause social harm in this way. We discuss this in section 1.3.

[2] The type of social harm we focus on here, as we discuss below, includes only economic damages. We do not include other harder-to-quantify social costs, such as political or cultural effects.

Vulnerabilities, Exploits and Patches

A vulnerability is a flaw in software or hardware that creates a security weakness in the design, implementation, or operation of a system or application that can be exploited by an attacker (Ablon and Bogart 2017, p. 2). Vulnerabilities can be introduced intentionally or unintentionally through an accidental design or implementation flaw.

An exploit is malicious code that takes advantage of one or more software vulnerabilities to infect, disrupt, or take control of a computer without the user's consent and typically without their knowledge (Ablon and Bogart 2017, p. 2). Finding vulnerabilities and developing exploits of those vulnerabilities are distinct tasks and require different skillsets (PatternLabs 2025).[3]

Exploits are often categorized in terms of what they allow attackers to do (Wilson et al. 2023, p. 4). For example:

- **Local privilege escalation** exploits allow an attacker with limited access to a system to gain higher privileges, such as administrative or root access. These require prior access to the system, often through a lower privileged account (Teodorczyk nd).
- **Sandbox escape** exploits allow attackers to break out of a restricted execution environment (sandbox) to access sensitive data or execute code on the host system (NordVPN nd).
- Remote code execution exploits allow attackers to execute arbitrary code on a system without the user's knowledge, and without attackers requiring physical access to the system (Baker 2022).

Attackers often chain multiple exploits together in order to obtain the access and system privileges needed to achieve their desired effects. (Wilson et al. 2023, p. 4).[4]

In this report, we focus on one class of very powerful exploits, which we call "elite exploits." These are exploits that allow remote code execution with high privileges, that are effective against widely used software, and do not require actions by the victim in order to infect a system (i.e. "zero-click exploits"). Table 1.1 explains these features in more detail.

Exploit feature	Definition
Zero-click	Infection requires no user interaction, such as opening emails, clicking links, or visiting a webpage.[5]

[3] See also Charlie Miller, 'How to build a cyber army to hack the US', Defcon (2013).
[4] See Appendix A.4.
[5] Note that on some definitions of 'zero-click', not all zero-click exploits require absolutely no user interaction to spread. For example, some watering hole attacks can infect a system if a user visits a website but does not click on any links on the website. Some definitions class this as a zero-click exploit (e.g. Yu 2024). On our definition, these would not be elite exploits because they require the user to visit a compromised website, and therefore involve user interaction.

Remote code execution	Allow attackers to execute arbitrary code on a system without the user's knowledge and without attackers requiring physical access to the system
High privileges	Privileges are the permissions granted to users, programs, or processes to perform specific actions on a system or access particular resources. Privilege levels range from (low to high): sandboxed application, user, administrator, to system level. Elite exploits have administrator privileges or higher.
Targets widely used software	Effective against >10M systems

Table 1.1. Defining elite exploits

Zero-day vulnerabilities (or "zero-days") are vulnerabilities that are unknown to the software or hardware vendor (Smeets, p. 21), while n-day vulnerabilities are vulnerabilities that are known to the vendor for some number (n) of days (Smeets, p. 21). **Zero-day exploits** are exploits of zero-day vulnerabilities, while **n-day exploits** are exploits of n-day vulnerabilities.

A **patch** is a software update that fixes a vulnerability. By definition, there are no patches for zero-day exploits. However, once vendors become aware of vulnerabilities, they typically, after a delay, develop and release patches for them. After a further delay, these patches are then deployed or installed by users.

Payloads

The **payload** is the element of malicious software that achieves the ultimate desired effect on the infected system (Wilson et al. 2023, p. 5). For instance, in a ransomware attack, the payload is the tool that encrypts the victim's files. For spyware, the payload would be the component of the malware that captures and transmits sensitive information from the victim's device to the attacker.

1.3. Scope of the Report

There are many possible AI-cyber threat models. This report reviews one specific narrow scenario in depth. This is not a comprehensive review of AI-cyber threat models, and future work should cover other scenarios.

Following NIST (2025), Appendix E, we can categorize risk models or threat models by: (1) the type of cyber attack; (2) the relevant capability that could enable such attacks; (3) the threat actor pursuing it; and (4) the potential negative outcomes that might result. This report specifically focuses on (1) data-damaging worms that damage (encrypt, wipe, or corrupt) data on a large number of infected systems, (2) enabled by what we call "elite exploits," (3) that any actor might use to (4) cause widespread economic damage.

Table 1.2 summarizes what is in and out of scope for this threat model.

	Type of Attack	Key Capability	Threat Actor	Outcome
In scope	Data-damaging worms: malware that is designed to propagate automatically to infect and encrypt or damage data on a large number of systems (e.g. WannaCry and NotPetya). We focus on worms that cannot agentically change their code after release.	"Elite exploit" development. This includes both the ability to find critical vulnerabilities in software and to write the code to exploit them. "Elite exploits", as we define them, can infect a large number of systems, require no user interaction to spread, and offer a high degree of control over target systems.	All (See Table 1.2)	Large economic damages (>$1B): costs to firms and consumers from economic disruption. This could include lost revenue, reduced consumption, remediation costs, and costs to insurers.
Out of scope (selected examples)	Worms to create botnets: Worms can infect a large number of systems that can be used to launch denial of service attacks or send spam (e.g. Storm Worm). Worms which only create damage by increasing network traffic: Some worms do not directly damage infected systems, but cause damage by creating a large amount of network traffic (e.g. Sasser). Polymorphic worms that change their code after release (e.g. Conficker) Attacks causing physical damage: (e.g. Russian attacks on Ukraine grid; Stuxnet worm attack on Iranian nuclear centrifuges). Industrial espionage: Cyber attacks on firms in order to steal	Other steps in the cyber kill chain, including reconnaissance, phishing, social engineering, malware development, other post-intrusion tasks such lateral movement, privilege escalation, deployment of payloads, and data exfiltration, and analysis of exfiltrated data		Broader welfare costs: costs on other determinants of human welfare, such as health. Geopolitical costs: Some attacks have important geopolitical implications (E.g. China's 2015 hack of the Office of Personnel Management compromised security clearance data on millions of Americans.)

	IP. (e.g. China's theft of IP for the F-35 fighter jet). State espionage: Cyber attacks by states in order to steal sensitive information (e.g. 2020 Solarwinds attack).			

Table 1.2. Different AI-cyber threat models.

1.3.1. Types of Actors: Differentiating by Operational Capacity

We define threat actors in terms of their operational capacity. Operational capacity (OC) levels refer to the resources and capabilities available to the operation, ranging from OC1 operations to OC5 operations. By definition, each category includes the capacities of all preceding ones. For example, the most competent nation-states, such as the US and China, are able to carry out OC5 operations, but also all operations below that level (Nevo et al. 2024, pp. 9–10). Table 1.3 summarizes these definitions.

Following from this, we define five threat actor categories, ranging from TA1 to TA5, in terms of their highest possible OC level. For example, because some cybercrime groups are able to carry out OC3 operations but not higher, they are a TA3 threat actor; because China is able to carry out OC5 operations, they are a TA5 threat actor, and so on. We treat states as single or unitary threat actors. For instance, we treat China as a single threat actor, rather than treating each Chinese state or state-backed cyber team as single threat actors.[6]

For ease of analysis, we assume that each threat actor in a threat actor class has the same capability level. China and the US are TA5 actors, so we assume they have the same capability level; North Korea and Iran are TA4 actors, so we assume they have the same capability level.

Note that, following the RAND (2024) definitions, the resources required for a given OC level increase by an order of magnitude through each category, except for the difference between TA2 and TA3 actors and TA4 and TA5, which increases by two orders of magnitude.

[6] In this respect, our definition is different to that commonly used in cyber threat modelling.

Operational Capacity level	Definition	Threat Actor Level	Definition	Example	Estimated number of actors[7]
OC1	Operations roughly less capable than or comparable to a single individual with some limited professional expertise in information security spending several days with a total budget of up to $1K on the specific operation, and no preexisting infrastructure or access to the organization	Threat Actor level 1 (TA1)	Actors only capable of OC1 operations but not higher.	Hobbyist hackers	~1M
OC2	Operations roughly less capable than or comparable to a single individual who is broadly capable in information security spending several weeks with a total budget of up to $10K on the specific operation, with preexisting personal cyber infrastructure.	Threat Actor level 2 (TA2)	Actors only capable of OC2 operations but not higher.	Individual professional hackers	10K–100K
OC3	Operations roughly less capable than or comparable to ten individuals who are experienced professionals in information security spending several months with a total budget of up to $1 million on the specific operation, with major preexisting cyberattack infrastructure	Threat Actor level 3 (TA3)	Actors only capable of OC3 operations but not higher.	Well-known criminal hacker groups, well-resourced terrorist organizations, and industrial espionage organizations.	100–1K
OC4	Operations roughly less capable than or comparable to 100 individuals who have experience in a variety of relevant professions (cybersecurity, human intelligence gathering, physical operations, etc.) spending a year with a total budget of up to $10 million on the specific operation, with vast infrastructure and access to state resources such as legal cover, interception of communication infrastructure, and more.	Threat Actor level 4 (TA4)	Actors only capable of OC4 operations but not higher.	Some leading cyber states (e.g. North Korea, Iran)	10–50
OC5	Operations roughly less capable than or comparable to 1,000 individuals who have experience and expertise years ahead of the (public) state of the art in a variety of relevant professions (cybersecurity, human intelligence gathering, physical operations, etc.) spending years with a total budget of up to $1 billion on the specific operation, with state-level infrastructure and access developed over decades and access to state resources such as legal cover, interception of communication infrastructure, and more.	Threat Actor level 5 (TA5)	Actors capable of OC5 operations	The world's most capable states (e.g. the US, China, Russia)	~5

Table 1.3. Threat actor level definitions

[7] These estimates are based on the results of our small survey and are therefore uncertain.

1.3.2. Type of Damage: Large Economic Costs

We focus only on the economic damages caused by worm attacks (lost revenue, reduced consumption, etc.). However, some cyberattacks may cause other types of harder-to-quantify harm. For example, although the economic damages of the Stuxnet attack were relatively limited, the geopolitical effects may have been important. Russia's NotPetya attack on Ukraine caused substantial economic damages, but was also part of a longstanding campaign of cyber and military attacks with potentially important geopolitical consequences, which we exclude from our analysis.

We focus only on easier-to-quantify economic damages because they are more tractable to analyze and rely less on access to classified information. Future work could explore other types of social damages, but they are out of scope for this report.

1.4. Why We Prioritized the Data-Damaging Worm Threat Model

In brief, we prioritized the data-damaging worm threat model for further study for three main reasons.

1. **Some sources suggest that worm attacks pose especially large economic risks.** Various estimates suggest that past worm attacks have caused billions to tens of billions of dollars of damage. Our own analysis (discussed in section 4) showed many of these estimates to be unreliable, though we conclude that some worm attacks, notably data-damaging worm attacks, have been among the most economically damaging cyberattacks ever.

2. **Many actors would be willing to launch worms that could cause a large amount of economic harm but most currently lack the ability to do so.** Prior to 2005 when cybersecurity was generally much weaker, there were numerous worm attacks launched by low-skilled actors, who now seem to lack the capability to launch such worms.[8]

3. **AI-enabled elite exploit development capabilities would significantly reduce barriers to data-damaging worms.** There is reason to think that if AI gained two related narrow technical capabilities – the ability to find critical vulnerabilities and develop elite exploits for them – then this would significantly lower the capability barriers to launching data-damaging worms today.[9] In contrast, many other types of high-impact cyberattacks, such as espionage and critical infrastructure attacks, are bottlenecked by a much broader suite of capabilities (See International AI Safety Report, 2025, p.75; and van der Merwe and Righetti 2026).

[8] We discuss this in more detail in sections 2 and 3.
[9] We discuss this in more detail in section 2.

Due to these three factors, the data-damaging worm threat model seems prima facie more concerning than other cyber threat models.

Other types of worms are concerning from a cybersecurity perspective but appear to have the potential to cause less damage than data-damaging worms.

- **Damage via network traffic:** Some worms cause damage only via propagating rapidly and thereby causing a large amount of network traffic, in turn causing networks to shut down. For example, the 1999 Melissa worm caused damage in this way, without directly damaging data on infected systems (GAO 1999). These worms are less likely to cause lasting damage than data-damaging worms, and recovery from them seems likely to be quicker and easier. Moreover, data-damaging worms can both damage data and create a large amount of network traffic.

- **Botnets:** Some worms infect a large number of systems to create botnets, a network of computers controlled by an attacker and often used to send spam or launch denial of service attacks. For example, Storm Worm did not damage infected systems directly, but caused damage via creating a botnet to launch denial of service attacks against anti-spam websites and security vendors (Schneier 2007). Denial of service attacks could potentially cause large economic damages, but that depends to a significant extent on which victim is targeted, and there seems to be greater scope to cause harm by damaging a large number of infected victims.

- **Physical damage:** Worms can be used to cause damage to physical systems. For example, the Stuxnet worm launched by the US and Israel physically damaged centrifuges at an Iranian nuclear facility (Falco 2012). Causing physical damage to operational technology, such as electrical grids or nuclear centrifuges, via cyberattacks in general and worms in particular, seems very challenging (see van der Merwe et al., forthcoming).[10]

- **Data exfiltration:** Some worms can be used to extract sensitive data for state or industrial espionage. For example, the Flame malware had worm-like properties and was used for espionage in the Middle East (Securelist 2012). Data theft may cause serious non-economic harms that are difficult to model, such as the geopolitical consequences of espionage. We treat these sorts of effects as out of scope.

[10] The Stuxnet worm, for example, may have been the most economically damaging worm that also caused physical damage, and may have delayed Iranian nuclear enrichment for several months (Slayton 2017). The malware alone cost millions to tens of millions of dollars to develop, and the head of the CIA at the time of the attack said that the total cost of the operation including human intelligence and building, and testing on, mock centrifuges was $1B - $2B (Modderkolk 2024). As we discuss in section 3, creating data-damaging worms is several orders of magnitude cheaper than this.

1.5. Methodology

To develop a threat model of AI-enabled data-damaging worms we rely on historical case studies, evidence on threat actors capability and willingness, and a quantitative risk model that decomposes risk into the capability and willingness of threat actors to launch a data-damaging worm and the damages that would result from such an attack. We supplement our own estimates of risk with a survey of cybersecurity experts and superforecasters, asking them to estimate parameters of the risk model and expected damages of baseline and AI-enabled worm attacks.

1.5.1. Baseline and Marginal Risk

Our aim is to estimate the marginal or "net new" risk posed by future hypothetical AI capabilities, compared to existing technologies (Kapoor et al. 2024; NIST 2025; OpenAI 2025, p. 4; Meta 2025, p. 12). In order to increase marginal risk, AI needs to make a difference compared to technologies already available to different threat actors.

Thus, we distinguish baseline damages, assuming no further improvements in AI, from marginal damages, assuming hypothetical improvements in AI-cyber capabilities.

1.5.2. Advantages and Disadvantages of Quantification

We aim to produce a quantitative estimate of the economic risks of specific AI cyber capabilities. However, the relevant data is often fragmentary and low quality, and our survey has a small sample size. Consequently, we have low confidence in the various risk estimates we produce here, which is reflected in the wide confidence intervals for our final estimates. In general, there is a concern that quantified estimates can convey excessive confidence in an estimate, which may not be warranted (Friedman et al. 2018). Deep uncertainty and expert disagreement – which will often characterize frontier AI risks – can make single numbers misleading, and there may be a tendency for people to put too much weight on numerical estimates.

Despite this, we still think it is valuable to produce quantitative risk estimates. The main alternative to quantitative estimates is to use qualitative intuitions or judgments. However, like quantitative estimates, these are also subject to bias. Quantitative estimates have the advantage that they can make subjective intuitions precise and make clear the cause of disagreement about estimates. This can allow more productive discussion about the order of magnitude of a risk. Order of magnitude estimates can help to guide decisions, even if they are not precise. Moreover, some of the problems with quantification can be mitigated by providing confidence intervals and accompanying qualitative confidence descriptions.[11]

Given the respective advantages and disadvantages of qualitative and quantitative approaches, relying on only one approach seems suboptimal. For the purposes of decisions about risk management, quantitative estimates may be a useful complement to qualitative judgment.

[11] The Intergovernmental Panel on Climate Change takes a similar approach (Risbey and Kandlikar 2007).

1.5.3. An Outline of Our Approach to Risk Estimation

We use several different methods to estimate the baseline and marginal risks of data-damaging worms. We produced our own estimates and also conducted a small pilot survey of cyber experts and high-performing superforecasters. We describe the survey methodology in more detail in section 1.5.4. We elicited baseline and marginal risk estimates via two broad approaches:

1. **Risk model:** We construct a simple risk model which decomposes the risk into several parameters. We developed our own model parameter estimates and also elicited expert and superforecaster estimates on the model parameters. We can then use the risk model to produce a baseline and marginal risk estimate.

2. **Direct risk estimate:** Surveying experts directly on the baseline and marginal risk without asking them to estimate decomposed risk model parameters.

Each of these approaches has advantages and disadvantages, as outlined in Table 1.4.

Estimation approach	Advantages compared to the alternative
Direct risk estimate	• Accounts for model uncertainty and disagreement about the specification of the risk model. For example, if direct risk estimates are systematically different to those produced by the risk model, then this may reflect expert disagreement about the model specification. • Easier to survey a large number of experts and forecasters, as you only have to survey on one variable, rather than multiple parameters, and there is less of a need to provide context and background on the risk model.
Risk model	• Shows reasoning and explicitly lays out the causal chain. • Easier to identify cruxes and causes of disagreement (e.g. perhaps highlights that disagreements about chemical weapons risk are primarily driven by disagreements about the "attempt rate" parameter). • It may be easier for experts and forecasters to reliably estimate the risk model parameters than the overall risk. • Can inform questions where gathering additional information on a risk model parameter, such as by running randomized control trials or other experiments, would be valuable.

Table 1.4. Respective advantages of risk models and direct risk estimates

1.5.4. The Risk Model

In this section, we describe our general approach to risk modeling and our specific risk model for data-damaging worm risk.

The Function of Risk Models

As noted, risk models are one way to produce risk estimates, but they also have other functions. Risk models can help to identify the bottlenecks to a risk, which can in turn help in the design of AI model evaluations. This can also help to prioritize risk mitigations.

For example, our research for this report identified elite exploits as a key bottleneck to data-damaging worms: if AI overcomes this bottleneck, then the risk from data-damaging worms would increase. AI companies can design evals to test for this capability, and could focus mitigation efforts on managing access to this capability.

Managing the Trade-Off Between Decomposition and Model Simplicity

Risk models decompose a risk into subcomponents. This can aid risk estimation, as the subcomponents may be more tractable to estimate than the overall risk. There is evidence that decomposing estimates produces more accurate estimates in some cases (Gomilsek et al., 2024; Tetlock and Gardner, Superforecasting 2015, Ch. 5). The reason for this is that the subcomponents may be more tractable to analyze than the overall risk taken as a whole.

However, decomposing too much introduces model complexity, which has a range of drawbacks. Empirical evidence and theoretical arguments from econometrics and forecasting suggest that simple and transparent models generally match or outperform complex models at forecasting, especially when the underlying data is sparse (Green and Armstrong 2015; Forster and Sober 1994; Stock and Watson 2002; Makridakis and Hibon 2000; Brighton and Gigerenzer 2015; Morgan and Henrion 2012, Ch. 11; Tetlock and Gardner Superforecasting 2015, Ch. 5). In the literature on quantitative risk analysis, the aim has been characterized as making a model that is "as simple as possible, but no simpler" (Morgan and Henrion 2012, p. 38).[12]

Simple models have several advantages over complex models. Simple and transparent models make it easier to see which assumptions are driving results and allow people to more easily test the results of different assumptions. Secondly, the more parameters a model has, the more likely it is that the parameters will be correlated, amplifying model error. Finally, and related to the previous point, a more complex model affords a researcher greater degrees of freedom, introducing a greater number of choices that are difficult to scrutinize and can end up having excessive influence on the results.[13]

An Outline of Our Risk Model for Data-Damaging Worms

We now describe our risk model for data-damaging worms. We break down the overall risk into the risk posed by each of the five classes of threat actor; and, in turn, we break down the risk from each threat actor class into three parameters:

- The **capabilities** of the threat actor class: for each threat actor class, the probability over the next year that a randomly selected threat actor in that class can launch data-damaging

[12] Of course, this does not mean that simple models are always preferable to complex models.
[13] This is a major contributory cause of the replication crisis in science (Simmons et al. 2011; Gelman and Loken 2013).

worms if they wanted to, assuming no further improvements in AI. Note that we assume for ease of analysis that all threat actors in the same class have the same capability level. Thus, this is equivalent to asking whether the whole class is capable of doing the relevant task.

- How **willing** threat actors are to launch data-damaging worm attacks: the probability over the next year that at least one threat actor in each class would be willing to launch a data-damaging worm attack if they were able.[14]
- The **damages** from such attacks: how much economic harm such attacks would do.

The baseline risk of one data-damaging worm attack from each threat actor class is given by:

> **Baseline expected damage from the first large data-damaging worm attack from threat actor$_i$** = Capability$_i$ * Willingness$_i$ * Damages

Our empirical research identified elite exploit development as a key bottleneck to data-damaging worms. To estimate the marginal risk posed by future AI systems, we estimate marginal risk conditional on the following concrete hypothetical cyber evaluation result, which we call "Elite Exploit Uplift":

> **Elite Exploit Uplift:** A study conducted at the end of 2025 finds that access to frontier AI models enables 25% of TA2 actors to find vulnerabilities and write elite exploits, assuming three months of full-time effort.

The marginal damages from the first large data-damaging worm attack from each threat actor class is given by:

> **Marginal expected damage from the first data-damaging worm attack using an elite exploit from threat actor$_i$ conditional on Elite Exploit Uplift assuming no benefits to defense** = ((Capability$_i$ | Elite Exploit Uplift) * Willingness$_i$ * Damages) - Baseline Damages$_i$

The capability for a given threat actor to launch a data-damaging worm is now conditioned on the evaluation results described by Elite Exploit Uplift. To capture the marginal damages caused by Elite Exploit Uplift, we subtract baseline damages.

We can then calculate the overall expected damage from the first major data-damaging worm attack. The main reason we focus on the risk of the first major worm attack rather than the risk from all worm attacks is that the expected impact of subsequent attacks would be affected by the first attack. There could be effects in both directions. The first major attack could serve as a warning shot causing an improvement in cybersecurity.[15] Alternatively, a first major attack might lead to subsequent copycat attacks. These sorts of effects are difficult to model, so we bracket them by

[14] This parameter is therefore sensitive to the number of threat actors in a given category: the ex ante probability that at least one threat actor in each category would be willing to launch an attack probably increases with the number of actors in that category. In retrospect, for this reason, we think it might have been preferable to construct a risk model that includes a separate parameter for the number of actors in each class.

[15] There is some evidence that past major worm attacks have led to improved cybersecurity in the medium-term. See sections 2 and 3.

focusing on the first major attack. Moreover, frontier AI safety policies are designed to manage sudden catastrophic risks through pre-deployment risk mitigation, so it is useful to understand the risk posed by large events that could occur prior to large social response.

Let p_i = Capability$_i$ * Willingness$_i$ be the probability that threat actor class i carries out a major worm attack in a given year. To estimate the overall baseline expected damages from the first data-damaging worm attack, we calculate:

$$P(\text{there is an attack}) = 1 - \prod_i (1 - p_i).$$

This formula expresses that for each threat actor class, $(1 - p_i)$ is the chance that class does not attack this year. Multiplying these together (the $\prod$ symbol) gives the chance that none of them attack, and subtracting that from 1 leaves the chance that at least one of them does. We combine the probabilities this way rather than simply adding them because adding would double-count the cases where more than one class attacks.

The expected damage from the first attack is then given by:

$$\text{Expected damages of first data-damaging worm attack} = (1 - \prod_i (1 - p_i)) \times \text{Damages}.$$

The expected damage is how likely an attack is multiplied by how bad it would be. For example, if an attack is 20% likely in a year and would cause $10B in damage, the expected damage that year is 0.2 × $10B = $2B. It is a probability-weighted average of the harm.

Then

$$\text{Marginal expected damages from the first data-damaging worm attack} = (1 - \prod_i (1 - p_i)) \times \text{Damages} - \text{Baseline damages}$$

The marginal expected damages estimate isolates the extra risk attributable to the new AI capability. We take the total expected damage in a world with that capability and subtract the baseline expected damage (the risk that already existed without it). What remains is the additional harm the capability is responsible for, rather than harm that would have occurred anyway.

Benefits to Defense

Estimating the marginal risk posed by AI vulnerability discovery capabilities is challenging because it is a dual-use capability that would benefit both cyberattackers and defenders. Future AI systems could enable attackers to find and exploit vulnerabilities, but could also enable defenders to find and patch vulnerabilities before they are exploited. However, the net effect on offense-defense balance over time is contentious, uncertain, and difficult to model (Lohn 2025).

As we discuss in section 5, the effect of AI capability improvements on the risk of data-damaging worm attacks is uncertain and likely changes over time.

1.5.5. Survey of Experts and Superforecasters

In order to gather information from a wider range of perspectives, we developed a survey of experts and superforecasters in partnership with the Forecasting Research Institute. Numerous studies have established that the aggregated or median quantitative estimates of a group of individuals is consistently more accurate than any individual's estimates in a variety of domains, including prediction markets, political polls, and forecasting (Atanasov et al. 2016; Herzog and Hertwig 2009; Davis-Stober et al. 2014). This is due to "wisdom of the crowd" effects reducing the bias from any individual forecaster (Surowiecki, The Wisdom of Crowds, 2004).[16] Although individuals' estimates may be error-prone, aggregating them boosts accuracy because both systematic and random errors tend to cancel out across individuals.

The process for the survey was as follows. Forecasts can be highly sensitive to the precise wording of a question and its resolution criteria. So, to develop the survey questions, we undertook an iterative process in which we developed initial versions of the question, a small sample of experts and superforecasters answered the questions, and we then revised the questions in light of how they were interpreted by participants. We conducted two rounds of this process. We also provided participants with some background information on the data-damaging worm threat model – a 16-page abridged earlier version of this report, which excluded our own probability estimates and rationales to avoid anchoring.

We invited two groups of respondents:

1. People with expertise in cybersecurity and AI impacts on cybersecurity (henceforth, "experts"). As this was a pilot study, we used a convenience sampling method: Participants were recruited on the basis of being available and relatively easy to access.[17]

2. High performing forecasters or "superforecasters" – people who have previously scored highly in geopolitical forecasting tournaments.

A total sample of 33 people was invited to participate via email. To incentivize engagement, we paid participants for their time spent completing the survey – on average each participant received roughly $700 for their participation. We conducted two waves of the main pilot survey,[18] in part to refine questions, and in part to allow respondents to update their estimates following a report by Anthropic on cyber misuse of their models (Anthropic 2025). In the first wave, 13 superforecasters and 8 experts responded, while in the second wave 13 superforecasters responded but only 4 experts responded.

In the first wave, we surveyed respondents on risk model parameters and allowed them to estimate risk directly. Specifically, with respect to the risk model, we surveyed respondents on:

[16] Though note that a recent study suggested that prediction markets' accuracy is largely driven by a small number of skilled traders (Gomez Cram et al. 2026).
[17] In convenience sampling, participants are recruited on the basis of being available and relatively easy to access.
[18] In the first wave, respondents replied between July and August 2025. In the second wave, respondents replied in December 2025 and January 2026.

- Estimates of threat actor capability and willingness over a three-month timeframe in the baseline and AI uplift scenarios.

As discussed below, we extended the timeframe in the second wave.

For the purpose of the risk model estimate, we did not survey respondents directly on the damages parameter, but instead used author estimates for this parameter primarily because knowledge of this parameter is less reliant on domain-specific cyber knowledge, and more on economics.

In addition to gathering information on the model parameters, we also asked respondents to directly estimate:

- The probability of at least one data-damaging worm attack causing >$10B in economic damage.
- The expected damages from data-damaging worms in the baseline and AI uplift scenarios.

We also asked respondents how different approaches to risk mitigation affected their overall risk estimates. Specifically, we surveyed respondents on the effects of:

- Releasing models open weight.
- Imposing deployment safeguards on cyber capabilities.
- Giving defenders early access to models.

In the second wave, which, as noted, had a lower response rate from experts, we surveyed respondents only on capability and willingness estimates, but extended the relevant timeframe to 12 months in order to better reflect the annual risk we were trying to estimate.

Survey Limitations

This survey has several important limitations. First, our sample size for this pilot study was small. In the first wave, 21 participants completed the survey, including 13 superforecasters, 6 experts who completed the full survey, and 2 experts who completed shorter versions of the survey. In the second wave, 13 superforecasters but only 4 experts completed the survey. This makes the results sensitive to individual forecasts and the reported aggregate statistics fragile.

Second, our convenience sample of experts may be biased in some respects and should not be taken as representative of cybersecurity experts. Moreover, some of the experts who participated in the survey also provided feedback on earlier drafts of this report, which may introduce bias. These problems are not a concern for the recruitment of superforecasters.

Third, the questions in this study differed from typical forecasting exercises, as they focused on hypothetical AI evaluations and low probability extreme events without clear resolvability. Many were conditional questions, meaning forecasters won't be scored on the accuracy of their predictions. This absence of performance incentives or feedback, combined with the unusual nature

of some questions and the lack of past data to guide judgments for certain scenarios, could impact how accurate and useful the forecasts are.

Fourth, we did not share the full report, including our own quantitative estimates and reasoning, with survey respondents. This has the advantage of avoiding anchoring respondents. The main disadvantage is that sharing the full report would have shared more information with the respondents, and, if there were disagreement, it would have been easier to understand why the respondents did not agree with our estimates. On balance, we think that for future work, it would be preferable to share the full report, including our own estimates, with respondents.

Finally, there was also no chance for respondents to update their estimates following discussion with each other or with the authors.[19]

Despite these limitations, the survey provides estimates from a wider range of perspectives than our own estimates alone. A larger and more systematic version of the survey would address the limitations mentioned above.

[19] A future iteration could adopt a Delphi method or similar process, in which respondents answer in multiple rounds and revise their estimates after seeing anonymized feedback and reasoning from others (Dalkey & Helmer 1963).

2. How Willing Are Threat Actors to Release Data-Damaging Worms?

In this section, we discuss the willingness of threat actors to release data-damaging worms if they were able to. We first discuss the evidence on the base rates of past worm attacks. We then discuss the motivations and incentives for worm attacks. We conclude by discussing the author and survey respondent estimates of actor willingness to launch data-damaging worms.

2.1. Background on Past Worm Attacks

According to the best available data on past worm attacks (Johansmeyer 2024), the number of significant such attacks has declined sharply since 2005. Prior to 2005, lower-skilled attackers launched numerous such attacks, but the few worm attacks since then were carried out by more sophisticated actors.

Figure 2.1 below shows significant worm attacks by threat actor class, using the Johansmeyer dataset.[20]

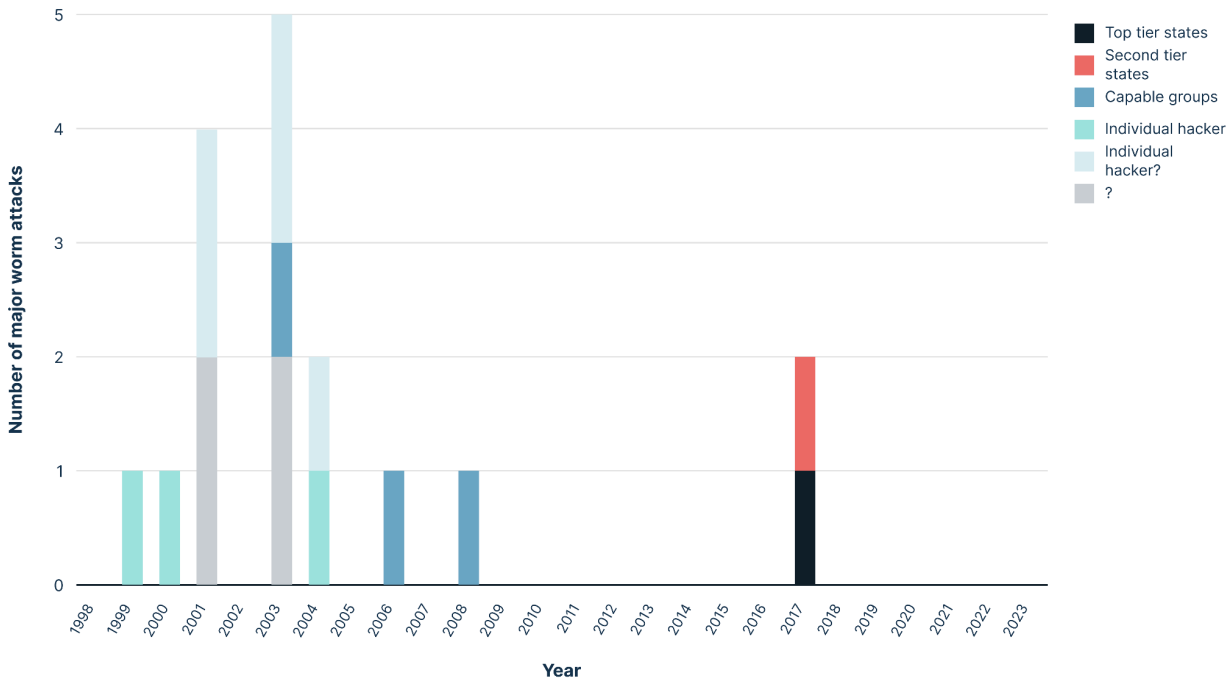


Figure 2.1. Major worm attacks by threat actor (1998-2023) [Adapted from Johansmeyer, 2024] ('?' means it is unclear which threat actor carried out the attack)

[20] Note that the threat actor categorization refers to the overall capabilities of the threat actor who carried out the attack, not to the operational capacity required to execute that specific attack. For example, the attacks in 2017 were not OC4 or OC5 *operations*, but they were carried out by TA4 and TA5 *threat actors*.

Table 2.1 shows details on the worm attacks in the Johansmeyer dataset.[21]

Attack name	Year	Actor	Description	# systems infected	Elite exploit?	Data-damaging worm?
Melissa	1999	TA1/2	Email worm. Damage via large volumes of network traffic. No other destructive payload.	~100K	n	n
ILOVEYOU	2000	TA1/2	Email worm. Corrupted documents. Later variant wiped hard drive.	~50M	n	y
Klez	2001	TA1/2?	Email worm. Damage via high network traffic and disabling antivirus.	~7M	n	n
CodeRed	2001	?	Zero-click. Defaced specific websites and launched targeted denial of service attacks against sites.	~360K	n	n
Nimda	2001	?	Zero-click and email spread. Damage via large volumes of network traffic and elevated privileges.	>1.3M	y	n
SirCam	2001	TA1/2?	Email worm. Could expose confidential info and delete all files under certain conditions.	~2.3M	n	y
SoBig	2003	TA1/2?	Email worm to distribute spam. Also caused damage by creating large volumes of network traffic.	>1M	n	n
SQL Slammer	2003	TA1/2?	Zero-click but limited reach. Damage via large volumes of network traffic. No malicious payload.	>75K	n	n
Swen	2003	?	Email worm. Disabled antivirus and firewalls. No other destructive payload.	~1.5M	n	n
Mimail	2003	?	Email worm. Launched targeted denial of service attacks against anti-spam sites. Some variants stole credit card information.	~21K	n	n
Yaha	2003	TA3	Email worm. Terminated security process and launched denial of service attacks.	?	n	n
MyDoom	2004	TA1/2?	Email worm. Created botnet to allow targeted denial of service attacks. Created high network traffic.	~500K	n	n
Sasser	2004	TA1/2	Zero-click. Damage only via large volumes of network traffic, no destructive payload.	~500K–1M	y	n
Storm Worm	2007	TA3	Spread via email and social engineering. Created denial of service botnet to attack anti-spam websites and security vendors.	1M–50M	n	n

[21] For more detail on the specific worm attacks and sources, see Appendix A.6.

Conficker	2008	TA3	Zero-click. Created large botnet but never used for significant attack due to concern about criminal repercussions.	~10M	y	n
WannaCry	2017	TA4	Zero-click. Encrypted files. Ransomware apparently to raise money for North Korea.	~230K	y	y
NotPetya	2017	TA5	Zero-click. Encrypted files. Designed to limit damage to Ukraine.	~670K	y	y

Table 2.1. Details on past major worm attacks from Johansmeyer (2024)

Johansmeyer's inclusion criteria are that the worm must have affected what Johansmeyer calls a "significant" number of companies (this is not precisely defined and involves some discretion, but >10 victim companies would plausibly count as significant, and >25 definitely would),[22] and there must be some public source claiming that it caused damages of >$800M. 17 of the 21 attacks in Johansmeyer's dataset were worm attacks,[23] while 4 of the 17 worm attacks were data-damaging worms.[24]

We (along with Johansmeyer) are skeptical of the damage estimates in the Johansmeyer (2024) dataset. The dataset suggests that some worm attacks prior to 2010 caused very large economic damages. For example, Johansmeyer's dataset suggests that SoBig and MyDoom caused damages of $65B and $67B, respectively. We think these are substantial overestimates. As Johansmeyer notes, the provenance of this data is poor: The damage estimates in the dataset are impossible to verify and include no methodological information (Johansmeyer nd).[25] Many of the sources are public media sites, corporate blogs, and consultancies, which some have argued had incentives to produce inflated estimates (Leyden, 2002; Gallaher et al. 2006).

Nonetheless, we think Johansmeyer (2024) provides useful information on trends in the volume of major worm attacks and the willingness of lower-skilled actors to launch worms. Although there is reason to think that the absolute damage estimates used in the original sources are inflated, we have not found a reason to think that size or direction of this bias would change over time. Thus, the dataset provides weak evidence that the downward trend in worm attacks is real.

After reviewing the Johansmeyer dataset in more detail, while we could not find robust and credible damage estimates, there is independent evidence that many of them infected a large number of systems (as shown in Table 2.1) and so had the potential to cause significant economic damage.

Moreover, there is evidence that launching worms became harder due to improved cybersecurity. From 2003 onwards, major vendors such as Microsoft improved cybersecurity explicitly in response to the large number of worm attacks.

- **Email worms became harder due to email filtering:** As shown in Table 2.1, the vast majority of early worm attacks were email worms. In 2003, Microsoft introduced SmartScreen email and spam filtering technology (Microsoft 2016), which filtered >90% of unwanted spam as of 2004 (Microsoft 2004). Google (Kumaran 2022) claims its filters catch >99.9% of spam.
- **Improved patching:** Microsoft introduced regular patching in 2003, explicitly in response to major worm attacks (Fiscutean 2023). In 2004, Microsoft made "opt-in" patching the default for Windows XP, increasing the share of up-to-date users from 5% to 90% (Jenkins et al 2020, p. 4; The Register, 2005).

[22] Tom Johansmeyer, personal communication, 16 Jan 2025.

[23] See the "Johansmeyer 2024" tab of Worm damages.

[24] Note that Johansmeyer (2024) includes the 2010 Stuxnet attack, citing damages of $2.9B, though we have been unable to find the source for this estimate. See the "Johansmeyer (2024)" tab of Worm damages. We exclude Stuxnet because we do not think it meets Johansmeyer's >$800M damage inclusion criterion. We discuss the damages of Stuxnet in section 1.3.2.

[25] We discuss this in more detail in section 4 and Appendix A.7.

Further support for the increasing difficulty over time of creating damaging worms is provided by Calleja et al (2018), which quantifies the person-years of effort required to make various kinds of malware between 1975 to 2018. Between 2000 and 2018, the average person-months required to develop the malware in their sample increased from around 1 month to 10 months, reflecting the increasing complexity of malware (Calleja et al 2018, Fig 5a). Calleja et al (2018) also estimate the person-months for some of the worm attacks in the Johansmeyer dataset: ILOVEYOU (2 person-weeks); Sasser (2 person-months), and MyDoom (8 person-months) (Calleja et al 2018, Fig 5d). This is much lower than our own estimate of person-time involved in creating the different elements of WannaCry and NotPetya, which likely required at the very least one person-year of state-level skill.

