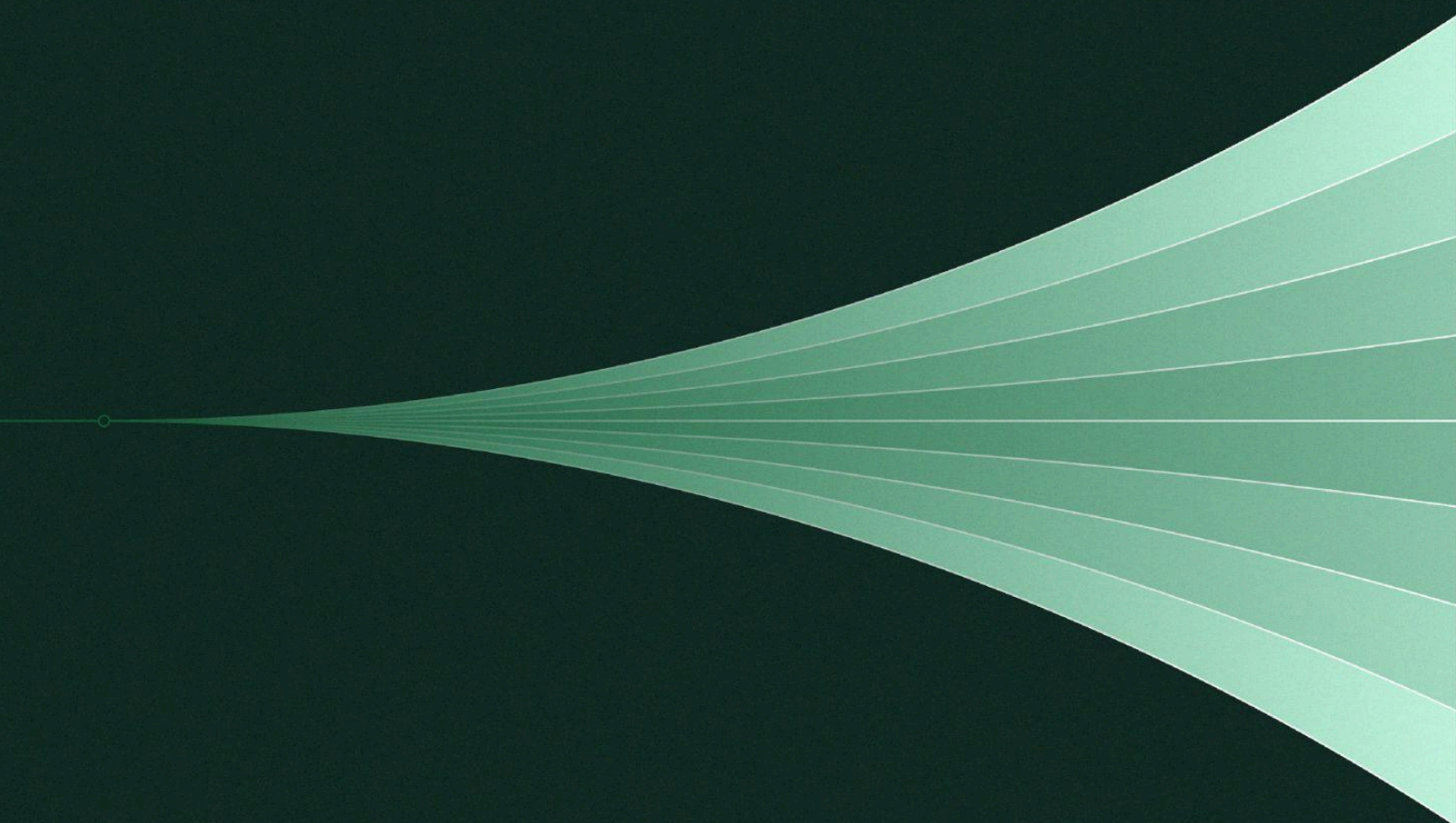




Forecasting AI Cyber Risks and Capabilities

Results of a 2025 Pilot Study

Authors: Rebecca Ceppas de Castro, Bridget Williams, John Halstead, Matthew van der Merwe,
Josh Connor



Authors and affiliations

Rebecca Ceppas de Castro¹, Bridget Williams^{1,2}, John Halstead³, Matthew van der Merwe³, Josh Connor¹

1 = Forecasting Research Institute, 2 = University of Oxford, 3 = Centre for the Governance of AI

Acknowledgments

This research would not have been possible without the support of Coefficient Giving or the thoughtful participation of our survey respondents. We are grateful to Luca Righetti for feedback on the survey and research design, and to Victoria Schmidt and Amory Bennett for assistance with survey design.

Disclaimers

This report is based on forecasts from a survey conducted primarily in July and August 2025, with follow-up responses collected between December 2025 and January 2026. Because AI-cyber capabilities are changing rapidly, the results should not be interpreted as a current assessment of frontier-model cyber capabilities. In particular, at the time of the survey, Claude Mythos Preview/Mythos 5, Claude Fable 5, GPT-5.3-Codex, GPT-5.5/GPT-5.6/GPT-5.5-Cyber and other cyber-specialized models had not been deployed.



Table of Contents

Executive Summary	03
Introduction	07
Methods	14
General results	16
Data-damaging worms: results	17
Electrical grid cyberattacks: results	28
Forecasts of AI progress on relevant capabilities	37
Limitations	39
Conclusion	40
References	41
Appendix A: Survey details	43
Appendix B: Follow-up survey	53
Appendix C: Additional results	63



Executive Summary

This pilot study, conducted primarily in July and August 2025, investigated how AI capabilities may affect near-term cybersecurity risk. Using structured forecasting with 21 participants—13 superforecasters and eight cybersecurity experts—we examined two high-impact cyberattack pathways in 2026:

- **Data-damaging worm attacks**, similar to WannaCry and NotPetya, that could cause at least \$10 billion in economic damages.
- **Cyberattacks against the U.S. electrical grid** that cause large-scale blackouts with at least \$10 billion or at least \$100 billion in economic damages.

Participants provided responses to the survey between **July 23rd and August 29th, 2025** and some participants responded to a follow-up survey between December 11th, 2025 and January 6th, 2026. Because the study was designed as a pilot, the results should be interpreted as initial evidence for areas that merit further research, not as definitive estimates. Even so, the forecasts show a consistent pattern: participants assessed baseline risks of catastrophic cyber harms in 2026 as low but non-negligible, and they expected some AI capabilities to substantially increase those risks, especially when AI lowers barriers for moderate-sophistication actors.

Key findings

Data-damaging worms were assessed as the higher-risk pathway.

Participants estimated a **5–8% probability** of at least one data-damaging worm attack causing at least \$10 billion in damages in 2026. The median participant's forecast of expected annual damages from data-damaging worms was approximately **\$10–15 billion**.

Electrical-grid cyberattack risks were assessed as substantially lower.

Participants estimated a **1% probability** that a cyberattack against the U.S. electrical grid would cause at least **\$10 billion** in damages in 2026, and a **0.1% probability** that such an attack would cause at least **\$100 billion** in damages. Expected annual damages were roughly **\$0.2–1 billion**, more than an order of magnitude lower than for worms.

AI-enabled elite exploit development produced the largest risk increase.

Under a hypothetical scenario in which AI models enable 25% of moderately-skilled individual hackers to find vulnerabilities and write elite exploits, and models with this capability are available open-weight, worm attack risk estimates **increase by 3–3.5x**. The median expert forecast of a data-damaging worm attack causing at least \$10 billion in damages rose from **8% to 41%**, while the median superforecaster forecast rose from **5% to 15%**.

AI-cyber progress has already outpaced participants' forecasts.

Participants generally expected the most relevant capabilities to emerge after 2026. Yet, since the survey closed in summer 2025, frontier models have very likely crossed the 90% threshold on Cybench, with one later reaching 100% on a subset of tasks. Recent model evaluations also suggest rapid movement towards AI-enabled exploit development capabilities, the capability that was of greatest concern to forecasters. These developments imply a shorter window for model testing, release safeguards, and defensive planning than the original forecasts suggested.



Current cyber benchmarks may be informative, but they are imperfect signals of real-world cyber risk.