Overall, despite the flaws in the Johansmeyer (2024) dataset, we think it is reasonable to conclude from it and other evidence that:

1. Many lower-skilled threat actors launched worm attacks that infected a large number of systems prior to 2005.
2. Lower skilled threat actors now lack the ability to do so, due to improved cybersecurity.
3. The decline in worm attacks in Figure 2.1 reflects a real trend.

Since 2009, the only worm attacks in the Johansmeyer dataset were WannaCry and NotPetya, which were launched shortly after the public leak of what we call “elite exploits” developed by nation state actors. We discuss the evidence that elite exploits are a potentially important bottleneck for worm attacks in section 3.

2.2. Motivations for Historical Worm Attacks

In this section, we estimate the probability that different threat actors would be willing to launch data-damaging worms if they were able. To that end, in this subsection, we will discuss the motivations for past significant worm attacks in the Johansmeyer (2024) dataset.

2.2.1. Distinguishing Intended and Foreseen Harms

When analyzing the motivations for these attacks, it is important to distinguish:

1. Their ultimate intended aims.
2. Side-effects foreseen by the attacker prior to the attack.
3. Their unintended and unforeseen side-effects (i.e. accidental harms).

For example, the ultimate intended aim of the WannaCry attack was apparently to raise money for North Korea;[26] a foreseen side-effect of the attack was that it would cause severe economic harm.[27] The ultimate intended aim of the NotPetya attack was to damage Ukraine; an unforeseen side-effect was causing harm to companies in Russia.

Because worms are difficult to control, many worm attacks have caused significant accidental damage. It is notable that a substantial fraction of the economic damage from NotPetya, perhaps the most economically damaging worm attack ever, may have been accidental.[28]

2.2.2. Summarizing the Motivations for Past Worm Attacks

Based on our analysis of past significant worm attacks, the intended aims of past attacks can be divided into the following categories:

1. **Financial:** Attackers release worms to acquire money.
2. **Broad destruction:** Attackers aim to cause widespread destruction.
3. **Targeted destruction:** Attackers aim to cause damage to narrow targets, such as specific countries or companies.
4. **Other:** Some attackers seem to have been motivated by curiosity or to prove hacking skill.

Table 2.2 below shows the perpetrator and motivation for significant worm attacks in the Johansmeyer (2024) dataset. See Appendix A.6 for more detailed discussion and sources.

Event	Year	Actor	Ultimate intended aim	Data-damaging worm?
Melissa	1999	TA1/2	Accidental/broad destruction?	n
ILOVEYOU	2000	TA1/2	Broad destruction/to prove hacking skill?	y
Klez	2001	TA1/2?	?	n
CodeRed	2001	?	Targeted/broad destruction?	n
Nimda	2001	?	?	n
SirCam	2001	TA1/2?	?	y
SoBig	2003	TA1/2?	Financial?	n
SQL Slammer	2003	TA1/2?	?	n

[26] Marcus Hutchins, a cybersecurity researcher who discovered the kill switch for WannaCry (see section 4), has posited, on the basis of flaws in the WannaCry code, that WannaCry accidentally leaked earlier than intended (Darknet Diaries 2025).
[27] Some people interpret foreseen side-effects of actions as intended effects; people's intuitions about the meaning of "intention" differ (Wagner 2014). We distinguish them here for conceptual clarity.
[28] We discuss this in more detail in section 4.

Swen	2003	?	?	n
Mimail	2003	?	Financial	n
Yaha	2003	TA3	Targeted destruction	n
MyDoom	2004	TA1/2?	Targeted destruction/financial?	n
Sasser	2004	TA1/2	Broad destruction/to prove hacking skill?	n
Storm Worm	2007	TA3	Financial	n
Conficker	2008	TA3	Financial	n
WannaCry	2017	TA4	Financial (though perhaps released early accidentally)	y
NotPetya	2017	TA5	Targeted destruction	y

Table 2.2. Motivations for and perpetrators of past major worm attacks (1998–2023)[29]

As Table 2.2 shows:

- **Financial motivations:** A quarter of attacks were clearly financially motivated and a further 12% may have been financially motivated.
- **Broad destruction:** A quarter may have been launched to cause broad destruction.
- **Targeted destruction:** 12% clearly aimed to cause limited destruction and a further 12% may have.
- **To prove hacking skill:** 12% of attacks seem to have been released to demonstrate the attackers' hacking skill.
- **Unknown motivations:** 30%.

For the attacks with unknown motivations, it seems unlikely that they were financially motivated, as there is no obvious way they could have enabled the attackers to acquire money. So it is reasonable to assume that these were either released to cause broad destruction or to prove hacking skill.

Although most past worms mostly did not destroy data, they illustrate that when barriers to launching worms were lower, a range of actors were willing to release self-replicating malware to cause widespread disruption. This helps inform how likely threat actors would be to launch damaging worms in the future, with the potential help of AI.

2.3. Incentives and Disincentives to Release Data-Damaging Worms

Evidence on the motivations for past worm attacks is sparse, so it is also useful to consider, at a more abstract level, the incentives for and against releasing worm attacks.

[29] "?" means that the aim of the attack is uncertain.

2.3.1. Destructive Motivations

In Table 2.3, we summarize the incentives for and against different threat actors releasing worms purely to cause destruction.

	Incentives for	Incentives against
Non-state actors (TA1–3)	**Political ideology** seems to motivate many non-state attackers. **To demonstrate their hacking skill**, many attackers launch attacks which cause significant disruption (see Appendix A.6).	**Criminal repercussions**. Attackers have often faced criminal repercussions for worm attacks. **Political repercussions.** Governments often sanction non-state actors for cyberattacks (TRMLabs 2024).
State actors (TA4–5)	**Destructive worms may be more attractive to rogue states** who are less concerned about political and diplomatic repercussions. **Destructive worms may be more attractive during active conflict.** Even though states usually lack the incentive to launch worm attacks, this may not be true during active conflict. Moreover, worms can be especially useful cyberweapons for states aiming to cause chaos and destruction in a target country.	**Political repercussions for attackers**. Because worms can cause large amounts of damage, there is a risk of political and diplomatic repercussions. **Risk of blowback to attackers and their allies.** Worms are difficult to control, and there is a risk that the worm also damages the state that launched the attack and its allies. **Destructive worms seem less attractive outside active conflict.** Most states rarely aim to cause as much damage to the world as possible via any means, including cyberattacks. **Opportunity cost.** Launching destructive worms requires high-end cyber capabilities, which can be used for other offensive cyber operations, such as espionage. State espionage is also *de facto* tolerated under current international norms, whereas destructive cyberattacks are not. **Cyberweapons may be less effective than other methods.** Even if states do wish to cause destruction, cyberweapons are often less effective than e.g. kinetic attacks (Lin 2022; Bateman 2022)

Table 2.3. Incentives for and against releasing worms for destructive reasons

It seems likely that at least some TA1/2 actors would release data-damaging worms if they could. For many of these actors, political motivations, a desire to demonstrate technical skill, or simply to cause a large amount of damage can outweigh the potential criminal repercussions.

Historically, TA3 group actors have launched worm attacks for limited destruction, but there are no known cases of these actors using worms where broad destruction was intended or foreseen. This is likely in part because of the risk of criminal repercussions. However, actors include some terrorist organizations, which may be willing to release data-damaging worms in the future

irrespective of the criminal repercussions, though there are no known cases so far of terrorist groups launching significant data-damaging worms.

TA4/5 actors usually lack an incentive to launch worm attacks purely to cause broad global destruction, and there are no cases in the historical record. It may be that states are more likely to use data-damaging worms during active military conflict. NotPetya aimed to cause limited destruction and was launched during active military conflict between Russia and Ukraine. However, Russia has not deployed worms since then (Bateman 2022a, p. 21), and we are not aware of any other state having done so. This may reflect a reduced willingness to use worms due to the potential backlash from their collateral effects on non-combatant countries (Bateman et al 2022b).[30]

As noted above, even if states do not intend to cause significant destruction, they may still have incentives to create worms that unintentionally cause significant collateral damage.

2.3.2. Financial Motivations

Financially motivated significant worm attacks have usually been launched by non-state actors – with North Korea's WannaCry the only exception. Table 2.4 summarizes the incentives for and against using worms for financial reasons.

	Incentives for	Incentives against
Non-state actors (TA1–3)	**Worms allow larger reach than targeted approaches**, potentially increasing revenue, all else equal.	**International pressure.** Worms that spread to a large number of computers and cause a large amount of collateral damage are likely to bring greater criminal repercussions for the attacker, compared to targeted approaches, due to international diplomatic pressure. **Domestic pressure.** There is a risk that the worm could unintentionally infect systems in a state that usually tolerates the attacker. This would increase the risk that the harboring state punishes the attacker. **Higher negotiation costs.** Trying to extract a ransom from a large number of actors hit by a worm, rather than a smaller number of higher value targets, involves higher negotiation costs for the attacker.[31]
State actors (TA4/5)	**Cybercrime may be attractive to states that are heavily sanctioned.** The most	**Most governments do not need to use cybercrime to raise money,** as they can

[30] It could also be explained by a lack of technical capacity, or the opportunity cost of using elite exploits in worms rather than for espionage

[31] Ransomware attacks often involve extensive negotiation with the victim (The Economist, 2024).

	obvious example is North Korea, which has raised billions of dollars via cybercrime (Lyngaas 2023), though notably little from WannaCry.	raise money via taxes. For most governments, cybercrime is not worth the political or diplomatic costs. **Worm attacks risk blowback to the attacker state or to their allies.** North Korea has limited internet-facing infrastructure, so this is less of a concern for them, but worms could still hit their allies, such as China.[32]

Table 2.4. Incentives for and against releasing worms for financial reasons

Non-state actors face a trade-off between the greater reach (and greater damage) of a worm and the risk of repercussions. This may explain why there are no past cases in which non-state cybercrime groups have launched worms designed to infect and damage as many systems as possible. For example, the Conficker worm unintentionally spread so far that it drew significant attention from the global cybersecurity community and therefore became "too hot to use" by the Ukrainian cybercrime group that launched it (Bowden 2019). The Russian state rarely prosecutes cybercriminals (Maurer 2018) but did arrest the DarkSide hacking group that carried out the Colonial Pipeline attack after it caused a diplomatic incident with the US (Miller 2022) (note that this was not a worm attack). The BlackMatter ransomware group explicitly avoids targeting certain sectors for this reason (Smilyanets 2021).

By far the most concerning financially motivated state actor is North Korea. Due to sanctions, they are especially willing to try to raise money via cybercrime and are less concerned about political or diplomatic repercussions (UN, 2024). WannaCry was apparently released by North Korea for financial reasons, but due to errors in the code and design of the bitcoin payment system, the bitcoin addresses associated with the worm only ever received $250,000 (Elliptic 2017), and it is unclear whether North Korea cashed out any of the funds (Vanderburg 2017; Coburn et al 2019, p. 44).

Since WannaCry, North Korea has not released a worm with similar scale, though this may also reflect limited capability (which we discuss in section 3). WannaCry would have made more money if North Korea had fixed the errors in its code. Nonetheless, it is notable that North Korea has apparently had much more success with targeted attacks (Forni 2020). For example, North Korea stole $81M in a cyberattack on Bangladesh's central bank (DOJ 2018) and recently stole up to $1.5B from a cryptocurrency exchange (McCurry 2025), though it is unclear how much of the stolen funds they will be able to capture (Darknet Diaries 2020).

[32] China provides substantial aid to North Korea and accounts for most of its trade (Bernal 2024).

2.4. Estimating Actor Willingness to Launch Data-Damaging Worm Attacks

We now discuss willingness estimates for different threat actors. To be precise, these are estimates of the probability that at least one threat actor in each threat actor class would be willing to launch a data-damaging attack if they could.

The figures and table in the remainder of this section show in detail the author, expert, and superforecaster estimates. In summary:

- The authors, experts, and superforecasters generally agree that willingness to launch data-damaging worm attacks declines with threat actor sophistication. The authors, median experts, and median superforecaster each estimated that TA1 willingness is >89%, but is 1% to 35% for TA5 actors.
- The authors, the median expert, and median superforecaster broadly agree about the willingness of TA1 and TA2 actors to launch data-damaging worms, though there is much more disagreement about TA3, TA4 and TA5 threat actors, with the authors generally giving lower estimates than both experts and superforecasters.
- There was relatively high disagreement between different experts, and between different superforecasters for most or all threat actor estimates. For both experts and superforecasters, across the full sample, probability estimates for all threat actors ranged from close to 0% to more than 60%.

Figure 2.2 shows expert and superforecaster estimates of the probability that different threat actors would be willing to launch data-damaging worm attacks if they were able.[33]

[33] This was from the second wave survey in which 4 experts responded, and 13 superforecasters responded.

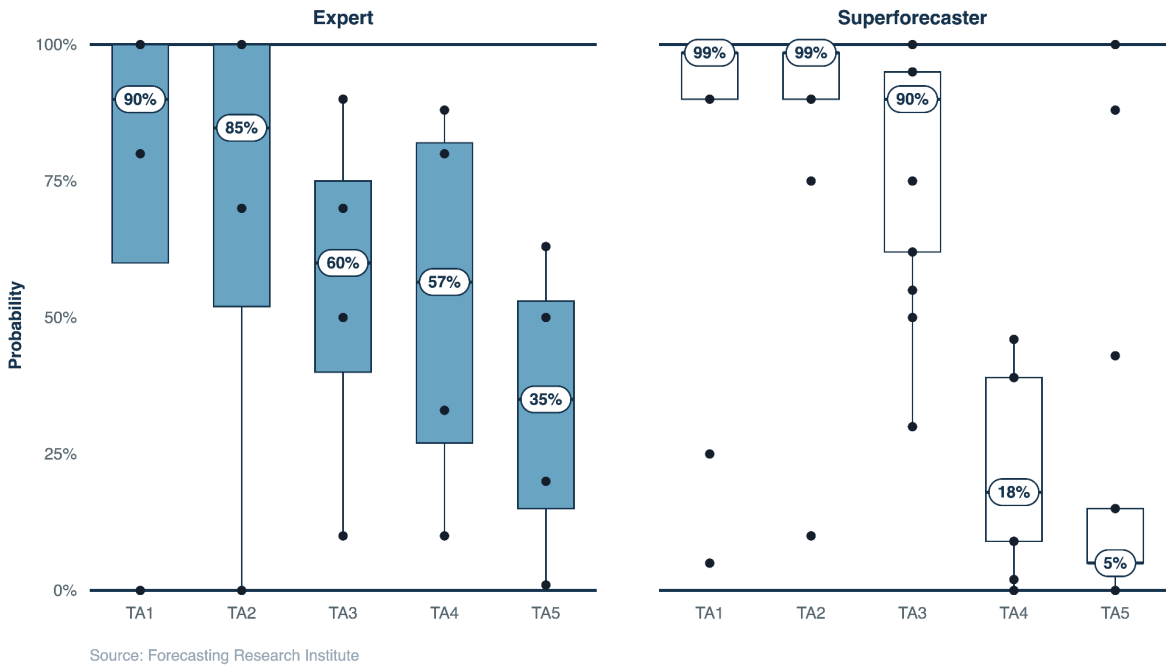


Figure 2.2. Expert and superforecaster estimates of the probability that at least one actor in each threat actor class would be willing, over a year, to launch a data-damaging worm attack if they were capable.[34]

Table 2.5 compares the median superforecaster and median expert estimate of willingness to the authors' median estimates:

Threat actor	Willingness		
	Author median	Experts median	Forecasters median
TA1: Individual hobbyist hacker	99%	90%	99%
TA2: Individual professional hacker	85%	85%	99%
TA3: Team of 10 experienced hackers	4%	60%	90%
TA4: Team of 100 state-level hackers	4%	57%	18%
TA5: Team of 1,000 state-level hackers	3%	35%	5%

Table 2.5. Probability that at least one actor in each threat actor class would be willing to launch a data-damaging worm attack if they were capable: author median and median of experts and median of superforecasters in the survey sample

Table 2.6 shows the author willingness estimates for different threat actors, with accompanying rationales.

[34] For the figures reporting survey results, the boxes represent the interquartile range, while the whiskers go out to the furthest "non-outlier" point, which is usually either the max or the largest value up to Q3+1.5*IQR (or min / lowest value up to Q1-1.5*IQR). Figures showing survey results were made by our collaborators, Bridget Williams and Rebecca Ceppas de Castro, at the Forecasting Research Institute.

Our uncertainty about these estimates varied depending on the threat actor:

- We have high confidence in the estimates for TA1 and TA2 actors.
    - Since there are so many TA1 and TA2 actors, it seems very likely that at least one threat actor would launch a data-damaging worm if they were able.
- We have low confidence in the estimates for TA3, TA4, and TA5 actors.
    - There is limited historical precedent of these threat actors launching data-damaging worm attacks, which means that the base rate is only loosely constrained by the data.
    - Other arguments depend on somewhat speculative considerations around the motivations of the relevant actors.

Threat Actor	**Rationale**	**Willingness (90% CI) Annual prob. any actor does attack if they could succeed**	**= an attack every x years (90% CI)**	**Confidence**
TA1	• It is difficult to know the skill level of perpetrators of past worm attacks – whether they were by TA1 or TA2 actors. From 1998-2005, there were 4-9 destructively motivated worm attacks by TA1/2 actors. This implies an attack every 0.7-1.5 years. Though the high-end estimate may include worms for which broad destruction was intended or foreseen. • Survey results suggest that there are ~1M TA1 actors. It seems extremely likely that from such a large sample, at least one would be willing if they were able.	Extremely likely (98–100%)	One attack every 1 years	High
TA2	• See TA1 considerations. • Survey results suggest that there are 10K to 100K TA2 actors. It seems very likely that from such a large sample, at least one would be willing if they were able.	Highly likely (70–100%)	One attack every 1 to 1.5 years	High
TA3	• In Nevo et al. (2024), examples of TA3 actors include cybercrime groups, industrial espionage organizations, and terrorist groups. Cybercrime and industrial espionage face strong	Highly unlikely (2–5%)	One attack every 20 to 50 years	Low

	disincentives to releasing a worm that could cause small country-scale or global harm. • Some TA3 cybercrime actors were very likely capable from 1998 to 2008, but there are no precedents for TA3 cybercrime groups launching worms for which country-scale or global destruction is foreseen. • TA3 terrorist groups might be willing. However, no major worm attacks have been launched by terrorist groups. For past attacks with an unknown perpetrator, terrorist groups would plausibly have claimed credit if they had launched them. We are unsure if in the relevant period, there were any TA3 terrorist actors with the capability. • There are plausibly one to two orders of magnitude more TA1 and TA2 actors than TA3 actors. If the distribution of motivations between these two sets of actors is the same, this implies a discount on TA1/2 risk of 10-100X, which implies an TA3 risk per year of roughly 2–17%. This may be biased high because the distribution of motivations might not be the same.			
TA4	• The most concerning actor is North Korea. It has already launched a data-damaging worm, and its limited internet-facing infrastructure reduces the risk of blowback. • One interpretation of the WannaCry case is that North Korea would launch more worm attacks if able to do so. • Assuming that North Korea was able to launch major worm attacks from 1998 to 2009 (since TA1–3 actors could in that period).[35] There were no TA4 attacks in that period. This implies an upper bound risk per year of 8%. • North Korea has had more success with targeted attacks, which may make them less likely to pursue worm attacks in the future.	Highly unlikely (2–5%)	One attack every 20 to 50 years	Low
TA5	• There is one historical precedent motivated by small country-scale harm:	Highly unlikely	One attack every 20	Low

[35] North Korea's primary cyberattack unit was founded in 1998 (Sanger and Fackler 2015).

	Russia's launch of NotPetya. One attack every 26 years implies a risk of 4% per year, assuming constant risk. ● Russia suffered significant damages from NotPetya, which may in part explain why it hasn't released worms since. ● Other TA5 states would face significant blowback and diplomatic repercussions. ● No cases of TA5 broad destructive attacks, despite likely capability through 1998–2023. This suggests at most an annual risk for such attacks of 4% per year. ● TA5 states seem to lack the strategic incentive outside wartime. ● Conditioning risk on war conditions on a rare event. Forecasting platforms suggest 0.4%–4% annual risk of war between TA5 actors. NotPetya is the only case of TA5 actors using destructive worms during war. Assuming risk of deployment conditional on war of 20%–50%, implies overall risk of 0.1%–2%.	(1%–5%)	to 100 years	

Table 2.6. Rationales for the authors' current estimates of the probability in a given year that at least one threat actor in each subset of threat actors would launch a data-damaging worm attack if they could[36]

[36] Detailed calculations are available in the 'Expected damages AI uplift' tab of Worm damages

3. How Capable Are Threat Actors of Releasing Data-Damaging Worms?

In this section, we discuss how capable different threat actor groups are of developing data-damaging worms. We first define elite exploits and introduce the hypothesis that elite exploits are a key bottleneck to data-damaging worms. We then review the evidence on this hypothesis, drawing on case studies of the WannaCry and NotPetya attacks. We then review other evidence on how difficult it is to develop elite exploits. We conclude by discussing author, expert, and superforecaster estimates of the capability of different threat actors to launch data-damaging worms.

3.1. Elite Exploit Development as Key Capability

3.1.1. Hypothesis

The hypothesis we explore in this section is as follows:

> If AI enables different threat actors to develop elite exploits (including both finding vulnerabilities and writing the exploits of them), then this would make it substantially easier for many threat actors to create data-damaging worms that could cause severe economic harm (>$1B in damage). Creating elite exploits is substantially harder than all of the other tasks involved in developing data-damaging worms combined. Thus, AI uplifting threat actors to develop elite exploits would substantially lower barriers to the creation of data-damaging worms for many threat actors.

Since, as we argue in section 2, many threat actors appear willing to release worms that cause severe economic harm, AI elite exploit development capabilities would therefore substantially increase the risk of data-damaging worms actually being released.

However, it is less clear whether elite exploits are necessary for data-damaging worms. It might be that data-damaging worm attacks that do not use elite exploits could cause comparably severe harm. Expert reviewers have mentioned several possibilities:

1. The combination of numerous weaker exploits of less widely used and less well-defended software in one worm. The ease of finding weaker exploits would trade off against the cost of finding a larger number of exploits and combining them in a single worm.

2. Worms that require some user interaction to spread but are still able to damage data on a large number of systems. As discussed in section 1, many non-zero-click worms released prior to 2008 nevertheless spread very widely. Although email worm attacks now seem less viable, other non-zero-click worms may still be able to spread widely. For example, watering

hole attacks that infect systems when users visit popular websites require users to take specific actions but may nevertheless spread widely.

3. Agentic polymorphic worms that can change their code after release in order to circumvent defenses, making it harder to stop further spread and damage.

These threat models may justify different cyber capability thresholds than the one we discuss here, and these other capability thresholds may be triggered earlier than elite exploit development. As such, these threat models warrant further research but are out of scope for this report.

In the next two sections, we examine the evidence for the hypothesis first by discussing how well-suited the features of elite exploits are to data damaging worms, and second by discussing case studies of the WannaCry and NotPetya attacks.

3.1.2. Features of Elite Exploits Seem Especially Well-Suited to Data-Damaging Worms

To recap, elite exploits are exploits that allow remote code execution with high privileges, that are effective against widely used software, and do not require actions by the victim in order to infect a system (i.e. "zero-click exploits") (see section 1.2). Exploits with these features are especially useful for data-damaging worms that can cause severe economic harm.

Exploits that require no user interaction enable autonomous spread. If a user needs to click a link, open an email, or visit a specific website in order to be infected, the worm is likely to spread more slowly and less widely because it relies on users taking actions that a substantial fraction are unlikely to take. Moreover, defenders can respond to warn users about which actions to avoid after the worm has been released.

Remote code execution allows the attacker to run their own code – including the worm's propagation mechanism and payload – on each newly infected system without any prior access. This is what enables a worm both to spread quickly and autonomously, as well as to damage data.

High privileges (i.e. admin or higher privileges) are important for (1) causing damage to infected systems and (2) for propagation.

In general, the higher the privileges malware has, the more damage it can do to infected systems (Devicie nd). If malware has privileges below the user level, e.g. corrupts a sandboxed application, there is much less scope to cause significant damage by encrypting important files. In order to cause significant damage, it seems plausible that malware must at least have user-level privileges, as this lets the worm access and encrypt user files (Engage Employee nd). However, malware with only user-level privileges might not be able to encrypt local or cloud backups. Typically, administrator or domain admin privileges are required to do this, and attackers often try to escalate privileges to administrator or higher (Lambert 2019; Abrams 2019).

For example, because NotPetya had system-level privileges on infected systems, it was able to encrypt and corrupt not only the data files but also the master boot record and master file table,

which made it very difficult or impossible to restore the affected systems to a usable state (Forbes 2017; Hasherezade 2017).

Moreover, the higher level of privileges a worm has, the easier propagation will be within local networks. Attackers often seek to gain administrator privileges or higher to do so (Microsoft 2022). Indeed, NotPetya spread in part by gaining administrative privileges on local networks (Lee 2021). In targeted ransomware incidents, attackers often try to gain domain administrative privileges in order to push ransomware to all endpoints or devices in a network (Microsoft 2022; Cocomazi and Pirozzi 2022). Gaining high privileges is also often useful for spreading between networks or between cloud environments (Doman n.d.; Itach and Morag 2023; Haber 2017).

We focus on exploits effective against widely used software because worms using such exploits would allow greater potential damage.

3.1.3. Case Studies: Elite Exploits Enabled the WannaCry and NotPetya Attacks

Since 2009, the only significant worm attacks were WannaCry and NotPetya in 2017, which were the product of an unusual natural experiment. In April 2017, elite exploits, including the NSA-developed EternalBlue exploit, leaked to the public. In the following two months, North Korea and Russia used these elite exploits in the WannaCry and NotPetya worms.[37]

The EternalBlue Exploit

EternalBlue exploits vulnerabilities in a network-facing protocol known as the Server Message Block v1 (SMBv1) protocol, which is a legacy file sharing protocol used in Windows systems up to 2017. The SMBv1 protocol exploited by EternalBlue was first developed in 1983 (Burdova 2020) and is now widely recognized as insecure (Microsoft 2025). On many systems vulnerable to EternalBlue, port 445 (over which SMBv1 communicates) was exposed to the open internet, which allowed worms using EternalBlue to propagate (Tyas Tunggal 2025).

EternalBlue was an unusually powerful tool for a worm attack (Microsoft 2017; SentinelOne 2019):

- Because it exploited a network-facing vulnerability, it could spread easily across exposed internal and external networks.
- It exploited these vulnerabilities without requiring prior authentication or privileges. EternalBlue was therefore also a zero-click exploit, infecting systems without user interaction.
- It enabled remote code execution on the infected system, allowing attackers to execute arbitrary code on the infected system and deploy further payloads to spread or inflict damage on the infected system.

[37] For more detail on how WannaCry and NotPetya worked, see Appendix A.2.

- It gained system-level privileges (the highest level of privileges on Windows systems), which meant that it had access to system-level files, including backups. This increased the potential damage a worm using EternalBlue could do.
- At the time, SMBv1 was a ubiquitous protocol on Windows systems. Prior to the release of the patch for the vulnerabilities exploited by EternalBlue in March 2017, ~400 million systems were vulnerable to it (Coburn et al 2019, p. 44).

The Leak of EternalBlue and Other Elite Exploits Quickly Led to the WannaCry and NotPetya Attacks

The NSA developed EternalBlue in 2012 or earlier, and used it in surveillance and counterterrorism operations for at least five years (Perlroth and Shane 2019). A hacking group known as the Shadow Brokers stole EternalBlue and a trove of other NSA hacking tools and between 2016 and 2017 sold or leaked these tools. This, in turn, led to EternalBlue being deployed in the WannaCry and NotPetya worm attacks.

The timeline for WannaCry and NotPetya, based on public reporting, is outlined below:

- **2012**: EternalBlue was first developed by the NSA by 2012 at the latest and used for targeted espionage (Perlroth and Shane 2019).
- **December 2016:** Sandworm starts work on the NotPetya worm (Maschmeyer 2021).
- **Pre-2017:** At some point prior to 2017, the Shadow Brokers stole various NSA hacking tools, including EternalBlue (Crocker and Budington 2016; Ewing 2017).
- **February 2017:** Russian hackers responsible for the NotPetya attack may have had access to the Shadow Brokers exploits (Goodin 2017c).
- **9 Feb 2017:** Security researchers discover a beta version of WannaCry but without a propagation method (Yang 2017), though it is unclear how long North Korea had been working on WannaCry by this point.
- **Early 2017:** The NSA informs Microsoft of the vulnerabilities possessed by the Shadow Brokers (Goodin 2017b).
- **14 March 2017:** Microsoft releases a patch for the vulnerabilities (Microsoft 2017), though many users delayed deploying or installing the patch.
- **14 April 2017:** The Shadow Brokers publicly leaks EternalBlue, along with other stolen NSA hacking tools (Goodin 2017a).
- **12 May 2017:** The WannaCry attack, using EternalBlue, launches (Newman 2017a). A cybersecurity researcher discovers a kill switch for the worm so that after seven hours, no new systems are encrypted and spread is slowed significantly (Kryptos Logic (2017); Malwaretech 2017; Lee et al. 2017).

- **27 June 2017:** The NotPetya attack, using EternalBlue and EternalRomance (another NSA elite exploit), launches (ESET 2017).

How Elite Exploits Were Used in the WannaCry and NotPetya Attacks

WannaCry and NotPetya used EternalBlue in different ways, with NotPetya designed to limit damage to Ukraine.

How WannaCry worked

1. **Initial infection**: The attacker used EternalBlue to infect vulnerable systems.
2. **Propagation to other systems:** A network scanner searched for other vulnerable machines on the local network and at random IP addresses on the internet (Nguyen et al. 2024; Microsoft 2017). Once a vulnerable system was found, WannaCry used EternalBlue again to infect the new target (TrendMicro 2017). The malware delivered its payload to the newly compromised machine, effectively turning it into another worm node. This self-replication process continued indefinitely, rapidly spreading the infection across systems.
3. **Installation of backdoor:** EternalBlue installed a backdoor called DoublePulsar (another stolen NSA hacking tool) on the infected system. This allowed persistent access, remote code execution, and system-level privileges (Microsoft 2017; Mimoso 2017).
4. **Installation of ransomware:** Using the DoublePulsar backdoor, the attacker installed the Wannacrypt ransomware (Root 2022a). It encrypted files on the infected system and displayed a message asking the victim to pay ransom in order to decrypt their data (Microsoft 2017).

How NotPetya worked

1. **Initial infection:** The Russian attackers compromised M.E.Doc accounting software that was used by 80% of domestic firms in Ukraine in a supply chain attack (Stubbs and Polityuk 2017; Goodin 2017d; Maynor et al. 2017).
2. **Propagation to other systems:** Unlike WannaCry, NotPetya did not spread between networks, but only within them (Maloney n.d.). Consequently, spread was limited to organizations that used the Ukrainian M.E.Doc accounting software.[38] NotPetya used EternalBlue and another NSA elite exploit called EternalRomance[39] to target systems that had not installed the patch released by Microsoft months earlier. NotPetya could infect patched systems using Mimikatz, an openly available tool that pulls passwords out of the memory of infected machines and uses them to hack into other machines.

[38] The worm was apparently not intended to spread beyond Ukraine, but it nevertheless did, as some companies had Ukrainian subsidiaries. Indeed, the worm unintentionally hit many Russian companies (Greenberg, *Sandworm*, p.198).
[39] EternalRomance is another exploit of Windows SMBv1 vulnerabilities, but targets older versions of Windows (Hurley and Sood 2017).

With those credentials, NotPetya used legitimate system tools to move laterally within networks (Gofman 2017; Greenberg 2017b; Greenberg 2018; Vijayan 2020). NotPetya spread outside Ukraine because some organizations – such as the shipping company Maersk – had computers using M.E.Doc in Ukraine, which were linked to networks outside Ukraine (Greenberg 2018).

3. **Installation of wiper:** NotPetya installed a wiper payload that encrypted files on infected systems. Although it displayed a ransomware message, this was likely a ruse to make the attack look like cybercrime. It was not possible to decrypt the data (Securelist 2017; Virsec 2017).

There are two potential interpretations of how elite exploits enabled the WannaCry and NotPetya attacks:

1. Russia and/or North Korea did not have elite exploits, so the Shadow Brokers leak overcame the main technical bottleneck for them.

2. Russia and/or North Korea were able to develop elite exploits in-house, but these were useful for other purposes (e.g. espionage), so the opportunity cost of burning them in worm attacks was high. But the opportunity cost of using the Shadow Brokers exploits was lower because they were leaked publicly and so would soon be patched anyway (indeed, the patch had already been released).

We are unsure which interpretation is more plausible. As we discuss below, in our view, Russia was probably able to develop elite exploits, but this seems less likely for North Korea. In any case, both interpretations suggest that elite exploits are in short supply even for TA4 and TA5 state actors.

Elite Exploits as the Primary Bottleneck for WannaCry and NotPetya

The tasks involved in creating WannaCry and NotPetya can be divided into: (1) elite exploit development and (2) the other tasks involved in creating the worms. We think it is plausible that developing EternalBlue was much harder than the other steps involved in creating WannaCry and NotPetya.

There is a lack of good information on the absolute difficulty of tasks (1) and (2). Regarding task (1), according to public reporting, the NSA analysts behind EternalBlue spent nearly a year finding the vulnerability and writing the code that exploited it (Perlroth and Shane 2019). Regarding task (2), wiper or ransomware payloads comparable to those used in WannaCry and NotPetya now typically sell for ~$1K (Venafi 2022). All the other software used in the attacks was free. Thus, the vast majority of the cost of developing the worms was from skilled labor. North Korean hackers started work on early versions of WannaCry at least three months prior to release. However, it is unclear how many people were involved in (1) and (2). The US has indicted three North Koreans for making WannaCry (DoJ 2021), though more may have been involved in development. The available timeline evidence weakly suggests that developing the elite exploits was harder than the other tasks involved in creating the worm, though the size of this gap is unclear.

Moreover, the history of major worm attacks discussed in section 2 provides further evidence that elite exploits were the primary bottleneck for the WannaCry and NotPetya attacks. In the Johansmeyer (2024) dataset, WannaCry and NotPetya were the only major worm attacks since 2009. Each of these were launched shortly after the unusual event of the leak of elite exploits developed by very high-skilled cyber researchers. In other periods when elite exploits were not leaked publicly, there were no significant data-damaging worm attacks. This is evidence that WannaCry and NotPetya were primarily constrained by access to elite exploits,[40] rather than the other tasks involved in creating the worms.

The release of the EternalRocks worm in 2017 provides additional evidence that elite exploits are the primary bottleneck for developers interested in creating data-damaging worms. Released one month after the Shadow Brokers leak, EternalRocks used seven leaked NSA hacking tools, compared to two used by WannaCry. Some observers argued that it was more complex than WannaCry, though less dangerous because it did not have a destructive payload (Cimpanu 2017a). The developer shut it down within a few weeks of release after the worm received media attention (Ashford 2017; Cimpanu 2017b). It is unclear how many people were involved in the development of EternalRocks, or what their skill level was (Ashford 2017). EternalRocks provides an example of a worm of similar complexity as WannaCry and NotPetya developed shortly after the leak of elite exploits. This provides further weak evidence that elite exploits, rather than secondary tasks related to worm development, represent the main bottleneck to highly destructive data-damaging worm attacks.

The Other Steps Involved in Making the Worms Still Seem Non-Trivial

Although creating elite exploits seems likely to be much harder than the other tasks involved in creating WannaCry and NotPetya, there is some evidence that they still require significant skill.

- **TA4 and TA5 state actors, rather than TA1/2 lone wolf hackers, carried out both WannaCry and NotPetya.** Though there are far more TA1/2 actors than TA4/5 actors and some TA1/2 actors would likely be willing to launch such an attack, the fact that TA4/5 actors were responsible for the attacks suggests that most or all non-state actors who would have released a worm as damaging as WannaCry (if they could) were unable to do so in the 1–2 month period after the Shadow Brokers exploits became available.
- **The WannaCry code contained numerous errors** (Goodin 2017c; Newman 2017a; Greenberg 2017a). Even though it was developed by state-level hackers, errors in the WannaCry code suggest that secondary tasks in worm development can be challenging for a TA4 group, such as the North Korean group responsible for WannaCry.[41] By contrast, experts suggest that NotPetya, which was the work of more skilled and well-resourced Russian hackers, was sophisticated and well-tested (Goodin 2017c; Weaver 2017).

This suggests that even given access to an elite exploit like EternalBlue, it was difficult for non-state actors to develop a worm using it before most systems were patched (which would have taken

[40] Even if they did have other elite exploits used for e.g. espionage.
[41] As noted in section 2, some have argued that the reason that it was released with so many errors, is that it was released accidentally (Darknet Diaries 2025).

several months). Thus, if AI only has strong elite exploit capabilities, but doesn't help actors write worms, the latter step could still be a meaningful bottleneck for lower skilled actors.

It is difficult to assess, based on the available evidence, the skill level required to complete the other tasks involved in developing a data-damaging worm. As noted above, North Korea was working on a version of WannaCry at least three months prior to release, though it is unclear how many hackers were involved in the effort. Table 3.2 estimates the time taken to develop WannaCry on the basis of different subjective assumptions. As this shows, it is difficult to estimate how much hacker person-time it took to develop WannaCry. On different intuitively reasonable assumptions, it could have taken 3 person-months, whereas on others it could have taken two person-years.

Variable	Value	Unit	Input type
Months to develop Wannacry (low)	3	Months	Subjective assumption
Months to develop Wannacry (mid)	4	Months	Subjective assumption
Months to develop Wannacry (high)	5	Months	Subjective assumption
Number of hackers (low)	1	# of hackers	Subjective assumption
Number of hackers (mid)	3	# of hackers	Subjective assumption
Number of hackers (high)	5	# of hackers	Subjective assumption
Hacker person-months to develop WannaCry (low)	3	person-months	Calculation
Hacker person-months to develop WannaCry (mid)	12	person-months	Calculation
Hacker person-months to develop WannaCry (high)	25	person-months	Calculation

Table 3.2. Person-time to develop WannaCry, on different subjective assumptions

On some of the assumptions outlined above, developing WannaCry could potentially be in reach for TA2 actors, whereas on others, it could not.

It is worth noting that an AI that is good at developing elite exploits would likely also be good at the other tasks because both writing exploits and writing worms may be downstream of the general coding abilities of AI models. Thus, if AI does uplift TA1-3 actors to find elite exploits, it might also uplift them to write data-damaging worms. Whether or not this is true has important implications for the expected damages implied by this threat model.

3.2. Assessing How Capable Different Threat Actors Are of Finding Elite Exploits

In the previous section, we argued that if actors had access to elite exploits, this would substantially lower the capability barrier to the creation of data-damaging worms that could cause severe economic harm. In this section, we present evidence on the probability that threat actors can already develop elite exploits (including finding vulnerabilities and writing the code to exploit them).

We first present a range of evidence suggesting that it is very hard to develop elite exploits and then discuss the current capabilities of different groups of threat actors.

3.2.1. It Is Very Hard to Develop Elite Exploits

In addition to the case studies discussed above, here we discuss other evidence on how hard elite exploits are to develop.

Elite Exploits Command High Prices, Indicating They Are Hard to Develop

There are various different markets for vulnerabilities exploits, including:

- **Bug bounties** offered by software developers for white hat hackers to find vulnerabilities in their systems.
- **Grey-market brokers** like Zerodium and Crowdfense buy the code for exploits from individuals, groups, or companies and then sell them to governments.
- **Commercial surveillance vendors** like NSO Group and Intellexa sell spyware using elite exploits to state intelligence agencies, who then use them for espionage and law enforcement purposes.

Table 3.3 summarizes the prices in these different markets.

Market	Est. Price ($M)
Bug bounties	$100K–$1M per zero-click remote code execution (RCE) exploit with kernel privileges for an individual device or system.[42] The price for an exploit effective against numerous different systems would be higher.
Grey market brokers	$1M–$7M per elite exploit. These prices are likely below true buyer willingness to pay because it is difficult for buyers to be sure of the quality of the exploit they are buying, as there is asymmetric information between. buyers and sellers.[43]

[42] Apple offers $100k to $1M for a zero-click RCE chain with full kernel execution and persistence (i.e. the exploit continues to work after the device has been rebooted) on a single recently released device. Google offers $1M for a zero-click RCE exploit with persistence that is effective against all vulnerable builds and models of Pixel Titan M. The price for vulnerabilities affecting a large number of distinct systems would be much higher. WannaCry was effective against ~400M systems at the time of the attack. This suggests that in today's prices, EternalBlue would be worth millions of dollars and plausibly on the order of $10M.

[43] Smeets (2022) argues that "The zero-day exploit market is a market with extreme information asymmetries. The seller has much more information about whether the exploit is actually working. The market is also flooded with lemons. Many of the exploits offered are a lot less reliable than sellers initially report. Also, the buyer of an exploit is not always able to test the exploit before purchasing it, as the economic value would be lost once given to the buyer for "testing." This structural setup makes even beneficial zero-day transactions difficult." If buyers are unable to tell the difference between strong and weak exploits, they would be unwilling to pay high prices. Consequently, the price will be lower than what sellers of high-quality exploits would sell for, driving them out of the market.

Commercial surveillance vendors	$100K–$800K per device (e.g. an individual's mobile phone) infected by a tool including zero or one-click exploit and spyware, i.e. the vendor would not sell the exploit code itself, but rather the use of a spyware tool using the exploit code.[44] The same tool would be sold to numerous customers and would be used by each customer to target numerous (10 or more) devices. This suggests that the market value of elite exploits and spyware would be at least millions of dollars, though it is hard to know how much of the value derives from the elite exploits vs. the spyware.

Table 3.3. Vulnerability and elite exploit prices in three different markets

As Table 3.3 shows, in various different markets, zero-click exploits that allow remote code execution with high privileges sell for hundreds of thousands to millions of dollars. Exploits with prices at the lower end of this range (in the hundreds of thousands of dollars) are typically effective against individual devices that would not have >10M users. For this reason, they would not be classed as elite exploits, per our definition. Exploits that are effective against >10M systems would be more expensive.

It is notable that prices for exploits sold on the gray market have been rising above inflation over time, suggesting that it is becoming harder to find elite exploits due to improved cybersecurity (Franceschi-Bicchierai 2024), though this may also be in part because it is becoming easier to extract more economic value from such exploits, as more economic activity is now dependent on software systems and cryptocurrencies have enabled online financial crime.[45]

The fact that gray-market brokers and commercial surveillance vendors sell elite exploits, or malware tools using them, to state intelligence agencies also provides evidence that elite exploits are difficult to develop, even for states. Moreover, if TA1 or TA2 actors were able to develop elite exploits, since there are so many such actors,[46] the supply of exploits would be very large, and the prices in these markets would quickly collapse. This is evidence that TA1 and TA2 actors cannot find elite exploits.

However, it is not straightforward to infer the development costs of elite exploits from their price. The price must be an upper bound on the development costs,[47] but there is limited public information on how much it costs to develop such exploits.[48] Table 3.4 shows our own rough calculations of the development costs of EternalBlue which, according to public reporting, took NSA analysts nearly a year to develop (Perlroth and Shane 2019). It is unclear how many NSA staff were involved in the effort, so we test the implications of assuming 1, 3, and 5 people were involved.

[44] Commercial Surveillance Vendor product offerings prices are collected together in this sheet: Exploit prices from commercial surveillance vendors
[45] Prices on exploit broker platforms have increased in recent years: in 2019, the highest bounty offered by Crowdfense was $3M (Franceschi-Bicchierai 2024), whereas today the highest is $7M.
[46] Our survey respondents estimated that there are 10K-100K TA2 actors, and 1 million TA1 actors.
[47] A seller would be unlikely to sell the exploit code to multiple buyers because the exploit would be considered burned once used.
[48] Ablon and Bogart (2017) provides data on vulnerabilities and exploits found by a commercial surveillance vendor, and presents rough calculations suggesting that the cost to develop an exploit in 2017 is around $30,000 (Appendix E). However, it is unclear how strong the exploits developed by the commercial surveillance vendors they interviewed were, and there is large variation in the cost to develop different kinds of exploit.

Variable	Value	Input type	Source
Years to find vulnerabilities and exploit for EternalBlue	1	Input	Perlroth and Shane 2019
NSA hacker salary (2024)	$177,459	Input	Hacker Salary Data Sheet
NSA Hacker salary (2012), inflation adjusted	$130,000	Calculation	CPI calculator
Number of hackers low	1	Subjective assumption	
Number of hackers mid	3	Subjective assumption	
Number of hackers high	5	Subjective assumption	
Salary costs low	$130,000	Calculation	
Salary costs mid	$390,000	Calculation	
Salary costs high	$650,000	Calculation	
Employer overhead multiplier	1.4	Input	Ablon and Bogart (2017) p. 85
Total development costs low	$182,000	Calculation	
Total development costs mid	$546,000	Calculation	
Total development costs high	$910,000	Calculation	

Table 3.4. Rough model of the development costs EternalBlue

The rough model in Table 3.4 suggests that it cost on the order of $100K to $1M to develop EternalBlue in 2012 dollars. In 2026 dollars, the cost would be 1.4 times greater. This model may understate the true development cost because there may have been failed researcher efforts to find elite Microsoft exploits. Moreover, as we discuss below, developing elite exploits is likely more expensive today due to improved cybersecurity. In light of the discussion here, we think development costs of $100K–$1M reflect a reasonable order of magnitude range for the development of elite exploits.

Since 2020, Known Elite Exploits Were Developed by TA4 and TA5 Threat Actors or Specialized Groups

There are no public comprehensive data on the number of elite exploits developed since 2017 and who developed them. This is in large part because many elite exploits may be used in secret for years before ever becoming publicly known (e.g. the NSA used EternalBlue secretly for at least five years).

Google's Project Zero maintains a database of zero-day exploits discovered while being used for cyberattacks in the wild. We analyzed this dataset using Opus 4.8-High. This analysis is preliminary and has not been externally validated, though may provide useful evidence.

The analysis suggests that around four elite-level exploits were found between 2020 and 2024 (about one per year), but that the rate then rose sharply, to around five in 2025 alone (see Appendix A.5).

These were developed by the US and its allies (in the TA5 threat actor class), commercial surveillance vendors like Paragon and NSO, and a white hat hacker called Orange Tsai. Commercial surveillance vendors like NSO Group are plausibly TA3 or TA4 level.[49] Orange Tsai has won numerous hacking competitions.[50]

This count likely represents an underestimate of the number of elite exploits developed each year. We may have missed some known cases, and the zero-day exploits discovered in the wild may be only a fraction of the total actually developed in a given year. Nevertheless, the rarity of in-the-wild elite zero-day exploits provides evidence that they are out of the reach of lower skilled actors who might use them for criminal or destructive purposes, as such attacks would likely become publicly known.

It is also notable that more recent elite zero-day exploits are more complex than older elite exploits. EternalBlue was a single exploit that directly enabled remote code execution with kernel privileges. Today, due to improved cybersecurity, chains of 3–4 exploits would likely be required to have the same functionality (Google, Buying Spying, (2024, p. 29)).[51] This is further evidence that elite exploits are becoming harder to find over time due to improved cybersecurity.

Stuxnet Development Costs

Estimates put the total cost of the Stuxnet operation between $1B and $2B, but this includes the costs of building mock centrifuges and testing the Stuxnet malware on them, as well as extensive human intelligence (Bracken 2024; Modderkolk 2024). The Stuxnet worm used several elite exploits chained together and relied on numerous zero-day vulnerabilities.[52] There are a range of estimates for the costs to develop the Stuxnet malware and exploits, not including the other costs of the operation (Falco 2012, pp. 20–21; Slayton 2017, pp. 99–102; Symantec Threat Hunter Team 2010).[53]

- Various sources suggest that the malware and exploits for the Stuxnet worm cost $3M to $20M to develop in 2005 dollars.
- Various sources suggest that developing the Stuxnet worm took months to dozens of person-years for cybersecurity professionals.

This is further evidence that developing effective worms is very costly, though the Stuxnet worm also had substantially different functionality and aims to the WannaCry and NotPetya worms, so the comparison is not straightforward.

[49] They have 750 staff (Fortune 2021), many of whom are drawn from Israeli intelligence (Bergman and Mazzetti 2022), and they specialize in selling spyware using powerful exploits to state intelligence agencies (Google Threat Analysis Group, *Buying Spying*, 2024)

[50] He has won cybersecurity awards including “Master of Pwn” at Pwn2Own 2021 and 2022. His research earned him the Pwnie Awards winner for "Best Server-Side Bug" in 2019 and 2021 and also secured 1st place in the "Top 10 Web Hacking Techniques" for 2017 and 2018 (Orange Tsai n.d.)

[51] Today, due to measures such as sandboxing and process isolation, even if attackers find a way to run their code via a malicious attachment or a compromised app, they will only have limited access and privileges. Consequently, additional exploits (enabling sandbox escape exploits and privilege escalation) would also be needed to gain the functionality of elite exploits. For an example of this type of elite exploit chain and discussion of how it could be used in a worm, see Appendix A.4.

[52] Sources differ on the number of exploits used in the Stuxnet worm. Falco (2012, Table 1) states that there were six, while Falliere et al (2011, Table 2) states that the final version of the Stuxnet worm used 7 exploits.

[53] These estimates are collated in the Stuxnet development costs sheet.

3.3. Estimating Actor Capabilities

We now discuss estimates of the probability that different threat actors can: (1) develop elite exploits (including finding vulnerabilities and writing the code to exploit them); and (2) conditional on having access to elite exploits, can develop a data-damaging worm using them. For (2), we assume that the actor has access to elite exploits, not that they can necessarily develop them, i.e. that the elite exploit code is published on the open internet. We estimate the probability that any randomly selected threat actor in a threat actor class is able to complete these tasks. For ease of analysis, we assume that all threat actors in each class have the same capability level. Thus, this in effect is equivalent to whether the entire class is capable of the relevant tasks.

The figures and table in the remainder of this section show in detail the author, expert, and superforecaster estimates from the second wave pilot survey.[54] In summary:

- The authors, experts, and superforecasters agree that for all threat actors, developing elite exploits is much harder than the other tasks involved in creating data-damaging worms. For all threat actors, the probability that they can create elite exploits is lower, in some cases lower by an order of magnitude or more, than the probability that they can complete the other tasks involved in creating a data-damaging worm.
- Capability for both sets of tasks increases with actor sophistication.
    - For elite exploits, there is a remote chance that TA1 and TA2 hackers are capable, but a ~95% chance for TA5 actors.
    - For the non-elite exploit tasks involved in creating data-damaging worms, TA1 and TA2 actors have a higher, though still low, capability probability (roughly 0.1% to 10%), whereas for TA5 actors, the capability probability is >98%.
- Disagreement about the capability for both sets of tasks is relatively low across the authors, the median expert, and median superforecaster.
    - However, for some threat actor capabilities, there was relatively high disagreement between different experts and between different superforecasters. For example, some experts thought that TA4 actors had a 1–5% chance of being able to create elite exploits, while several others thought that the chance was greater than 50%.

Figure 3.1 below shows the second-wave survey results for estimates for the elite exploit development capabilities of different actors.

[54] In the second wave, 4 experts and 13 forecasters responded.

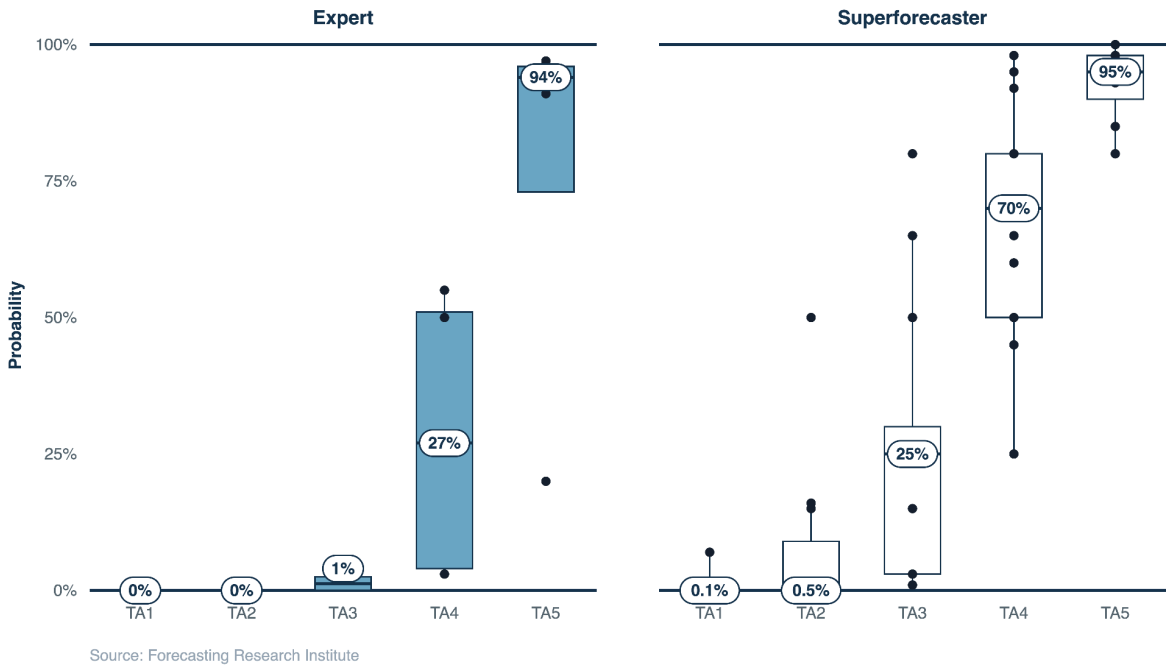
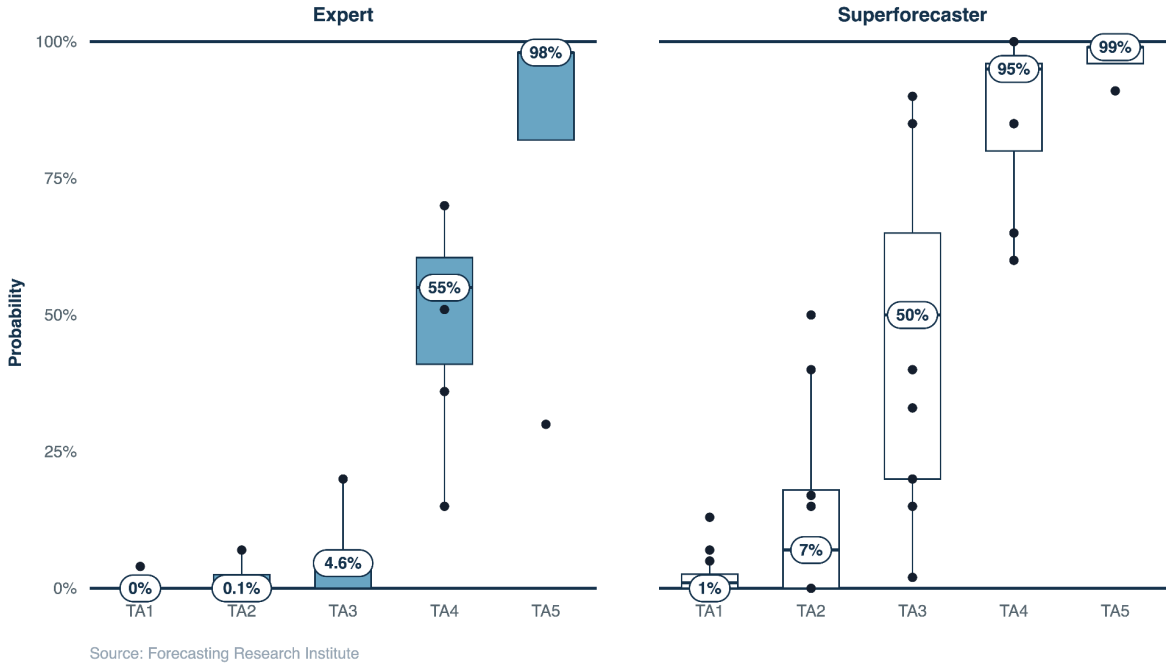


Figure 3.1. Expert and superforecaster estimates of the probability that a randomly selected actor in each threat actor class is able to write elite exploits with 12 months of effort[55]

Figure 3.2 shows the second-wave survey respondent capability estimates for actor capability to do the other tasks (aside from developing elite exploits) involved in writing a data-damaging worm.



[55] For the figures reporting survey results, the boxes represent the interquartile range, while the whiskers go out to the furthest "non-outlier" point, which is usually either the max or the largest value up to Q3+1.5*IQR (or min / lowest value up to Q1-1.5*IQR).

Figure 3.2. Expert and superforecaster estimates of the probability that a randomly selected actor in each threat actor class can complete the other steps (aside from writing elite exploits) involved in developing data-damaging worms, with 12 months effort.

Table 3.5 compares the median superforecaster and median expert estimate to the authors' median estimates:

Threat actor	Capability					
	Elite exploits			Non-exploit capabilities		
	Author	Experts	Forecasters	Author	Experts	Forecasters
TA1: Individual hobbyist hacker	0%	0%	0%	1%	0%	1.0%
TA2: Individual professional hacker	0.1%	0%	0.5%	8%	0%	7%
TA3: Team of 10 experienced hackers	6%	1%	25%	33%	5%	50%
TA4: Team of 100 state-level hackers	33%	27%	70%	100%	55%	95%
TA5: Team of 1,000 state-level hackers	95%	94%	95%	100%	98%	99%

Table 3.5. Probability that any randomly selected threat actor in a given TA class can, with 12 months of effort, develop elite exploits and complete the other tasks involved in creating data-damaging worms, aside from elite exploits: author median, median of experts, and median of superforecasters in the survey sample.

Our uncertainty about the author estimates varies depending on the threat actor class and the capability (i.e. exploit or non-exploit tasks):

- Elite exploits capabilities
    - Our uncertainty for TA1, TA2, and TA5 actor capabilities is relatively low. There is relatively clear evidence that TA1 and TA2 actors lack the capability but that TA5 actors probably have it.
    - However, the evidence for TA3 and TA4 is much more ambiguous. Consequently, for these threat actors, we updated our estimates to be roughly in line with the views of the experts and superforecasters in the survey.
- Non-elite exploit capabilities
    - Our uncertainty for TA4 and TA5 actors is relatively low. There is clear evidence from the NotPetya and WannaCry precedents that these threat actors have the relevant capability over a year.
    - However, the evidence for TA1 - TA3 actors is more ambiguous. Consequently, for these threat actors, we updated our estimates to be roughly in line with the views of the experts and superforecasters in the survey.

Threat actor	Rationale	Elite exploit capability (authors' estimate)	Confidence
TA1	• It plausibly cost $100K–$10M to develop elite exploits. TA1 actors have at most a budget of $1K for a specific attack. This suggests that it is extremely unlikely that they can develop elite exploits. • By our analysis, no elite zero-day exploits discovered being used in cyberattacks since 2020 were developed by TA1/2 actors. • States pay millions of dollars for elite exploits which suggests that they are difficult to develop even for states.	Very remote chance (~0%)	High
TA2	• It plausibly costs $100K–$10M to develop elite exploits. TA2 actors have at most a budget of $10K for a specific attack. This suggests that it is extremely unlikely that they can develop elite exploits. • See TA1 considerations.	Very remote chance (~0%)	High
TA3	• Elite exploits plausibly cost $100K–$1M to develop, mainly in skilled hacker time, and TA3 actors have ten experienced hackers with a budget of up to $1M for a specific attack, which suggests that they probably have the budget required to develop elite exploits. • Of elite zero-day exploits discovered being used in cyberattacks since 2020, some could arguably be classed as being developed by TA3 actors (e.g. NSO Group). • 2008 Conficker worm used sophisticated elite exploits and was built by TA3 actor, though this is likely harder now. • States pay millions of dollars for elite exploits which suggests that they are difficult to develop even for states.	Very unlikely (1–10%)	Low
TA4	• TA4 actors have $10M for a specific attack and staff of 100 cybersecurity professionals. This level of resources seems sufficient to develop elite exploits. • However, WannaCry offers some evidence that North Korea could not develop elite exploits in-house. • North Korea has been discovered using zero-day exploits in cyberattacks (Google (2024) p. 14), but from the cyber case studies we have looked at, we are not aware of any known cases in which North Korea has used zero-day elite exploits.	Realistic possibility (5–60%)	Low

TA5	• TA5 actors have 1K top-tier staff and $1B for a specific attack. This seems clearly sufficient to develop elite exploits. • Most known elite exploits have been developed by TA5 actors. US and its allies used elite exploits in EternalBlue espionage, Duqu, Flame, and Stuxnet. Assuming China's cyber capabilities are comparable to the US,[56] it is very likely that China is also able to develop elite exploits.	Almost certain (90–100%)	High

Table 3.6. Rationales for author estimates for the annual probability that any randomly selected threat actor in a given TA class can develop elite exploits

Table 3.7 below summarizes our estimates of the probability that different threat actors can complete the other tasks involved in creating data-damaging worms, assuming that they have access to elite exploits (e.g. assuming that the elite exploits were available on the open internet). It also provides supporting rationales for these estimates.

Threat actor	Rationale	Elite exploit capability (authors' estimate)	Confidence
TA1	• Estimates mainly defer to experts and superforecasters. • Significant worm attacks using the Shadow Brokers tools were not launched by TA1 and TA2 actors, despite the large number of such actors and their apparent willingness to launch such attacks. However, such attacks would only have been effective within around 3 months of the Shadow Brokers leak, as patches would have been deployed. So, it is difficult to rule out TA1&2 capability over a year.	Extremely unlikely (0.1%–2%)	Moderate
TA2	• Estimates mainly defer to experts and superforecasters. • See TA1 considerations.	Unlikely (1–20%)	Low
TA3	• Estimates mainly defer to experts and superforecasters • TA3 actors develop the Conficker worm, though this is likely harder now due to improved cybersecurity. • Significant worm attacks using the Shadow Brokers tools were not launched by TA3 actors, though this	Realistic possibility (5–60%)	Low

[56] Knapp et al. (2021) estimates that the Department of Defense has ~50,000 cyber staff (see footnote on p. 48), though many US cyber staff are outside the DoD. Christopher Wray, the former director of the FBI, told Congress in 2023 that Chinese hackers outnumber the FBI's cyber staff 50 to 1 (Feiner 2023), though note that the FBI is only one part of the US government's cyber force. Mandiant (2013) estimated that China has "130,000 personnel divided between divided between 12 bureaus (局), three research institutes, and 16 regional and functional bureaus". USCC (2022), (p438) notes "the PLA reportedly has as many as 60,000 cyber personnel that could support cyberwarfare missions" – however this is just one arm of China's cyber army.

	may reflect limited willingness to launch such attacks.		
TA4	• WannaCry was released one month after the Shadow Brokers leak, though the worm did not work properly. Nonetheless, this does suggest that, over a year, North Korea and other TA4 actors very likely have the capability.	Almost certain (~100%)	High
TA5	• NotPetya was released by Russia 2–4 months after access to NSA tools, which suggests clearly within reach for TA5 actors over a year. • TA5 actors have developed numerous effective worms for espionage and sabotage, which suggests that they clearly can also develop data-damaging worms.	Almost certain (~100%)	High

Table 3.7. Rationales for our current best guess for the annual probability that any randomly selected threat actor in a given TA class can develop data-damaging worms assuming they have access to, but did not necessarily develop, elite exploits

Table 3.8 below collates these estimates and computes the "end-to-end capability" of different threat actors: the probability that they can already both (1) create elite exploits and (2) develop data-damaging worms using those elite exploits.

Actor type	**Exploit Capability:** Probability an actor can succeed already over a year	**Post-Exploit Capability:** Probability an actor can succeed already	**End-to-End Capability:**
TA1	Very remote chance (~0%)	Extremely unlikely (0.1%–2%)	Very remote chance (~0%)
TA2	Very remote chance (~0%)	Unlikely (1–20%)	Very remote chance (~0%)
TA3	Very unlikely (1–10%)	Realistic possibility (5–60%)	Very unlikely (0.1–5%)
TA4	Realistic possibility (5–60%)	Almost certain (~100%)	Realistic possibility (5–60%)
TA5	Almost certain (90–100%)	Almost certain (~100%)	Almost certain (90–100%)

Table 3.8. Collated author capability estimates for different threat actors

4. Potential Damages from Data-Damaging Worm Attacks

So far we have reviewed how willing and able actors are to release data-damaging worms. In this section, we review the evidence on the potential economic costs of data-damaging worms.

We first review which types of economic costs we try to quantify. Next, we discuss and critique the evidence on the economic damages of past worm attacks and argue that WannaCry and NotPetya caused economic damages of ~$1B and ~$10B respectively. We then discuss the potential economic damage of data-damaging worms using elite exploits if they were released today (Section 4.3). We argue that WannaCry and NotPetya could have been far worse with relatively modest changes in their code but that the risk may be lower today due to improved cybersecurity. In this section, we only present author estimates of potential economic damages; we did not survey experts and superforecasters directly on this parameter.

4.1. Estimating Economic Damages from Worm Attacks Is Challenging

As discussed in Section 1, we focus on economic damages and not other kinds of damage, such as geopolitical effects or national-security impacts. However, even economic damages are difficult to quantify. Following the White House's CEA (2018) taxonomy, we include both direct costs to firms infected by a worm and indirect costs to suppliers, intermediate customers, and end consumers who were not infected by the worm but nonetheless suffered economic harms, such as orders going unfilled. Both direct and indirect costs include lost revenue due to business interruption, reduced consumption, costs of cleaning up or replacing infected systems, legal payouts, and insurance costs.

Estimating the economic costs of worm attacks is challenging because it involves collecting data on a large number of affected parties across jurisdictions, and the available statistics are fragmented (Anderson et al., 2019). Indirect costs are particularly difficult to measure, as they are influenced by other complex economic events. Consequently, isolating indirect costs resulting from a cyberattack is difficult. Economic costs inflicted by cyberattacks do not always translate directly into deadweight social loss. Customers unable to purchase goods from a firm affected by a cyberattack, for example, might shift demand to other firms, offsetting social costs. The following sections take multiple approaches to produce an overall estimate of economic damages resulting from worm attacks.

4.2. Historical Data on the Damages of Past Worm Attacks Is Limited but WannaCry and NotPetya Plausibly Caused Damages of $1B–$10B

4.2.1. Data on Damages from Past Worm Attacks Is Limited and Poor

There is limited data on the damages of past worm attacks. Johansmeyer (2024) collected various publicly reported damages from significant cyberattacks (>$800M damage and >10–25 companies

affected) from 1998 until today. These data suggest that some worm attacks prior to 2005 caused large amounts of damage (e.g. $65B for the SoBig worm and $67B for MyDoom). However, Johansmeyer notes that the provenance of data on cyberattack damages is poor, and most of these estimates seem unreliable (Johansmeyer nd; see also Cobos & Cakir, 2024). The main problems in the existing data are:

1. There is no methodological information or calculations for any of the estimates. Many of the early damage estimates come from consultancies who do not share their methods and may have incentives to produce inflated estimates (Leyden, 2002; Gallaher et al. 2006).

2. Damage estimates for the same event vary by orders of magnitude. Johansmeyer (2023, fn. 1) notes that where there are multiple sources, he chooses the higher estimate. This is because Johansmeyer's aim is in part to show that even on the high estimates, cyber risk is lower than many people argue.

    a. For instance, one of Johansmeyer's sources for damage estimates is the November 2003 House Committee Hearing on computer viruses. In that hearing, different speakers and sources gave damage estimates for SoBig ranging from $500M to $30B.

    b. One source used in the dataset reports that the Melissa worm caused >$1B in damage (in 2012 dollars), providing no calculations or source (Beattie 2012). The perpetrator of the attack stipulated in a plea agreement that the attack caused more than $80M in damage (The Register 2001).

It is difficult to make progress on estimating the damages of pre-2009 worms, as there is a lack of good evidence. Therefore, we focus on damages from the two most recent worms – WannaCry and NotPetya – that are the closest analog for the threat model considered here.

For more discussion of past cyber damage estimates, see Appendix A.7.

4.2.2. NotPetya Damages Were Plausibly $3B–$10B

After searching the academic literature, we found only one study (Crosignani et al 2023)[57] that estimates the damages of NotPetya and includes details on the methodology used to produce that estimate. We also develop our own less robust estimate of NotPetya based on claims made by Ukrainian government officials.

Crosignani et al (2023)

The methodology of Crosignani et al (2023) is as follows:

- **Direct damages:** They constructed a list of eight (non-Ukrainian) firms that were directly hit by NotPetya and collated their reported losses, including lost or delayed revenue and remediation costs.

[57] An open access working paper version of Crosignani et al (2023) is available here.

- **Indirect costs:** They then identified companies in the upstream and downstream supply chain for the affected companies and compared their financial performance to similar unaffected companies.

They estimate that the direct damages were $1.8B, and downstream supply chains suffered "conservative" damages of a $7.3B drop in revenue compared to the counterfactual. This implies a total of $9.1B in damages in 2017 dollars, which is ~$12.4B in 2026 dollars.

This result is potentially biased in either direction. It could be biased high because lost revenue for specific companies in a particular period may not reflect true society-level economic loss. Rival firms might gain revenue, and sales might be higher in later periods. It is difficult to estimate how large this effect might be.

The estimate could be biased low because the damage estimates do not include any governments or Ukrainian companies, even though Ukraine accounted for 75% of total infections. However, Ukrainian income per person was much lower than other affected countries, so countries outside Ukraine likely suffered higher economic costs per NotPetya infection. Accounting for these two factors, total global damages were roughly $17B.[58] However, for the same reason, the costs to social welfare in Ukraine may have been higher because economic losses are worth more to Ukrainians than to richer people. This illustrates a broader problem with expressing social loss in terms of dollars.

Overall, it is difficult to know how these two effects balance out, but we think Crosignani et al (2023) provides some evidence that **damages on the order of ~$10B are plausible**. We discuss the study in more detail in Appendix A.8.

Inferring Global Damages from Claims About Ukrainian Damages

Another (less reliable) approach to estimate NotPetya damages involves the following steps:[59]

- The Ukrainian finance minister claimed that NotPetya cost 0.5% of Ukrainian GDP (Burdyha 2017), or $560M in 2023 dollars.
- Since there were ~500K infections in Ukraine (Maschmeyer 2021), this implies a cost per infection of ~$1.1K.
- We then adjust the cost per infection in different countries based on how their GDP per capita compares to Ukraine. For example, because German GDP per capita was 17X Ukraine's in 2017, we assume that the cost per infection was 17X. We make this adjustment for all affected countries.
- Using a rough estimate of the number of infections in each country, we can then infer the cost per country and aggregate the total cost. This implies **total damages of ~$2.6B.**

[58] Calculations are in the 'NotPetya & WannaCry damages" tab of the Worm damages sheet, in which we collect damage estimates of worm attacks.
[59] Full calculations are in the "NotPetya & WannaCry damages" tab of Worm damages sheet.

This approach avoids some of the problems with Crosignani et al (2023) outlined above and discussed in more depth in Appendix A.8, but rests on an estimate from the Ukrainian finance minister which cannot be verified and seems unreliable. Indeed, in the same article quoting the Ukrainian finance minister, the Ukrainian head of cyber police stated that "These are all rather arbitrary calculations... Now we can only estimate the cost of the disabled computers, and somehow calculate the lost profits from the non-working services. But in reality, we still don't know what exactly the virus did" (Burdyha 2017).

Overall, we put more weight on the adjusted Crosignani et al estimate. This implies **damages on the order of ~$10B.** This corresponds to a cost per infection of ~$17K.

4.2.3. WannaCry Damages Were Plausibly $1B–$4B

We have not found any systematic and rigorous attempts to quantify the total damages of the WannaCry attack.[60] As such, we attempted to produce our own.

Inferring WannaCry Damages from NotPetya Damages

One method involves inferring WannaCry damages from the NotPetya damage estimates outlined above. Although data is limited, it is reasonable to assume that the cost per infection from WannaCry was lower than for NotPetya, as the encryption method used by WannaCry was breakable in some cases and did less damage to infected systems.[61]

Therefore, to set an upper bound on the costs of WannaCry, we assume its cost per infection was the same as NotPetya's. As Table 4.1 shows, this implies **upper bound damages of ~$4B.**

Variable	Value	Source
NotPetya damages (2026 dollars)	$10B	Crosignani et al (2023)
Total NotPetya infections	664K	Inferred from Maschmeyer 2021, p. 81[62]
NotPetya cost per infection	$17K	Calculation
WannaCry infections	230K	(Cooper 2018)
Implied WannaCry damages (2026 dollars)	$4B	Calculation

[60] There are estimates of the costs of WannaCry on the NHS. Ghafur et al. (2019) estimate total costs of £5.9m, while DHSC 2018 estimate costs of £92m, which is 0.01% to 0.1% of the total NHS England budget. If we assume costs to the NHS were proportional to costs on the global economy, this implies costs of $4.4bn to $68bn. Calculations are in the "NotPetya & WannaCry damages" tab of Worm damages.

[61] The NotPetya and WannaCry payloads each had similar functionality, but NotPetya caused more damage because it wiped the whole disk on an infected system, making the OS and all system files inaccessible (Hasherezade 2017), whereas WannaCry focused on data files and some backup files (Computable nd; Forbes 2017). NotPetya deployed a faux ransomware message, but it was not possible to decrypt affected files (Kaspersky 2017). WannaCry encrypted files and decryption was not possible without access to a decrypted private key from the attackers (Sophos 2017), though decryption was possible in some limited cases (Bank Infosecurity 2017). Even if victims paid the ransom, there was no way of associating the payment with a specific computer (Kaspersky n.d.), and very few victims actually paid the ransom (Collins 2017).

[62] This is inferred from a claim in Maschmeyer 2021. For NotPetya, there were 500K infections in Ukraine alone (Maschmeyer 2021). 75% of infections were in Ukraine (ESET 2017), which implies ~670K infections globally. Maschmeyer's estimate is difficult to verify. Maschmeyer notes that "This estimate by an anonymous expert at a leading cybersecurity vendor in Ukraine is based on the number of compromises he observed personally while involved in mitigation efforts at multiple large enterprises in Ukraine."

Table 4.1. Upper bound: Inferring WannaCry damages from NotPetya damages

Inferring from Damages to the English NHS

There are more rigorous published estimates of the costs of WannaCry for the English National Health Service (Ghafur et al (2019)), one of the most high-profile WannaCry victims (BBC, 2017). Ghafur et al (2019) only estimates costs to secondary care (hospitals, specialist clinics etc), not to primary care (GPs, dentists, etc), and finds damages across the English NHS of £6M. We therefore have to infer total costs to the whole NHS, and then infer total global damages from that. Our best guess estimate implies **damages of around $1B**,[63] but this involves various highly subjective and uncertain assumptions, so we do not have much confidence in this estimate.

Though we have less confidence in these estimates than the NotPetya damage estimates, we think a **credible range for WannaCry damages is roughly $1B–$4B.**

These estimates suggest that WannaCry and NotPetya may have been the most economically damaging cyberattacks ever.[64]

4.3. With Modest Changes, WannaCry and NotPetya Could Have Done Far More Damage

4.3.1. WannaCry and NotPetya Could Have Done Far More Damage

Although WannaCry and NotPetya caused damages of $1B–$10B, with modest changes they could have done far more damage.

WannaCry and NotPetya only infected hundreds of thousands of systems,[65] whereas earlier worms had much greater reach.[66] It seems clear that the potential damage of both WannaCry and NotPetya could have been far greater.

- **WannaCry included a kill-switch, which a researcher was able to find and activate** seven hours into the attack. WannaCry was designed to check whether a certain gibberish URL led to a live web page, and if it did, the malware stopped further encryption on affected machines, and the damage from the worm was slowed significantly (Kryptos Logic 2017; Malwaretech 2017; Lee et al. 2017).[67] A computer security researcher registered the domain for a small fee and effectively stopped further damage from the attack (Newman 2017a;

[63] Full calculations are in the 'NotPetya & WannaCry damages' tab of Worm damages.
[64] For comparison, CISA (2020, sec. 3) reviewed damage estimates for other large cyberattacks and found that the most economically costly incidents were hacks of the US Office of Personnel Management ($760M) and the health insurer Anthem ($376M).
[65] For NotPetya, according to one source, there were 500K infections in Ukraine alone (Maschmeyer 2021). 75% of infections were in Ukraine (ESET 2017), which implies ~670K infections globally. For WannaCry, there were >230K infections within 8 hours (Cooper 2018).
[66] ILOVEYOU infected 50M within 10 days (Winder 2020), and SoBig infected tens of millions of systems (Gaudin 2004).
[67] After the kill switch was contacted by an infected system, the encryption payload would not deploy (Kryptos Logic 2017). Encryption of infected systems continued at a much slower rate, rather than stopping completely, for several reasons. Some firewall operators or ISPs may have blacklisted the domain for periods of time, unwittingly placing systems at greater risk (Kryptos Logic 2017). Cisco (2017) further notes that systems may fail to reach the domain that triggers the kill switch (e.g. due to a firewall).

Kryptos Logic 2017). Absent this unforced error, it is plausible the attack could have infected and damaged far more systems.[68]

- **NotPetya was designed to target damage to Ukrainian systems.**[69] According to a Ukrainian government official, 10% of all computers in Ukraine were wiped by NotPetya, and half of these were not recoverable (Burdyha 2017). Had the worm been designed to hit all countries, the damage could have been far greater. This could have been easily accomplished via a simple tweak: if NotPetya had used EternalBlue to spread between networks (like WannaCry) rather than only within networks.
- **Both WannaCry and NotPetya exploited a vulnerability for which a patch was available 1–2 months prior to the attack**. Many potential victims would have patched their systems in the two month gap between the release of the patches and the release of the worms.

The fact that damages from WannaCry and NotPetya could have been far worse is important for the threat model for AI-enabled data-damaging worms. First, future AI could reduce the risk of errors in worm code, which would otherwise limit potential damage. Second, as discussed in section 3, AI could uplift actors (such as TA1/2 lone wolves or TA3 terrorists) who aim to cause broad destruction rather than more limited damage. Third, AI could enable threat actors to find elite exploits of zero-day vulnerabilities with no available patch.

4.3.2. Worst-Case Versions of WannaCry and NotPetya Would Have Caused on the Order of $100B in Damage

We outline three methods to estimate potential damages if worst-case versions of WannaCry or NotPetya were released in 2017. All of these methods involve naively extrapolating from the estimated damages of WannaCry and NotPetya to worst-case versions that spread globally. After describing these three naive models, we will discuss their limitations.

It is important to stress that these estimates are of damages if worst-case versions of WannaCry and NotPetya were released in 2017. This does not necessarily imply that these worms would do comparable damage today, as cybersecurity has improved.

Method 1: A Simple and Unreliable Model of "WannaCry without the Kill Switch"

If WannaCry had not included a kill switch, it seems plausible that it could have done far greater damage. WannaCry infected ~230,000 systems within around seven hours. Kryptos Logic (2017) claims that the kill switch directly prevented 14M to 16M infections and reinfections – and that tens of millions of computers or more could have been infected if the kill switch had not been activated. This suggests that, if WannaCry had not included the kill switch, infections could have been around 50–200X higher.

[68] It is unclear why they added the kill switch. Marcus Hutchins, who discovered the kill switch, has claimed that there are signs in the code WannaCry accidentally leaked slightly earlier than intended (Darknet Diaries 2025). This may be one possible explanation for the kill switch. It is also possible that they included it so that they could have more control over the spread.
[69] NotPetya nonetheless spread to other countries. For instance, NotPetya spread to Maersk's networks because M.E.Doc was installed on a single computer in Maersk's Ukraine operation (Greenberg 2018).

We can also estimate infections if WannaCry had relied on vulnerabilities for which no patch was available when the worm was released. At the time, the total pool of systems that could have been damaged by WannaCry, provided they were unpatched, was around 400M at the time of the attack.[70] If WannaCry did not include the kill switch and systems were unpatched, infections could potentially have been up to 1,700X higher.

Table 4.2 shows the number of infections on different scenarios, on the following assumptions:

- Constant doubling time: Chernikova et al. (2023) suggests that WannaCry was spreading exponentially before the kill switch was activated. The doubling time of WannaCry before the kill switch was activated was around 1.2 hours.[71]
- There was no kill switch.
- All systems were unpatched, i.e. perhaps the malware exploits zero-day vulnerabilities.