Participants generally saw AI performance on Cybench as informative of technical progress, but not as the best standalone indicator of whether capabilities most relevant to catastrophic cyber outcomes, such as elite exploit development or real-world grid attacks, had been achieved. Respondents generally viewed **current benchmarks as incomplete measures of operational constraints such as zero-day discovery, targeting, coordination, and persistence**. Current cyber evaluations may track important technical progress without fully measuring the forms of AI-enabled uplift most relevant to real-world cyber risks.

Model access controls were seen as meaningful but incomplete mitigations.

Nearly half of the participants thought that keeping the relevant models—those that could enable moderate-sophistication actors to develop elite exploits—**proprietary and protected by anti-jailbreaking measures would at least halve the risk** of a large-scale worm attack. However, participants also noted that such measures may be less effective against sophisticated actors or exploit-as-a-service markets.

In a follow-up survey conducted four months later, forecasts remained largely unchanged.

A subset of participants revisited some of their forecasts between December 2025 and January 2026, after reviewing the original study results and Anthropic’s report on an AI-assisted cyber espionage campaign [1]. The median forecast for a data-damaging worm attack causing at least \$10 billion in damages in 2026 **remained unchanged**. Many participants thought the Anthropic report suggested that **actors were more capable** than they had originally expected, but most did not view it as direct evidence that AI could develop elite exploits or execute a large-scale data-damaging worm by 2026.

Table 1: Summary of 2026 risk forecasts by threat model

Threat model	Baseline ≥\$10B probability	Baseline expected damages	Main AI impact result
Data-damaging worm	Experts: 8% Superforecasters: 5%	Experts: ~\$15B Superforecasters: ~\$10B	Capability 1 (AI enables moderately-skilled individual hackers to develop elite exploits) raises the risk to 41% for experts and 15% for superforecasters. Expected damages increase by ~3–5x.
U.S. electrical grid cyberattack	Experts: 1% Superforecasters: 1%	Experts: \$0.2B Superforecasters: \$1B	Uplift in an ICS/OT capture-the-flag has only modest effect. An AI-enabled \$100M warning shot raises risk to 15% for experts and 4% for superforecasters.



Data-damaging worms: main takeaways

The median expert forecasted an 8% probability that a large-scale data-damaging worm attack would cause at least \$10 billion in economic damages in 2026, while the median superforecaster forecasted a 5% probability. These baseline forecasts were anchored by the low historical frequency of cyberattacks of this magnitude. However, participants noted that the 2017 WannaCry and NotPetya worm attacks demonstrate that worms can spread rapidly, and that future attacks could have large impacts in a more digitally dependent economy. Expected damages were approximately \$10–15 billion.

The main bottleneck identified in this threat model was the development of elite exploits; powerful software exploits that can spread without user interaction, enable high-privilege remote code execution, and work against widely used software. In agreement with this, AI-enabled elite exploit development was seen as a significant driver of increased risk. If an open-weight AI model enabled 25% of moderately-skilled individual hackers to find vulnerabilities and write elite exploits, the median expert forecast rose to 41% and the median superforecaster forecast to 15%. Participants viewed this as an important risk pathway because individual hackers are relatively numerous and may be more willing to cause damage, but they are usually capability-constrained; more sophisticated actors are generally more capable but constrained by escalation and retaliation risks.

Participants expected model-access controls and safeguards to reduce this risk, especially for actors below state-level sophistication, but not to eliminate it. In particular, having models with the relevant capabilities protected with anti-jailbreak measures was believed to significantly decrease the probability of a large-scale data-damaging worm attack. Model-weight theft, exploit-as-a-service markets, and slow patching of vulnerabilities remained important limitations.

Probability of a data-damaging worm attack causing $\geq$ \$10 billion in economic damages in 2026

We show baseline forecasts and forecasts conditional on three scenarios. Forecasts for P1 and P2 are also conditional on C1 having been met.

C1 (Capability 1): AI enables 25% of moderately-skilled individual hackers to develop elite exploits.

P1: Proprietary models with anti-jailbreak measures.

P2: Temporary protections with early access to defenders.

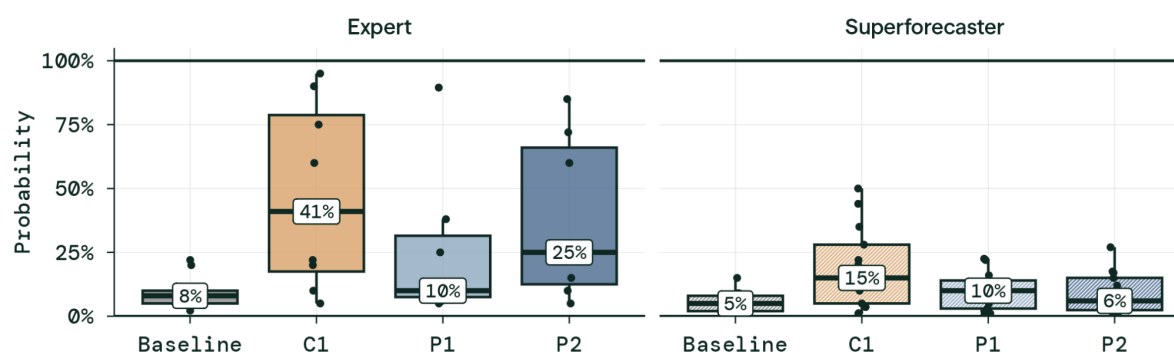


Figure 1: Forecasts of the probability of at least one data-damaging worm attack causing at least \$10 billion in economic damages in 2026, under baseline assumptions, and conditional on AI capabilities and potential mitigations.



Electrical grid cyberattacks: main takeaways

Participants assessed large-scale cyberattacks against the U.S. grid in 2026 as substantially less likely than data-damaging worm attacks. The median expert and superforecaster both put the probability of a cyberattack against the grid causing at least \$10 billion in damages at 1%; for a larger attack, causing at least \$100 billion, both medians were 0.1%. Expected annual damages were also much lower, at approximately \$0.2–1 billion.

Participants emphasized that grid attacks face additional barriers beyond cyber capability, including expertise in industrial control systems (ICS), operational coordination, physical infrastructure constraints, and geopolitical escalation risks. They generally viewed a large-scale grid attack as more likely in the context of war, or a state-level conflict, than as ordinary cybercrime.

Forecasts of large-scale cyberattacks against the U.S. electrical grid were less sensitive to AI capability scenarios than forecasts in the data-damaging worm threat model. AI uplift on an ICS/OT (operational technology) capture-the-flag style competition only modestly increased the median forecast from 1% to 1.5–2%. A real-world AI-enabled warning shot causing at least \$100 million in damages produced a larger increase in risk, raising the median forecast to 15% among experts and 4% among superforecasters. Overall, results from this pilot suggest that AI may increase risk associated with this threat model, but large-scale grid attacks remain constrained by operational complexity and geopolitical considerations.

Probability of a cyberattack against the U.S. electrical grid causing a blackout with $\geq$ \$10 billion in economic damages in 2026

Capability 3: AI enables individual hobbyist hackers to match a group of roughly ten well-known criminal hackers in an OT-specific CTF-style competition.

Capability 4: A real-world warning shot incident, where AI enables an individual hacker to launch a cyberattack against the U.S. grid, causing at least \$100M in damages.

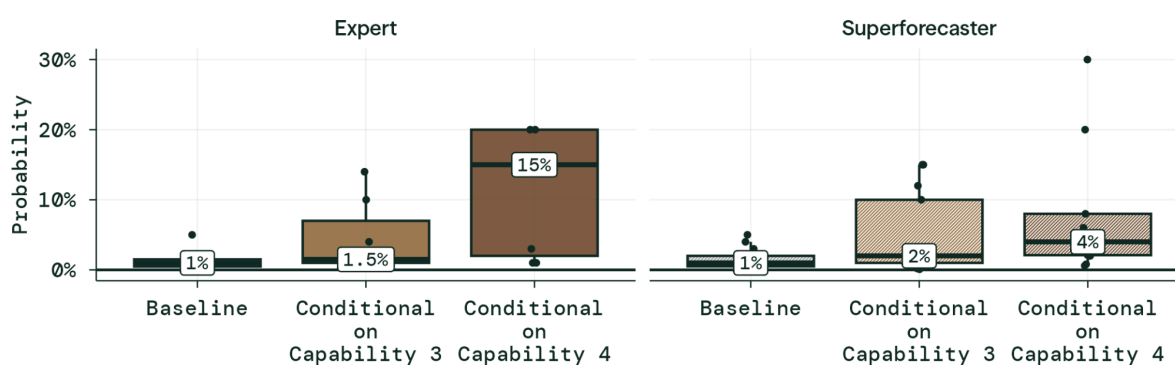


Figure 2: Forecasts of the probability of at least one cyberattack against the U.S. electrical grid causing at least \$10 billion in economic damages in 2026, under baseline assumptions, and conditional on Capability 3 (AI enables TA1 actors to perform like a TA3 actor in an OT-specific capture-the-flag style competition) and Capability 4 (a real-world warning shot incident).