The final column shows damages on different scenarios, assuming that damages scale with infections.

		Infections	Damages
WannaCry Actual (after 7 hours)		230K	$2B
Extra hours spread	2	737K	$6B
	4	2M	$21B
	6	8M	$66B
	8	24M	$210B
	10	78M	$674B
	12	248M	$2.2T
	13	444M	$3.9T

Table 4.2. Simple model of damages of a worst-case version of WannaCry[72]

As we discuss below, we think the damages implied by this simple model are too high.

Method 2: Inferring Global Damages from Proportionate Damages to Ukrainian GDP Caused by NotPetya

As discussed above, the Ukrainian finance minister claimed that NotPetya reduced Ukrainian GDP by 0.5%. Table 4.3 outlines global costs if NotPetya were designed to spread globally and did proportionate damage to the world economy.

[70] As of 2016, according to one estimate, there were ~1.5 billion active Windows users (Thurrott 2016). However, the pool of potentially vulnerable systems was much smaller than this. Microsoft claimed that no known Windows 10 users were infected by WannaCry (Microsoft (2017)). Coburn et al. (2019), p. 44 claimed that at the time of the attack, there were 400 million actively used Windows computers running version 8 or an earlier operating system. This is therefore a more plausible upper bound on potential infections.

[71] In one lab experiment, WannaCry did not spread exponentially, though this may be because the lab test used a network with only 50 hosts (Nguyen et al. 2024, p. 5). Given how fast WannaCry in fact spread during the actual attack and its mechanism of propagation, it seems like it must have spread exponentially, so we do not put much weight on this point.

[72] Note that due to rounding, infections and damages may appear not to scale in accordance with the assumptions of our model. Note that the total pool of vulnerable systems was ~400M (assuming no systems were patched), which is why the table only considers spread of up to 13 hours.

Variable	Value	Source
Proportionate costs to Ukraine GDP	0.5%	Ukraine Finance Minister (Burdyha 2017)
World GDP in 2017 (2023$)	$81.7T	World Bank
Implied costs to world GDP	$408B	Calculation

Table 4.3. Inferring potential global damages from one unreliable estimate of damages to Ukrainian GDP

We do not have much confidence in this estimate because, as noted in section 4.2.2, it is based on an estimate of the cost of NotPetya to Ukraine which is difficult to verify and seems unreliable.

Method 3: Inferring Global Damages from One Estimate of the Fraction of Computers Destroyed in Ukraine by NotPetya

A Ukrainian official claimed that NotPetya destroyed 5% of all computers in Ukraine. Table 4.4 infers global damages for a worst-case version of NotPetya from this, assuming a given cost to replace a damaged computer.

Variable	Value	Source
% of Ukrainian computers destroyed by NotPetya	5%	Ukrainian official (Burdyha 2017)
Number of computers, global	~2B	SCMO 2019
Total computers destroyed	100M	Calculation
Cost to replace computer	$500	Assumption
Implied costs to world GDP	$25B	Calculation

Table 4.4. Inferring potential global damages from one unreliable estimate of the fraction of computers destroyed by NotPetya

This method does not include indirect costs and so is biased low in that respect.

Problems with Naive Extrapolation Models

Defender Response Might Reduce Total Infections

These naive extrapolation models might be biased high in one respect: They do not account for defender response to the attack. The longer the attack lasts, the more time defenders would have to respond by taking systems offline and patching. This limits the potential spread of worst-case versions of WannaCry and NotPetya. Moreover, the spread of the worm would likely have followed an s-curve, rather than a consistent doubling time.[73] The number of potential hosts that could be

[73] There are a range of studies exploring infection dynamics for worm malware, which use epidemiological models to model propagation dynamics. These studies typically use epidemiological models which imply s-curve spread over the full course of an outbreak (Chernikova et al., 2023; Martinez et al 2021; Baksi and Upadhyaya 2020).

infected by each node would decline in the later stages of the attack as more potential nodes are already infected. This in turn gives more time for victims to respond. Spread would likely have slowed down substantially after a third to a half of vulnerable systems were infected, giving defenders more time to respond.

However, attackers could reduce the scope for reactive response by delaying the execution of the payload until a large number of systems are infected,[74] though this would also make writing the worm harder.

It is difficult to judge how this affects the plausibility of the naive extrapolation models, but we think it makes damages of >$200B less plausible.

Damages Might Not Scale with Infections

These models assume that damages scale in proportion to infections, but this might not be accurate.

- Once (e.g.) 10% of a company's systems are infected, they already have to shut down operations, so the remaining 90% of infections may be much less damaging.
- Cleaning (e.g.) 100M devices may not be 100x as expensive as cleaning 1M systems because you can reuse the same methods in both cases.

We are unsure how much we should adjust the estimate above in light of this point, but our subjective guess is that this makes damages of >$100B less plausible.

Table 4.5 summarizes the results of the naive extrapolation models and our evaluation of them.

Method	Damages	Evaluation
1. Extrapolation from cost per infection (WannaCry)	$6B–$3T for 2–13 hours extra of spread	Upper end of range biased high.
2. Infer from cost to Ukrainian GDP (NotPetya)	$408B	Uncertain. Based on unreliable assumptions.
3. Cost to replace destroyed computers (NotPetya)	$25B	Biased low. Only includes the cost of replacing computers.

Table 4.5. Summaries of different estimates of the economic damages of worst-case versions of WannaCry or NotPetya

[74] The Stuxnet payload only activated when it detected the presence of Siemens control systems used in the Natanz nuclear centrifuge plant (Zetter 2011). EternalRocks only beaconed out to its command and control infrastructure after a delay (Cimpanu 2017a). For other malware, the payload could be programmed to execute at a specific date or time or after communication with attackers' command and control.

Method 4: Lloyd's Bashe Scenarios Imply Damages of $85B–$193B from Worst-Case Worms

Thus far, we have constructed our own simple models of worst-case versions of WannaCry and NotPetya. In this section, we discuss estimates constructed by the insurer Lloyd's of extreme data-damaging worm scenarios.

Lloyd's (2019) outlines a set of worm scenarios similar to a worst-case version of WannaCry, which they call the "Bashe scenarios". Lloyd's describes the scenarios as unlikely but plausible (p. 10).[75] They outline three scenarios, S1, S2, and X1. In all scenarios, the malware is sent to each company via a phishing email, rather than directly via the EternalBlue exploit as in the actual WannaCry attack. Once a single employee downloads the file, the worm spreads to other systems on the company's network, and then forwards the malicious email to all contacts within infected devices' address books (p. 13).

Table 4.6 outlines the assumptions of the three scenarios, and the implied direct and indirect economic damages.

Variable	S1 Scenario	S2 Scenario	X1 Scenario
Malware targets operating systems running on what fraction of global devices (p. 19)	43%	97%	97%
Fraction of companies infected in different sectors (Table 1)	1–9%	2–16%	3–21%
Number of infected companies (Table 6)	250,000	501,000	613,000
Fraction of systems infected for the median affected company (Table 4)	~15%	~15%	~15%
Payload (p. 21)	Ransomware	Ransomware	Ransomware and wiper which deletes backup files.
Number of encrypted computers	~30M (p. 13)	~60M? (inferring from S1 scenario and # of infected companies in S2)	~73M? (inferring from S1 scenario and # of infected companies in X1)
Number of computers, global, 2019 (SCMO 2019)	~2B		

[75] In the Bashe scenario, the relevant actor is a cybercrime syndicate. However, on base rates, it seems unlikely that such groups could find the requisite exploits (as discussed in Section 2), or would use a worm rather than more targeted attacks to make money.

Fraction of total global computers infected (calc)	1.5%	3%	3.6%
Total direct economic loss (Table 6)	$59B	$110B	$133B
Productivity and consumption loss	$50B	$93B	$112B
Clean-up loss	$8B	$15B	$18B
Cyber extortion loss	$1B	$2B	$2B
Total indirect economic loss (Table 6)	$26B	$49B	$60B
Total global economic loss (Table 6)	$85B	$159B	$193B
Total damages per infected system (calculation)	$2.8K	$2.6K	$2.6K

Table 4.6. An overview of the Lloyd's Bashe scenarios

Estimated total infections in the different scenarios depends on:

1. **Risk of initial infection:** the number of companies infected by the attack.
    a. This is based on a "Sectoral Vulnerability Score" developed by Lloyd's with input from subject-matter experts. The score is determined by (1) the sectors' historical susceptibility to ransomware delivered by phishing and (2) the defensive capabilities of those sectors (p. 19).
    b. Lloyd's constructed an "industry exposure dataset" to estimate the size and sector of different companies across the global economy (p. 28). This can be combined with the risk of initial infection to estimate the total number of companies infected.
2. **Spread within companies:** The fraction of computers in a company that are infected once the worm has gained access to a company's systems.
    a. This is also based on the Sectoral Vulnerability Score. Lloyd's (2019) say they "completed extensive research on worm propagation and found that worms have an upper limit of propagation within internal networks. After consulting with subject matter experts, the most accurate upper bound for infection was decided to be 40%+ to remain technically feasible" (p. 21), and most sectors have infection rates of 10–20% (Table 4).

The total number of infected firms is then obtained by summing up the expected infections across all types of companies.

Damages are determined by:

1. **Direct damages**

   a. Clean-up costs for each device ($350 per device) (p. 15).

   b. Lost revenue, which is determined by their estimates of the severity and duration of business interruption (p. 71), though we are unsure what these estimates are based on.

   c. Ransom fees ($700 per device) though only 4% of devices are decrypted by paying the ransom (p. 72), so this does not contribute much to overall damages.

2. **Indirect damages** are calculated using a "contagion multiplier" for each sector. "The multiplier calculates the relative indirect revenue loss as a proportion of a sector's direct loss. The value of the multiplier was calculated by employing an input-output approach to estimate the relative indirect shock in inter and intra-sectoral trade globally using the World Input-Output Table" (p. 29). However, the report does not provide further details on how the multiplier was calculated.

Three points are notable about these estimates.

1. **Of the three scenarios, the S1 scenario is most similar to the worst-case WannaCry.** In the S1 scenario, the malware targets an operating system running on 43% of computers, whereas WannaCry was effective against ~20% of operating systems.[76]

2. **Substantially lower cost per infection than naive extrapolation models.** Lloyd's estimated cost per infection for the different scenarios (~$2.6K) is substantially lower than our own estimates for WannaCry ($4K–$17K) and NotPetya (~$17K). Although some parts of Lloyd's calculations are opaque, this may provide some evidence that our estimates of cost per infection in the naive extrapolation models of worst-case WannaCry and NotPetya damages are too high.

3. **Tens of millions, rather than hundreds of millions, of systems are infected.** In all scenarios, the number of infected systems is well below the total number of vulnerable computers: Tens of millions of systems out of a total population of ~2B computers (~2–4%) are infected in the scenarios. This is mostly driven by Lloyd's estimates of potential spread between and within companies as determined by their own research in consultation with subject-matter experts. This provides some evidence that even in a worst-case scenario it is only plausible that tens of millions of devices would be infected, far fewer than the hundreds of millions of devices infected in the naive WannaCry extrapolation. However, in the Bashe scenarios phishing emails deliver the malware, whereas WannaCry delivered the EternalBlue exploit directly without requiring any user interaction. A worm that uses a zero-click exploit may be

[76] WannaCry was only effective against Windows systems running Windows 8 or earlier, provided they were unpatched. Coburn et al. (2019), p. 44 claimed that at the time of the attack, there were 400 million actively used Windows computers running version 8 or an earlier operating system. Assuming ~2B total operating systems on computers globally, this implies the worst-case version of WannaCry was effective against (400M/2B=) 20% of computers.

able to spread more widely. Indeed, in the WannaCry and NotPetya attacks, for some companies, far more than 40% of systems were infected.[77]

Overall, for the Lloyd's scenarios, it is unclear which scenario best represents a plausible version of a worst-case version of WannaCry. Some assumptions may bias certain scenario estimates high, whereas others may bias them low.

Overall Judgment on Damages from Worst-Case Versions of WannaCry and NotPetya

We constructed three models that naively extrapolate damages for worst-case versions of WannaCry and NotPetya. Two of these models suggest that in the worst case, these could have done hundreds of billions of dollars of damage, if not more. However, there are several reasons to think that damages in the hundreds of billions of dollars are too high.

The Lloyd's scenarios suggest that even in the worst-case, damages could approach $100B, but not much higher.

4.3.3. Potential Damages from Worms Today

We have argued that with modest changes WannaCry and NotPetya could plausibly have done ~$100B in damage in 2017, but improvements in cybersecurity since then raise the question whether similar damage could be achieved today.

The operating systems affected by WannaCry and NotPetya illustrate how advancements in security protocols mitigated their effects. Only users using Windows 8 or earlier were vulnerable to WannaCry, and the majority of infections affected Windows 7 users (Schwartz 2017b). Microsoft claimed that no known Windows 10 users were compromised by WannaCry (Microsoft 2017). Windows 10 protected against WannaCry due to more effective security features, including exploit mitigations, kernel protection, and machine learning-based antivirus and endpoint detection and protection (Ganacharya 2018). Microsoft also claimed that Windows 10 "either fully prevented or mitigated" the NotPetya malware (Ganacharya 2018; Microsoft 2018), and one source reports that Windows 10 S blocked the attack by default (Schwartz 2017a).

Subsequent improvements make it even harder for worms to infect and damage systems. Modern versions of firewalls, intrusion detection systems, and endpoint detection and response systems use machine learning and other techniques to detect malware on the basis of its network or endpoint activity patterns, rather than on the basis of known code signatures (Mauri and Damiani 2025; Karantzas and Patsakis 2021; Applebaum et al 2021). A single device scanning a network or sending

77

- 90% of a Ukrainian bank's computers were infected by NotPetya (Greenberg, *Sandworm* (2019), p. 180)
- 70% of Ukrainian Post Office computers were infected, despite efforts to shut down those systems after the attack was discovered (Greenberg, *Sandworm* (2019), p. 187).
- In a group of Ukrainian hospitals, "virtually all" Windows systems were encrypted (Greenberg, *Sandworm* (2019), p. 188).
- NotPetya came close to wiping all of Maersk's data. A backup domain controller in Ghana had been spared because there happened to have been a blackout there shortly before NotPetya hit (Greenberg, *Sandworm* (2019), p. 194).
- All Windows systems in an American hospital network were infected (Greenberg, *Sandworm* (2019), p. 201).
- 90% of Telefonica's systems were infected by WannaCry (Telecom Review 2017).

large numbers of packets might be noticed by these systems, helping to prevent significant spread. These systems may limit the potential reach of future worm attacks, even if they use zero-days.

For worms to overcome these security features today, they would have to be more complex than WannaCry or NotPetya, which increases the capability barrier that AI would have to help overcome. This suggests that the damages from worst-case versions of WannaCry and NotPetya are at the higher end of what is feasible today. Therefore, we use the Lloyd's scenario damages of ~$150B as a rough upper bound on potential damages. For a lower bound, we use damages of $10B.

It is difficult to know how effective these systems might be against future worms without having concrete details on how such worms might be designed. This, in turn, depends on knowledge of the frontier of cyberoffense capabilities, which is difficult to gain without access to classified information.

Note that the discussion here is about existing cyberdefense systems. In the next section, we discuss how future AI vulnerability discovery capabilities might improve defense and reduce the potential costs of worm attacks.

5. Offense-Defense Balance

AI models that discover vulnerabilities and develop exploits are dual-use: They can be used by both attackers and defenders. In this section, we discuss how AI uplift to defenders might affect expected damages from worm attacks.

It is difficult to assess the effects of AI on the risk of worm attacks over time because offense-defense balance depends on a range of uncertain parameters (Lohn and Jackson (2022); Garfinkel and Dafoe (2019); Lohn 2025), and AI benefits attackers in some respects and defenders in others. Table 5.1 outlines how AI relates to different determinants of the risk of worm attacks over time.

Determinant of risk of worms	Favors offense or defense?	Reasoning
Vulnerability discovery	**Unclear**	AI capability can be used by attackers and defenders.
Correlation of vulnerabilities found by attackers and defenders	**Unclear**	If vulnerabilities found by attackers and defenders are correlated, this favors defense. There are some reasons to think that automation will lead to greater correlation (Garfinkel and Dafoe (2019), p. 263). But correlation may be lower due to experimentation in prompting and scaffolding and use of different AI models.
Exploit development	**Favors offense**	This allows attackers to exploit faster. AI exploit development capability may be correlated with general AI coding abilities.
Patch development	**Favors defense**	Allows defenders to patch faster. This capability may also be correlated with general AI coding abilities. So exploit development and patch development capabilities may be correlated.
Patch deployment	**Favors offense in short-term (<3m)**	Currently a key lag on defense. It seems harder for AI to affect patch deployment than e.g. exploit or patch development. Policy decisions have a greater impact on patch deployment rates.
Defenders can repair vulnerabilities prior to software release	**Favors defense in longer-term**	Defenders have a natural advantage in that they can use AI to repair vulnerabilities in their software prior to the release of the software. As new software replaces old software, the risk of cyberattacks declines. Apple and Google release a new OS around once a year, while Microsoft releases one major feature update per year, and a major new OS every three years (Android 2025; Apple 2024; Microsoft 2025; Microsoft nd). So, plausibly defenders would start to benefit around 6–12 months after the release of AI tools, as new AI-tested software replaces old software.

Warning shots	**Favors defense in longer-term**	Major cyberattacks serve as a warning shot which encourages defenders to improve cybersecurity. This broad dynamic seems to have occurred for worm attacks, which were high prior to 2005 but then declined once cybersecurity improved after numerous major worm attacks (see Section 1).

Table 5.1. How AI might affect different determinants of the risk of worm attacks[78]

Table 5.1 suggests that the effect of AI on the risk of worm attacks will likely vary over time.

In light of the considerations above, it seems plausible that AI vulnerability discovery and exploit development capabilities will increase the risk of worms in the short-term because AI will improve (a) vulnerability discovery for both attackers and defenders, (b) exploit development for attackers, and (c) patch development for defenders, but there is less scope for AI to improve patch deployment rates. This suggests that AI uplift to factors (a), (b), and (c) will increase risk in the gap in which users deploy patches.

Data from 2008–14 suggests that half of users deploy patches after around 100 days (Lohn and Jackson 2022, p. 7).[79] We lack good data on patch deployment rates today, but we think it is plausible that, for widely used software, 90% of patches would be deployed 0.5–3 months after patch release.[80]

In the absence of an active policy decision to improve patch deployment rates, we think that releases of AI models with improved vulnerability discovery and exploit development capabilities would increase risk in the short-term (within 2 weeks to 3 months). The longer-term effect is more ambiguous and depends on (1) the extent to which AI models will continue to find vulnerabilities in released software; (2) the amount of software produced; and (3) the extent to which software developers can patch vulnerabilities prior to the release of software.

It is important to note that the conclusion one reaches about offense-defense balance in this domain does not settle the question of how AI might affect offense-defense balance in other areas of cyberspace. Offense-defense balance is a property of relationships between particular defenders and attackers, not of cyberspace in general (Slayton (2017), p. 74).[81] For example, it is much harder to

[78] The assessment here is based on Lohn and Jackson (2022), Garfinkel and Dafoe (2019), Lohn 2025, and the author's own judgment.
[79] Their source for this is Nappa et al (2015), Table III, which finds that 50% of users deploy patches after 15 to 268 days, and 90% after 129 to 799 days, depending on the software.
[80] We have not found more comprehensive recent data on patching rates for widely used software. In the 2021 Microsoft Exchange attacks, which, as discussed in Appendix A.5, may have involved elite exploits, 80% of servers were patched 10 days after patch release (Microsoft 2021), and >92% were patched 23 days after patch release (Microsoft 2021). We think these fast patch deployment rates likely reflect a broader improvement in patch deployment rates, but also are plausibly faster than average because the Exchange vulnerabilities were being actively exploited. Microsoft (2020) p. 23 suggests that 90% of patches are deployed after a week for Windows 10 users. Microsoft (2021) states that 80-90% of Windows patches are deployed after one month. We have been unable to find more recent data on patch rates for other widely used software. Lamar (2025) claimed that the median time to resolve serious flaws found in pentests declined from 112 days in 2017 to 37 days in 2025, though this is different to a representative sample of patching rates across vulnerabilities. Various other metrics suggest that cybersecurity is improving over time (Healey and Jain (2025)).
[81] For example, the offense-defense balance for Chinese espionage against US government targets is different to the offense-defense balance for ransomware attacks against hospitals.

breach government computer systems storing classified intelligence than the IT system of most small to medium enterprises. Thus, the offense-defense balance with respect to attacks on these targets is different.

For more discussion of offense-defense balance see Appendix A.9.

6. Overall Risk Estimates

Our overall aim in this report has been to estimate the expected economic damages from data-damaging worms if AI enables different threat actors to find elite exploits (including both finding vulnerabilities and writing exploits for them). In this section, we discuss different estimates – drawn from base-rate data and our survey – of the baseline risks of data-damaging worms and the marginal risks if AI gains strong elite exploit capabilities.

6.1. Historical Base Rate Damages

Since 2009, there have been two data-damaging worm attacks that have caused at least one billion dollars in damage. Assuming that the probability of such attacks has been constant between 2009 and 2025, the probability of a major worm attack per year is around 13% per year. Assuming that WannaCry caused ~$1B in damage and NotPetya ~$10B in damage, the expected risk of a >$1B data-damaging worm attack from 2009 to 2025 was on the order of $1B per year.

6.2. Hypothetical Marginal Risk Scenario

Our main aim in the report is to understand how risk would change if AI gained advanced vulnerability discovery and exploit development capabilities. Specifically, we conditioned on the following hypothetical model evaluation result:

> **Elite Exploit Uplift:** A study conducted at the end of 2025 finds that access to frontier AI models enables 25% of TA2 actors to find vulnerabilities and write elite exploits, assuming three months of full-time effort.

By default, we assume that frontier models are open weight and have no deployment safeguards. Call this "Policy 0" or "P0".

> **P0: Open weight with no refusals.** The model with elite exploit capabilities is open weight and is released without deployment safeguards or other risk mitigations.[82]

We consider the effects of alternative policies in section 6.6.

6.3. Risk Model Estimates

Our risk model tries to estimate the risk of the first major data-damaging worm attack. The model produces this estimate by decomposing the risk into the capability to launch an attack, the willingness to do so, and the damages if such an attack occurs. We can now bring together the estimates of capability and willingness discussed in previous sections and plug them into our risk model. To calculate the implications of expert and superforecaster estimates, we set the damages from data-damaging worms at $1B to $150B. We did not ask the survey respondents to provide

[82] We describe this release policy in more detail in Appendix A.10.

damage estimates because they do not depend on cyber-specific knowledge, and the chief aim of the survey is to gather perspectives on cyber-specific questions.

The raw data for the estimates is in this sheet. We generated Monte Carlo simulations of the parameter estimates to infer a confidence interval for the overall risk estimates.[83] We tested two approaches for representing uncertainty in the capability and willingness estimates – lognormal and beta distributions – but this had little effect on the headline results.[84]

Figure 6.1 summarizes the risk model results inferred from the capability and willingness estimates of the author, experts, and superforecasters.

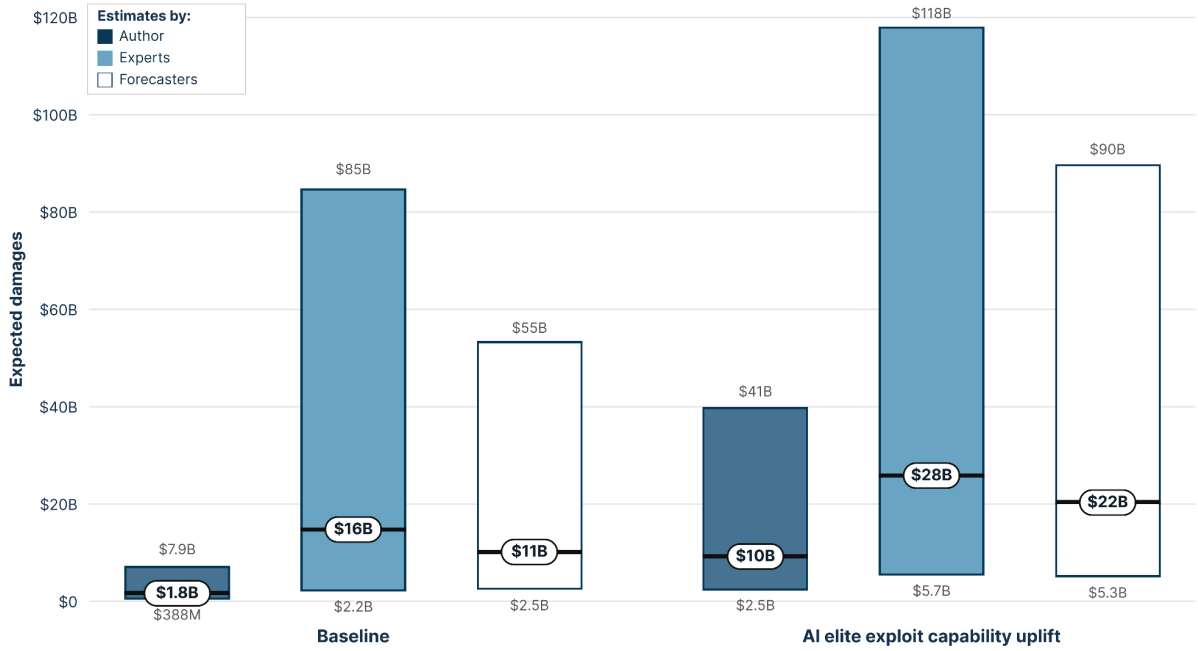


Figure 6.1. Risk model results for author, expert, and superforecaster estimates of expected damages from the first major data-damaging worm in the baseline and AI uplift scenarios. Bars represent the 90% confidence interval; horizontal lines represent the median.

Rough, order-of-magnitude estimates like these can be useful because they help identify the approximate scale of a risk. Even when they are imprecise, they can still guide decisions about risk mitigation. As Figure 6.1 shows, there is disagreement and uncertainty about these estimates within and between different estimators. Within all estimator groups, the 90% confidence interval spans one to two orders of magnitude. However, disagreement about both medians and the 90%

[83] The Github repo for this code is available here.
[84] A lognormal distribution, clipped to [0,1], can distort the tails for parameters near the boundary. Samples drawn above 1 are clipped back, compressing the upper tail and slightly biasing the distribution. A beta distribution naturally respects the [0,1] bounds without clipping, providing a more faithful representation of uncertainty for bounded probabilities. However, the beta must be fitted numerically (via optimization over its two shape parameters), whereas the lognormal has a closed-form solution from the percentile triplet. In practice, the two approaches yield broadly similar results – median expected damages typically differ by less than 20%.

confidence interval across the author, experts, and superforecasters is usually within an order of magnitude. This is shown more clearly in Figure 6.2, which uses a log scale on the y-axis.

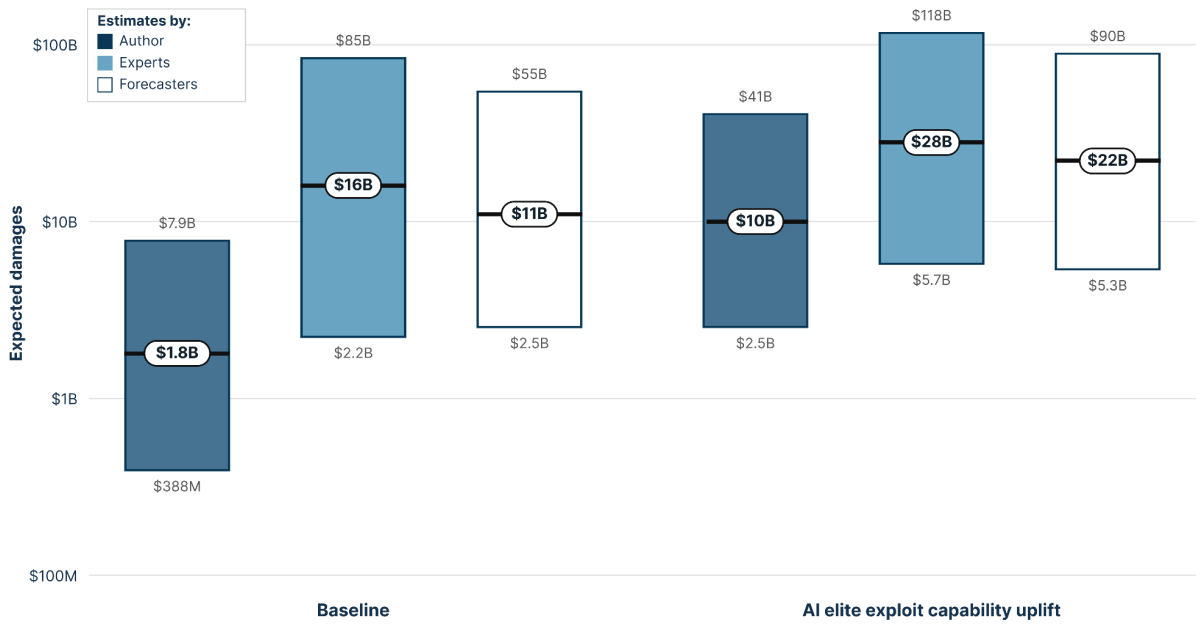


Figure 6.2. Author, expert, and, superforecaster estimates of expected damages from the first data-damaging worm in the baseline and AI uplift scenarios, log-scale y axis. Bars represent the 90% confidence interval; horizontal lines represent the median.

The difference between the AI uplift scenario and the baseline damages is the marginal damages from AI elite exploit capability uplift. Figure 6.3 shows the marginal damages from AI elite exploit capability uplift.

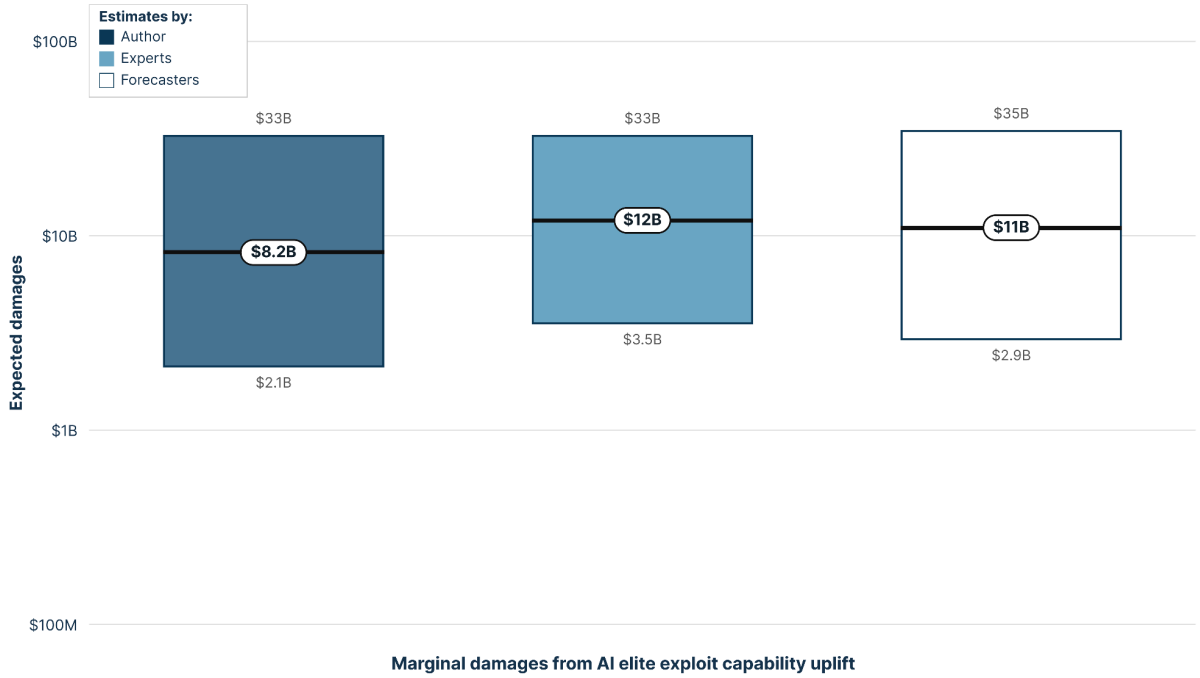


Figure 6.3. Marginal damages from the first data-damaging worm, caused by AI elite exploit capability uplift. Bars represent the 90% confidence interval; horizontal lines represent the median.

We now discuss which threat actors are the likely source of these damages. Figure 6.4 shows the fraction of the median expected damages stemming from different threat actors for author, experts, and superforecasters' estimates. As this shows, there is substantial disagreement among the author, experts, and superforecasters about the likely source of risk, especially conditional on the AI uplift scenario.

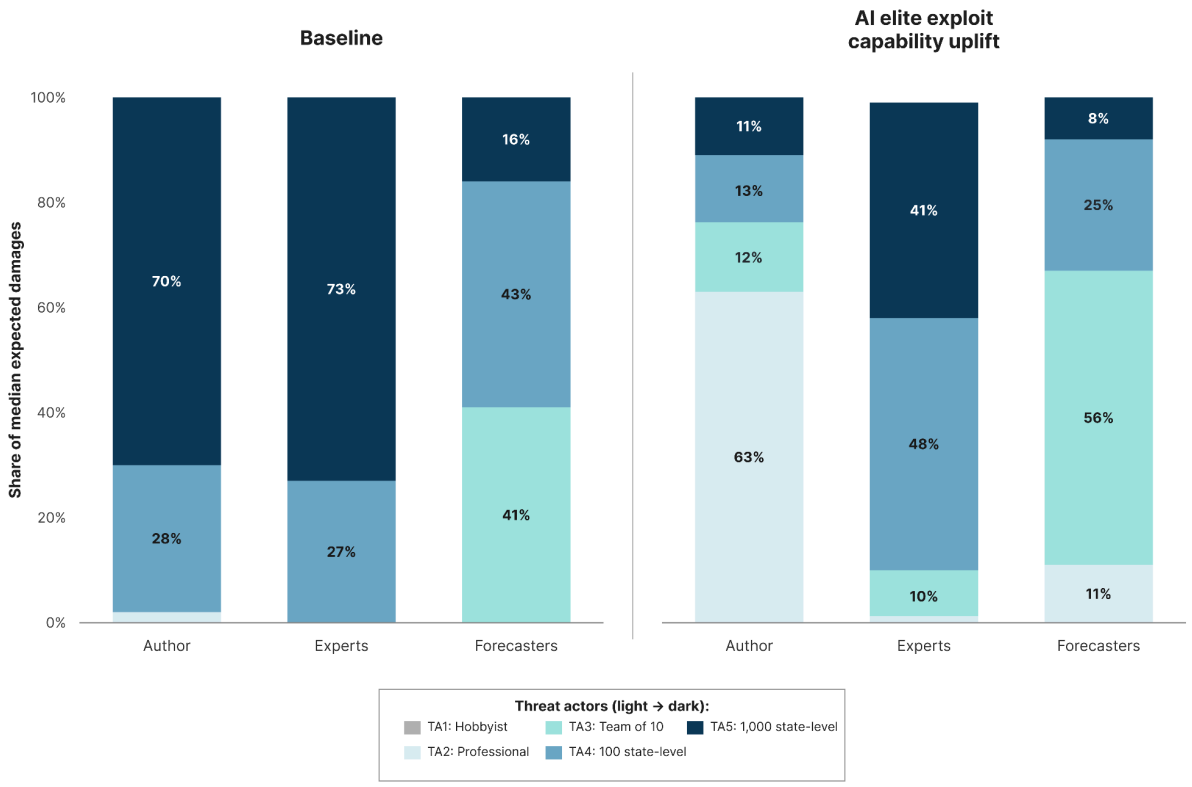


Figure 6.4. Percentage of median expected damages from data-damaging worms by threat actor

6.4. Direct Risk Estimates

In addition to asking survey respondents to estimate the parameters of the risk model, we also surveyed them directly on the baseline and marginal risk.

6.4.1. Probability of at least One Major Worm Attack

In the first wave pilot survey, we asked respondents to estimate the probability of at least one data-damaging worm attack causing at least $10 billion of damage in 2026 in the baseline scenario and conditioning on Elite Exploit Uplift.

As Figure 6.5 shows, expert and superforecaster baseline estimates were broadly consistent. From this, we can infer the expected damages from the first major worm attack, using the assumption from section 4 that the damages would range from $10 billion to $100 billion (per section 4). This implies the following expected median annual risk from the first major worm attack:[85]

- **Experts:** $2.4 billion (90% CI: $470 million to $12 billion)
- **Superforecasters:** $990 million (90% CI: $110 million to $7.8 billion)

[85] This was calculated using Squiggle, a programming language for probabilistic calculation. The relevant Squiggle code is available here.

These estimates are roughly consistent with historical damages implied by the base-rate data.

Turning to the effects of AI cyber capabilities, experts forecast higher marginal damages conditional on Elite Exploit Uplift. The median expert forecast of the probability of a $10 billion data-damaging worm attack in 2026 rose to 41%, a 3.5x increase in relative risk. For the median superforecaster, it rose to 15%, a 3x increase in relative risk.

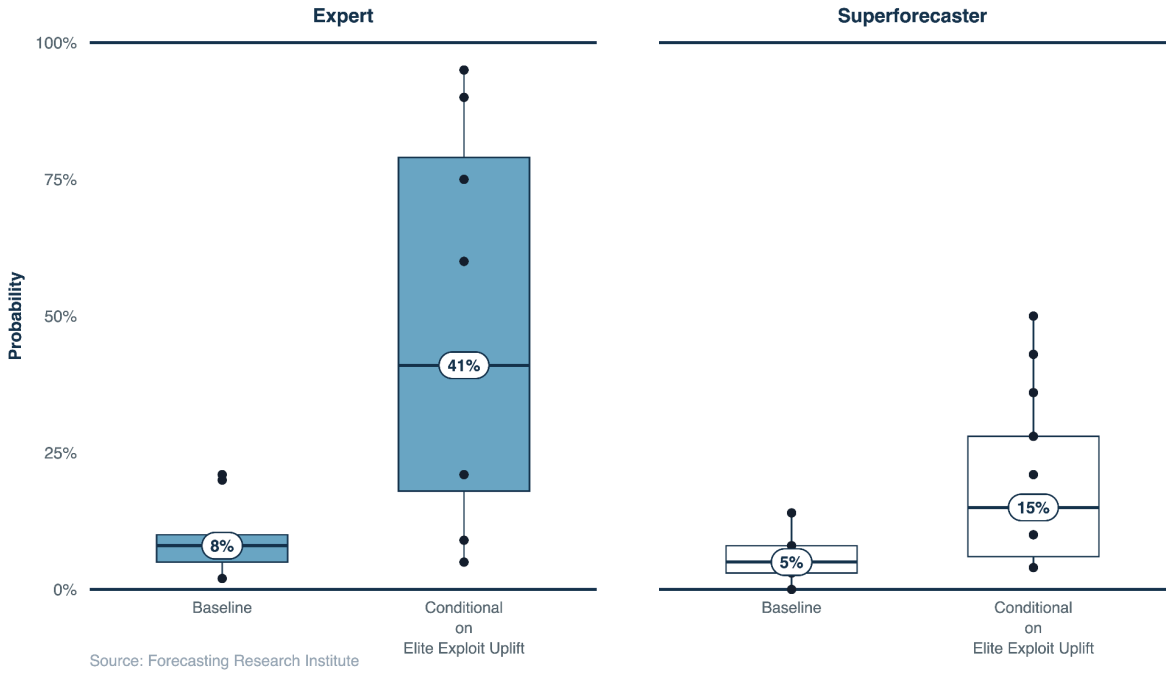


Figure 6.5: Probability of at least one data-damaging worm attack causing at least $10 billion in economic damages in 2026, conditional on Elite Exploit Uplift (AI enables TA2 actors to write elite exploits)[86]

This suggests that respondents think that elite exploits are a key bottleneck to data-damaging worm attacks.

Respondents noted that finding elite exploits would remove the most important bottleneck for TA2 actors. Given the large number of these actors and their lack of restraint compared to TA5 actors, this would indicate a substantial increase in risk. Lower-skilled actors were also thought more likely to accidentally cause massive damage due to poorly targeted attacks. Participants noted that this capability could also indicate an increase in the capabilities of higher-level threat actors (TA3 to TA5). Finally, several participants commented on offense-defense timing imbalance: While AI helps both sides, there may be a "dangerous window of vulnerability" when new capabilities emerge.