Introduction

Rapid advances in artificial intelligence capabilities have introduced new dimensions of uncertainty into the cybersecurity landscape. As frontier AI systems become increasingly sophisticated in their ability to identify vulnerabilities, write code, and automate complex tasks, questions arise about how these capabilities might alter the cyber threat environment. AI could strengthen defenders by supporting vulnerability discovery, patch development, threat detection, and incident response. It could also benefit attackers by reducing the skill, time, and coordination required to identify vulnerabilities, develop exploits, and scale attacks. A central uncertainty is whether AI will strengthen defensive capabilities overall, disproportionately benefit malicious actors, or affect different threat actors in different ways.

This pilot study is an initial effort to systematically forecast how AI might affect large-scale cyber risks over the near term, with a particular focus on 2026. We used structured forecasting methods to elicit judgments about two high-impact cyber threat scenarios: data-damaging worm attacks and cyberattacks against the United States (U.S.) electrical grid.

Background

Data-damaging worms

A “worm” is malware that can spread autonomously between systems without an attacker needing to infect each system individually. Worms can cause damage in several ways. In this survey, we focused on “data-damaging worms”: worms that directly damage—by wiping, encrypting, or corrupting it—on a large number of systems. Relevant historical examples include [WannaCry](#) and [NotPetya](#), both released in 2017, which are estimated to have caused roughly \$1B–\$10B in damage [2, 3] after infecting and damaging hundreds of thousands of systems.

Worm attacks can exploit **vulnerabilities**: bugs in software or hardware that create security weaknesses in the design, implementation, or operation of a system or application. 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. We define exploits (or exploit chains) as **“elite” if they satisfy all of the following criteria**:

- **Zero-click**: Infection requires no user interaction, such as opening emails, clicking links or visiting a webpage
- **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**: Have administrator privileges or higher.
- **Targets widely used software**: Effective against more than 10 million systems.

Elite exploits are especially well-suited to worm attacks as they enable autonomous spread and significant damage on a large number of systems. The leak of elite exploits initially developed by the NSA in 2017 quickly led to the two major data-damaging worm



attacks cited above: WannaCry and NotPetya. Developing elite exploits may require an order of magnitude more skilled researcher time than the other tasks involved in developing a data-damaging worm.

Table 2 below uses data adapted from Johansmeyer (2024) [3] and provides details on past major worm attacks. Here, “major worm attacks” are attacks that affected more than 10–25 companies, and for which at least one source claims damages of more than \$800M.

Where this table and later results refer to TA1–TA5 actors, these labels denote classes of threat actors in terms of sophistication, or *operational capacity* (OC). These categories are ordered by ascending capability: TA1 actors are assessed as able to carry out OC1 operations, TA2 actors are able to carry out OC2 operations, and so on up to TA5. These categories are based on the five levels of cyberattack operational capacity defined by RAND [23, pp. 9–10]. By definition, each category includes the capacities of all preceding categories. For example, the most capable nation-states, such as the U.S. and China, are able to carry out OC5 operations and all operations below that level. [Appendix A](#) provides the full definitions.

Table 2: Past major worm attacks: descriptions, approximate infections, exploit type, and data-damaging status.

Attack name	Year	Actor	Description	# Infections	Elite exploit?	Data-damaging worm?
Melissa	1999	TA1/2	Email worm. Damage via high network traffic, no other destructive payload	~100K	☐	☐
ILOVEYOU	2000	TA1/2	Email worm. Corrupted documents, later variant wiped hard drive.	~50M	☐	☒
Klez	2001	TA1/2?	Email worm. Damage via high network traffic and disabling antivirus	~7M	☐	☐
CodeRed	2001	?	Zero-click. Defaced specific websites and launched targeted denial of service attacks against sites	~360K	☐	☐
Nimda	2001	?	Zero-click + email spread. Damage via high network traffic and elevated privileges	>1.3M	☒	☐



SirCam	2001	TA1/2?	Email worm. Could expose confidential info, and delete all files in certain conditions	~2.3M	☐	☒
SoBig	2003	TA1/2?	Email worm to distribute spam. Also caused damage by creating high network traffic	>1M	☐	☐
SQL Slammer	2003	TA1/2?	Zero-click, but limited reach. Damage via high network traffic, no malicious payload	>75K	☐	☐
Swen	2003	?	Email worm. Disabled antivirus and firewalls, no other destructive payload	~1.5M	☐	☐
Mimail	2003	?	Email worm. Launched targeted denial of service attacks against anti-spam sites. Some variants stole credit card information	~21K	☐	☐
Yaha	2003	TA3	Email worm. Terminated security processes and launched denial of service attacks	?	☐	☐
MyDoom	2004	TA1/2?	Email worm. Created botnet to allow targeted denial of service attacks. Created high network traffic	~500K	☐	☐
Sasser	2004	TA1/2	Zero-click. Damage only via large volumes of network traffic, no malicious payload	~500K-1M	☒	☐
Storm Worm	2007	TA3	Spread via email and social engineering. Created denial of service botnet to attack groups	1M-50M	☐	☐
Conficker	2008	TA3	Zero-click. Created large botnet, but never used for significant attack due to concern about criminal repercussions	~10M	☒	☐



WannaCry	2017	TA4	Zero-click. Encrypted files. Ransomware apparently to raise money for North Korea	~230K	☒	☒
NotPetya	2017	TA5	Zero-click. Encrypted files. Designed to limit damage to Ukraine	~670K	☒	☒

In recent years, the number of major worm attacks has steeply declined, and worm attacks have been released by more sophisticated actors. It is plausible that these trends were driven by improvements in cybersecurity (for example in email filtering technology, patching practices, and malware detection systems) over time. These ongoing improvements may make it harder to cause substantial damage with worms today.

The effect of AI on the risk of worm attacks is uncertain and likely to vary over time. AI systems that can discover vulnerabilities, develop exploits, or automate attack steps are dual-use: the same capabilities could help both defenders and attackers. The net effect depends on several uncertain factors, including which actors gain access to the most capable models, how quickly vulnerabilities are disclosed and patched, and whether patch deployment keeps pace with discovery.

Cyberattacks against the U.S. electrical grid

The power grid consists of generators, transmission and distribution networks, and substations. Grid operations, like other infrastructure and industrial processes, rely on two broad types of computer systems:

- **Information technology (IT) systems:** systems that handle most business operations like billing, email, and administration, and are typically connected to the internet.
- **Operational technology (OT) systems:** programmable systems that interact with the physical environment or manage devices that do. In grid operations, these systems will monitor and control equipment like generators, circuit breakers, and transformers.

Grid cyberattacks will involve compromising OT systems, since this is a precondition for interfering directly with grid behavior. OT environments, when compared to IT environments, present additional challenges. They typically are—or should be—segmented from IT and internet-facing networks [7], and they use more niche software and protocols requiring more specialized knowledge; OT devices may have relatively individualized configurations to a given environment.

In this study, we focused on major blackouts: large-scale outages affecting hundreds of thousands of customers or more. There is less historical data on large-scale grid cyberattacks than on many other cyber threats. Examining databases of prior cyberattacks, including [CSIS](#), [CFR](#), and [CISSM](#), identified only three instances in which cyberattacks against the grid caused power outages, all of which occurred in Ukraine. There are many more cases of grid cyberattacks that did not result in outages but did



cause other disruptions, such as temporary loss of operator visibility or control over grid equipment. Between 2010 and 2022, U.S. utilities reported roughly four disturbances per year due to cyberattacks—to our knowledge, none of these have resulted in outages [8]. Table 3 gives more details on past grid cyberattacks.

Table 3: Selected past grid cyberattacks

Location	Year	Actor	Details	Blackout
Ukraine	2015	TA5	Attack on distribution substations caused ~3.5h outage for 225k customers	☒
Ukraine	2016	TA5	Attack on transmission substation caused ~1h outage of similar scope	☒
US	2018	?	Brief loss of control/view over generation assets [9; 10]	☐
US	2019	?	Control centre IT network offline for 12–24h after a ransomware attack [11]	☐
Norway	2019	?	Reported loss of \$67–84m; operators resorted to manual controls [12]	☐
Ukraine	2022	TA5	Attempted attack early in the Russia–Ukraine war, thwarted by defenders	☐
Germany	2022	TA5	Attack on Viasat modems in Ukraine led to operator's loss of control and view of 5,800 wind turbines (11GW) in Germany [13]	☐
Ukraine	2022	TA5	Attack on substation caused outage coinciding with missile attack [14]	☒
Denmark	2023	TA4/5	One operator lost visibility into assets in three remote locations; the incident had “no material impact to energy operations” [15; 16]	☐

Cyber-induced blackouts have been both infrequent and modest in scale. The three blackouts in Ukraine following grid cyberattacks resulted in outages lasting a few hours for fewer than one million people.¹ We estimate these blackouts caused roughly \$5 million in economic damages altogether.

In these cyberattacks, attackers compromised Ukrainian utility companies, gained access to OT environments, and opened circuit breakers at substations, stopping power from flowing through the substation and causing outages. In the 2015 attack, attackers also disrupted communications between utility control centers and substations, forcing utilities to send field crews to affected substations to manually re-close the breakers

¹ Note that customers≠people. In the US, there are ~2.1 people per electricity customer [17]. In Ukraine, the ratio is ~2.3 people per customer —17.7m electricity customers [18] with a population of 41.2m.



before power could be restored. These attacks did not cause widespread physical damage to grid equipment or cascading failures — their impacts were fairly localized.

By contrast, the worst accidental blackouts have caused much greater impacts. For example, the 2003 Northeast Blackout in the U.S. is estimated to have cost roughly \$12 billion in 2024 dollars, affected about 50 million people, and lasted up to two days. In economic-damage terms, this represents roughly four orders of magnitude greater impact than the 2015 Ukraine attack.

Table 4: Historical accidental blackouts and grid cyberattacks

Blackout	Type	Duration (hours)	Scope (millions of customers)	Damages (2024 USD)
2003 Northeast US	Accident	43	24	\$12B
2003 Italy/Switzerland	Accident	15	20	\$2B
2015 Ukraine	Cyberattack	3.5	0.2	~\$0.1–10M
2016 Ukraine	Cyberattack	1	0.3	<\$0.1M
2022 Ukraine	Cyberattack	unknown	unknown	unknown

Beyond the historical record, several scenario estimates suggest that deliberate U.S. blackouts could cause damages of \$100 billion or more under certain assumptions (Table 5). These estimates vary substantially in their assumptions, and only Lloyd’s scenario explicitly describes a cyberattack.

Table 5: Selected estimates for worst-case deliberate blackouts

Source	Damages	Region	Type	Scenario details
Lloyd’s Business Blackout (2015) [19]	\$80–300B economic costs ²	Eastern US	Cyberattack	• Cyberattack causes a blackout over Eastern US. • 3 scenarios ranging in severity, with time to restore 90% of power ranging from 2–4 weeks (lost load 19–68 TWh).
UK Risk Register (2025) [20]	£10–100B	UK	Accident or deliberate	• A total failure of the national transmission system leading to nationwide loss of power—full restoration could take up to 7 days [20, p. 90]. • Note that the UK economy is ~8x

² They also estimate a cumulative GDP impact of \$240–1,000bn over five years.



				smaller than US
Rose et al. (2007) [21]	\$2–15B	LA County	Kinetic attack	• Kinetic terrorist attack causes two-week blackout. • Estimates losses using 3 different assumptions about level of resilience. • Assuming linear recovery, damages of ~\$2–15B, i.e. 13–94% loss of regional economic output. • Note that LA County is ~1/30 of the US economy.³
National Academy of Sciences 2012 [22, p.16]	\$100B	US (unspecified scope)	Attack (unspecified cyber/kinetic)	• “A systematically designed and executed terrorist attack [against the grid] could cause disruptions considerably more widespread and of much longer duration than the largest power system disruptions experienced to date could lead to costs of hundreds of billions of dollars—that is, perhaps as much as a few percent of the U.S. GDP.”

The most relevant prior OT cyberattacks have required substantial time, resources, and specialized expertise, often from state-level actors. Beyond the significant resource and time investments, prior OT cyberattacks have required a diverse range of capabilities, including long-term reconnaissance, knowledge of the target OT environments, tailored malware, realistic test environments, and stealth.

These requirements make the role of AI difficult to assess. AI could plausibly assist with some components of a grid attack, such as reconnaissance, code generation, vulnerability discovery, and operator decision support. It is less clear how much AI would help with other important bottlenecks, such as obtaining access to segmented OT systems, understanding highly specific grid configurations, avoiding detection, and coordinating an operation over time.

³ <https://fred.stlouisfed.org/series/REALGDPALLO6037>; <https://fred.stlouisfed.org/series/GDPCA>



Methods

To develop the survey, we used an iterative process in which the research team drafted forecasting questions, a small sample of experts and superforecasters answered them, and the team revised the questions in light of how respondents interpreted them. We conducted two rounds of this process. Because forecasts can be highly sensitive to question wording and resolution criteria, we revised the questions to improve clarity and focus on the most important aspects of AI capability progress. We focused the survey on two possible pathways to large-scale harms from AI: data-damaging worm attacks and attacks against the U.S. electrical grid. These pathways are only a subset of the ways that AI could be used to perpetrate cyberattacks. In addition to the survey questions, we also provided participants with some background information, which can be viewed in [Appendix A](#).

Because this was designed as a pilot study, we used convenience sampling and aimed for a sample of around 15 to 30 people.⁴ We invited two groups of respondents: 1. people with expertise in cybersecurity and AI impacts on cybersecurity (“experts”) and 2. superforecasters, who are 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 the time they spent completing the survey; the average payment was \$700. Participants responded to the main survey between July 23rd and August 29th, 2025.

A few months after the main survey, we recontacted participants with a short, one-hour follow-up survey to capture any significant changes in their views. In this survey, we presented a summary of the main results from the original survey (aggregate numerical results and a description of the main arguments given by participants), as well as a description of Anthropic’s cybersecurity incident report [1]. Participants reflected on these and were given the opportunity to update a subset of their forecasts between December 11th, 2025 and January 6th, 2026. For more details, see [Appendix B](#).

Survey structure

We asked participants to consider two cyberattack pathways. For the first scenario, in which a data-damaging worm attack causes at least \$10 billion in economic damages in 2026, participants were asked to provide:

1. **Baseline** (unconditional) forecasts of this risk
 - a. Including holistic forecasts and forecasts of actor capability and willingness
2. Forecasts **conditional on the outcomes of a hypothetical AI capability scenario**:
 - a. Capability 1: Vulnerability and elite exploits uplift
 - i. A study conducted at the end of 2025 finds that access to frontier AI models enables 25% of moderately-skilled individual hackers to find vulnerabilities and write elite exploits, assuming three months of full-time effort.

⁴ In convenience sampling, participants are recruited on the basis of being available and relatively easy to access.



- ii. P0: All models that meet Capability 1 are open-weight
 - b. Including holistic forecasts and forecasts of actor capability
- 3. Forecasts **conditional on the implementation of two mitigation measures**
 - a. P1: Proprietary models with refusals and anti-jailbreak measures
 - b. P2: Temporary protections with early access for defenders

For the second scenario, a cyberattack against the U.S. electrical grid causing a blackout with at least \$10 billion (or at least \$100 billion) in economic damages in 2026, participants provided:

1. **Baseline** (unconditional) forecasts of this risk
 - a. Including holistic forecasts and forecasts of actor capability and willingness
 - b. For both the \$10 billion and \$100 billion versions
2. Forecasts **conditional on the outcomes of hypothetical scenarios:**
 - a. Capability 3: Uplift in an OT-specific capture-the-flag style competition
 - i. A study is conducted at the end of 2025 to measure AI uplift on a CTF-style competition focused on industrial control systems (ICS) and OT cybersecurity, finds that, with access to AI, individual TA1 actors complete as high a proportion of tasks as a team of ten experienced cybersecurity professionals (equivalent to a TA3 actor) without access to AI, when both teams are given the same amount of time.
 - b. Capability 4: Real-world warning shot
 - i. A cyberattack against the U.S. grid causes a blackout, leading to at least \$100 million in economic damages, and is subsequently confirmed by credible sources to have been perpetrated by low- or moderately-skilled individual hackers using AI.
3. Comments about relevant mitigation measures

Questions around actor capability and willingness were intended to distinguish between an actor's capability to produce a specified outcome and their willingness to attempt such an attack, conditional on being capable. As such, financial damage values were present in the wording of both questions. It is possible that respondents had different interpretations of the capability questions, either as asking whether an actor could achieve the specific damage threshold, or more loosely as asking whether the actor could mount a serious attack of that general type. We discuss implications of this ambiguity where relevant as we present results.