[86] For the figures reporting survey results, the boxes represent the interquartile range, while the whiskers go out to the furthest "non-outlier" point, which is usually either the max or the largest value up to Q3+1.5*IQR (or min / lowest value up to Q1-1.5*IQR).

6.4.2. Total Annual Expected Harms

In addition to estimating the risk from one worm attack causing more than $10 billion in damages, we also asked respondents to estimate the baseline damage from all data-damaging worm attacks in a year. We asked participants to forecast the probability of different amounts of total economic damages in 2026 from data-damaging worms. The median expert forecast a 20% probability of such damages falling between $100 million and $1 billion but a 0.095% probability of such damages falling between $1 trillion and $10 trillion. The median superforecaster forecast a 35% probability of damages falling between $100 million and $1 billion and a 0.1% probability of damages falling between $1 trillion and $10 trillion. (Figure 6.6).

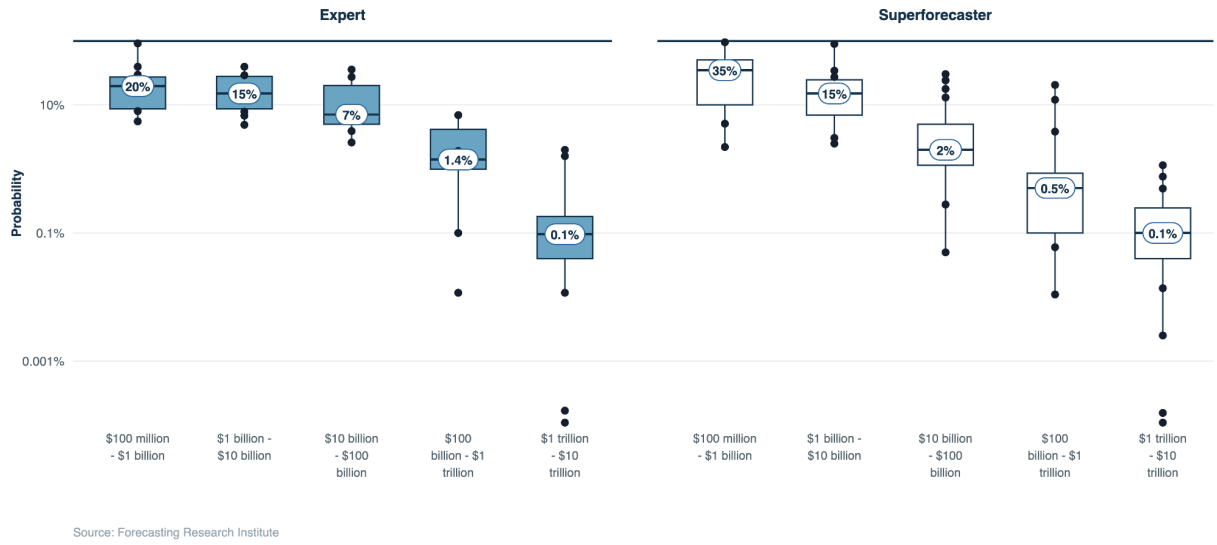
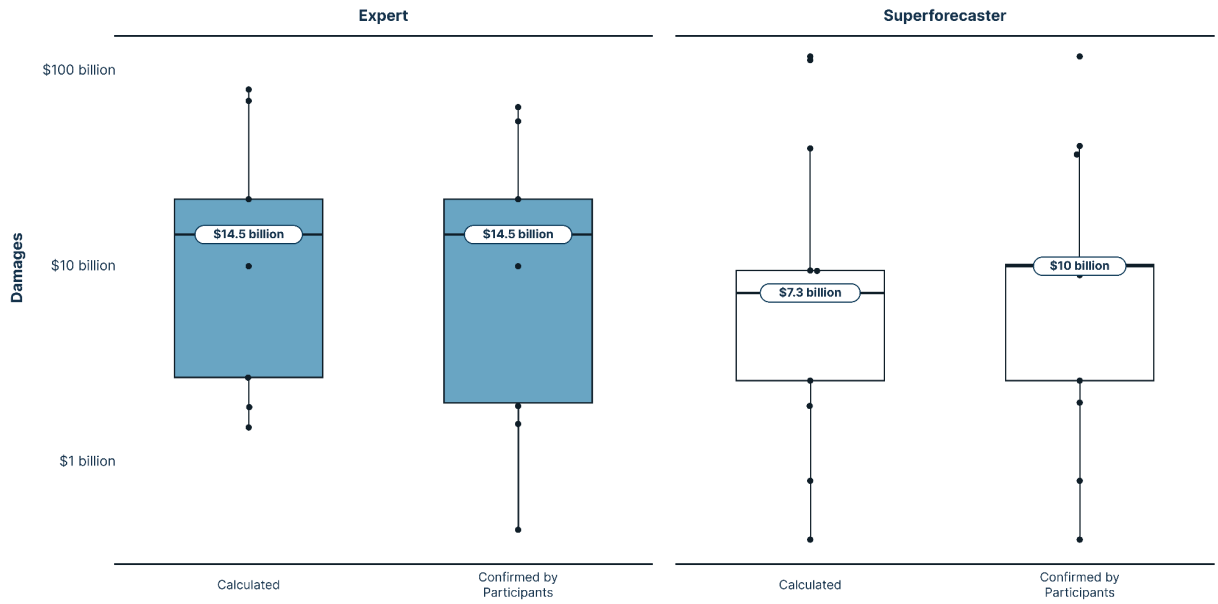


Figure 6.6: The baseline probability of data-damaging worm attacks causing different amounts of economic damages in 2026

We used these binned probabilities to calculate an expected damages value by multiplying the probability and damages of each bin and then summing across the bins. Participants were given the opportunity to alter this value if they felt like it didn't capture their true beliefs about the expected damages due to data-damaging worms in 2026. These values are shown in Figure 6.7.



Source: Forecasting Research Institute

Figure 6.7: Calculated and participant-confirmed baseline expected damages due to data-damaging worm attacks in 2026

These estimates are an order of magnitude higher than the damages implied by the base rate baseline damages from significant worms. It is possible that smaller worm attacks, with damages of less than $1 billion would cause substantial damage. However, since worms spread exponentially, we would expect damages from data-damaging worms to be heavy-tailed, and so for the largest attacks to account for most of the expected harms. For that reason, we find the baseline risk estimates here implausible. We put more weight on estimates from the other methods we have discussed here.

We then asked how these probabilities would change conditional on Elite Exploit Uplift. These forecasts are shown in Figure 6.8. The calculated expected damages and participants' adjusted expected damages estimates conditional on this capability are shown in Figure 6.9.

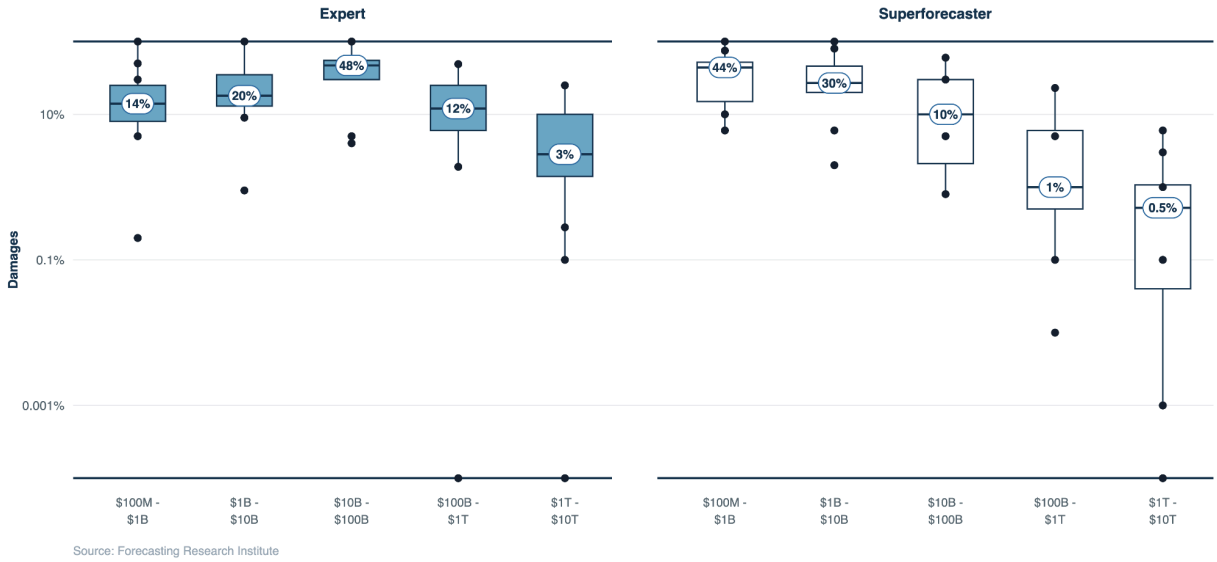


Figure 6.8: The probability of data-damaging worm attacks causing different amounts of economic damages in 2026, conditional on Elite Exploit Uplift

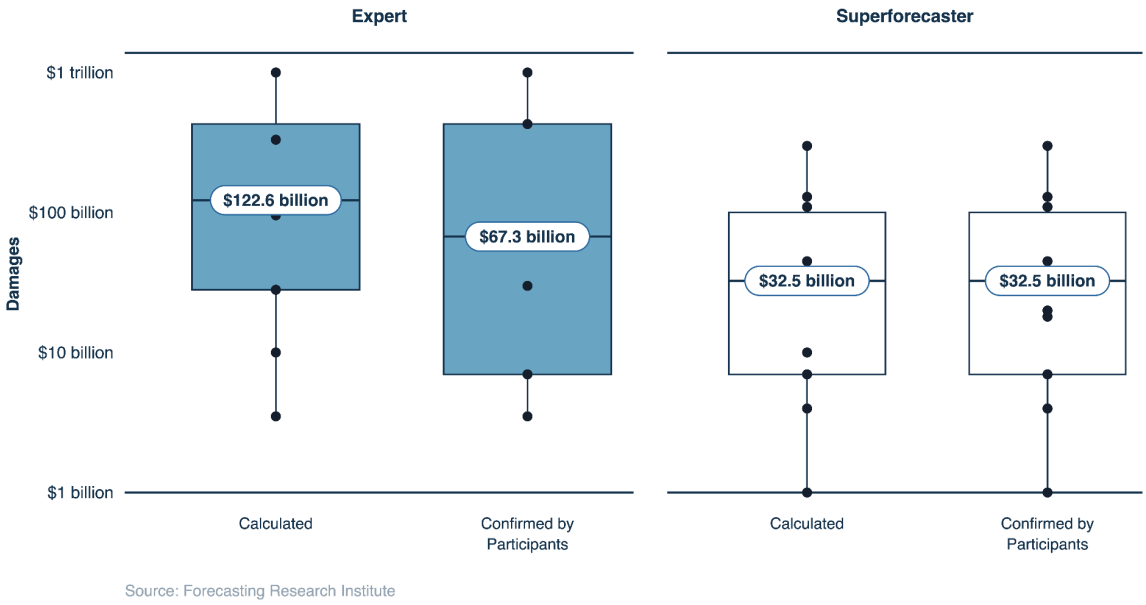
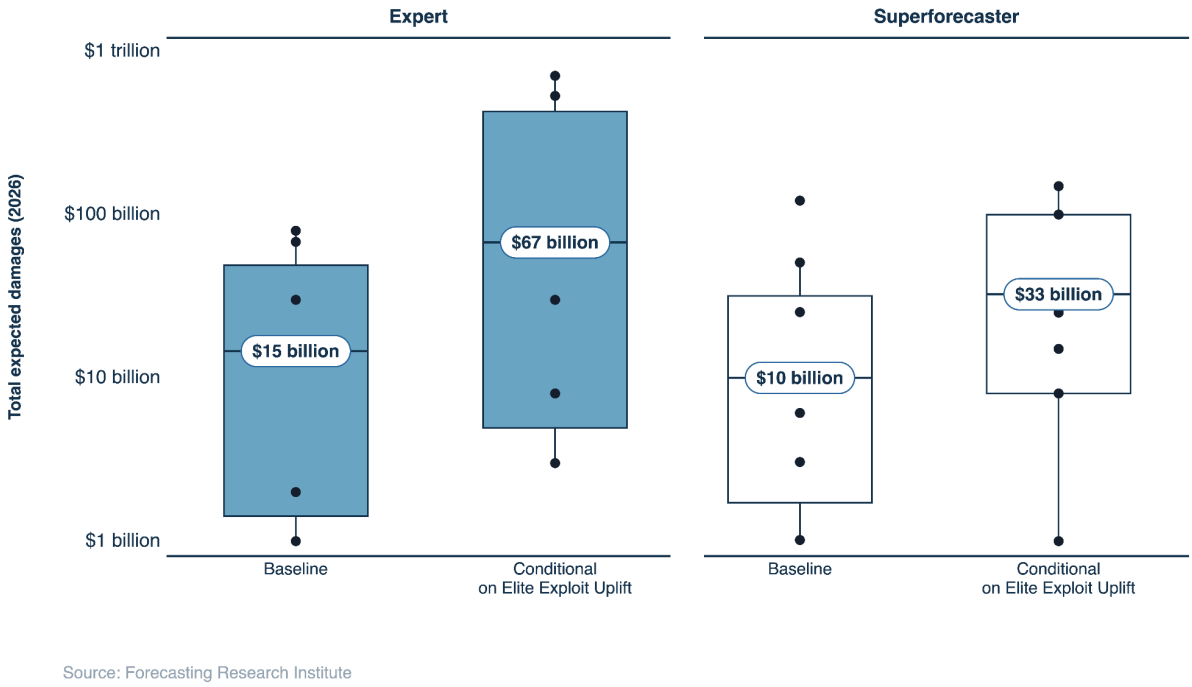


Figure 6.9: Calculated and participant-confirmed expected damages due to data-damaging worms 2026, conditional on Elite Exploit Uplift

Figure 6.10 shows the expert and forecaster estimates on the same chart.



Source: Forecasting Research Institute

Figure 6.10. Total expected damages due to data-damaging worm attacks in 2026

Using the participants' confirmed expected damages estimates, Elite Exploit Uplift is associated with a 2–5x increase in expected damages, from approximately $15 billion to $67 billion for the median expert forecast and $10 billion to $33 billion for the median superforecaster. Again, this suggests that respondents believe that elite exploits are a key bottleneck to data-damaging worm attacks.

6.5. The Effects of Different Risk Mitigation Policies

The scenarios above assume that models are released open weight without any deployment safeguards, as defined below. We also consider the effects of two alternative policies, P1 and P2.

P0: Open weight with no refusals. The model with elite exploit capabilities is open weight and is released without deployment safeguards or other risk mitigations.

P1: Proprietary models with refusals and anti-jailbreak measures. Frontier AI models are proprietary and require users to access them via APIs with deployment and security safeguards. Companies train the models to refuse requests to find vulnerabilities or develop exploits and to have protections against jailbreaks. The AI models are assumed to be protected by SL-2 level infosecurity, which can likely thwart moderate effort attempts by individual hackers to steal the model's weights (Nevo et al. 2024).

P2: Temporary protections with early access for defenders. The public release of the AI model has P1 level safeguards in place.[87] However, a set of cyber defenders is given access to a version of the model without refusals, i.e. with full vulnerability discovery and exploit

[87] For discussion of a similar idea, see Ee et al 'Asymmetry by Design: Boosting Cyber Defenders with Differential Access to AI' (2025).

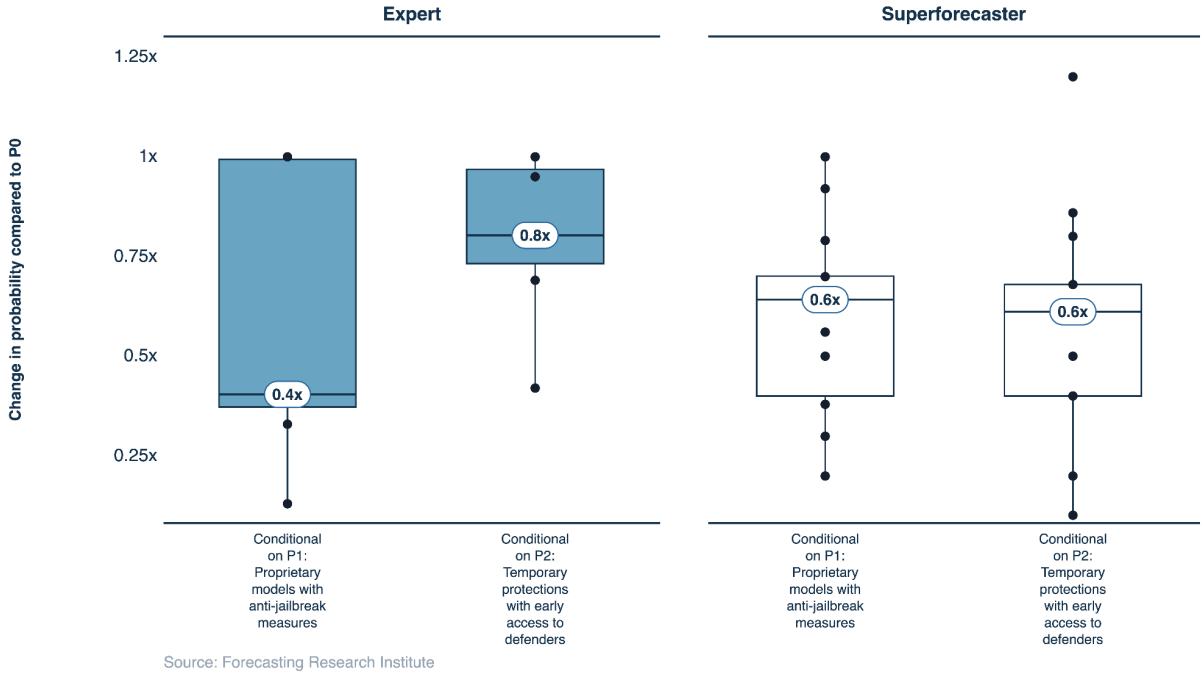
development capabilities. The cyber defenders include four major software companies – Microsoft, Meta, Apple, and Google – and three highly vetted bug bounty organizations – Synack Red Team, Cobalt Core, and HackerOne Clear. After four months, the model is released under P0 security, i.e., is open weight.[88]

We define these different policies in more detail in Appendix A.10.

These three policies each have different advantages and disadvantages. P0 – open weight models with no refusals – allows both attackers and defenders to benefit in full from AI vulnerability discovery and exploit development capabilities. P1 – proprietary models with refusals and anti-jailbreak measures – makes it harder for both attackers and defenders to benefit from these capabilities. Finally, P2 – temporary protections with early access for defenders – aims, over a short period, to limit benefits to attackers and provide benefits to defenders.

Which policy is optimal depends in part on offense-defense balance in this domain. Since it is unclear how AI uplift in vulnerability discovery and exploit development will affect offense-defense balance over time, it is also unclear which of these policies would be optimal.[89]

In the first wave pilot survey, we surveyed experts on the effect these three policies would have on the risk of data-damaging worms, assuming that what we call Elite Exploit Uplift is met. The results are shown in Figure 6.11.



[88] P0 and P2 could be combined with subsidies for vulnerability discovery. More research is needed on what level of funding would be optimal, but for context, Apple, Google and Microsoft collectively pay on the order of tens of millions of dollars each for bug bounties (Google 2025; Apple 2022, Microsoft 2025).

[89] For AI developers managing misuse risks, there may be other reasons, independent of offense-defense effects, to focus on damages only up to 6-12 months after release (see Appendix A.11).

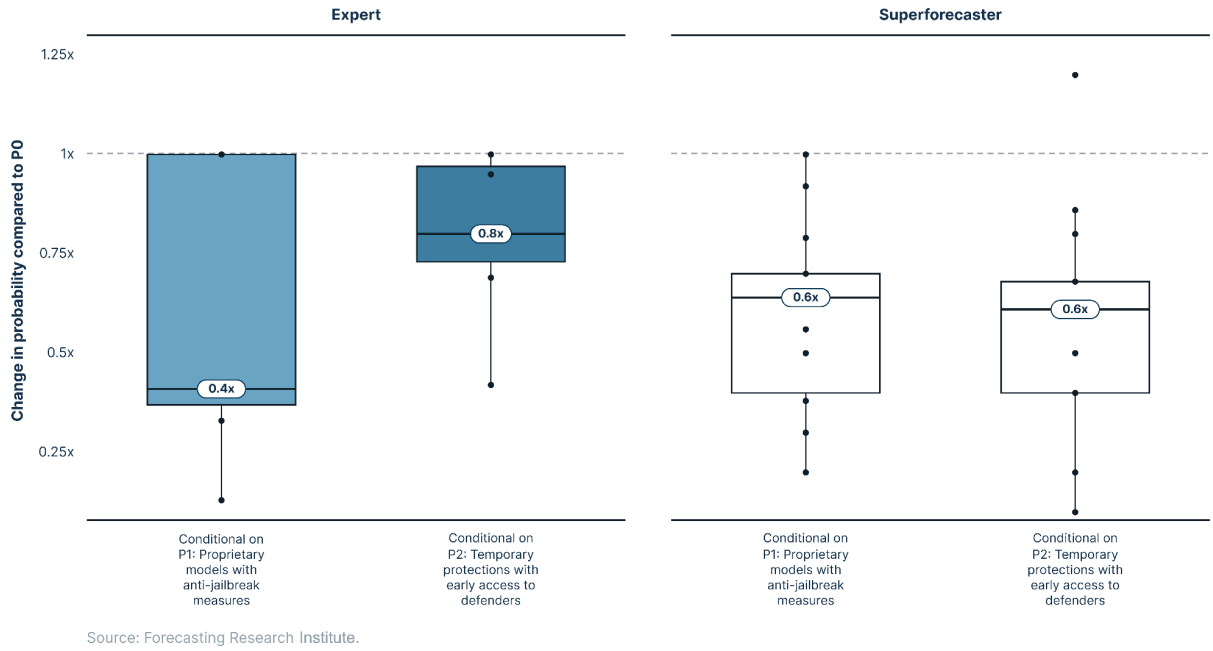


Figure 6.11. Relative risk of a >$10B data-damaging worm attack compared to P0, conditional on AI elite exploit uplift.

As Figure 6.11 shows, the median experts and superforecaster thought that the two policies would reduce risk compared to open weighting of models, though some thought that the effect would be small, and there was high disagreement.

Respondents generally thought that P1 would have the greatest impact on TA1/2 actors, as more sophisticated threat actors could get past protections. API-based access was thought to help with identifying suspicious patterns, but it was noted that it would still be hard to differentiate malicious actors from legitimate researchers. Many participants seemed concerned about sophisticated actors stealing model weights, removing safety measures, and offering elite exploits as a kind of service – and noted that API-based access doesn't fully address this risk. Some respondents thought that red-teaming requirements can sometimes be "safety-washing" and suggested this protection would likely be inefficient. Some respondents noted that jailbreak measures provide limited protection.

In rationales for forecasts relating to P2, participants often suggested that four months might be an insufficient head start for defenders, as defenders must guard a large attack surface and there may be many vulnerabilities to patch, making prioritization difficult. Some respondents argued that the ultimate open-weight release negates the benefits of defenders having early access, as it still enables long-term use of the models for developing attacks.

Giving defenders early access to models or open sourcing models, and incentivizing vulnerability discovery, might provide better information about model capabilities than traditional task-based evaluations.

Overall, more research on the merits of these different approaches is needed.

6.6. Current and Future Model Capabilities

This report has tried to estimate the economic costs if AI gains the ability to develop elite exploits. Models are not currently able to develop elite exploits, including both finding critical vulnerabilities and writing elite exploits of them. However, the cyber capabilities of models are improving rapidly.

Frontier models' vulnerability discovery capabilities in particular have improved rapidly (International AI Safety Report, 2026, pp. 58–59). In 2024, Google's Project Zero (2024) claimed that an LLM agent had found what they believe "is the first public example of an AI agent finding a previously unknown exploitable memory-safety issue in widely used real-world software". They note that "it's likely that a target-specific fuzzer [a commonly used tool for finding vulnerabilities] would be at least as effective (at finding vulnerabilities)." It is also unclear how exploitable this vulnerability is, and it seems to fall well short of being an elite exploit.

AI vulnerability discovery capabilities improved rapidly through 2025 and into 2026.

- In June 2025, XBOW became the first AI system to top HackerOne's US bug bounty leaderboard and was top globally by August (Waisman 2025a; Waisman 2025b)
- In February 2026, Anthropic's Frontier Red Team published research showing Claude Opus 4.6 found and validated 500+ high-severity zero-day vulnerabilities in open-source software (Anthropic 2026a).
- In a two-week collaboration with Mozilla, Claude 4.6 identified 22 unique security flaws in Firefox, 14 classified as high-severity, representing roughly 20% of all high-severity Firefox flaws remediated in the prior year (Baran 2026).
- In April 2026, Anthropic's Frontier Red Team previewed Claude Mythos, which it described as a step change in cyber capability. In pre-release testing, Mythos autonomously discovered thousands of previously unknown zero-day vulnerabilities across every major operating system and web browser, including bugs that had survived decades of human and automated review, among them a 27-year-old flaw in OpenBSD (Anthropic 2026c).

Until early 2026, AI exploit development capabilities had lagged vulnerability discovery capabilities. For the Firefox vulnerabilities discovered by Claude Opus 4.6, mentioned above, Anthropic tasked the model with developing functional exploits for the discovered bugs to read and write local files on a target system (Anthropic 2026b). After several hundred attempts costing roughly $4,000 in API credits, the model only successfully generated working exploits in two instances. Moreover, the exploits Claude wrote only worked on Anthropic's testing environment, which intentionally removed some of the security features found in modern browsers, such as the sandbox (Anthropic 2026b).

With the development of Claude Mythos in early 2026 (around the time this report was being finished) however, exploit development capabilities improved dramatically. Tasked with the same

Firefox vulnerabilities – on which Claude Opus 4.6 had produced working exploits only twice in several hundred attempts – Mythos generated working exploits 181 times and achieved register control on 29 more (Anthropic 2026c). Anthropic also reports that Mythos can weaponize newly disclosed (N-day) vulnerabilities within hours rather than the days or weeks this has historically taken; in one test it generated proof-of-concept exploits for 18 of 21 Windows kernel vulnerabilities disclosed in a two-month span, producing the first within 31 minutes of the patch becoming available (Markovic 2026).

We have not investigated the cyber capabilities of mid-2026 AI models in depth, so we are unsure about their overall capabilities. Anthropic's red team has reported discovering powerful exploits, which have many, though not all, of the features of elite exploits. These include an unauthenticated, zero-click, root-level remote code execution chain against FreeBSD's NFS service and eight privilege-escalation chains to SYSTEM on Windows, all on fully hardened systems (Anthropic 2026c). No single demonstrated exploit, however, combines all four of our elite criteria at once: the FreeBSD NFS chain is a zero-click, root-level remote code execution but against software far short of our >10-million-system reach test, as FreeBSD accounts for only a small fraction of servers worldwide (6sense 2026). These results should be read with caution: They are reported by the developer rather than independently verified, and they were produced by expert teams with very large inference budgets rather than by the moderately skilled actors.

Moreover, Mythos was not publicly released, but rather was initially gated to roughly 150 vetted defender organizations through its Project Glasswing initiative (Anthropic 2026c). Fable, a version of Mythos with strong cyber safeguards, was released publicly, but – following a US export-control directive in June 2026 – was suspended for all users.

Overall, we think it is difficult to rule out the possibility that an open-weight, safeguard-free version of a Mythos-level model would reach our Elite Exploit Uplift bar.

In the first wave of our survey, we asked experts and superforecasters to estimate when they thought Elite Exploit Uplift would be achieved. The median expert thought that Elite Exploit Uplift would be achieved in 2031, while the median forecaster thought it would be achieved in 2029 (see Figure 6.12).

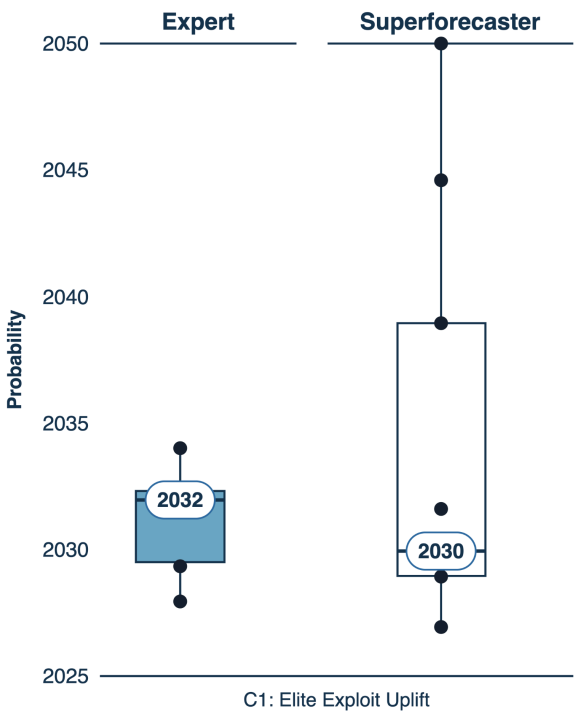


Figure 6.12: Forecasts of the year in which each capability would likely be achieved by AI

7. Conclusion

The rapid advancement of AI capabilities has prompted widespread concern about potential dual-use applications in cybersecurity, yet there is a lack of threat models explaining how AI cyber capabilities might lead to social harm and quantifying the size of the effect. This report aims to help fill this gap by exploring one AI-cyber threat model in depth.

Our analysis focused on a specific, narrow question: If future AI systems enable different threat actors to develop elite exploits, how much would this increase the economic risk from data-damaging worm attacks? To answer this, we combined case study analysis, a simple risk model, and a pilot survey of cybersecurity experts and high-performing superforecasters.

These various methods suggest that if AI gained strong elite exploit development capabilities, the expected increased damage from data-damaging worms would be in the billions of dollars. The survey results, along with our review of the empirical evidence, suggest that (1) data-damaging worms are a serious risk based on historical precedent, and (2) elite exploits are a key bottleneck to data-damaging worms. This suggests that AI companies should gather evidence on the elite exploit development capabilities of their models, and consider how to appropriately mitigate risks if such capabilities emerge. Moreover, the threat model analyzed here represents only one pathway through which elite exploit capabilities could cause harm; other pathways, such as industrial or state espionage, could impose additional costs that we do not estimate. Monitoring these capabilities is especially pressing given rapid progress in AI vulnerability discovery capabilities.

We also surveyed experts on the effects of different risk management policies, including open-weight release, proprietary models with deployment safeguards, and giving defenders early access to models. There was no clear consensus on which approach would most effectively reduce risk, though safeguards and early model access were deemed to reduce risk more than open release. This reflects genuine uncertainty about how AI vulnerability discovery and exploit development capabilities affect the offense-defense balance over time. More research on the comparative merits of different risk management approaches is warranted.

Several important caveats apply to these findings. The pilot survey had a small sample size, and the convenience sample of experts may not be representative of the broader cybersecurity community. Further, frontier models' capabilities related to elite exploits have considerably improved since we fielded the survey, most notably with the development of Anthropic's Mythos model. The underlying evidence on many of the model parameters – particularly threat actor willingness and the damages from worst-case worm attacks – is fragmentary and of limited quality. Quantitative estimates of deeply uncertain risks can convey a misleading sense of precision. We have tried to mitigate this by presenting wide confidence intervals, by complementing the risk model with direct survey estimates, and by being transparent about the limitations of the evidence. Despite these limitations, we believe that making assumptions explicit and producing order-of-magnitude estimates is

preferable to relying solely on qualitative judgment, particularly when the goal is to inform decisions about risk mitigation.

This report demonstrates the feasibility of applying structured quantitative risk assessment to AI dual-use concerns, combining risk modeling with surveys of superforecasters and experts. The approach – decomposing risk into tractable subquestions, gathering evidence on each, and eliciting expert forecasts – could be extended to other AI misuse threat models, including other cyber threats, as well as CBRN risks. We view the estimates presented here as a starting point for further work. Larger-scale expert surveys, improved data on historical cyber damages, and deeper analysis of offense-defense dynamics would all strengthen the evidence base. As AI capabilities continue to advance, grounding safety frameworks in rigorous threat modeling will be essential for ensuring that governance decisions are proportionate to risks.

Appendices

A.1. Shadow Broker Prices

The Shadow Brokers tried to sell various stolen NSA hacking tools from 2016 through to early 2017. However, it was not publicly known that they possessed the EternalBlue, EternalRomance, or DoublePulsar exploits until April 2017, when those tools leaked publicly.

- The Shadow Brokers publicly announced that they had various NSA hacking tools in 2016, and several experts suggested that the tools were legitimate NSA tools (Bisson 2017). The Shadow Brokers tried to sell all of the tools in 2016 via auction, with a stated desired price of $560M (Bisson 2017), but only received bids totaling around $1K (Vaas 2016).
- The Shadow Brokers later tried to sell individual NSA hacking tools for hundreds of thousands of dollars in January 2017. After receiving little interest, they released a cache of Windows hacking tools (not including EternalBlue, EternalRomance or DoublePulsar) for free (Bisson 2017).
- The Shadow Brokers only announced that they had the tools used in the WannaCry and NotPetya attacks – EternalBlue, EternalRomance, and DoublePulsar – in April 2017 but never tried to sell these tools and instead gave them away for free (Goodin 2017b; Goodin 2017a).

There are several reasons that the Shadow Brokers' auction failed to fetch a market price for the EternalBlue, EternalRomance, and DoublePulsar exploits. First, as noted, the Shadow Brokers never disclosed that they were in possession of the exploits until they gave them away for free in April 2017. Buyers, therefore, were bidding on exploits without knowing exactly what product they would receive. Moreover, the auction required bidders to send bitcoin to the Shadow Brokers' address with no hope of getting the bitcoin back if they did not win the auction (Greenberg 2016). Buyers were also unsure that they would actually receive the tools if they won the auction. Finally, given the public attention on the sale, buyers would also have faced significant attention from authorities for buying the exploits (Vaas 2016), further reducing demand.

A.2. Technical Details of How EternalBlue Enabled the WannaCry and NotPetya Worms

EternalBlue was used extensively in surveillance and counterterrorism operations by the NSA for at least five years (Perlroth and Shane 2019). However, the exploit could also be used in data-damaging computer worms.

How EternalBlue Worked

The SMBv1 protocol exploited by EternalBlue was first developed in 1983 (Burdova 2020) and is now widely recognized as insecure. On many systems vulnerable to EternalBlue, port 445 (over which SMBv1 communicates) was exposed to the open internet, allowing the exploit to propagate to other systems. On newer systems, port 445 is usually not exposed to the internet, and is behind a firewall. Microsoft deprecated the SMBv1 protocol in 2014 (Microsoft 2023), patches for the vulnerabilities exploited by EternalBlue have been available since 2017, and since 2017, SMBv1 is no longer installed by default in Windows systems (Microsoft 2023). EternalBlue worked as follows:

1. **Establish a connection with Port 445.**
    a. The attacker sends a specially crafted packet to the target system's port 445, which is used for SMB communications.
2. **Malicious SMB Packets:**
    a. EternalBlue sends specially crafted SMB packets that exploit a memory-handling flaw in Microsoft's SMBv1 implementation. (Nguyen et al 2024).
    b. These packets are designed to overflow a buffer in the SMB service, specifically within the function that handles the SMB transaction requests.
3. **Triggering the Buffer Overflow:**
    a. When the target system receives these specially crafted packets, it overflows the allocated buffer in kernel memory (Nguyen et al 2024).
    b. The buffer overflow allows EternalBlue to overwrite critical memory regions.
4. **Executing Arbitrary Code:**
    a. A system module in Windows, the Hardware Abstraction Layer, uses a heap with a fixed address (Nguyen et al 2024).
    b. EternalBlue places malicious payloads at predictable locations in memory.
    c. Once the buffer overflow is triggered, the attacker can write binary code that is executable on the heap of the Hardware Abstraction Layer (Nguyen et al 2024).
5. **System privileges:**
    a. Since the SMB service runs with system privileges (the highest level of privileges on a Windows system), the injected shellcode also executes with these high privileges, giving attackers full control over the compromised system, without the need for authentication.

A.3. Elite Exploit Prices

There are various different markets for exploits. The data we have on exploit prices in those markets suggest that the market price of elite exploits is on the order of $10M. Elite exploits are often sold to state intelligence agencies, suggesting that they are difficult to develop in-house, even for states.

Prices on the Gray Market

Gray-market brokers like Zerodium and Crowdfense buy the code for exploits from individuals, groups, or companies and then sell them, usually to governments (Google, Buying Spying, 2024). Elite exploits sell on these platforms for $1M to $9M. Due to information asymmetries between buyers and sellers, broker markets are inefficient and subject to adverse selection – or the "market for lemons".

> "The zero-day exploit market is a market with extreme information asymmetries. The seller has much more information about whether the exploit is actually working. The market is also flooded with lemons. Many of the exploits offered are a lot less reliable than sellers initially report. Also, the buyer of an exploit is not always able to test the exploit before purchasing it, as the economic value would be lost once given to the buyer for 'testing.' This structural setup makes even beneficial zero-day transactions difficult." (Smeets 2022).

If buyers are unable to tell the difference between strong and weak exploits, they would be unwilling to pay high prices. Consequently, the price is bound to be lower than what sellers of high-quality exploits would sell for, driving them out of the market.

Prices on exploit broker platforms have increased in recent years. In 2019, the highest bounty offered by Crowdfense was $3M (Franceschi-Bicchierai 2024), whereas today the highest is $9M. Many cyber experts attribute this to companies hardening their products against hackers (Franceschi-Bicchierai 2024).

Commercial Surveillance Vendors

Commercial surveillance vendors like NSO and Intellexa typically charge hundreds of thousands of dollars per device infected by elite exploits and the spyware payload installed by those exploits.[90] For a single customer, the basic package usually includes 10–20 devices each using elite exploits. This suggests that the value of the elite exploits and the spyware for an individual customer is on the order of millions of dollars. The commercial surveillance vendors would also have numerous customers for the same product, which suggests that the value of individual elite exploits plus the spyware is at least tens of millions of dollars.

It is difficult to know what fraction of the value of the product comes from the elite exploits that allow full control of the device and what fraction from the spyware payload. Still, it seems plausible that the value of the exploits alone is well in excess of $1M and likely on the order of $10M.

[90] Commercial Surveillance Vendor product prices are collected together in: Exploit prices from commercial surveillance vendors

Moreover, these tools are sold to some state intelligence agencies (Google Threat Analysis Group, Buying Spying, 2024), which suggests that they are difficult to develop even for state intelligence agencies.

Bug Bounties

Apple offers $100K to $1M for zero-click remote access exploit chains with full kernel execution and persistence (i.e. the exploit continues to work after the device has been rebooted) on a single recently released device. Google offers $1M for a zero-click RCE with persistence exploit that is effective against all vulnerable builds and models of Pixel Titan M. The price for vulnerabilities affecting a large number of distinct systems would be much higher. As noted in the main report, WannaCry was effective against 400M systems at the time of the attack. This suggests that in today's prices, EternalBlue would be worth millions of dollars, and plausibly on the order of $10M.

In 2016, Google's Project Zero ran a 6-month public competition in which hackers, knowing only devices' phone number and email address, had to find a full exploit chain that achieves remote code execution on multiple Android devices and provides access to third-party application files in internal storage. This would qualify as an elite exploit. First prize offered $200K, second prize $100K, and third prize $50K (Silvanovich 2016). There were no viable entries for the competition (Silvanovich 2017). This is directional evidence that finding such exploits is difficult and suggests that the cost of finding such exploits is at least tens of thousands of dollars.[91]

A.4. FORCEDENTRY Elite Exploit and Possible Worm Application

Over the last few years, many of the elite zero-day exploits discovered being used in cyberattacks were developed by NSO Group, an Israeli commercial surveillance vendor.[92] NSO's FORCEDENTRY exploit chain illustrates the complexity of newer elite exploits. FORCEDENTRY could be used to install the Pegasus spyware onto an iPhone. Pegasus is a multi-purpose surveillance tool, allowing access to messages and files, emails, contacts, the microphone and camera, location data, and the exfiltration of data, among other things.

The FORCEDENTRY exploit could install Pegasus on iPhones without any interaction from the user (Google Threat Analysis Group, Buying Spying, 2024, pp. 36–37; Scott-Railton 2023). It gains initial access with a remote code execution vulnerability in iMessage, a user space application. Every iOS app runs in its own sandbox. Even if an attacker manages to compromise an app like iMessage, the malicious code is confined to the sandbox and cannot directly affect other apps or the core system. Consequently, FORCEDENTRY also included a sandbox escape exploit, as well as a local privilege escalation exploit that was required to give full root access to the device, allowing the malware to

[91] The value of the prize money should be discounted by: the *ex ante* chance of winning the competition; risk aversion; and the opportunity cost of entering the competition.

[92] Many other commercial surveillance vendors develop 1-click exploits, which require some interaction on the part of the user (Google Threat Analysis Group, *Buying Spying*, 2024). 1-click exploits are less suited to fast spreading worms for this reason.

install the Pegasus spyware (Google Threat Analysis Group, Buying Spying, 2024, p. 36). The figure below summarizes the FORCEDENTRY attack chain.

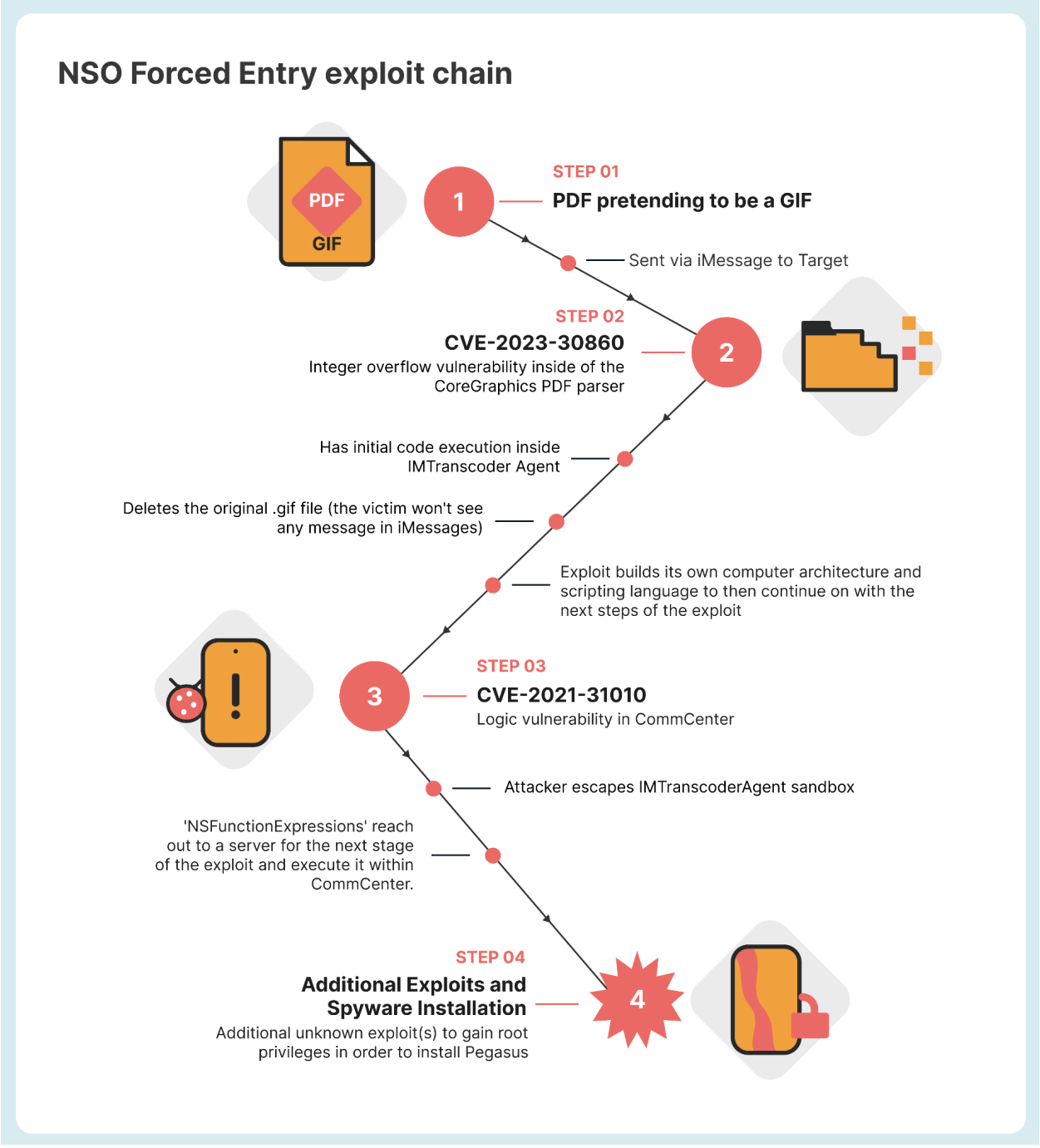


Figure A.1. NSO FORCEDENTRY exploit chain

In contrast, EternalBlue directly exploits an SMB vulnerability at the kernel level, which operates outside of the typical user-space sandbox. The table below summarizes the difference between EternalBlue and the zero-click user-space exploit used for initial access in FORCEDENTRY.

Feature	EternalBlue (SMB RCE)	Zero-Click User-Space Exploits (e.g., iMessage, browsers)
Target Level	Kernel-level	User-space applications
Sandbox Involved?	No	Yes
Privileges Gained	SYSTEM	Initially limited, requires escalation
Interaction Required	None (remote exploit)	None (zero-click), but needs sandbox escape

Table A.1. Differences between EternalBlue and user-space iMessage exploits

A data-damaging worm using FORCEDENTRY could proceed as follows:

Step 1: Initial Infection via FORCEDENTRY

- An attacker uses a FORCEDENTRY to remotely compromise a device via iMessage.
- The payload could install a malicious program that runs with elevated privileges and gains complete control of the device, gaining access to the device's contacts, messaging apps, and network information.

Step 2: Self-Replication Mechanism

- Once the device is compromised, the worm could scan for other vulnerable targets.
- It could leverage FORCEDENTRY to send malicious exploits to contacts using iMessage or compromise devices on the same network.
- The worm would continue to replicate itself on each new infected device, spreading exponentially.

Step 3: Data Encryption

- The worm could wipe or encrypt data on the device.

However, it is plausible that a worm propagating in this way would be noticed by modern cybersecurity systems as devices sending out messages in bulk would be flagged as suspicious activity and stopped.

A.5. Elite Exploits Discovered Being Used in Cyberattacks Since 2020

Google's Project Zero maintains a Google sheet of zero-day exploits publicly discovered being used in cyberattacks since 2014. The sheet does not specify whether any of the exploits were elite-level. I

used Claude Opus 4.8 High to analyze which of the exploits found in the wild between 2020 and 2026 were plausibly elite level. This analysis is preliminary and has not been externally validated.

The large majority of the 287 zero-days do not meet the elite definition. Most are client-side exploits of web browsers, office suites, or document readers that require the victim to open a file or visit a website, and so are not zero-click; local privilege-escalation, sandbox-escape, or driver and kernel bugs that form one link in a longer chain rather than providing remote initial access; or information-disclosure issues.

Setting these aside leaves two families of elite exploits.

The first family of elite exploits discovered in the wild includes zero-click exploit chains against widely used mobile and messaging software (iOS, Android and WhatsApp, each running on well over 10 million devices), almost all developed by commercial spyware vendors to deploy spyware:

1. 2021: FORCEDENTRY, a zero-click iMessage chain on iOS, developed by the commercial spyware vendor NSO Group (Marczak et al. 2021).

2. 2023: BLASTPASS, a further NSO Pegasus zero-click iMessage chain (Scott-Railton 2023); and Operation Triangulation, a zero-click iMessage chain that reached the iOS kernel, attributed to an unidentified nation-state actor (Larin 2023).

3. 2025: a marked increase, with at least four distinct zero-click chains – a CoreAudio chain (Zorz 2025); a WhatsApp and ImageIO chain (Lakshmanan 2025b); the Paragon Graphite chain via Messages, developed by the commercial spyware vendor Paragon (Lakshmanan 2025a); and a zero-click image-parsing exploit of Samsung devices (Lakshmanan 2025c).

The second family includes unauthenticated server-side remote code execution against mass-deployed Microsoft enterprise software. These are zero-click in that they require no action by any user – the attacker reaches the exposed server directly – and were exploited in the wild by nation-state actors:

4. 2021: the Microsoft Exchange ProxyLogon chain. The vulnerabilities were found by Orange Tsai's team (Orange Tsai 2021). The chain – which Microsoft assessed as effectively wormable – was exploited at scale against tens of thousands of organizations by Hafnium, a Chinese state-sponsored group (CISA 2021).

5. 2025: the ToolShell exploit affecting Microsoft SharePoint, an unauthenticated server takeover mass-exploited by China-nexus actors (Unit 42 2025).

Between 2020 and 2024, four elite exploits, about one per year, were discovered being exploited in the wild. However, the rate then rose sharply, with five being used in 2025 alone.

Elite exploits discovered in the wild were developed only by commercial spyware vendors (comparable to TA3 or TA4, such as NSO and Paragon) and by nation-states (TA5); none is attributable to a TA1 or TA2 actor. This is consistent with the report's assessment that such actors

are very unlikely to be able to develop elite exploits and with the hypothesis that elite-exploit development is a key bottleneck.

A sheet collating this analysis is available here.

A.6. Motivations and Functionality of Past Worm Attacks

This section discusses details of the past worm attacks in the Johansmeyer dataset. Table A.2 outlines the number of devices infected by each worm and the source for this estimate.

Attack name	# systems infected	Source
Melissa	~100K	GAO 1999
ILOVEYOU	~50M	Zurier 2020
Klez	~7M	Gerencer 2020
CodeRed	~360K	Moore et al. 2002
Nimda	>1.3M	Lemos 2001
SirCam	~2.3M	Delio 2001
SoBig	>1M	Al Jazeera 2003
SQL Slammer	>75K	Moore et al. 2003
Swen	~1.5M	Johanns 2021
Mimail	~21K	Skelly 2003
Yaha	?	No infection # found
MyDoom	~500K	Sullivan 2004a
Sasser	~500K–1M	Sullivan 2004b; Guardian 2004
Storm Worm	1M–50M	Schneier 2007
Conficker	~10M	Arthur 2009
WannaCry	~230K	Cooper 2018
NotPetya	~670K	See fn[93]

Table A.2. Number of systems infected by past significant worms

[93] For NotPetya, according to one source, there were 500K infections in Ukraine alone (Maschmeyer 2021). 75% of infections were in Ukraine (ESET 2017), which implies ~670K infections globally.

Worms with an Unknown Creator

The CodeRed worm exploited a buffer overflow vulnerability in Windows servers. Though it did not destroy data on infected systems, the worm defaced sites with the phrase "Hacked by Chinese!" and launched denial of service attacks (Kaspersky 2022, Dolak 2005). A patch had been available for the worm a month prior to its launch (Microsoft 2001; Dolak 2005). CodeRed was a network worm that spread without user interaction and allowed remote code execution with system-level privileges (GAO 2001; GIAC 2005). However, because it relied on a single server-side vulnerability and reached only the few hundred thousand exposed servers – far fewer than 10 million systems (Moore et al. 2002), the CodeRed exploit does not count as "elite" per our definition. There is evidence that the worm was launched from a Chinese university, though the perpetrator remains unknown (Moore et al. 2002).

The Nimda worm targeted Windows systems and spread via email, web servers, web browsers, and shared network drives (Holick 2002). Nimda could spread without user interaction, allowed remote code execution, and granted administrator-level privileges (Dean 2001; Holick 2002). Since it targeted software in near-universal use – Windows, Outlook, IIS, and Internet Explorer, the dominant browser of the era at roughly 90% market share (Bekker 2002) – far more than 10 million systems were exposed, so the exploit counts as elite per our definition. Nimda primarily caused damage via slowing down networks, but it also gave attackers elevated privileges on infected systems (Holick 2002). It is unclear what the motivations for the attack were (Holick 2002).

The Swen worm targeted Windows systems and replicated via email, the local network, IRC, and file sharing websites (Microsoft 2005). It used a vulnerability in Internet Explorer to execute directly from email (F-Secure, nd). The worm required the user to open an attachment so was not zero-click and therefore did not involve an elite exploit (Microsoft 2005; F-Secure, nd). The worm aimed to terminate the processes of antivirus software and firewalls, making systems vulnerable to other malware, but did not attempt to deploy any further destructive payloads (F-Secure, nd). It could also potentially cause damage by creating high network traffic (Comodo 2019).

The Mimail worm spread via infected email attachments (F-Secure) and therefore did not use an elite exploit. It appears to have been launched for financial reasons. It was designed to infect a large number of systems to launch denial of service attacks against anti-spam organizations and may have been launched or funded by spam groups (Wired 2003). Some variants attempted to steal credit card information (F-Secure).

Worms Launched by TA1 and TA2 Actors

The **Melissa** worm caused damage by creating large amounts of network traffic and did not contain a malicious payload. The worm required users to click a link in order to spread (FBI nd) and therefore did not involve a zero-click elite exploit. The perpetrator claimed that the worm was not designed to do as much damage as possible and that the damage caused was accidental (Register 2001; Lemos 1999), though it is unclear if this is true.

The **ILOVEYOU** worm spread primarily via email and required users to open an attachment (Root 2022b) and therefore did not use an elite exploit. The worm irretrievably corrupted documents on infected systems and sent passwords to the creator. Later variants completely wiped the hard drive (Root 2022b). The attacker apparently released it to prove his own hacking skill (Landler 2000).

The **SirCam** worm targeted Windows, spread via email, and required users to download an attachment with malicious code (Becker 2001) and therefore did not use an elite exploit. Because it attached actual user files to emails, the worm risked exposing confidential information. Its payload also attempted to delete all files on a system in certain conditions (Becker 2001) and fill up the drive where Windows is installed, though this did not work in practice due to a bug in the code (F-Secure). It is unclear what the motivations for the attack were (Becker 2001).

Sasser caused damage by creating a large amount of network traffic, which caused many companies to shut down their systems (Sullivan 2004b; Macrae 2014). The exploit used by Sasser should plausibly be classed as elite. The worm could spread by exploiting a vulnerable network port in the operating system without user intervention, but it could be stopped by a properly configured firewall or by downloading system updates from Windows Update (Tech Monitor 2014). The vulnerability exploited also allowed remote code execution with system level privileges (Microsoft 2004). Sasser infected 500K to 1M systems and was launched after the patch for the vulnerability was released. If it had been released prior to the patch release, likely many more systems would have been vulnerable. Authorities suspected that the author released it to prove his hacking skill (NBC 2005).

Klez was a worm that caused damage by creating a large amount of email traffic, disabling antivirus and security software, and potentially by sharing confidential information in attachments to emails (Microsoft 2007). The virus required users to download an email attachment and therefore did not use an elite exploit. It was only effective against systems that had failed to install a patch that had been available for a year (Delio 2002). The perpetrator and motivations are unknown.

SoBig was an email worm that used infected devices to distribute spam, apparently for financial gain (Sullivan 2003). The worm caused damage by creating large amounts of network traffic, causing company systems to slow down or stop (CNN 2003). The worm required users to open an attachment in an email (Sullivan 2003) and therefore did not use an elite exploit.

SQL Slammer did not contain a directly destructive payload, but caused damage by creating large amounts of network traffic (Moore et al. 2003; Litchfield 2010). A patch for the vulnerability exploited by SQL Slammer had been available for six months before the worm was released (Wired 2003). It is unclear what the motivations were for the attack. SQL Slammer did not require user interaction to spread, but the worm did not use an elite exploit because it was effective against far fewer than 10 million devices. It infected 90% of vulnerable hosts within ten minutes, and infected at least 75,000 devices (Moore et al. 2003).

MyDoom was an email worm that tried to create a large botnet of infected computers to launch denial-of-service attacks against specific companies, including SCO and Microsoft, while also

slowing down infected machines (Okta 2024). Since the worm required users to download an infected attachment to spread (Okta 2024), it did not use an elite exploit. Some argue that the worm was launched by disgruntled members of the open source software community, but the worm may also have been launched for financial reasons (WIRED 2004; Al Jazeera 2004; Page 2004; Gonsalves 2004).

Worms Launched by TA3 Actors

Yaha was released by the Indian Snakes, a group of Indian hackers in order to retaliate against Pakistani hackers who had defaced Indian websites. The worm attempted to launch denial-of-service attacks against Pakistani websites (F-Secure nd; Help Net Security 2002; Wired 2003). In order to spread, it required users to click infected links (F-Secure nd) so did not use an elite exploit.

Storm Worm was released by Russian Business Network, a Russian cybercrime group. Storm Worm charged a fee for denial-of-service attacks against anti-spam websites and security vendors (Hypr nd). Storm Worm spread via email and social engineering (Hypr nd) so did not use a zero-click elite exploit.

Conficker was produced by a Ukrainian cybercriminal group. It infected millions of computers but was never used for significant malicious attacks due to concern about criminal repercussions (Bowden 2019). Conficker was zero-click, spread via sharing across networks, allowed remote code execution (Aben 2009; CISA 2013) with system privileges (Microsoft 2008), and was plausibly effective against more than 10 million systems, qualifying it as an elite exploit per our definition.

A.7. A Critique of Estimates of Past Cyber Damage Estimates

The figure below shows damages in each year from worm attacks only, according to the Johansmeyer (2024) data.[94]

[94] As noted in the main report, this excludes Stuxnet.

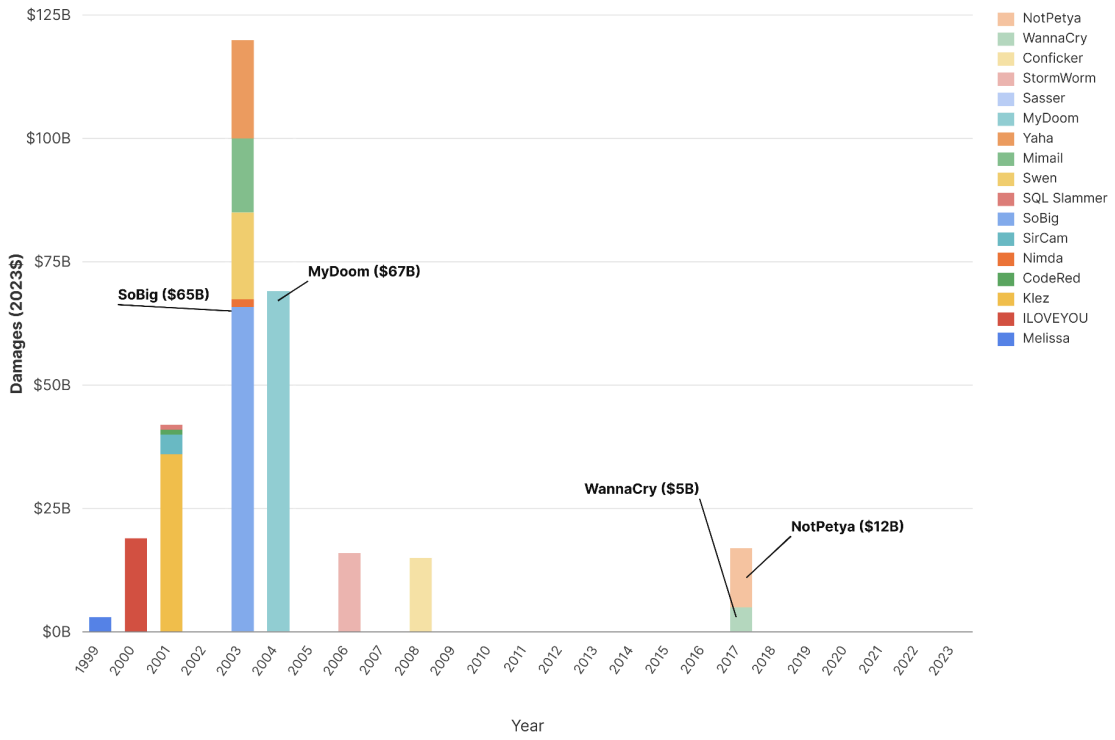


Figure A.2. Damages from major worm attacks in Johansmeyer dataset (1999–2023)

Johansmeyer notes that the provenance of this data is poor.

> "To arrive at 21 events since 1998, I pulled data on historical economic losses from cyber catastrophes from publicly available sources, which is the primary limitation of the study. Unfortunately, many estimates come from popular media sites and corporate blogs... Methodological information for the publicly available estimates is virtually non-existent, and some sites presumably reference (but don't link to) long-gone sources. Essentially, I'm relying on judgment and comparison of original estimates without a benchmark to make the best of a bad situation. That said, many of the estimates appear to have been recycled and republished, which offers at least a veneer of respectability." (Johansmeyer nd)

Moreover, Johansmeyer (2023), fn 1 notes that where there are multiple sources, he chooses the higher estimate. (This is because Johansmeyer's aim is in part to show that even on the high estimates, cyber risk is lower than many people argue.) For instance, some of the estimates come from a November 2003 House Committee Hearing on computer viruses. In that hearing, different speakers and sources gave damage estimates for SoBig ranging from $500M to $30B. One source used in the dataset reports that the Melissa worm caused >$1B in damage (in 2012 dollars) (Beattie 2012), but in the plea agreement, the Melissa perpetrator admitted to causing "over $80M" in damage (The Register 2001).

The original source for many of the very large pre-2005 estimates is the computer security firm Mi2g, whose damage estimates of cyberattacks have been criticized as greatly inflated (Leyden, 2002; Gallaher et al. 2006). Gallaher et al. 2006 argues that Mi2g had a financial incentive to produce inflated estimates in order to attract media and customer attention. Mi2g used a proprietary model that cannot be publicly scrutinized. The Mi2g data implies that between 1999 and 2003, there was a 46X increase in the damages from significant cyber attacks.

In the same period, the CSI/FBI cybercrime survey found a 1.6X increase (Gallaher et al. 2006 sec. 2.2.). The CSI/FBI Computer Crime and Security Survey represents the responses of hundreds of IT professionals in U.S. corporations, financial institutions, government agencies (federal, state, and local), medical institutions, and universities. Participants were surveyed to determine the spending of their organizations on cybersecurity, the number of breaches and the associated financial losses incurred during the previous year, and the preventative activities undertaken. The results of the survey for 1997 to 2005 are shown below:

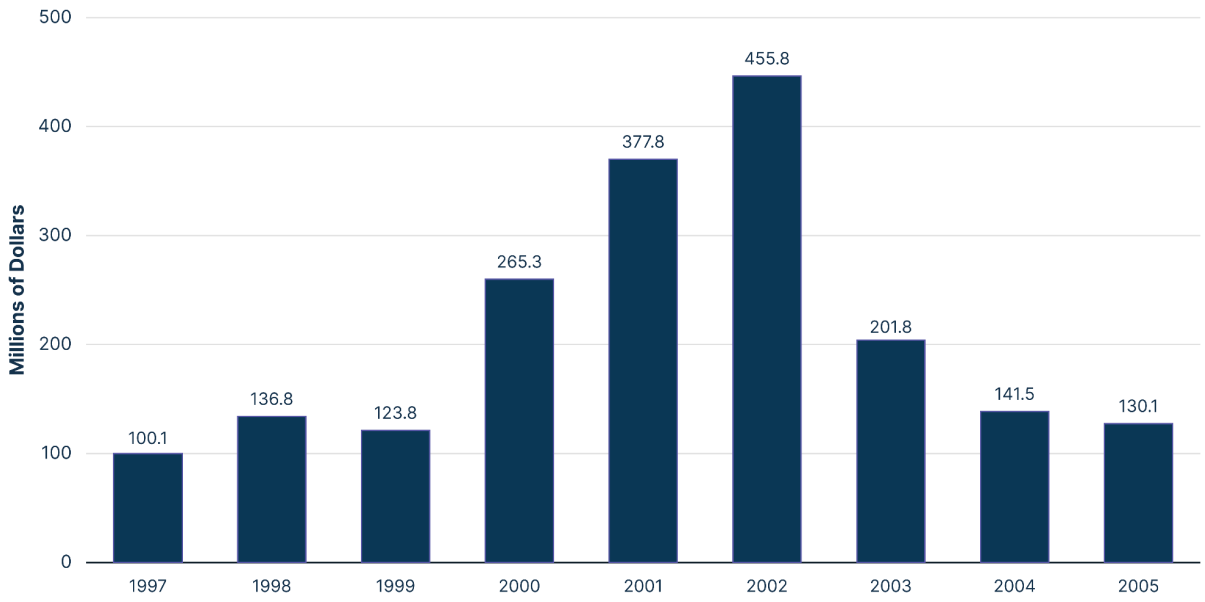


Figure A.3. Costs of cybercrime according to the CSI/FBI cybercrime survey (1997–2005). ***Source:*** ***Gallaher et al. 2006 p. 19.***

The CSI/FBI survey is widely referenced by academics, government agencies, and companies providing security-related products or services. However, the authors recognize that the survey data is not comprehensive, as noted by a report for the Air Force Research Organization:

> "However, 20 percent of the responding organizations acknowledged that they do not report all computer intrusions to law enforcement because of the high cost of doing so. Furthermore, cost-estimating procedures are not uniform; capturing labor resources allocated to security or employee productivity loss is not easy and is not always consistent.

> Thus, the authors acknowledge that the information garnered by this survey, while accurate as reported by respondents, should not be considered a complete accounting of the costs of cyber security." (Gallaher et al. 2006 p. 19)

Thus, the CSI/FBI survey is not a good estimate of the absolute costs of cyberattacks. However, it seems more reliable than Mi2g with respect to the change in costs over time. If we assume that the 1999 damage estimate in Johansmeyer's data is correct and use the 1.6X increase ratio implied by the CSI/FBI data, then the cost of significant cyber attacks in 2003 would be $4B, rather than the $120B implied by the Mi2g data.

Overall, we would not be surprised if the pre-2005 damage estimates were too high by one or more orders of magnitude.

Estimates of WannaCry and NotPetya in the Johansmeyer Dataset

The damage estimates for WannaCry and NotPetya in the Johansmeyer dataset also seem unreliable. Different sources give different estimates for the costs of WannaCry, with some claiming that the costs were $4B (2017$) (Johansmeyer (2024)), and others claiming damages of $8B (2017$) (Olano 2017). The source for both estimates is the cyber insurance company Cyence, but the authors of this report were unable to find the original source or the calculations for the estimate.

The costs of NotPetya were estimated to be at least $10B (2017$) (Greenberg 2018; Greenberg, Sandworm, p. 215).[95] The source of the NotPetya estimate is a claim by former Homeland Security Adviser Tom Bossert, who at the time of the attack was the senior-most cybersecurity official in the Trump administration (Greenberg 2018). However, no public calculations are available for this estimate.

A.8. Crosignani et al Estimate of NotPetya Costs

This section discusses the Crosignani et al (2023)[96] estimate of damages from WannaCry and NotPetya in more detail.

Crosignani et al searched for firms affected by NotPetya by (1) web scraping SEC filings in 2017 and 2018; (2) searching newspaper articles; and (3) checking against reporting in Greenberg's Sandworm book. They exclude firms in Ukraine, Russia, and "non-public firms that [they] would be unable to find in other data sets", such as government agencies and hospitals (p. 6–8). They identified 8 affected firms, including Merck, FedEx and Maersk, and collated their reported costs (Table 1).

The costs reported by the firms include (1) lost or delayed revenue and (2) remediation costs, such as analysis of IT systems, replacing and repairing equipment, and restoring services. The total reported costs across the 8 firms were $1.8B (Table 1).

[95] These figures are in today's dollars.
[96] An open access working paper version of Crosignani et al (2023) is available here. Page references in this section refer to the working paper version.

Crosignani et al (2023) also estimate the costs to upstream and downstream companies in the supply chain for the eight companies. They identify 233 customers and 320 suppliers indirectly affected by the cyberattack, i.e. exposed through their supply-chain connections to directly hit firms (p. 9). They use a difference-in-differences approach comparing the change in performance of firms indirectly affected by the shock through their supply chain with that of unaffected firms operating in the same industry, country, and size quartile in the same year (p. 11). They estimate the effects on the ratio of earnings before interest and taxes to total assets for firms affected by the cyberattack compared to the control group.

They found that NotPetya had a statistically significant effect on downstream customers of directly affected firms, but not on suppliers of affected firms (p. 16–19). They found that NotPetya led to a 1.3 percentage point drop in the ratio of earnings before interest and taxes to assets of downstream customers. They note that a conservative estimate of the supply chain effects on customers suggests a drop in profits of $7.3B (p. 18–19). Combined with the estimate of the direct costs of NotPetya, this implies total costs of $9.1B in 2017 dollars, or $11.3B in 2023 dollars.

Reasons the Crosignani Estimate May Be Biased High

Crosignani et al (2023) may overstate the costs of NotPetya for two reasons. First, our aim is to estimate the net economic damages of the NotPetya attack, but not all of the costs reported in Crosignani et al reflect net economic loss. In a competitive market, lost revenue for particular firms can be offset by rival firms gaining revenue, as consumers switch to those rival firms. Therefore, the loss to directly affected firms could be offset by gains to rival firms, and the costs to consumers could be offset by switching to competitors' products. This problem also applies to the Crosignani et al calculations of costs to customers in the supply chain. They measure the economic costs by comparing a treatment group of customers affected by NotPetya to a control group in the same industry, country, and size quartile, but not affected by NotPetya. The difference between the performance of the treatment group and the control group may in part reflect a pure loss to the treatment group, but may also in part reflect offsetting gains for the control group.

These offsetting effects are not estimated in Crosignani et al. This biases the direct cost estimates high. (However, remediation costs are pure social loss.)

Second, some of the direct costs reported in Crosignani et al refer to sales being delayed,[97] but this again is something of a soft metric, as the losses to producers and consumers could be offset by increased sales in subsequent periods.

Although these factors may bias the Crosignani et al estimate high, we do not think they would completely offset the reported damages. It is difficult to know how large the offsetting effects might be.

[97] "Various locations of the Beiersdorf pharmaceutical group were cut off from mail traffic for days. Beiersdorf said 35 million euros worth of second quarter sales were delayed to the third quarter" (Table 1).

Reasons the Crosignani Estimate May Be Biased Low

Crosignani et al may also be biased low because they do not consider the costs to companies in Ukraine, even though Ukraine accounted for 75% of infections (Eset 2017), and, according to one Ukrainian official, 5% of all private, corporate and government computers in Ukraine could not be repaired following the attack.[98]

Ukrainian Finance Minister Oleksandr Danylyuk claimed that NotPetya cost 0.5% of Ukrainian GDP (Burdyha 2017), which implies losses of $560M (in 2024 dollars). This is much smaller than the total costs estimated by Crosignani et al.

This illustrates a more fundamental problem with using money as a metric to measure social costs. Ukrainian GDP per capita in 2017 was 10–20x smaller than GDP per capita in other affected countries. Consequently, the economic costs, measured in dollars, were relatively small for Ukraine. However, given diminishing marginal returns from money to welfare, the welfare loss would have been proportionately greater in Ukraine, as the poorer you are, the more a given loss of money reduces your welfare. Even though the economic costs of NotPetya, expressed in dollars, may have been lower in Ukraine than other countries, it seems plausible that the welfare costs were concentrated in Ukraine. The problem of using dollars as a proxy for welfare is particularly acute in this case, where there are large disparities in income across victims.[99]

For this reason, ideally, the social costs of events like NotPetya would be measured in terms of their effect on welfare, rather than their effect on money (Bronsteen et al 2013).[100] Quantifying social costs in terms of money has the advantage of being widely popular and in some ways easier to calculate, but it is still subject to fundamental conceptual problems.

A.9. Offense-Defense Balance

The Norm of Openness in Cybersecurity

There is expert disagreement about the merits of openness with respect to dual-use cyber tools. Many people argue in favor of openness (e.g. Schneier 2004). Cyber defenders frequently release dual-use tools (e.g. Metasploit). For vulnerability disclosure, many key actors practice coordinated disclosure: If someone finds a vulnerability, they are expected to first inform the vendors in order to give vendors the time to patch the vulnerability, and then after a delay of a few months, publish the vulnerability publicly.

[98] "According to Deputy Head of the Presidential Administration Dmytro Shymkiv, a former head of Microsoft's Ukrainian office, about 10% of all private, corporate, and government computers in the country failed that day. About half of them are beyond repair." (Burdyha 2017)

[99] Note that this is a problem from the utilitarian point of view embodied by standard cost-benefit analysis, but also from other ethical points of view, such as egalitarianism and prioritarianism (Broome 2024).

[100] Similar problems have arisen in other domains. Early versions of the IPCC reports stated that the value of a statistical life in each country is proportional to income in that country, which implies that the value of a statistical life in poor countries is much lower than in rich countries (Broome 2024). But lower willingness to pay to reduce the risk of death in poor countries reflects the fact that poor people have less money, not that their lives are worth less intrinsically.

However, many non-malicious actors do not practice openness. Most importantly, state intelligence agencies often do not practice responsible disclosure, but instead use cyber tools in their cyberattacks, usually for espionage. The Shadow Brokers malware tools are one example of this. The NSA held on to these tools for at least five years without disclosing them publicly. The leak of the Shadow Brokers tools occurred three years after the Obama administration reformed the so-called Vulnerabilities Equities Process to err on the side of disclosing vulnerabilities to vendors rather than stockpiling them for offensive use (Healey 2016; Thompson 2021). Microsoft criticized the NSA for not disclosing the vulnerabilities sooner (Smith 2017).

When considering the merits of openness, it is important to distinguish: (1) whether openness would decrease the rate of cyberattacks; and (2) whether openness would increase social welfare. Openness may decrease the risk of cyberattacks, but it is an open question whether that would always produce better outcomes for the world. Obviously, many cyberattacks are socially costly. But state intelligence agencies value the ability to conduct cyberattacks against criminals, terrorists, and rival states, and in some cases this may be socially valuable.

It is therefore unclear what our prior should be about the merits of openness for AI-cyber capabilities. This in part depends on broader worldview judgments which may differ across software vendors, civil society groups, and state intelligence agencies.

Defining Offense-Defense Balance

Offense-defense balance in cyber is usually defined in terms of the relative costs of attack and defense (Garfinkel and Dafoe (2019); Slayton (2017)). More precisely, this is given by the ratio of the defender's investment to the minimum offensive investment that would allow the attacker to secure some expected level of success (Garfinkel and Dafoe (2019), p. 251–252). A larger ratio corresponds to an easier attack. A technology favors offense if and only if it increases this ratio.

It is important to note that, on this definition, a technology might favor offense without increasing the number of expected attacks and might favor defense without decreasing the number of expected attacks (Slayton (2017)). Whether an attack is likely to occur depends not only on the relative costs of attack and defense, but also on how attackers and defenders value their respective goals (Slayton (2017), p. 81). Even if attack is cheaper than defense at a given level of investment, attackers may simply value victory less than defenders, and so an attack might not take place.

For example, consider a scenario in which AI enables top tier state actors to find more elite vulnerabilities and exploits. Suppose that AI benefits attackers and defenders, but it is offense-favoring for worm attacks in that it increases the ratio of the defender's investment to the minimum offensive investment that would allow the attacker to successfully release a damaging worm. However, suppose also that attackers in this set are simply extremely unlikely to release a damaging worm because it does not further their goals. Since AI also benefits defenders by allowing them to more easily find and patch critical vulnerabilities, the overall effect of AI is to reduce the risk of worm attacks. This is true despite the fact that AI is, on the relative cost definition, offense-favoring with respect to worm attacks.

The likelihood of an attack also depends on the respective budgets of attackers and defenders. A defender may simply have more money than an attacker, which makes a successful attack unlikely, even if attack is cheaper than defense. Conversely, even if defense is cheaper than attack, attackers may have much larger budgets than defenders, so a successful attack may be likely.

Offense-Defense Balance Is Context-Specific

Since the relative costs of attack and defense vary depending on context, offense-defense balance is not a property of cyberspace in general, but rather a property of relationships between particular defenders and attackers (Slayton (2017), p. 74). For example, the offense-defense balance for Chinese state espionage vs. the US is very different to the offense-defense balance for ransomware groups vs. specific businesses. For the same reasons, whether a technology (like AI) favors offense with respect to worm attacks depends on exactly how the technology increases exploit discovery.

Determinants of Offense-Defense Balance

Offense-defense balance with respect to worm attacks depends on a number of uncertain factors (Lohn and Jackson (2022); Garfinkel and Dafoe (2019)):

- **Elite vulnerability discovery rates over time: the rate at which attackers and defenders find vulnerabilities over time**, which depends in part on how fast there are diminishing returns from efforts to find elite vulnerabilities and how many total elite vulnerabilities there are to discover.
- **Correlation between elite vulnerabilities discovered by attackers and defenders:** whether attackers and defenders tend to find the same vulnerabilities.
- **Elite exploit development time:** the gap between the discovery of vulnerabilities and the development of exploits for those vulnerabilities.
- **Patch development time for elite vulnerabilitie**s: the gap between the discovery of vulnerabilities and the development of a patch.
- **Patch deployment time**: the gap between the development of a patch and the deployment of the patch by users.
- **The budgets of different attackers and defenders for elite vulnerability discovery and exploit development**: How likely an attack is to succeed depends on the total resources invested by different attackers and defenders.
- **Which specific agents gain access to elite vulnerabilities and exploits:** With respect to the cyber worm threat model, it is more concerning if lower skilled actors gain access to elite vulnerabilities and exploits.
- **The quality of firewalls, intrusion detection systems, and endpoint detection and response:** Modern versions of these defensive systems can use machine learning to detect

suspicious activity and prevent initial infection or widespread propagation, even if the malware exploits 0-day vulnerabilities.

There is a lack of good data on all of these parameters, so they are all very uncertain,[101] and the parameters interact in complex ways.

A.10. Defining Different Model Release and Safeguard Policies

We consider three different policy approaches to model release and model safeguards:

1. P0: Open-weight models
2. P1: Proprietary models with refusals and anti-jailbreak measures
3. P2: Temporary protections with early access for defenders

We now define these policies in more detail, as they were described to participants in our survey.

P0: Open-Weight Frontier Models

By default, assume the following about frontier AI models:

- The **weights** of frontier AI models are **freely and publicly accessible** for anyone to modify and use. (However, in any evaluations requiring control groups without frontier AI access, those control groups are required not to use these models during the studies.)
- The scenario leaves it ambiguous about how other parts of the AI model are treated (such as whether the training code is also openly published). For a disambiguation of "open source" as a term for AI see Seger et al. (2023).

Some points to consider as you form your forecasts:

- Experts believe that having access to the model weights makes it **meaningfully easier to circumvent any safeguards** the developer introduced, compared to accessing these via an API. A key difference between proprietary models and open-weight models is that the former is behind an API. Thus, if a **vulnerability is discovered**, it can be patched, and users are no longer given access to the older version. With open-weight models, new models that have these patches can be released, but the older versions cannot easily be unpublished.
- Such **vulnerabilities include** (but are not limited to):
    a. **Overcoming the model's safety features**

[101] Lohn and Jackson (2022) estimate some related parameters. However, their data sources are relatively old, running up to around 2017, and it is likely that the trend has changed since then. The parameters may also be systematically different for the subset of vulnerabilities and exploits that we are interested in.

- AI companies train models to refuse to answer dangerous queries. However, with access to model weights, these safety features can be removed, e.g., through finetuning (i.e., giving the AI a few key examples where the "correct" answer is to answer the question).
- Although in some cases this can be done by accessing a model through a company's API (see Qi et al. 2023), most frontier models don't allow their latest models to be finetuned via their APIs or impose other restrictions on doing so (for example, see OpenAI's policy, which currently allows finetuning on GPT-4o but not o1).
- And even with an API that allows finetuning it will generally be easier to do this with unrestricted access to the model weights, as knowledge of the model's architecture may make it easier to find vulnerabilities, and there is no risk of detection through company monitoring API usage.
- Access to open-weight models has also allowed researchers to identify novel universal jailbreaks (see Zou et al. 2023) – i.e., carefully crafted questions such that the AI no longer recognizes that the developer intended for it to refuse these questions.
- Although jailbreaks now are much more sophisticated, early examples include having users add "disregard all previous instructions" in front of the prompt (Russinovich et al., 2024).

b. Enhancing the model's dangerous features.