See [Appendix A](#) for the detailed survey questions and resolution criteria.

General results

The sample size for this study was small: 21 participants completed the original survey, including 13 superforecasters, six experts who completed the full survey, and two experts who completed shorter versions. We invited all respondents who completed the survey to participate in a follow-up study. Table 6 shows completion rates.

Table 6: Survey sample description.

	Total	Superforecasters	Experts
Original survey	21	13	8
Follow-up survey	17	13	4
<i>Proportion of original participants who completed the follow-up survey</i>	81%	100%	50%

Threat actors

Some forecasting questions referred to the TA1–TA5 threat actor categories introduced earlier in the report. For estimating the number of actors in each category, we treated states as single, unitary threat actors. For instance, we treated China as a single threat actor, rather than treating each Chinese state or state-backed advanced persistent threat (APT) group as a separate actor.⁵

Figure 3 shows estimates of the number of threat actors in each category, and Table 7 breaks these estimates down by participant group.

Table 7: Estimates of the number of threat actors in each TA category. Values are group medians, with interquartile ranges shown in parentheses.

	TA1	TA2	TA3	TA4	TA5
Experts	1.2m (115k, 3.75m)	24k (8.3k, 71k)	300 (169, 500)	13 (9, 18)	4 (4, 5)
Superforecasters	1m (300k, 5.5m)	50k (20k, 250k)	500 (300, 4.5k)	50 (20, 75)	6 (5, 10)

⁵ In this respect, our definition is different from that commonly used in cyber threat modelling.

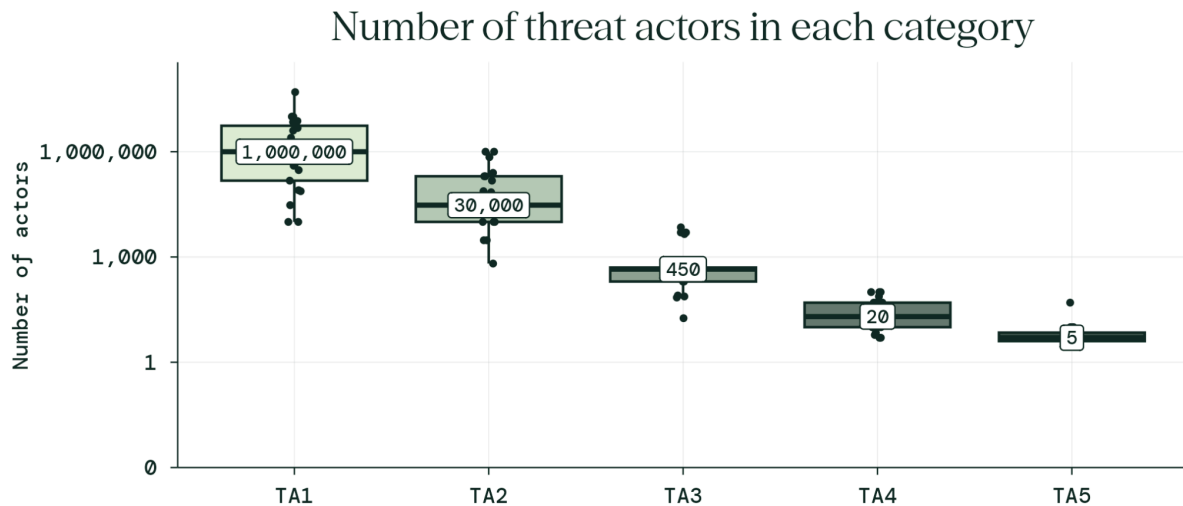


Figure 3: Participant forecasts of the number of threat actors in each TA category. Labels show medians from the full sample.

Data-damaging worms: results

Baseline forecasts

Participants first provided a baseline forecast for the probability that a data-damaging worm attack would cause at least \$10 billion in economic damages in 2026. The median expert forecast was 8% (IQR: 5–10%), and the median superforecaster forecast was 5% (IQR: 2–8%). Figure 4 shows participants' responses to this question.

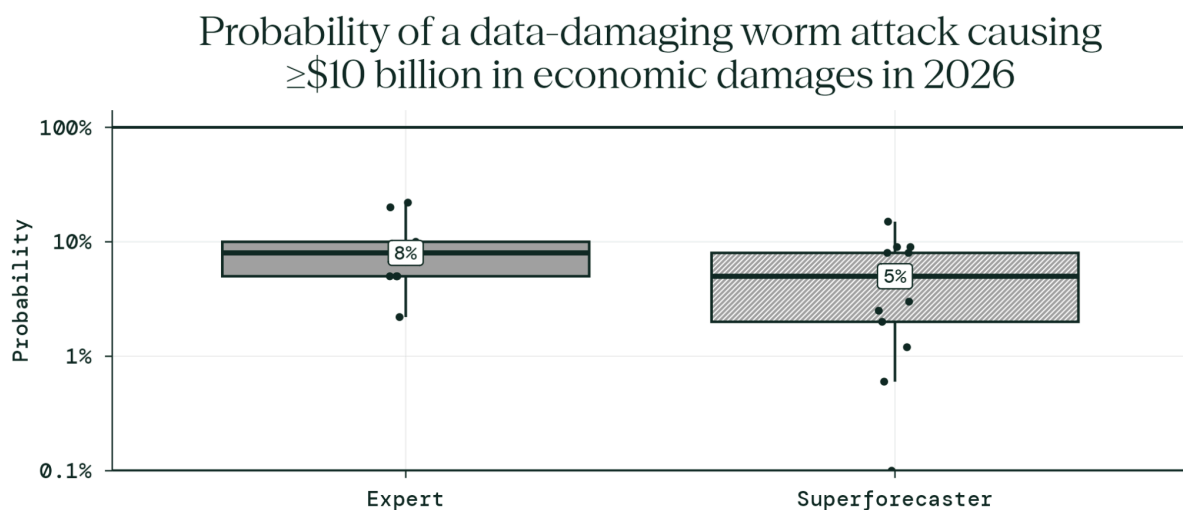


Figure 4: Forecasts of the probability of at least one data-damaging worm attack causing at least \$10 billion in economic damages in 2026.

Participants' rationales commonly referenced the very low base rate for events of this kind, noting that \$10 billion would be a damage level with few precedents for this type of attack. Many noted that only TA4 and TA5 actors would have the resources to conduct this scale of attack. Although some participants noted that there are several active conflicts or areas of geopolitical tension globally (Ukraine/Russia, Iran/Israel,



China/Taiwan) they also noted that there is generally little motivation for major powers to use cyberattack capabilities. Some respondents suggested that AI could cause future outcomes to depart from historical patterns, but thought that 2026 would probably be too soon for this effect to dominate. There were mixed views on whether improved cyberdefense outweighs potential for greater harm in a more digitally dependent society.

Conditional on an event of this magnitude occurring in 2026, participants thought TA4 and TA5 actors were the most likely causes. For both experts and superforecasters, these actors accounted for roughly 80% of the probability (see Figure 5).