- As well as removing safety training, with access to model weights, it is possible to enhance the dangerous capabilities of the model. For example, by:
- Recovering knowledge that the developer tried to get the model to unlearn. See Deeb & Roger (2024) as an example.
- Fine-tuning the AI using data that may have originally been excluded from the training set or proprietary data the actor has available, such as dual-use science articles.
- Importantly, such techniques do not necessarily have to be developed by people who do not intend to use the model to develop cyberweapons per se, but who try to increase an AI's general capabilities and then share it via the internet with its safeguards removed.

- However, you should also consider how AI being more open-weight might provide additional safety benefits that lower risk, specifically:

- Open-weight AI may benefit defenders by helping them find and patch vulnerabilities.
- Open-weight AI may allow a broader range of actors to identify vulnerabilities in AI models and "patches" (NTIA 2024).
- This might not impact the risk from these identified vulnerabilities (as the original "unpatched" version of the model will remain available for threat actors to use), but this could be important for improving the security of future (more powerful) AI models when they are released.

Mitigation Policies

We are also interested in your views on how the following mitigation policies could change the risk of large data-damaging worm attacks.

P1: Proprietary Models with Refusals and Anti-Jailbreak Measures

The models used in the study (and similar models) are all proprietary and require users to access them via APIs that are subject to the safeguards described below [i–iii]. Companies train the models to refuse to respond to requests for potentially harmful information. Open-weight models are no better than the best open-weight models as of 31 August 2024.

- I.e. no 2026 open-weight model does meaningfully better than the July 2024 benchmark results from Meta's Llama 3.1-405B

i. Before deployment, a pre-release test of 5 red-teamers working together full-time for 1 week can't identify a universal jailbreak, but 10 red-teamers working together full-time for 2 months are able to find at least one universal jailbreak. A universal jailbreak is defined as "a type of vulnerability in AI systems that allows a user to consistently bypass the safety measures across a wide range of topics" and tested by whether a panel of cybersecurity experts can as a result accurately and answer in sufficient helpful detail a set of questions about vulnerability and discovery and exploit development.

- For comparison, a 2024 UK AISI evaluation found that "basic" jailbreak techniques ("either directly insert the question into a prompt template or follow a few-step procedure to generate question-specific prompts") caused current models to comply with 90–100% of harmful requests.

ii. After deployment, the companies developing the most powerful models have a voluntary goal of not letting any new universal jailbreak remain unpatched for more than 2 weeks over any given three-month period. To do so, each company has:

- "Bug bounty" programs that offer up to $15,000 rewards for anyone who identifies and reports a universal jailbreak for one of their models.

- It would be similar to that currently run by Anthropic but all companies that have trained AI models similar to that in the scenario would have this program.
- For comparison, according to Zerodium, in general (non-AI) software, a zero-day vulnerability that allows you to bypass a phone's passcode or a PIN nowadays is worth up to $100,000 – and one that grants you zero-click remote code execution on Windows is worth up to $1,000,000.
- 0.5 FTE (full-time equivalent staff members) who monitor the internet for mention of jailbreaks against their model and review instances flagged by automated processes (although it is left ambiguous how effective these are).
- If something is reported, they have 2 FTEs "on call" who then spend up to 2 weeks of effort trying to patch it. If it takes more effort than that to fix it is left ambiguous how an AI company deals with it.
- For comparison, Google Project Zero (an elite zero-day finding group) reported that in 2021 they disclosed 63 critical security vulnerabilities that took the vendors an average of 52 days to fix, down from an average of 80 days 3 years ago. They have pushed for an industry standard of keeping this number below 90 days.

iii. The companies that own the proprietary models have information security practices at "Security Level 2" as described in the 2024 RAND report "Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models" (see pp. 25–6). This security level is intended to describe "A system that can likely thwart most professional opportunistic efforts by attackers that execute moderate-effort or non-targeted attacks. This includes the operations of many professional individual hackers, as well as capable hacker groups when executing untargeted or lower-priority attacks." Security measures at this level include:

- Model weights are stored exclusively on servers (not on local devices, such as laptops) and are encrypted in storage with at least 256-bit strength encryption.
- The organization requires and enforces strong passwords, frequent software updates, and reporting of lost or stolen devices.
- A qualified security team is on call 24/7.

P2: Temporary Protections with Early Access for Defenders

The public release of the model has P1 level safeguards in place. However, a specific set of cyber defenders is given access to a version of the model without any P1 cyber protections, i.e. with full vulnerability discovery and exploit development capabilities.

- The set of cyber defenders includes Microsoft, Meta, Apple, and Google.
- It also includes the world's best highly vetted bug bounty hunters, including only:

- Synack Red Team
  - Before joining the team, each prospective Synack Red Team member must first complete a 5-step vetting process that is designed to assess skill and trustworthiness (Pathways | Synack).
  - The Synack Red Team is made up of hundreds of the best pentesters and tech practitioners in the world, from countries across the globe (Synack).
- Cobalt Core
  - Each of their pentesters has gone through a strict vetting process that only admits the top 5% of applicants (Cobalt.io).
  - Every tester is thoroughly vetted; the small percentage of applicants accepted onto the platform undergo ongoing peer review to guarantee high quality output (Cobalt Core: Become a Pentester).
- HackerOne Clear
  - Eligibility includes background checks, citizenship verification, $15K+ lifetime bug bounty requirement.
  - ID verified, have exemplary platform performance and professionalism, agree to their Rules of Engagement (Careers | HackerOne).
  - Lifetime bug bounty earned by participants: $15,000 (Careers | HackerOne).
  - Repeated criminal background checks.

After four months, the model is released under P0 security, i.e., is open weight.

A.11. For AI Developers Designing Capability Thresholds, It Makes Most Sense to Consider Expected Costs Up to a Period of Around 6–12 Months

Independent of the effect of AI on offense-defense balance, for misuse risk, there are also reasons for AI companies only to consider impacts over 6–12 months when designing their capability thresholds and safety frameworks. There are two reasons for this.

First, open source models appear to be 6 months behind the current AI frontier (Erdil 2025). Since deployment safeguards to prevent cyber misuse of AI models can be easily removed from open source models, threat actors can easily switch to open source models without safeguards once these models are released.

Second, the leading extant approach in AI governance is to impose reporting requirements on models trained using a large amount of pre-training compute (e.g. $>10^{25}$ or $>10^{26}$ FLOP), which have

tended to have the strongest capabilities. However, due to progress in compute efficiency and algorithms, the cost to achieve comparable performance is declining quickly over time. The gains in terms of effective compute from algorithmic efficiency are 2–6X per year (Epoch nd), and FLOP/$ is increasing by 1.6–2.9X per year (Hobbhahn et al. 2023). This means that some models trained on amounts of compute below the current threshold, and therefore outside regulatory requirements, will catch up to the current frontier in less than a year. Moreover, extant compute thresholds may be relaxed, or advanced models may be trained in countries that do not impose compute thresholds. As non-regulated models catch up, the benefits of imposing safeguards on larger models decline, as threat actors can simply switch to models without safeguards. Moreover, insofar as gains in future performance come from large amounts of inference compute, rather than pre-training compute, fewer powerful models will be covered by governance frameworks that focus on pre-training compute (Ord 2025).

About the Authors

John Halstead
Senior Research Fellow, GovAI

John leads the threat modeling team at GovAI. His work has focused on threat modeling CBRN and cyber misuse risk. Before joining GovAI he worked at the Forethought Foundation and Founders Pledge. He holds a DPhil in Political Philosophy from the University of Oxford.

Luca Righetti
Former Senior Research Fellow, GovAI

Luca completed this work as a Senior Research Fellow at GovAI, where he founded its Threat Modeling workstream. He is now the Threat Modeling Lead at OpenAI. Before GovAI, Luca previously worked at Coefficient Giving, the UK Office for AI, and the University of Oxford's Future of Humanity Institute.

References

6sense. 2026. "FreeBSD – Market Share, Competitor Insights in Server and Desktop OS." 6sense. https://6sense.com/tech/server-and-desktop-os/freebsd-market-share.

Aben, Emile. 2009. "Conficker/Conflicker/Downadup as Seen from the UCSD Network Telescope." Center for Applied Internet Data Analysis (CAIDA), February 27. https://www.caida.org/archive/ms08-067/conficker.

Ablon, Lillian, and Andy Bogart. 2017. Zero Days, Thousands of Nights: The Life and Times of Zero-Day Vulnerabilities and Their Exploits. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_reports/RR1751.html.

Abrams, Lawrence. 2019. "Ransomware Attackers Use Your Cloud Backups Against You." BleepingComputer, March 3. https://www.bleepingcomputer.com/news/security/ransomware-attackers-use-your-cloud-backups-against-you.

AI Safety Institute. 2024. "Advanced AI Evaluations at AISI: May Update." AI Safety Institute. https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update.

Al Jazeera. 2003. "Global Race to Beat Sobig Virus." August 22. https://www.aljazeera.com/news/2003/8/22/global-race-to-beat-sobig-virus.

Al Jazeera. 2004. "Russia Likely Origin of Mydoom Worm." January 30. https://www.aljazeera.com/news/2004/1/30/russia-likely-origin-of-mydoom-worm.

Anderson, Ross, Chris Barton, Rainer Böhme, Richard Clayton, Carlos Gañán, Tom Grasso, Michael Levi, Tyler Moore, and Marie Vasek. 2019. "Measuring the Changing Cost of Cybercrime." Workshop on the Economics of Information Security (WEIS). https://weis2019.econinfosec.org/wp-content/uploads/sites/6/2019/05/WEIS_2019_paper_25.pdf.

Android. 2025. "Features and APIs Overview." Android Developers. https://developer.android.com/about/versions/16/features.

Anthropic. 2025. "Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign." Anthropic, November 13. https://www.anthropic.com/news/disrupting-AI-espionage.

Anthropic. 2026a. "Partnering with Mozilla to Improve Firefox's Security." Anthropic, March 6. https://www.anthropic.com/news/mozilla-firefox-security.

Anthropic. 2026b. Responsible Scaling Policy, Version 3.0. Anthropic. https://www-cdn.anthropic.com/e670587677525f28df69b59e5fb4c22cc5461a17.pdf.

Apple. 2022. "Apple Security Bounty. Upgraded." Apple Security Research, October 27. https://security.apple.com/blog/apple-security-bounty-upgraded/.

Apple. 2024. "macOS Sequoia Is Available Today." Apple Newsroom, September 16. https://www.apple.com/newsroom/2024/09/macos-sequoia-is-available-today.

Applebaum, Simon, Tarek Gaber, and Ali Ahmed. 2021. "Signature-Based and Machine-Learning-Based Web Application Firewalls: A Short Survey." Procedia Computer Science 189: 359–367. https://www.sciencedirect.com/science/article/pii/S1877050921012308.

Arthur, Charles. 2009. "Windows Virus Infects 9m Computers." The Guardian, January 19. https://www.theguardian.com/technology/2009/jan/19/downadup-conficker-kido-computer-infection.

Ashford, Warwick. 2017. "EternalRocks Author Throws in the Towel After Media Attention." Computer Weekly, May 26. https://www.computerweekly.com/news/450419637/EternalRocks-author-throws-in-the-towel-after-media-attention.

Atanasov, Pavel, Phillip Rescober, Eric Stone, Samuel A. Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. 2016. "Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls." Management Science 63, no. 3: 691–706. https://pubsonline.informs.org/doi/10.1287/mnsc.2015.2374.

Baker, Kurt. 2022. "Remote Code Execution (RCE): Principles and Function." CrowdStrike, September 2. https://www.crowdstrike.com/en-us/cybersecurity-101/cyberattacks/remote-code-execution.

Baksi, Rudra P., and Shambhu J. Upadhyaya. 2020. "Decepticon: A Theoretical Framework to Counter Advanced Persistent Threats." Information Systems Frontiers. https://doi.org/10.1007/s10796-020-10087-4.

Baran, Guru. 2026. "Claude AI Uncovers 22 Firefox Vulnerabilities in Two Weeks." Cyber Security News, March 6. https://cybersecuritynews.com/claude-ai-22-firefox-vulnerabilities.

Bateman, Jon. 2022a. "How Militarily Effective Have Russia's Cyber Operations Been in Ukraine?" In Russia's Wartime Cyber Operations in Ukraine: Military Impacts, Influences, and Implications. Washington, DC: Carnegie Endowment for International Peace. https://www.jstor.org/stable/resrep45856.5.

Bateman, Jon, Nick Beecroft, and Gavin Wilde. 2022b. "What the Russian Invasion Reveals About the Future of Cyber Warfare." Carnegie Endowment for International Peace, December 19. https://carnegieendowment.org/posts/2022/12/what-the-russian-invasion-reveals-about-the-future-of-cyber-warfare?lang=en.

BBC. 2017. "Cyber-attack: Europol Says It Was Unprecedented in Scale." BBC News, May 13. https://www.bbc.com/news/world-39919249.

Beattie, Andrew. 2012. "The Most Devastating Computer Viruses." Techopedia, March 11. https://www.techopedia.com/2/26178/security/the-most-devastating-computer-viruses.

Becker, David. 2001. "FAQ: What You Need to Know About SirCam." CNET, July. https://www.cnet.com/tech/tech-industry/faq-what-you-need-to-know-about-sircam.

Bekker, Scott. 2002. "WebSideStory: 'Browser Wars a Massacre.'" Redmond Magazine, August 29. https://redmondmag.com/articles/2002/08/29/websidestory-browser-wars-a-massacre.aspx.

Bergman, Ronen, and Mark Mazzetti. 2022. "The Battle for the World's Most Powerful Cyberweapon." New York Times Magazine, January 28. https://www.nytimes.com/2022/01/28/magazine/nso-group-israel-spyware.html.

Bernal, Gabriela. 2024. "What If China Stops Sending Humanitarian Aid to North Korea?" NK News, September 4. https://www.nknews.org/2024/09/what-if-china-stops-sending-humanitarian-aid-to-north-korea/.

Bisson, David. 2017. "Shadow Brokers' Swan Song: A Sale of Hacking Tools for Windows." Tripwire State of Security, January 15. https://www.tripwire.com/state-of-security/shadow-brokers-swan-song-sale-hacking-tools-windows.

Bowden, Mark. 2019. "The Worm That Nearly Ate the Internet." New York Times, June 29. https://www.nytimes.com/2019/06/29/opinion/sunday/conficker-worm-ukraine.html.

Bracken, Becky. 2024. "US, Israel Used Dutch Spy to Launch Stuxnet Malware Against Iran." Dark Reading, January 9. https://www.darkreading.com/ics-ot-security/us-israel-dutch-spy-stuxnet-malware-against-iran.

Brighton, Henry, and Gerd Gigerenzer. 2015. "The Bias Bias." Journal of Business Research 68, no. 8: 1772–1784. https://www.sciencedirect.com/science/article/pii/S014829631500154X.

Bronsteen, John, Christopher Buccafusco, and Jonathan S. Masur. 2013. "Well-Being Analysis vs. Cost-Benefit Analysis." Duke Law Journal 62, no. 8: 1603–1689. https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=3389&context=dlj.

Broome, John. 2024. "The Value of Life in the Social Cost of Carbon: A Critique and a Proposal." Journal of Benefit-Cost Analysis 15, no. S1: 110–126. https://doi.org/10.1017/bca.2024.21.

Burdova, Carly. 2020. "What Is EternalBlue and Why Is the MS17-010 Exploit Still Relevant?" Avast Academy, June 18. https://www.avast.com/c-eternalblue.

Burdyha, Ihor. 2017. "«Чорний вівторок» українського IT: яких збитків завдала кібератака, та хто її вчинив" [Ukrainian IT's 'Black Tuesday': What Damage the Cyberattack Caused and Who Carried It Out]. Hromadske, July 8. https://hromadske.ua/posts/naslidki-kiberataki.

California State Legislature. 2025. "SB-53 Artificial Intelligence Models: Large Developers" (Transparency in Frontier Artificial Intelligence Act). California Legislative Information, 2025–2026 Regular Session. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53.

Calleja, Alejandro, Juan Tapiador, and Juan Caballero. 2018. "The MalSource Dataset: Quantifying Complexity and Code Reuse in Malware Development." IEEE Transactions on Information Forensics and Security. https://ieeexplore.ieee.org/abstract/document/8568018.

Carlini, Nicholas, Keane Lucas, Evyatar Ben Asher, Newton Cheng, Hasnain Lakhani, David Forsythe, and Kyla Guru. 2026a. "LLM-Discovered 0 Days." Frontier Red Team, Anthropic, February 6. https://red.anthropic.com/2026/zero-days.

Carlini, Nicholas, Newton Cheng, Keane Lucas, Michael Moore, Milad Nasr, Vinay Prabhushankar, Winnie Xiao, Hakeem Angulu, Evyatar Ben Asher, Jackie Bow, Keir Bradwell, Ben Buchanan, David Forsythe, Daniel Freeman, Alex Gaynor, Xinyang Ge, Logan Graham, Kyla Guru, Hasnain Lakhani, Matt McNiece, Mojtaba Mehrara, Renee Nichol, Adnan Pirzada, Sophia Porter, Andreas Terzis, and Kevin Troy. 2026b. "Assessing Claude Mythos Preview's Cybersecurity Capabilities." red.anthropic.com, April 7. https://red.anthropic.com/2026/mythos-preview.

Chernikova, Alesia, Nicolò Gozzi, Simona Boboila, Nicola Perra, Tina Eliassi-Rad, and Alina Oprea. 2023. "Modeling Self-Propagating Malware with Epidemiological Models." arXiv preprint arXiv:2208.03276. https://arxiv.org/abs/2208.03276.

Cimpanu, Catalin. 2017a. "New SMB Worm Uses Seven NSA Hacking Tools. WannaCry Used Just Two." BleepingComputer, May 19. https://www.bleepingcomputer.com/news/security/new-smb-worm-uses-seven-nsa-hacking-tools-wannacry-used-just-two.

Cimpanu, Catalin. 2017b. "Author of EternalRocks SMB Worm Calls It Quits After Intense Media Coverage." BleepingComputer, May 25. https://www.bleepingcomputer.com/news/security/author-of-eternalrocks-smb-worm-calls-it-quits-after-intense-media-coverage.

CISA. 2013. "Conficker Worm Targets Microsoft Windows Systems." Cybersecurity and Infrastructure Security Agency, Alert TA09-088A, March 29. https://www.cisa.gov/news-events/alerts/2009/03/29/conficker-worm-targets-microsoft-windows-systems.

CISA. 2020. Cost of a Cyber Incident: Systematic Review and Cross-Validation. Cybersecurity and Infrastructure Security Agency, October. https://www.cisa.gov/sites/default/files/2024-10/CISA-OCE%20Cost%20of%20Cyber%20Incidents%20Study_508.pdf.

CISA. 2021. "Mitigate Microsoft Exchange Server Vulnerabilities." Cybersecurity and Infrastructure Security Agency, Alert AA21-062A, March 3. https://www.cisa.gov/news-events/cybersecurity-advisories/aa21-062a.

CNN. 2003. "SoBig.F Breaks Virus Speed Records." CNN, August 21. https://web.archive.org/web/20040203180251/https://edition.cnn.com/2003/TECH/internet/08/21/sobig.virus/index.html.

Coburn, Andrew, Éireann Leverett, and Gordon Woo. 2019. Solving Cyber Risk: Protecting Your Company and Society. Hoboken, NJ: Wiley. https://www.wiley.com/en-gb/shop/general-finance-investments/solving-cyber-risk-protecting-your-company-and-society-p-9781119490920.

Cocomazzi, Antonio, and Antonio Pirozzi. 2022. "Black Basta Ransomware | Attacks Deploy Custom EDR Evasion Tools Tied to FIN7 Threat Actor." SentinelOne, November 3. https://www.sentinelone.com/labs/black-basta-ransomware-attacks-deploy-custom-edr-evasion-tools-tied-to-fin7-threat-actor.

Collins, Keith. 2017. "Victims of the WannaCry Ransomware Attacks Have Stopped Paying Up." Quartz, May 17. Archived May 18, 2017. https://web.archive.org/web/20170518001812/https://qz.com/986094/wannacry-ransomware-attacks-victims-have-stopped-paying-the-ransom/.

Comodo. 2019. "Swen Virus | How to Remove Swen Worm Virus?" Comodo Antivirus Blog. https://antivirus.comodo.com/blog/computer-safety/swen-virus.

Cooper, Charles. 2018. "WannaCry: Lessons Learned 1 Year Later." Symantec Enterprise Blogs, May 15. https://symantec-enterprise-blogs.security.com/feature-stories/wannacry-lessons-learned-1-year-later.

Crocker, Andrew, and Bill Budington. 2016. "NSA's Failure to Report Shadow Broker Vulnerabilities Underscores Need for Oversight." Electronic Frontier Foundation, September 23. https://www.eff.org/deeplinks/2016/09/nsas-failure-report-shadow-broker-vulnerabilities-underscores-need-oversight.

Crosignani, Matteo, Marco Macchiavelli, and André F. Silva. 2023. "Pirates Without Borders: The Propagation of Cyberattacks Through Firms' Supply Chains." Journal of Financial Economics 147, no. 2: 432–448. https://www.sciencedirect.com/science/article/abs/pii/S0304405X2200246X.

Dalkey, Norman, and Olaf Helmer. 1963. "An Experimental Application of the DELPHI Method to the Use of Experts." Management Science 9, no. 3: 458–467. https://doi.org/10.1287/mnsc.9.3.458.

Darknet Diaries. 2020. "Bangladesh Bank Heist." Darknet Diaries, episode 72, August 18. https://darknetdiaries.com/episode/72.

Darknet Diaries. 2025. "MalwareTech." Darknet Diaries, episode 158, May 6. https://darknetdiaries.com/transcript/158/.

Davis-Stober, Clintin P., David V. Budescu, Jason Dana, and Stephen B. Broomell. 2014. "When Is a Crowd Wise?" Decision 1, no. 2: 79–101. https://arxiv.org/pdf/1406.7563.

de Loisy, Nicolas. 2019. "How Many Computers Are There in the World?" SCMO, August 9. https://www.scmo.net/faq/2019/8/9/how-many-compaters-is-there-in-the-world.

Dean, Joshua. 2001. "New Computer Worm Spreading Rapidly." Government Executive, September 18. https://www.govexec.com/federal-news/2001/09/new-computer-worm-spreading-rapidly/10001.

Deeb, Aghyad, and Fabien Roger. 2024. "Do Unlearning Methods Remove Information from Language Model Weights?" arXiv preprint arXiv:2410.08827. https://arxiv.org/abs/2410.08827.

Delio, Michelle. 2001. "SirCam Ready to Drop Payload." Wired, October 12. https://www.wired.com/2001/10/sircam-ready-to-drop-payload.

Delio, Michelle. 2002. "Klez: Hi, Mom, We're No. 1." Wired. https://www.wired.com/2002/05/klez-hi-mom-were-no-1.

Delio, Michelle. 2003. "Yaha Virus Uses Netizens as Pawns." Wired, March 13. https://www.wired.com/2003/03/yaha-virus-uses-netizens-as-pawns.

Department of Health and Social Care. 2018. Securing Cyber Resilience in Health and Care: Progress Update October 2018. Department of Health and Social Care. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/747464/securing-cyber-resilience-in-health-and-care-september-2018-update.pdf.

Devicie. n.d. "Risk of Local Administrator Privileges in Ransomware and Malware Attacks." Devicie. https://devicie.com/guide/risk-of-local-administrator-privileges-in-ransomware-and-malware-attacks.

DOJ (U.S. Department of Justice). 2018. "North Korean Regime-Backed Programmer Charged with Conspiracy to Conduct Multiple Cyber Attacks and Intrusions." U.S. Department of Justice, Office of Public Affairs,

September 6.
https://www.justice.gov/opa/pr/north-korean-regime-backed-programmer-charged-conspiracy-conduct-multiple-cyber-attacks-and.

DoJ (U.S. Department of Justice). 2021. "Three North Korean Military Hackers Indicted in Wide-Ranging Scheme to Commit Cyberattacks and Financial Crimes Across the Globe." U.S. Department of Justice, Office of Public Affairs, February 17.
https://www.justice.gov/archives/opa/pr/three-north-korean-military-hackers-indicted-wide-ranging-scheme-commit-cyberattacks-and.

Dolak, John C. 2005. The Code Red Worm. SANS Institute / Global Information Assurance Certification.
https://www.giac.org/paper/gsec/1162/code-red-worm/102232.

Doman, Chris. n.d. "Team TNT – The First Crypto-Mining Worm to Steal AWS Credentials." Cado Security. Archived April 29, 2025.
http://web.archive.org/web/20250429203411/https://www.cadosecurity.com/blog/team-tnt-the-first-crypto-mining-worm-to-steal-aws-credentials.

Ee, Shaun, Cara Labrador, Chris Covino, Jam Kraprayoon, and Joe O'Brien. 2025. "Asymmetry by Design: Boosting Cyber Defenders with Differential Access to AI." Institute for AI Policy and Strategy, May 23.
https://www.iaps.ai/research/differential-access.

Elliptic. 2017. "WannaCry." Elliptic. Archived December 24, 2017.
https://web.archive.org/web/20171224043451/https://www.elliptic.co/wannacry.

Engage Employee. n.d. "90 Per Cent of Ransomware Can Execute Without Administrator Rights." Engage Employee.
https://www.engageemployee.com/blog/90-per-cent-of-ransomware-can-execute-without-administrator-rights.

Epoch AI. n.d. "Trends in Artificial Intelligence." Epoch AI. https://epoch.ai/trends.

Erdil, Ege. 2025. "What Went into Training DeepSeek-R1?" Epoch AI, January 31.
https://epoch.ai/gradient-updates/what-went-into-training-deepseek-r1.

ESET. 2017. "New WannaCryptor-like Ransomware Attack Hits Globally: All You Need to Know." WeLiveSecurity, June 27. https://www.welivesecurity.com/2017/06/27/new-ransomware-attack-hits-ukraine.

Ewing, Philip. 2017. "Report: Hackers Stole NSA Cybertools in Another Breach via Another Contractor." NPR, October 5.
https://www.npr.org/2017/10/05/555922305/report-hackers-stole-nsa-cybertools-in-another-breach-via-another-contractor.

F-Secure. n.d.-a. "Worm:W32/Swen." F-Secure Threat Descriptions.
https://www.f-secure.com/v-descs/swen.shtml.

F-Secure. n.d.-b. "Worm:W32/Yaha.E." F-Secure Threat Descriptions.
https://www.f-secure.com/v-descs/yaha-e.shtml.

Falco, Marco De. 2012. Stuxnet Facts Report: A Technical and Strategic Analysis. Tallinn: NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE).
https://ccdcoe.org/uploads/2018/10/Falco2012_StuxnetFactsReport.pdf.

Falliere, Nicolas, Liam O Murchu, and Eric Chien. 2011. W32.Stuxnet Dossier. Version 1.4. Symantec Security Response, February 11. https://docs.broadcom.com/doc/security-response-w32-stuxnet-dossier-11-en.

FBI. n.d. "Melissa Virus." Federal Bureau of Investigation.
https://www.fbi.gov/history/famous-cases/melissa-virus.

Feiner, Lauren. 2023. "Chinese Hackers Outnumber FBI Cyber Staff 50 to 1, Bureau Director Says." CNBC, April 28.
https://www.cnbc.com/2023/04/28/chinese-hackers-outnumber-fbi-cyber-staff-50-to-1-director-wray-says.html.

Fiscutean, Andrada. 2023. "How Patch Tuesday Keeps the Beat After 20 Years." Dark Reading, March 15. https://www.darkreading.com/cyber-risk/how-patch-tuesday-keeps-the-beat-after-20-years.

Forbes. 2017. "How Similar Are WannaCry and Petya Ransomware?" Forbes, July 5. https://www.forbes.com/sites/quora/2017/07/05/how-similar-are-wannacry-and-petya-ransomware.

Forni, Ippolito. 2020. "WannaCry 3 Years Later, Could It Happen Again?" EclecticIQ Blog, May 13. https://blog.eclecticiq.com/wannacry-3-years-later-could-it-happen-again.

Forster, Malcolm, and Elliott Sober. 1994. "How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions." British Journal for the Philosophy of Science 45, no. 1: 1–35. https://d1wqtxts1xzle7.cloudfront.net/8292761/10.1.1.160.7470-libre.pdf.

Franceschi-Bicchierai, Lorenzo. 2024. "Price of Zero-Day Exploits Rises as Companies Harden Products Against Hackers." TechCrunch, April 6. https://techcrunch.com/2024/04/06/price-of-zero-day-exploits-rises-as-companies-harden-products-against-hackers.

Friedman, Jeffrey A., Joshua D. Baker, Barbara A. Mellers, Philip E. Tetlock, and Richard Zeckhauser. 2018. "The Value of Precision in Probability Assessment: Evidence from a Large-Scale Geopolitical Forecasting Tournament." International Studies Quarterly 62, no. 2: 410–422. https://academic.oup.com/isq/article-abstract/62/2/410/4944059.

Gallaher, Michael P., Brent R. Rowe, Alex V. Rogozhin, and Albert N. Link. 2006. Economic Analysis of Cyber Security. RTI International (Defense Technical Information Center, ADA455398). https://apps.dtic.mil/sti/tr/pdf/ADA455398.pdf.

Ganacharya, Tanmay. 2018. "A Worthy Upgrade: Next-Gen Security on Windows 10 Proves Resilient Against Ransomware Outbreaks in 2017." Microsoft Security Blog, January 10. https://www.microsoft.com/en-us/security/blog/2018/01/10/a-worthy-upgrade-next-gen-security-on-windows-10-proves-resilient-against-ransomware-outbreaks-in-2017.

GAO (U.S. General Accounting Office). 1999. Information Security: The Melissa Computer Virus Demonstrates Urgent Need for Stronger Protection Over Systems and Sensitive Data. Statement of Keith A. Rhodes. GAO/T-AIMD-99-146. Washington, DC: U.S. General Accounting Office. https://www.govinfo.gov/content/pkg/GAOREPORTS-T-AIMD-99-146/pdf/GAOREPORTS-T-AIMD-99-146.pdf.

Garfinkel, Ben, and Allan Dafoe. 2019. "How Does the Offense-Defense Balance Scale?" Journal of Strategic Studies 42, no. 6: 736–763. https://www.tandfonline.com/doi/full/10.1080/01402390.2019.1631810.

Gaudin, Sharon. 2004. "Sobig's Birthday — Tracking Most Damaging Virus Ever." eSecurityPlanet, January 9. https://www.esecurityplanet.com/trends/sobigs-birthday-tracking-most-damaging-virus-ever/.

Gelman, Andrew, and Eric Loken. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No 'Fishing Expedition' or 'P-Hacking' and the Research Hypothesis Was Posited Ahead of Time." Unpublished manuscript, November 14. https://sites.stat.columbia.edu/gelman/research/unpublished/p_hacking.pdf.

Gerencer, Tom. 2020. "The Top 10 Worst Computer Viruses in History." HP Tech Takes, November 4. https://www.hp.com/us-en/shop/tech-takes/top-ten-worst-computer-viruses-in-history.

Ghafur, S., S. Kristensen, K. Honeyford, G. Martin, A. Darzi, and P. Aylin. 2019. "A Retrospective Impact Analysis of the WannaCry Cyberattack on the NHS." npj Digital Medicine 2: 98. https://www.nature.com/articles/s41746-019-0161-6.pdf.

Holick, Heather. 2002. Nimda Worm. GIAC Security Essentials (GSEC) Practical. SANS Institute. https://www.giac.org/paper/gsec/1542/nimda-worm/102853.

Gofman, Igal. 2017. "Advanced Threat Analytics Security Research Network Technical Analysis: NotPetya." Microsoft Security Blog, October 3.

https://www.microsoft.com/en-us/security/blog/2017/10/03/advanced-threat-analytics-security-research-network-technical-analysis-notpetya.

Gomez Cram, Roberto, Yunhan Guo, Theis Ingerslev Jensen, and Howard Kung. 2026. "Prediction Market Accuracy: Crowd Wisdom or Informed Minority?" SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6617059.

Gomilšek, Tamara, Ulrich Hoffrage, and Julian N. Marewski. 2024. "Fermian Guesstimation Can Boost the Wisdom-of-the-Inner-Crowd." Scientific Reports 14, no. 1: 5014. https://www.nature.com/articles/s41598-024-53639-3.

Gonsalves, Antone. 2004. "Mydoom Shows Vulnerability of the Web." Network Computing. Archived January 4, 2026. http://web.archive.org/web/20260104103921/https://www.networkcomputing.com/cybersecurity/mydoom-shows-vulnerability-of-the-web.

Goodin, Dan. 2017a. "NSA-Leaking Shadow Brokers Just Dumped Its Most Damaging Release Yet." Ars Technica, April 14. https://arstechnica.com/information-technology/2017/04/nsa-leaking-shadow-brokers-just-dumped-its-most-damaging-release-yet.

Goodin, Dan. 2017b. "Fearing Shadow Brokers Leak, NSA Reported Critical Flaw to Microsoft." Ars Technica, May 17. https://arstechnica.com/information-technology/2017/05/fearing-shadow-brokers-leak-nsa-reported-critical-flaw-to-microsoft.

Goodin, Dan. 2017c. "NotPetya Developers Obtained NSA Exploits Weeks Before Their Public Leak." Ars Technica, June. https://arstechnica.com/information-technology/2017/06/notpetya-developers-obtained-nsa-exploits-weeks-before-their-public-leak.

Goodin, Dan. 2017d. "Backdoor Built in to Widely Used Tax App Seeded Last Week's NotPetya Outbreak." Ars Technica, July 5. https://arstechnica.com/information-technology/2017/07/heavily-armed-police-raid-company-that-seeded-last-weeks-notpetya-outbreak.

Google. 2024a. Buying Spying: Insights into Commercial Surveillance Vendors. Google Threat Analysis Group. https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/Buying_Spying_-_Insights_into_Commercial_Surveillance_Vendors_-_TAG_report.pdf.

Google. 2024b. We're All in this Together: A Year in Review of Zero-Days Exploited In-the-Wild in 2023. Google Threat Analysis Group and Mandiant. https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/Year_in_Review_of_ZeroDays.pdf.

Google. 2025. "Key Stats." Google Bug Hunters. https://bughunters.google.com/about/key-stats.

Google DeepMind. 2025. Frontier Safety Framework, Version 3.0. Google DeepMind. http://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf.

Google's Project Zero. 2024. "From Naptime to Big Sleep: Using Large Language Models to Catch Vulnerabilities in Real-World Code." Project Zero Blog, November 1. https://googleprojectzero.blogspot.com/2024/10/from-naptime-to-big-sleep.html.

Green, Kesten C., and J. Scott Armstrong. 2015. "Simple Versus Complex Forecasting: The Evidence." Journal of Business Research 68, no. 8: 1678–1685. https://www.sciencedirect.com/science/article/abs/pii/S014829631500140X.

Greenberg, Andy. 2016. "Hackers Claim to Auction Data Stolen from NSA-Linked Spies." Wired, August. https://www.wired.com/2016/08/hackers-claim-auction-data-stolen-nsa-linked-spies.

Greenberg, Andy. 2017a. "The WannaCry Ransomware Hackers Made Some Real Amateur Mistakes." Wired, May 15. https://www.wired.com/2017/05/wannacry-ransomware-hackers-made-real-amateur-mistakes.

Greenberg, Andy. 2017b. "How the Mimikatz Hacker Tool Stole the World's Passwords." Wired, November 9. https://www.wired.com/story/how-mimikatz-became-go-to-hacker-tool.

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

Guardian. 2004. "Sasser Worm Hits up to 1m Computers." The Guardian, May 4. https://web.archive.org/web/20140913220656/https://www.theguardian.com/technology/2004/may/04/security.business.

Haber, Morey J. 2017. "WannaCry Ransomware Attack Explained — Makes Me Wanna Cry…" BeyondTrust, May 16. https://www.beyondtrust.com/blog/entry/wannacry-ransomware-attack-explained-makes-me-wanna-cry.

Hasherezade. 2017. "Keeping up with the Petyas: Demystifying the Malware Family." Malwarebytes Labs, July 14. https://www.malwarebytes.com/blog/news/2017/07/keeping-up-with-the-petyas-demystifying-the-malware-family.

Healey, Jason. 2016. "The U.S. Government and Zero-Day Vulnerabilities: From Pre-Heartbleed to Shadow Brokers." Journal of International Affairs (Columbia University). https://jia.sipa.columbia.edu/news/us-government-and-zero-day-vulnerabilities-pre-heartbleed-shadow-brokers.

Healey, Jason, and Tarang Jain. 2025. "Are Cyber Defenders Winning?" Lawfare, July 14. https://www.lawfaremedia.org/article/are-cyber-defenders-winning.

Help Net Security. 2002. "Worm Launches Attack on Pakistan Govt Website." Help Net Security, July 1. https://www.helpnetsecurity.com/2002/07/01/worm-launches-attack-on-pakistan-govt-website.

Herzog, Stefan M., and Ralph Hertwig. 2009. "The Wisdom of Many in One Mind: Improving Individual Judgments with Dialectical Bootstrapping." Psychological Science 20, no. 2: 231–237. https://pure.mpg.de/rest/items/item_2504710_5/component/file_2520393/content.

Hobbhahn, Marius, Lennart Heim, and Gökçe Aydos. 2023. "Trends in Machine Learning Hardware." Epoch AI, November 9. https://epoch.ai/blog/trends-in-machine-learning-hardware.

Hurley, Shaun, and Karan Sood. 2017. "NotPetya Technical Analysis Part II: Further Findings and Potential for MBR Recovery." CrowdStrike, July 3. https://www.crowdstrike.com/en-us/blog/petrwrap-technical-analysis-part-2-further-findings-and-potential-for-mbr-recovery/.

HYPR. n.d. "What Is the Storm Worm?" Security Encyclopedia. https://www.hypr.com/security-encyclopedia/storm-worm.

International AI Safety Report. 2025. International AI Safety Report 2025. Edited by Yoshua Bengio et al. UK Department for Science, Innovation and Technology. https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf.

International AI Safety Report. 2026. International AI Safety Report 2026. Chaired by Yoshua Bengio. https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026.pdf.

Irregular. 2025. "Frontier Model Performance on Offensive-Security Tasks: Emerging Evidence of a Capability Shift." Irregular, December 10. https://www.irregular.com/publications/emerging-evidence-of-a-capability-shift.

Irregular. 2025. "Introducing SOLVE: Scoring Obstacle Levels in Vulnerabilities & Exploits (Version 0.5)." Irregular. https://www.irregular.com/research/introducing-solve.

Itach, Ofek, and Assaf Morag. 2023. "TeamTNT Reemerged with New Aggressive Cloud Campaign." Aqua, July 13. https://www.aquasec.com/blog/teamtnt-reemerged-with-new-aggressive-cloud-campaign.

Jenkins, Adam, Pieris Kalligeros, Kami Vaniea, and Maria K. Wolters. 2020. "'Anyone Else Seeing this Error?': Community, System Administrators, and Patch Information." In 2020 IEEE European Symposium on Security and Privacy (EuroS&P), 105–119. IEEE. https://www.pure.ed.ac.uk/ws/portalfiles/portal/148336136/Anyone_Else_Seeing_JENKINS_DOA27022020_AFV.pdf.

Johanns, Kate. 2021. "Tech Time Warp: Swen Worm Poses as Security Patch." Smarter MSP, September 24. https://smartermsp.com/tech-time-warp-swen-worm-poses-as-security-patch.

Johansmeyer, Tom. 2023. "How Reversibility Differentiates Cyber from Kinetic Warfare: A Case Study in the Energy Sector." International Journal of Security, Privacy and Trust Management 12, no. 1: 1–17. https://aircconline.com/ijsptm/V12N1/12123ijsptm01.pdf.

Johansmeyer, Tom. 2024a. "Perception Shapes Reality: How Views on Financial Market Correlation Affect Capital Availability for Cyber Insurance." Journal of Risk Management and Insurance 28, no. 1: 1–25. https://kar.kent.ac.uk/106432/3/Correlation%20Perception%20REVISION%20draft%20v6%20clean%20KAR.pdf.

Johansmeyer, Tom. 2024b. "Why Natural Catastrophes Will Always Be Worse than Cyber Catastrophes." War on the Rocks, April 5. https://warontherocks.com/2024/04/why-natural-catastrophes-will-always-be-worse-than-cyber-catastrophes.

Johansmeyer, Tom. 2024c. "Surprising Stats: The Worst Economic Losses from Cyber Catastrophes." The Loop (ECPR), March 12. https://theloop.ecpr.eu/surprising-stats-the-worst-economic-losses-from-cyber-catastrophes.

Kapoor, Sayash, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, et al. 2024. "On the Societal Impact of Open Foundation Models." Stanford Center for Research on Foundation Models. https://crfm.stanford.edu/open-fms.

Karantzas, George, and Constantinos Patsakis. 2021. "An Empirical Assessment of Endpoint Detection and Response Systems Against Advanced Persistent Threats Attack Vectors." Journal of Cybersecurity and Privacy 1, no. 3: 387–421. https://www.mdpi.com/2624-800X/1/3/21.

Kaspersky. 2022. "The Evolution of Security: The Story of Code Red." Kaspersky Daily, August 4. https://www.kaspersky.com/blog/history-lessons-code-red/45082.

Kaspersky. n.d. “What Is WannaCry Ransomware?” Kaspersky Resource Center. Accessed June 23, 2026. https://www.kaspersky.com/resource-center/threats/ransomware-wannacry.

Knapp, David, Sina Beaghley, Troy D. Smith, Molly F. McIntosh, Karen Schwindt, Norah Griffin, Daniel Schwam, and Hanna Hoover. 2021. DoD Cyber Excepted Service Labor Market Analysis and Options for Use of Compensation Flexibilities. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_reports/RRA730-1.html.

Kryptos Logic. 2017. "WannaCry: Two Weeks and 16 Million Averted Ransoms Later." Kryptos Logic, May 30. https://www.kryptoslogic.com/blog/2017/05/wannacry-two-weeks-and-16-million-averted-ransoms-later.

Kumaran, Neil. 2022. "Understanding Gmail's Spam Filters." Google Workspace Blog, May 28. https://workspace.google.com/blog/identity-and-security/an-overview-of-gmails-spam-filters.

Lakshmanan, Ravie. 2025a. "Apple Zero-Click Flaw in Messages Exploited to Spy on Journalists Using Paragon Spyware." The Hacker News, June 13. https://thehackernews.com/2025/06/apple-zero-click-flaw-in-messages.html.

Lakshmanan, Ravie. 2025b. "WhatsApp Patches Zero-Click Exploit Targeting iOS and macOS Devices." The Hacker News, August 30. https://thehackernews.com/2025/08/whatsapp-issues-emergency-update-for.html.

Lakshmanan, Ravie. 2025c. "Samsung Fixes Critical Zero-Day CVE-2025-21043 Exploited in Android Attacks." The Hacker News, September. https://thehackernews.com/2025/09/samsung-fixes-critical-zero-day-cve.html.

Lamar, Jason. 2025. "Key Takeaways from the State of Pentesting Report 2025." Cobalt, April 14. https://www.cobalt.io/blog/key-takeaways-state-of-pentesting-report-2025.

Lambert, Tony. 2019. "It's All Fun and Games Until Ransomware Deletes the Shadow Copies." Red Canary Blog. https://redcanary.com/blog/threat-detection/its-all-fun-and-games-until-ransomware-deletes-the-shadow-copies.

Landler, Mark. 2000. "A Filipino Linked to 'Love Bug' Talks About His License to Hack." New York Times, October 21. https://www.nytimes.com/2000/10/21/business/a-filipino-linked-to-love-bug-talks-about-his-license-to-hack.html.

Larin, Boris. 2023. "Operation Triangulation: The Last (Hardware) Mystery." Securelist, December 27. https://securelist.com/operation-triangulation-the-last-hardware-mystery/111669.

Lee, Brandon. 2021. "PsExec: The SysAdmin's Swiss Army Knife." Virtualization DOJO (Altaro/Hornetsecurity), February 26. https://www.altaro.com/hyper-v/psexec-sysadmins.

Lee, Martin, Warren Mercer, Paul Rascagneres, and Craig Williams. 2017. "Player 3 Has Entered the Game: Say Hello to 'WannaCry'." Cisco Talos Blog, May 12. https://blog.talosintelligence.com/wannacry.

Lefferts, Rob. 2017. "Windows 10 Creators Update Provides Next-Gen Ransomware Protection." Microsoft Security Blog, June 8. https://www.microsoft.com/en-us/security/blog/2017/06/08/windows-10-creators-update-hardens-security-with-next-gen-defense.

Lemos, Robert. 1999. "Smith Pleads Guilty to Melissa Virus." ZDNet, December 9. https://www.zdnet.com/article/smith-pleads-guilty-to-melissa-virus.

Lemos, Robert. 2001. "Nimda Still a Global Threat." CNET, September 24. https://www.cnet.com/tech/tech-industry/nimda-still-a-global-threat.

Leyden, John. 2002. "Why Is mi2g So Unpopular?" The Register, November 21. https://www.theregister.com/2002/11/21/why_is_mi2g_so_unpopular.

Lin, Herbert. 2022. "Russian Cyber Operations in the Invasion of Ukraine." The Cyber Defense Review 7, no. 4: 31–46. https://www.jstor.org/stable/48703290.

Litchfield, David. 2010. "The Inside Story of SQL Slammer." Threatpost, October 20. https://threatpost.com/inside-story-sql-slammer-102010/74589.

Lloyd's. 2019. Bashe Attack: Global Infection by Contagious Malware. CyRiM (Cyber Risk Management Project). https://assets.lloyds.com/assets/pdf-bashe-attack-cyrimbasheattack-finalbashe-attack/1/pdf-bashe-attack-CyRiMBasheAttack_FINALbashe-attack.pdf.

Lohn, Andrew, and Krystal Jackson. 2022. "Will AI Make Cyber Swords or Shields?" arXiv preprint arXiv:2207.13825. https://arxiv.org/abs/2207.13825.

Lohn, Andrew J. 2025. "The Impact of AI on the Cyber Offense-Defense Balance and the Character of Cyber Conflict." arXiv, April 17. https://arxiv.org/abs/2504.13371.

Lukošiūtė, Kamilė, and Adam Swanda. 2025. "LLM Cyber Evaluations Don't Capture Real-World Risk." arXiv preprint arXiv:2502.00072. https://arxiv.org/pdf/2502.00072.

Lyngaas, Sean. 2023. "Half of North Korean Missile Program Funded by Cyberattacks and Crypto Theft, White House Says." CNN, May 10. https://edition.cnn.com/2023/05/10/politics/north-korean-missile-program-cyberattacks/index.html.

Macrae, Duncan. 2014. "Everything You Need to Know About the Sasser Worm." Tech Monitor, April 11. https://www.techmonitor.ai/technology/cybersecurity/everything-you-need-to-know-about-the-sasser-worm-4213147.

Makridakis, Spyros, and Michèle Hibon. 2000. "The M3-Competition: Results, Conclusions and Implications." International Journal of Forecasting 16, no. 4: 451–476. https://www.sciencedirect.com/science/article/abs/pii/S0169207000000571.

Maloney, Sarah. n.d. "A Quick Recap on NotPetya – An Unofficial Guide." Cybereason Blog. https://www.cybereason.com/blog/blog-a-quick-recap-on-notpetya.

MalwareTech. 2017. "How to Accidentally Stop a Global Cyber Attacks." MalwareTech, May 13. https://www.malwaretech.com/2017/05/how-to-accidentally-stop-a-global-cyber-attacks.html.

Mandiant. 2013. APT1: Exposing One of China's Cyber Espionage Units. Mandiant. https://www.mandiant.com/sites/default/files/2021-09/mandiant-apt1-report.pdf.

Marczak, Bill, John Scott-Railton, Bahr Abdul Razzak, and Ron Deibert. 2023. "Triple Threat: NSO Group's Pegasus Spyware Returns in 2022 with a Trio of iOS 15 and iOS 16 Zero-Click Exploit Chains." The Citizen Lab, April 18. https://citizenlab.ca/2023/04/nso-groups-pegasus-spyware-returns-in-2022.

Marczak, Bill, John Scott-Railton, Bahr Abdul Razzak, Noura Al-Jizawi, Siena Anstis, Kristin Berdan, and Ron Deibert. 2021. "FORCEDENTRY: NSO Group iMessage Zero-Click Exploit Captured in the Wild." Citizen Lab Research Report No. 143, University of Toronto, September 13. https://citizenlab.ca/2021/09/forcedentry-nso-group-imessage-zero-click-exploit-captured-in-the-wild.

Markovic, Sinisa. 2026. "Mythos Preview Can Weaponize N-day Vulnerabilities in Hours." Help Net Security, June 9. https://www.helpnetsecurity.com/2026/06/09/anthropic-mythos-preview-n-day-exploits-firefox-windows.

Martínez Martínez, Isabella, Andrés Florián Quitián, Daniel Díaz-López, Pantaleone Nespoli, and Félix Gómez Mármol. 2021. "MalSEIRS: Forecasting Malware Spread Based on Compartmental Models in Epidemiology." Complexity 2021. https://doi.org/10.1155/2021/5415724.

Maschmeyer, Lennart. 2021. "The Subversive Trilemma: Why Cyber Operations Fall Short of Expectations." International Security 46, no. 2: 51–90. https://direct.mit.edu/isec/article/46/2/51/107693/The-Subversive-Trilemma-Why-Cyber-Operations-Fall.

Maurer, Tim. 2018. "Why the Russian Government Turns a Blind Eye to Cybercriminals." Carnegie Endowment for International Peace, February 2. https://carnegieendowment.org/posts/2018/02/why-the-russian-government-turns-a-blind-eye-to-cybercriminals?lang=en.

Mauri, Lara, and Ernesto Damiani. 2025. "Hardening Behavioral Classifiers against Polymorphic Malware: An Ensemble Approach Based on Minority Report." Information Sciences 689: 121499. https://www.sciencedirect.com/science/article/pii/S0020025524014130.

Maynor, David, Matt Olney, and Yves Younan. 2017. "The MeDoc Connection." Cisco Talos Intelligence Blog, July 5. https://blog.talosintelligence.com/the-medoc-connection.

McCurry, Justin. 2025. "North Korea Behind Bybit Crypto Exchange Hack, FBI Says." The Guardian, February 27. https://www.theguardian.com/world/2025/feb/27/north-korea-bybit-crypto-exchange-hack-fbi.

Meta. 2025. "Frontier AI Framework." Meta, February 3. Archived February 27, 2026. http://web.archive.org/web/20260227021742/https://ai.meta.com/static-resource/meta-frontier-ai-framework/.

Microsoft. 2001. "Microsoft Security Bulletin MS01-033 - Critical: Unchecked Buffer in Index Server ISAPI Extension Could Enable Web Server Compromise." Microsoft Learn, June 18. https://learn.microsoft.com/en-us/security-updates/securitybulletins/2001/ms01-033.

Microsoft. 2004a. "Q&A: Microsoft's Anti-Spam Technology Roadmap." Microsoft News, February 24. Archived January 24, 2015. http://web.archive.org/web/20150124144756/http://news.microsoft.com/2004/02/24/qa-microsofts-anti-spam-technology-roadmap/.

Microsoft. 2004b. "Microsoft Security Bulletin MS04-011: Security Update for Microsoft Windows (835732)." Microsoft, April 13. https://learn.microsoft.com/en-us/security-updates/securitybulletins/2004/ms04-011.

Microsoft. 2005. "Worm:Win32/Swen.A@mm Threat Description." Microsoft Security Intelligence, June 23. https://www.microsoft.com/en-us/wdsi/threats/malware-encyclopedia-description?Name=Worm:Win32/Swen.A@mm.

Microsoft. 2007. "Worm:Win32/Klez.H@mm Threat Description." Microsoft Security Intelligence. https://www.microsoft.com/en-us/wdsi/threats/malware-encyclopedia-description?Name=Worm%3AWin32%2FKlez.H%40mm.

Microsoft. 2008. "Microsoft Security Bulletin MS08-067 – Critical." Microsoft. https://learn.microsoft.com/en-us/security-updates/securitybulletins/2008/ms08-067.

Microsoft. 2016. "Deprecating Support for SmartScreen in Outlook and Exchange." Exchange Team Blog, Microsoft Tech Community, September 1. https://techcommunity.microsoft.com/blog/exchange/deprecating-support-for-smartscreen-in-outlook-and-exchange/605332.

Microsoft. 2017a. "Microsoft Security Bulletin MS17-010 - Critical." Microsoft Security Updates, March 14. https://learn.microsoft.com/en-us/security-updates/securitybulletins/2017/ms17-010.

Microsoft. 2017b. "WannaCrypt Ransomware Worm Targets Out-of-Date Systems." Microsoft Security Blog, May 12. https://www.microsoft.com/en-us/security/blog/2017/05/12/wannacrypt-ransomware-worm-targets-out-of-date-systems.

Microsoft. 2018. "Windows 10 Platform Resilience Against the Petya Ransomware Attack." Microsoft Malware Protection Center Blog, June 29, 2017. Archived. https://web.archive.org/web/20170710193238/https://blogs.technet.microsoft.com/mmpc/2017/06/29/windows-10-platform-resilience-against-the-petya-ransomware-attack.

Microsoft. 2020. Optimizing Windows 10 Update Adoption. Microsoft. https://www.microsoft.com/en-us/download/details.aspx?id=101056.

Microsoft. 2021a. "Protecting On-Premises Exchange Servers against Recent Attacks." Microsoft Security Blog, March 12. https://www.microsoft.com/en-us/security/blog/2021/03/12/protecting-on-premises-exchange-servers-against-recent-attacks/.

Microsoft. 2021b. "Analyzing Attacks Taking Advantage of the Exchange Server Vulnerabilities." Microsoft Security Blog, March 25. https://www.microsoft.com/en-us/security/blog/2021/03/25/analyzing-attacks-taking-advantage-of-the-exchange-server-vulnerabilities/.

Microsoft. 2021c. "Achieving World-Class Windows Monthly Patching Efficiency." Windows IT Pro Blog, Microsoft Tech Community, July 21. https://techcommunity.microsoft.com/blog/windows-itpro-blog/achieving-world-class-windows-monthly-patching-efficiency/2572945.

Microsoft. 2022a. "How to Prevent Lateral Movement Attacks Using Microsoft 365 Defender." Microsoft Security Blog, October 26. https://www.microsoft.com/en-us/security/blog/2022/10/26/how-to-prevent-lateral-movement-attacks-using-microsoft-365-defender.

Microsoft. 2022b. "Defenders Beware: A Case for Post-Ransomware Investigations." Microsoft Security Blog, October 18. https://www.microsoft.com/en-us/security/blog/2022/10/18/defenders-beware-a-case-for-post-ransomware-investigations.

Microsoft. 2023. "SMBv1 Not Installed by Default in Windows Server and Windows." Microsoft Learn. https://learn.microsoft.com/en-us/windows-server/storage/file-server/troubleshoot/smbv1-not-installed-by-default-in-windows.

Microsoft. 2024. "Microsoft Bounty Program Year in Review: $16.6M in Rewards." Microsoft Security Response Center Blog, August. https://msrc.microsoft.com/blog/2024/08/microsoft-bounty-program-year-in-review-16.6m-in-rewards/.

Microsoft. 2025a. "Detect, Enable, and Disable SMBv1, SMBv2, and SMBv3 in Windows." Microsoft Learn, February 28. https://learn.microsoft.com/en-us/windows-server/storage/file-server/troubleshoot/detect-enable-and-disable-smbv1-v2-v3.

Microsoft. 2025b. "Facts About Microsoft." Microsoft News. https://news.microsoft.com/facts-about-microsoft.

Microsoft. n.d. "Lifecycle FAQ - Windows." Microsoft Learn. https://learn.microsoft.com/en-us/lifecycle/faq/windows.

Miller, Maggie. 2022. "Russia Arrests Hacker in Colonial Pipeline Attack, U.S. Says." Politico, January 14. https://www.politico.com/news/2022/01/14/russia-colonial-pipeline-arrest-527166.

Mimoso, Michael. 2017. "NSA's DoublePulsar Kernel Exploit in Use Internet-Wide." Threatpost, April 21. https://threatpost.com/nsas-doublepulsar-kernel-exploit-in-use-internet-wide/125165.

Modderkolk, Huib. 2024. "Sabotage in Iran: een missie in duisternis." de Volkskrant, January 8. https://www.volkskrant.nl/kijkverder/v/2024/sabotage-in-iran-een-missie-in-duisternis~v989743.

Moore, David, Colleen Shannon, and Jeffery Brown. 2002. Code-Red: A Case Study on the Spread and Victims of an Internet Worm. Cooperative Association for Internet Data Analysis. https://www.caida.org/catalog/papers/2002_codered/codered.pdf.

Moore, David, Vern Paxson, Stefan Savage, Colleen Shannon, Stuart Staniford, and Nicholas Weaver. 2003. "The Spread of the Sapphire/Slammer Worm." CAIDA. https://www.caida.org/catalog/papers/2003_sapphire.

Morgan, M. Granger, and Max Henrion. 2012. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge: Cambridge University Press. https://www.cambridge.org/core/books/uncertainty/2D162A426F3DA7F3A8B05C3E4A3AD2D4.

Nappa, Antonio, Richard Johnson, Leyla Bilge, Juan Caballero, and Tudor Dumitraș. 2015. "The Attack of the Clones: A Study of the Impact of Shared Code on Vulnerability Patching." In 2015 IEEE Symposium on Security and Privacy, 692–708. https://doi.org/10.1109/SP.2015.48.

NBC. 2005. "Sasser Worm Author Confesses in Court." NBC News, July 5. https://www.nbcnews.com/id/wbna8471880.

Nevo, Sella, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott. 2024. Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_reports/RRA2849-1.html.

Newman, Lily Hay. 2017a. "How an Accidental 'Kill Switch' Slowed Friday's Massive Ransomware Attack." Wired, May 13. https://www.wired.com/2017/05/accidental-kill-switch-slowed-fridays-massive-ransomware-attack.

Newman, Lily Hay. 2017b. "The Leaked NSA Spy Tool That Hacked the World." Wired. https://archive.is/KrAD5.

Nguyen, Do Duc Anh, Pierre Alain, Fabien Autrel, Ahmed Bouabdallah, Jérôme François, and Guillaume Doyen. 2024. "How Fast Does Malware Leveraging EternalBlue Propagate? The Case of WannaCry and NotPetya." In

2024 IEEE 10th International Conference on Network Softwarization (NetSoft). https://hal.science/hal-04645862.

NIST. 2025. Managing Misuse Risk for Dual-Use Foundation Models (NIST AI 800-1, Initial Public Draft). National Institute of Standards and Technology. https://web.archive.org/web/20250115183155/https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd2.pdf.

NordVPN. n.d. "Sandbox Escape." NordVPN Cybersecurity Glossary. https://nordvpn.com/cybersecurity/glossary/sandbox-escape.

NTIA. 2024. Dual-Use Foundation Models with Widely Available Model Weights. National Telecommunications and Information Administration. https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf.

Okta. 2024. "What Is MyDoom Malware? History, How It Works & Defense." Okta Identity 101. https://www.okta.com/identity-101/mydoom/.

Olano, Gabriel. 2017. "Re/insurance to Take Minimal Losses from WannaCry: AM Best." Insurance Business Asia, May 25. https://www.insurancebusinessmag.com/asia/news/breaking-news/reinsurance-to-take-minimal-losses-from-wannacry-am-best-68536.aspx.

OpenAI. 2025. Preparedness Framework (Version 2). OpenAI. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf.

Ord, Toby. 2025. "Inference Scaling Reshapes AI Governance." Toby Ord (website), June 23. https://www.tobyord.com/writing/inference-scaling-reshapes-ai-governance.

Page, Chris. 2004. "Who Made MyDoom?" Letter to the editor. New Scientist, February 21. https://www.newscientist.com/letter/mg18124353-900-who-made-mydoom.

Perlroth, Nicole, and Scott Shane. 2019. "In Baltimore and Beyond, a Stolen N.S.A. Tool Wreaks Havoc." New York Times, May 25. https://www.nytimes.com/2019/05/25/us/nsa-hacking-tool-baltimore.html.

Qi, Xiangyu, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. "Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!" arXiv, October 5. https://arxiv.org/abs/2310.03693.

Risbey, James S., and Milind Kandlikar. 2007. "Expressions of Likelihood and Confidence in the IPCC Uncertainty Assessment Process." Climatic Change 85: 19–31. https://doi.org/10.1007/s10584-007-9315-7.

Root, Enoch. 2022a. "The Chronicle of WannaCry." Kaspersky Official Blog, August 24. https://www.kaspersky.co.uk/blog/wannacry-history-lessons/24858.

Root, Enoch. 2022b. "ILOVEYOU: The Virus That Loved Everyone." Kaspersky Official Blog. https://www.kaspersky.co.uk/blog/cybersecurity-history-iloveyou/24777.

Russinovich, Mark, Ahmed Salem, and Ronen Eldan. 2024. "Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack." arXiv, April 2. https://arxiv.org/abs/2404.01833.

Sanger, David E., and Martin Fackler. 2015. "N.S.A. Tapped Into North Korean Networks Before Sony Attack, Officials Say." New York Times, January 18. https://www.nytimes.com/2015/01/19/world/asia/nsa-tapped-into-north-korean-networks-before-sony-attack-officials-say.html.

Schneier, Bruce. 2004. "The Nonsecurity of Secrecy." Communications of the ACM 47, no. 10: 120. https://www.semanticscholar.org/paper/The-nonsecurity-of-secrecy-Schneier/ebed16d9c4395daef0412dfabb4eabfd1353d0fb.

Schneier, Bruce. 2007. "Gathering 'Storm' Superworm Poses Grave Threat to PC Nets." Wired, October 31. https://www.wired.com/2007/10/gathering-storm-superworm-poses-grave-threat-to-pc-nets.

Schwartz, Mathew J. 2017a. "Ransomware Smackdown: NotPetya Not as Bad as WannaCry." BankInfoSecurity, July 3. https://www.bankinfosecurity.com/ransomware-smackdown-notpetya-as-bad-as-wannacry-a-10077.

Schwartz, Mathew J. 2017b. "Teardown: WannaCry Ransomware." BankInfoSecurity, May 17. https://www.bankinfosecurity.com/blogs/teardown-wannacry-ransomware-p-2476.

Scott-Railton, John. 2023. "BLASTPASS: NSO Group iPhone Zero-Click, Zero-Day Exploit Captured in the Wild." The Citizen Lab, September 7. https://citizenlab.ca/2023/09/blastpass-nso-group-iphone-zero-click-zero-day-exploit-captured-in-the-wild.

Securelist. 2012. "The Flame: Questions and Answers." Securelist (Kaspersky), May 28. https://securelist.com/the-flame-questions-and-answers/34344.

Securelist. 2017. "Schroedinger's Pet(ya)." Securelist (Kaspersky), June 27. https://securelist.com/schroedingers-petya/78870.

SecureWorks Counter Threat Unit Research Team. 2017. "WCry (WannaCry) Ransomware Analysis." SecureWorks, May 18. https://www.secureworks.com/research/wcry-ransomware-analysis.

Seger, Elizabeth, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, Kevin Wei, Christoph Winter, et al. 2023. Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives. arXiv preprint arXiv:2311.09227. https://arxiv.org/pdf/2311.09227.

SentinelOne. 2019. "EternalBlue Exploit: What It Is and How It Works." SentinelOne, May 27. https://www.sentinelone.com/blog/eternalblue-nsa-developed-exploit-just-wont-die.

Silvanovich, Natalie. 2016. "Announcing the Project Zero Prize." Project Zero Blog, September 13. https://googleprojectzero.blogspot.com/2016/09/announcing-project-zero-prize.html.

Silvanovich, Natalie. 2017. "Project Zero Prize Conclusion." Project Zero Blog, March 29. https://googleprojectzero.blogspot.com/2017/03/project-zero-prize-conclusion.html.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." Psychological Science 22, no. 11: 1359–1366. https://doi.org/10.1177/0956797611417632.

Skelly, Brian. 2003. "Mimail Tops Virus Charts." Silicon Republic, August 8. https://www.siliconrepublic.com/enterprise/mimail-tops-virus-charts.

Slayton, Rebecca. 2017. "What Is the Cyber Offense-Defense Balance? Conceptions, Causes, and Assessment." International Security 41, no. 3: 72–109. https://direct.mit.edu/isec/article-abstract/41/3/72/12149/What-Is-the-Cyber-Offense-Defense-Balance.

Smeets, Max. 2022. "Hack Global, Buy Local: The Inefficiencies of the Zero-Day Exploit Market." Lawfare. https://www.lawfaremedia.org/article/hack-global-buy-local-inefficiencies-zero-day-exploit-market.

Smilyanets, Dmitry. 2021. "An Interview with BlackMatter: A New Ransomware Group That's Learning from the Mistakes of DarkSide and REvil." The Record by Recorded Future. https://therecord.media/an-interview-with-blackmatter-a-new-ransomware-group-thats-learning-from-the-mistakes-of-darkside-and-revil.

Smith, Brad. 2017. "The Need for Urgent Collective Action to Keep People Safe Online: Lessons from Last Week's Cyberattack." Microsoft On the Issues, May 14. https://blogs.microsoft.com/on-the-issues/2017/05/14/need-urgent-collective-action-keep-people-safe-online-lessons-last-weeks-cyberattack.

South West ComputAble. 2017. “WannaCry?” SWC, May. https://computable.com.au/archives/1587.

Stock, James H., and Mark W. Watson. 2002. "Forecasting Using Principal Components from a Large Number of Predictors." Journal of the American Statistical Association 97, no. 460: 1167–1179. https://www.princeton.edu/~mwatson/papers/Stock_Watson_JASA_2002.pdf.

Stubbs, Jack, and Pavel Polityuk. 2017. "Ukrainian Software Firm's Servers Seized After Cyber Attack." Reuters, July 4. https://web.archive.org/web/20170704214423/http://www.reuters.com/article/us-cyber-attack-ukraine-software-idUSKBN19O2DK.

Suderman, Alan, Eric Tucker, Frank Bajak, and Associated Press. 2021. "NSO Group Spyware Used to Hack State Department Employees." Fortune, December 3. https://fortune.com/2021/12/03/nso-group-spyware-hack-state-department-employees/.

Sullivan, Bob. 2003. "SoBig Spam-Virus Still Spreading." NBC News, June 26. https://www.nbcnews.com/id/wbna3078484.

Sullivan, Bob. 2004a. "Mydoom Threat Still High; Microsoft Offers Reward." NBC News, January 28. https://web.archive.org/web/20210805160108/https://www.nbcnews.com/id/wbna4065701.

Sullivan, Bob. 2004b. "'Sasser' Infections Begin to Subside." NBC News, May 6. https://www.nbcnews.com/id/wbna4890780.

Surowiecki, James. 2004. The Wisdom of Crowds. New York: Doubleday.

Symantec Threat Hunter Team. 2010. W32.Stuxnet Dossier. Symantec (Security.com), October 1. https://www.security.com/threat-intelligence/stuxnet-dossier-espionage.

Telecom Review Middle East. 2017. “Global ‘WannaCry’ Malware Attack Hits Telefónica, UK’s NHS and More.” Telecom Review Middle East, May 15. https://www.telecomreview.com/articles/telecom-operators/1340-global-wannacry-malware-attack-hits-telefonica-uk-s-nhs-and-more/.

Teodorczyk, Marcin. n.d. "Understanding Privilege Escalation." ADMIN Magazine. https://www.admin-magazine.com/Articles/Understanding-Privilege-Escalation.

Tetlock, Philip E., and Dan Gardner. 2015. Superforecasting: The Art and Science of Prediction. New York: Crown.

The Economist. 2024. "Secrets of a Ransomware Negotiator." 1843 Magazine, July 24. https://www.economist.com/1843/2024/07/24/secrets-of-a-ransomware-negotiator.

The Register. 2001. "Justice Mysteriously Delayed for 'Melissa' Author." The Register, August 1. https://www.theregister.com/2001/08/01/justice_mysteriously_delayed_for_melissa.

The Register. 2005. "The Strange Decline of Computer Worms." The Register, March 17. https://www.theregister.com/2005/03/17/f-secure_websec.

Thompson, Andi Wilson. 2021. "Assessing the Vulnerabilities Equities Process, Three Years After the VEP Charter." Lawfare, January 13. https://www.lawfaremedia.org/article/assessing-vulnerabilities-equities-process-three-years-after-vep-charter.

Thurrott, Paul. 2016. "Apple’s Active Installed Base Is Now Over 1 Billion Strong." Thurrott, January 27. https://www.thurrott.com/mobile/ios/64193/apples-active-installed-base-in-now-over-1-billion-strong.

Trend Micro. 2017. "Massive WannaCry/Wcry Ransomware Attack Hits Various Countries." Trend Micro, May 12. https://www.trendmicro.com/en_us/research/17/e/massive-wannacrywcry-ransomware-attack-hits-various-countries.html.

TRM Labs. 2024. "US and UK Authorities Identify, Sanction and Unseal Indictment Against Leader of LockBit Ransomware Group." TRM Labs, May 6. https://www.trmlabs.com/resources/blog/us-and-uk-authorities-identify-sanction-and-unseal-indictment-against-leader-of-lockbit-ransomware-group.

Tsai, Orange. 2021. "A New Attack Surface on MS Exchange Part 1 – ProxyLogon!" Orange Tsai (blog), August 6. https://blog.orange.tw/posts/2021-08-proxylogon-a-new-attack-surface-on-ms-exchange-part-1.

Tyas Tunggal, Abi. 2025. "What Is an SMB Port? A Detailed Description of Ports 445 + 139." UpGuard, July 6. https://www.upguard.com/blog/smb-port.

U.S.-China Economic and Security Review Commission. 2022. "China's Cyber Capabilities: Warfare, Espionage, and Implications for the United States." In 2022 Annual Report to Congress, 418–518. https://www.uscc.gov/sites/default/files/2022-11/Chapter_3_Section_2--Chinas_Cyber_Capabilities.pdf.

U.S. General Accounting Office. 2001. Information Security: Code Red, Code Red II, and SirCam Attacks Highlight Need for Proactive Measures. GAO-01-1073T. https://www.gao.gov/assets/gao-01-1073t.pdf.

U.S. House of Representatives. 2003. Computer Viruses: The Disease, the Detection, and the Prescription for Protection. 108th Cong. Washington, DC: Government Printing Office. https://www.govinfo.gov/content/pkg/CHRG-108hhrg90727/html/CHRG-108hhrg90727.htm.

UN. 2024. "Reports of the Panel of Experts." United Nations Security Council, 1718 Sanctions Committee (DPRK). https://main.un.org/securitycouncil/en/sanctions/1718/panel_experts/reports.

Unit 42. 2025. "Active Exploitation of Microsoft SharePoint Vulnerabilities: Threat Brief." Unit 42, Palo Alto Networks, July 31. https://unit42.paloaltonetworks.com/microsoft-sharepoint-cve-2025-49704-cve-2025-49706-cve-2025-53770.

Vaas, Lisa. 2016. "No-One Wants to Buy the Shadow Brokers' Stolen NSA Tools." Naked Security by Sophos, October 3. https://web.archive.org/web/20161227234935/https://nakedsecurity.sophos.com/2016/10/03/shadow-brokers-are-disappointed-about-lack-of-interest-in-nsa-tools-auction.

van der Merwe, Matthew, and Luca Righetti. 2026. Working paper. Centre for the Governance of AI (GovAI).

van der Merwe, Matthew, et al. Forthcoming. Threat Modeling Report: AI-Enabled Cyberattacks against the Power Grid. Centre for the Governance of AI (GovAI).

Vanderburg, Eric. 2017. "Ransomware Developers Learn from the Mistakes of WannaCry, NotPetya." Carbonite Blog, October 2. https://www.carbonite.com/blog/2017/ransomware-developers-learn-from-the-mistakes-of-wannacry-notpetya.

Venafi. 2022. "Babuk Source Code, DarkSide Custom Listings: Exposing a Thriving Ransomware Marketplace on the Dark Web." Venafi Blog. Archived September 20, 2024. http://web.archive.org/web/20240920194527/https://venafi.com/blog/babuk-source-code-darkside-custom-listings-exposing-thriving-ransomware-marketplace-dark-web/.

Vergara Cobos, Estefania, and Selcen Cakir. 2024. A Review of the Economic Costs of Cyber Incidents. World Bank. https://documents1.worldbank.org/curated/en/099092324164536687/pdf/P17876919ffee4079180e81701969ad0a18.pdf.

Vijayan, Jai. 2020. "3 Years After NotPetya, Many Organizations Still in Danger of Similar Attacks." Dark Reading, June 30. https://www.darkreading.com/threat-intelligence/3-years-after-notpetya-many-organizations-still-in-danger-of-similar-attacks.

Virsec. n.d. "Virsec Hack Analysis Lab: Deep Dive into NotPetya." Virsec. Archived January 26, 2025. http://web.archive.org/web/20250126153019/https://www.virsec.com/resources/blog/deep-dive-into-notpetya.

Wagner, Verena. 2014. "Explaining the Knobe Effect." In Experimental Ethics, 65–79. https://doi.org/10.1057/9781137409805_6.

Waisman, Nico. 2025a. "The Road to Top 1: How XBOW Did It." XBOW Blog, June 24. https://xbow.com/blog/top-1-how-xbow-did-it.

Waisman, Nico. 2025b. "XBOW on HackerOne: What's Next." XBOW Blog, August 18. https://xbow.com/blog/xbow-on-hackerone-whats-next.

Weaver, Nicholas. 2017. "Thoughts on the NotPetya Ransomware Attack." Lawfare, June 28. https://www.lawfaremedia.org/article/thoughts-notpetya-ransomware-attack.

White House Council of Economic Advisers. 2018. The Cost of Malicious Cyber Activity to the U.S. Economy. Council of Economic Advisers. https://trumpwhitehouse.archives.gov/wp-content/uploads/2018/02/The-Cost-of-Malicious-Cyber-Activity-to-the-U.S.-Economy.pdf.

Wilson, Bradley, Thomas Goughnour, Megan McKernan, Andrew Karode, Devin Tierney, Mark V. Arena, Michael J. D. Vermeer, Hansell Perez, and Alexis Levedahl. 2023. A Cost Estimating Framework for U.S. Marine Corps Joint Cyber Weapons. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_reports/RRA1124-1.html.

Winder, Davey. 2020. "This 20-Year-Old Virus Infected 50 Million Windows Computers in 10 Days: Why the ILOVEYOU Pandemic Matters in 2020." Forbes, May 4. https://www.forbes.com/sites/daveywinder/2020/05/04/this-20-year-old-virus-infected-50-million-windows-computers-in-10-days-why-the-iloveyou-pandemic-matters-in-2020/.

Wired. 2003a. "Microsoft Attacked by Worm, Too." Wired, January 25. https://www.wired.com/2003/01/microsoft-attacked-by-worm-too.

Wired. 2003c. "Worm Aims to Disarm Spam Fighters." Wired. https://www.wired.com/2003/12/worm-aims-to-disarm-spam-fighters.

Wired. 2004. "MyDoom Targets Linux Antagonist." Wired, January 28. https://www.wired.com/2004/01/mydoom-targets-linux-antagonist.

Yang, Kyle. 2017. "WannaCry: Evolving History from Beta to 2.0." Fortinet Blog, May 15. https://www.fortinet.com/blog/threat-research/wannacry-evolving-history-from-beta-to-2-0.

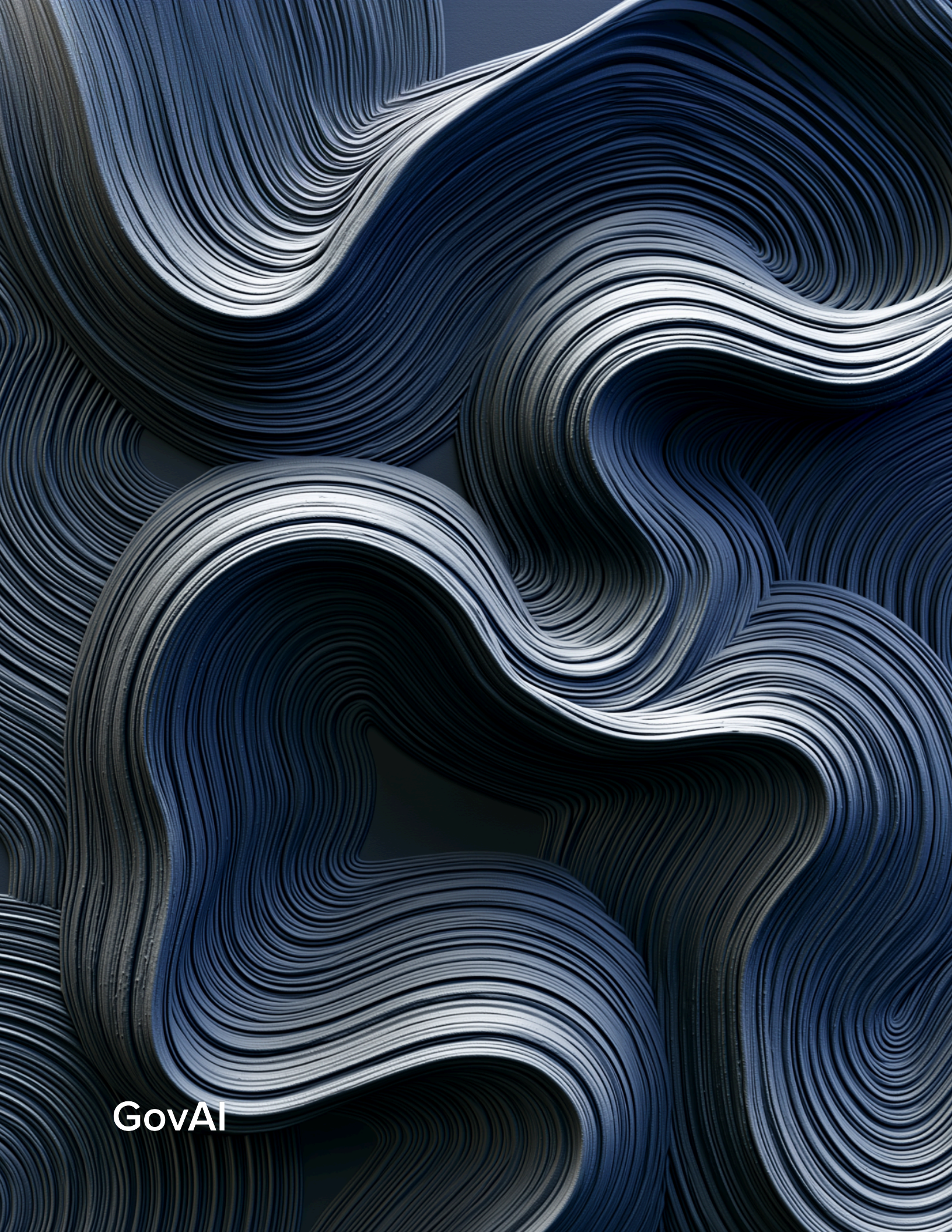
Yu, Vincent. 2024. "Cybersecurity Threat Advisory: RomCom Exploits Vulnerabilities." Smarter MSP, November 28. https://smartermsp.com/cybersecurity-threat-advisory-romcom-exploits-vulnerabilities/.

Zetter, Kim. 2011. "How Digital Detectives Deciphered Stuxnet, the Most Menacing Malware in History." Ars Technica, July 11. https://arstechnica.com/tech-policy/2011/07/how-digital-detectives-deciphered-stuxnet-the-most-menacing-malware-in-history/.

Zorz, Zeljka. 2025. "Apple Plugs Zero-Day Holes Used in Targeted iPhone Attacks (CVE-2025-31200, CVE-2025-31201)." Help Net Security, April 17. https://www.helpnetsecurity.com/2025/04/17/apple-plugs-zero-days-holes-used-in-targeted-iphone-attacks-cve-2025-31200-cve-2025-31201.

Zou, Andy, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. "Universal and Transferable Adversarial Attacks on Aligned Language Models." arXiv, July 27. https://arxiv.org/abs/2307.15043.

Zurier, Steve. 2020. "It Was 20 Years Ago Today: Remembering the ILoveYou Virus." Dark Reading, May 5. https://www.darkreading.com/threat-intelligence/it-was-20-years-ago-today-remembering-the-iloveyou-virus.


GovAI