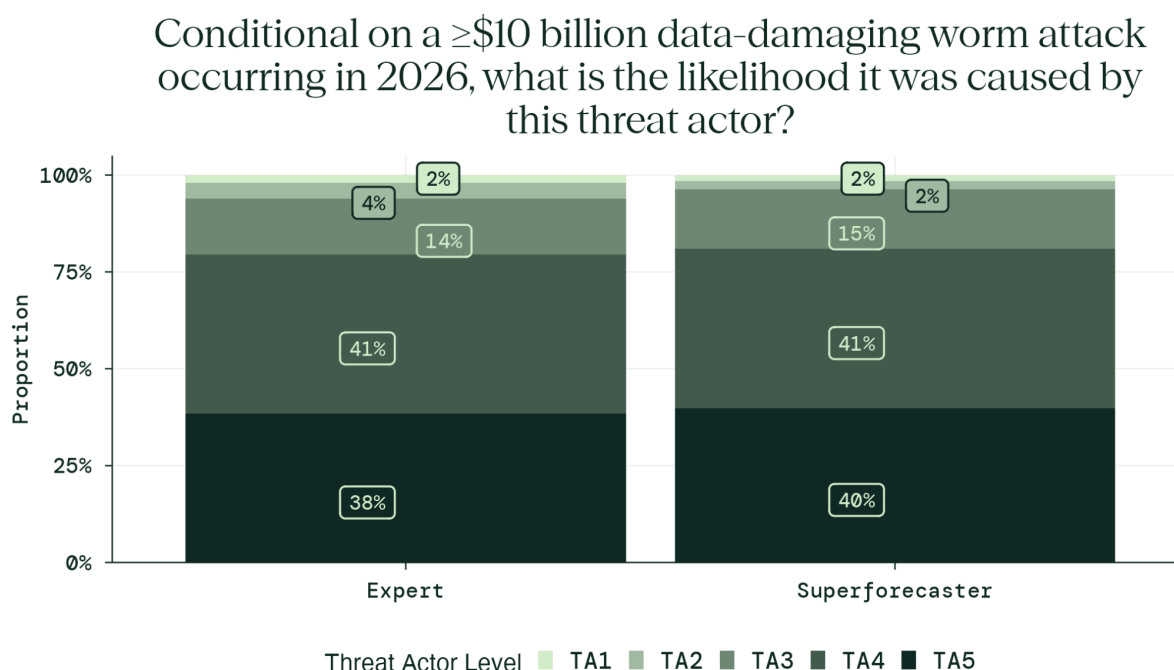


Figure 5: Average probability assigned to each threat actor type as the primary cause of a data-damaging worm attack, conditional on such an attack occurring in 2026 and causing at least \$10 billion in damages.

Rationales for this question noted that historical attacks of comparable magnitudes were attributed to TA4/5 actors. TA4 actors were considered slightly more likely than TA5 actors because they were perceived as more motivated to conduct such an attack and less constrained by escalation concerns. Many participants noted that it seemed close to impossible for a TA1 or TA2 actor to succeed at such an attack. A TA3 actor would be unlikely but could perhaps succeed with state support or access to leaked exploits.

Forecasts conditional on AI capabilities and mitigations

We then asked participants how their forecast would change under a hypothetical AI-capability scenario. This scenario, labeled Capability 1, involved AI enabling moderately-skilled individual hackers (TA2 actors) to find vulnerabilities and write elite exploits. It was described as:

BOX 1: CAPABILITY 1 DESCRIPTION

Capability 1: AI enables TA2 actors to write elite exploits.

A study conducted at the end of 2025 finds that access to frontier AI models enables 25% of moderately-skilled individual hackers (TA2 actors) to find vulnerabilities and write elite exploits, assuming three months of full-time effort.

Participants then assumed that this capability had been achieved and that models with this capability were available open-weight without any cyber safeguards (mitigation scenario PO). Under these assumptions, the median expert forecast of the probability of a data-damaging worm attack causing at least \$10 billion in 2026 rose to 41% (IQR: 17.5–78.8%). For the median superforecaster, it rose to 15% (IQR: 5–28%). (See Figure 6.) The median relative increase in risk was 3.5x (IQR: 2–10.8x) for experts and 3x (IQR: 2–6x) for superforecasters.

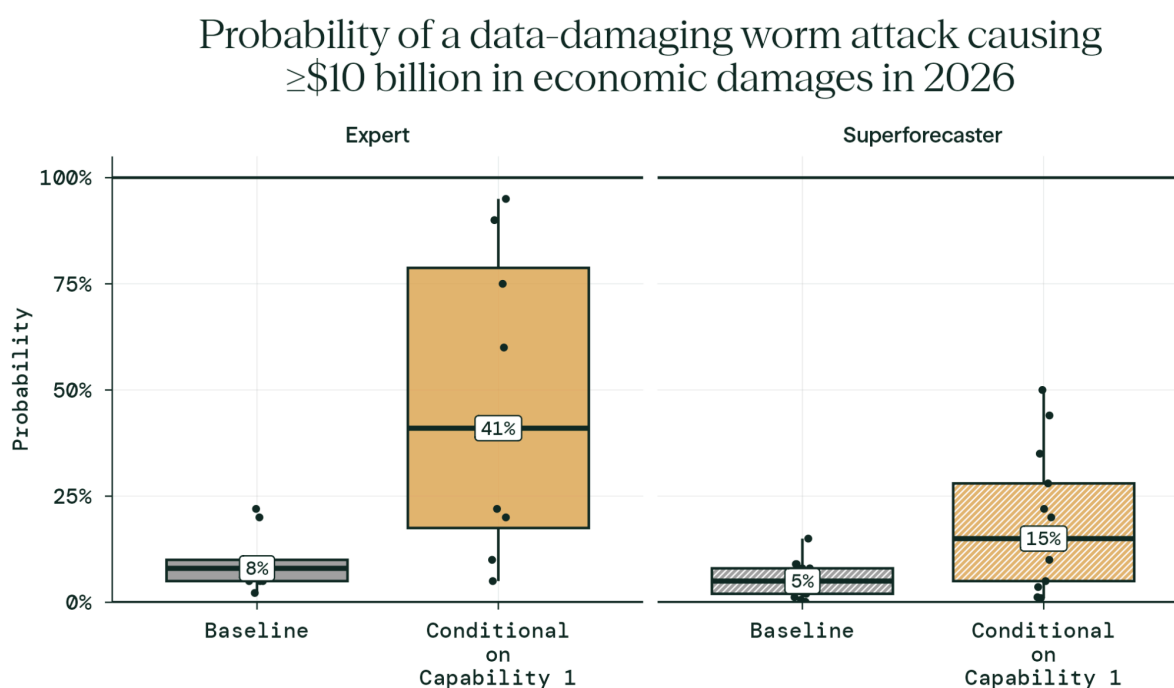


Figure 6: Forecasts of the probability of at least one data-damaging worm attack causing at least \$10 billion in economic damages in 2026, conditional on Capability 1 (AI enables 25% of moderately-skilled individual hackers (TA2) to write elite exploits).

There was general consensus that this scenario would represent a substantial increase in risk. Rationales noted that finding elite exploits would remove the most important bottleneck for TA2 actors, and given the large number of these actors and their lack of restraint compared to TA5 actors, this would indicate a substantial increase in risk. Actors newly enabled by this capability were also thought more likely to cause massive damage accidentally through 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 emphasized a defense–offense timing imbalance: although AI



helps both sides, there may be a “dangerous window of vulnerability” when new capabilities first emerge.

We then asked participants to say how their forecasts would change conditional on mitigations being put in place (still assuming Capability 1 had been met). We asked about two mitigation scenarios, described below.

BOX 2: POLICY 1 DESCRIPTION

P1: Proprietary models with refusals and anti-jailbreak measures

- The models used in the study (and similar models) are all proprietary and 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 August 31, 2024.
- Companies require users to access them via APIs that are subject to the following safeguards:
- Pre-deployment red-teaming to identify jailbreaks
- A voluntary goal of not letting any new universal jailbreak remain unpatched for more than 2-weeks over any given three-month period.
- A “bug bounty” program that offers up to \$15,000 rewards for anyone who identifies and reports a universal jailbreak for one of their models.
- 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).

BOX 3: POLICY 2 DESCRIPTION

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 only Microsoft, Meta, Apple, and Google and the world’s best highly vetted bug bounty hunters.
- After four months, the model is released under PO security, i.e., is open weight.



The detailed description of these scenarios, which was provided to participants, is available in [Appendix A](#).

Both groups of participants generally thought that these measures would meaningfully reduce risk. The median expert believed that mitigation scenario P1 (proprietary models with refusals and anti-jailbreak measures) would more than halve the probability of a data-damaging worm causing at least \$10 billion in damages, conditional on Capability 1 (Figure 7). Experts generally thought that P1 offered better protection than P2 (temporary protections with early access for defenders), with the median expert respondent believing that P2 offered half the risk reduction of P1. Conversely, the superforecaster participants generally thought that P2 offered a similar risk reduction to P1.

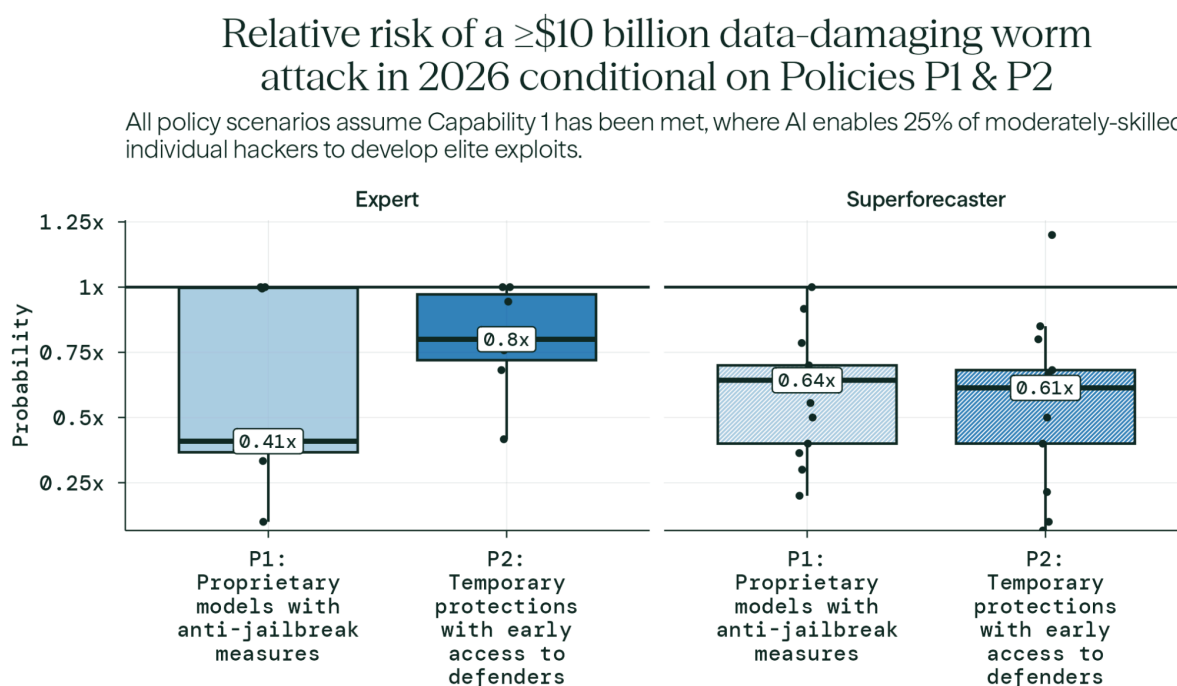


Figure 7: Relative risk of a $\geq$ \$10 billion data-damaging worm conditional on Capability 1 and the two mitigation scenarios.

Respondents generally thought P1 would have the greatest impact on TA1 and TA2 actors, because actors with higher operational capacity might bypass the protections. Some respondents noted that being able to quickly patch vulnerabilities matters more than working against specific jailbreaks. 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 were concerned that sophisticated actors could steal model weights, remove safeguards, and offer elite exploits as a service. They noted that API-based access would not 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 eventual open-weight release would reduce the value of early defender access because attackers would still be able to use the models over the long term.

Actor capability to conduct large-scale worm cyberattacks⁶

We asked participants how capability and willingness to launch a data-damaging worm attack causing at least \$10 billion in damages vary across threat actor categories. Both groups of participants generally thought that TA1 to TA3 actors were very unlikely to be able to develop a data-damaging worm and that almost all TA5 actors would have this capability (see Figure 8). Opinions were more mixed about the capabilities of TA4 actors, with the median expert suggesting that a randomly selected TA4 actor has just a 5% (IQR: 1–25%) probability of being able to create a data-damaging worm, and the median superforecaster putting the probability at 40% (IQR: 25–53%).

Actor end-to-end capability to develop data-damaging worms

Values represent the probability of a randomly selected actor in each actor class having the capability to perform all the steps required to develop and deploy a large-scale data-damaging worm with 3 months of effort.

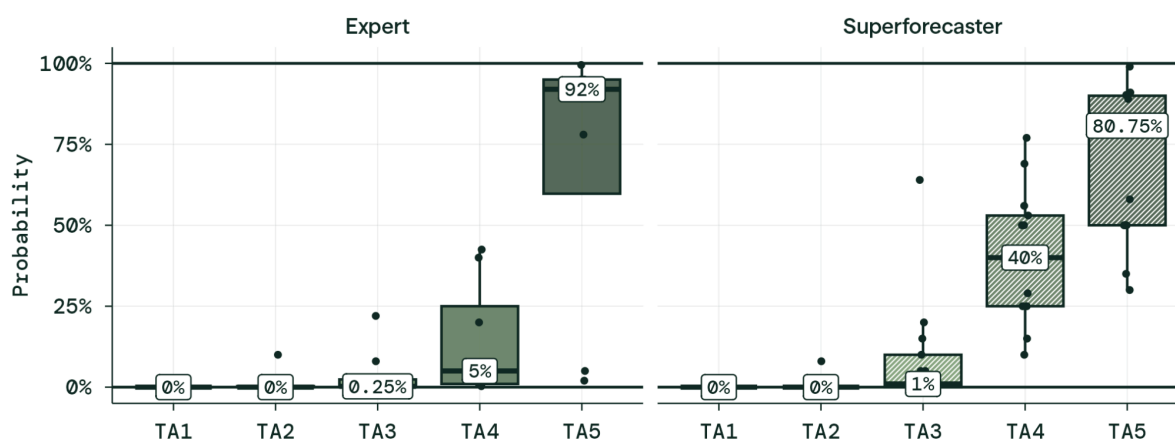


Figure 8: Probability that a randomly selected actor in each threat actor class is able to develop a data-damaging worm given 3 months of effort.

Finding elite exploits was seen as the most difficult part of developing this type of worm attack. Although the median expert thought that a randomly selected TA3 actor would have a 1% (IQR: 1–5%) chance of finding elite exploits, they would have a 16.5% (IQR: 2.1–52.5%) chance of performing the other steps necessary to develop a large-scale data-damaging worm. (See [Appendix C](#) for more detail.)

⁶ We asked again about capability and willingness in the follow-up survey. It was completed by a subset of participants, including only 4 experts from the original sample, and took place several months after the initial survey. Participants had access to additional information and were given the opportunity to update their views. In [Appendix C](#) we present results for similar capability questions, where we ask about a longer 12-month time horizon.



When asked to condition on Capability 1 (access to frontier AI models enables 25% of TA2 actors to find vulnerabilities and write elite exploits), participants, particularly experts, increased their estimate of the probability that a randomly selected actor could develop a data-damaging worm. In this hypothetical scenario, the median expert put a 51.5% (IQR: 14.5–86.3%) probability of a TA3 actor succeeding at this task, and a 13.8% (IQR: 3.5–20.8%) probability of a TA2 actor succeeding (Figure 9). Participants noted that damage at this scale requires rapid development and coordination, so organized groups were still favored over individuals even with the emerging AI capabilities.

Actor end-to-end capability to develop data-damaging worms, conditional on Capability 1

Values represent the probability of a randomly selected actor in each actor class having the capability to perform all the steps required to develop and deploy a large-scale data-damaging worm with 3 months of effort. Forecasts are conditional on Capability 1: AI enables 25% of moderately-skilled individual hackers to develop elite exploits.

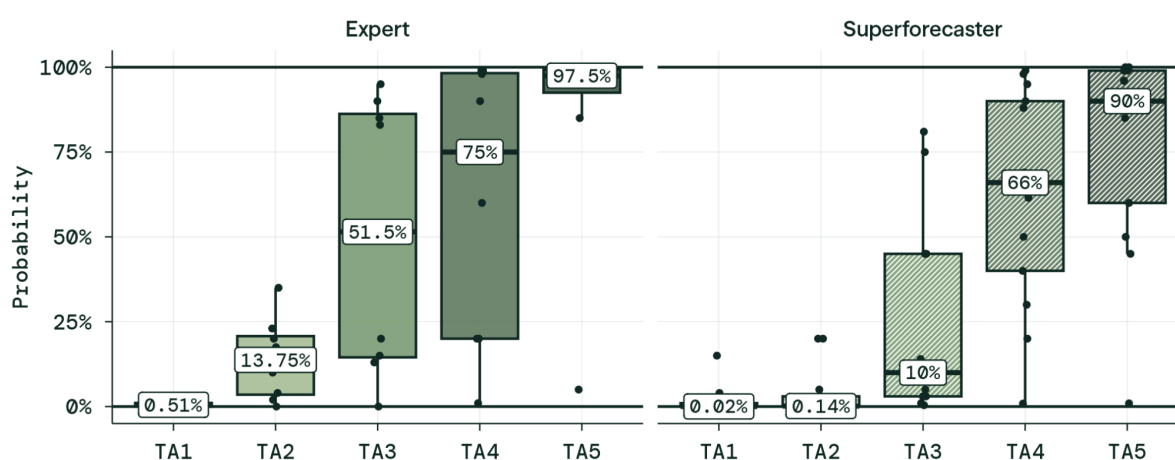


Figure 9: Probability that a randomly selected actor in each threat actor class is able to develop a data-damaging worm with 3 months of effort, conditional on access to frontier AI models enabling 25% of TA2 actors to find vulnerabilities and write elite exploits.

Actor willingness to engage in large-scale worm cyberattacks

We asked participants to estimate the probability that at least one actor in a class would be willing to launch a data-damaging worm attack if they were capable of doing so. Both groups gave high estimates—greater than 95%—for TA1 and TA2 actors, suggesting that the large number of such actors makes it almost certain that at least one would be willing to conduct such an attack. Some respondents noted that there is a roughly inverse relationship between willingness and capability, as more capable actors are constrained by serious geopolitical consequences.



Actor willingness to launch data-damaging worm attacks

Values represent the probability that at least one actor in each actor class would spend at least 3 months actively attempting to launch a data-damaging worm attack if they were capable.

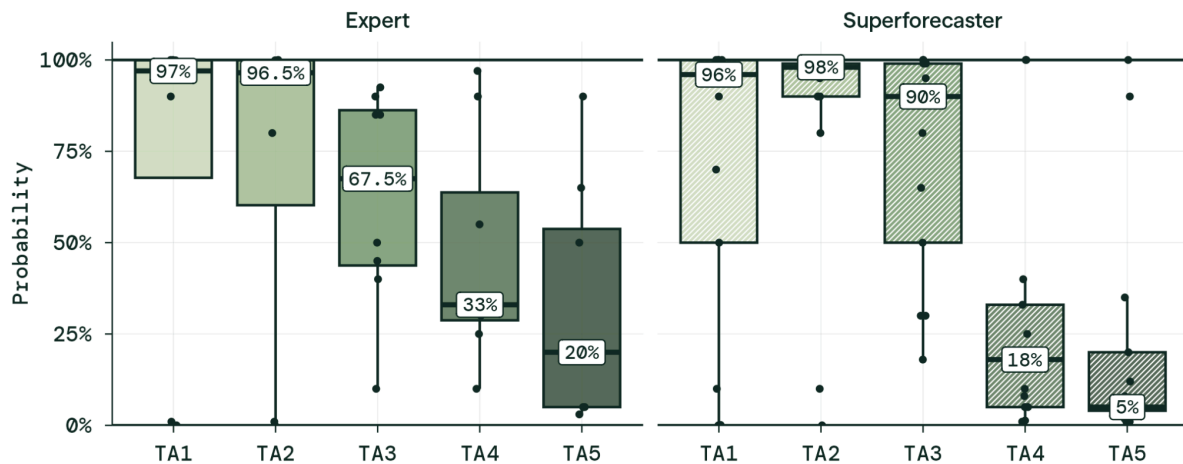


Figure 10: Probability that at least one actor in each threat actor class would spend 3 or more months actively attempting to launch a data-damaging worm in 2026, assuming the actor had the capability to do so.

Expected damages

We asked participants to forecast the probability that data-damaging worms would cause different ranges of total economic damages in 2026. The median expert assigned a 20% probability to damages between \$100 million and \$1 billion (IQR: 8–35%), but a 0.095% probability to damages between \$1 trillion and \$10 trillion (IQR: 0.01–0.1%). Experts and superforecasters broadly had similar estimates across the magnitude scale (see Figure 11), but differed in their estimates of the mid-range bin (\$10 billion–\$100 billion) ($p = 0.04$).

Likelihood that data-damaging worm attacks cause \$X in total economic damages in 2026

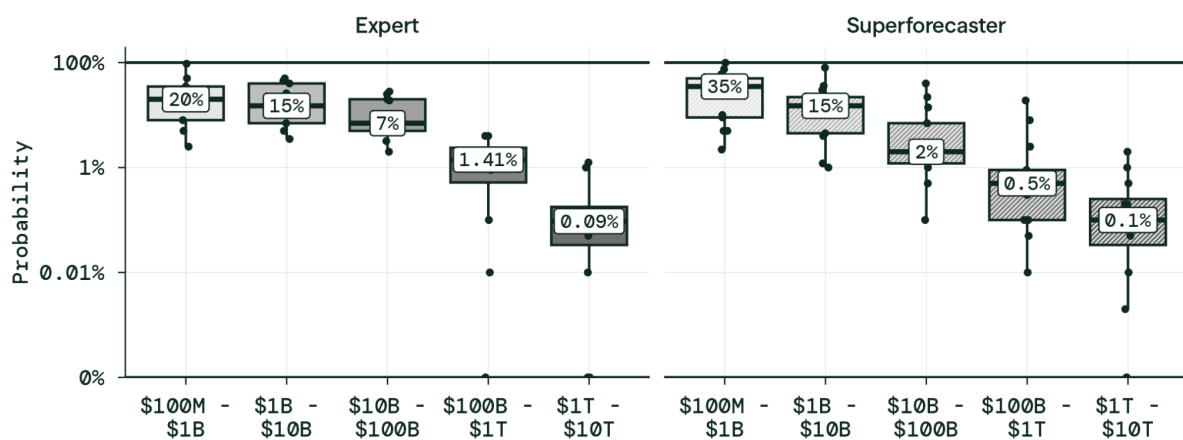


Figure 11: Forecasted probability that data-damaging worm attacks cause total economic damages in each range in 2026.



We used these binned probabilities to calculate expected damages. Participants could revise the calculated value if they felt like it did not capture their beliefs about expected damages from data-damaging worms in 2026. A few participants modified the values, but the aggregate results, especially for experts, remained consistent (see Figure 12).

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

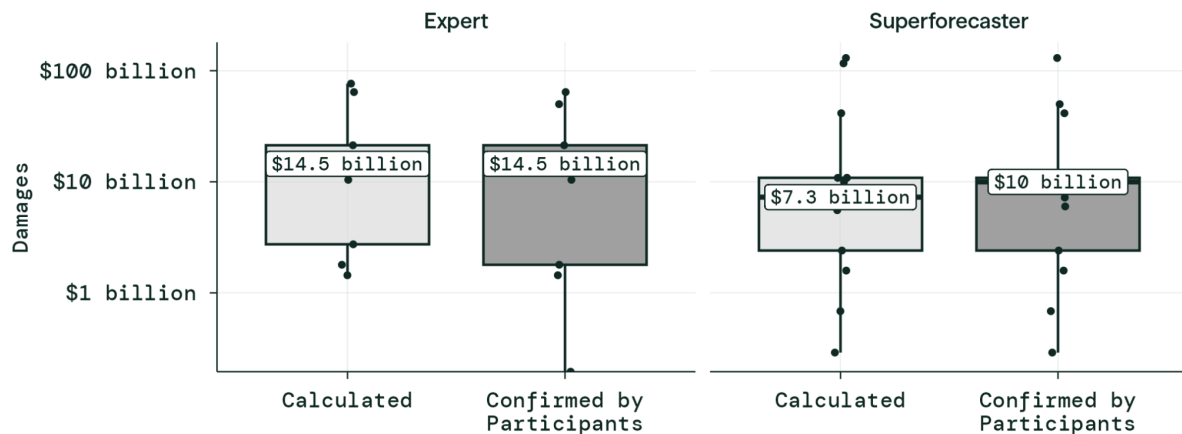


Figure 12: Calculated and participant-confirmed expected damages from data-damaging worm attacks in 2026.

We then asked how these probabilities would change conditional on AI reaching Capability 1. Figure 13 shows these forecasts, and Figure 14 shows both the calculated expected damages and participants' adjusted expected damages estimates under this condition. As before, both groups had similar estimates across the bins, with a significant difference only for the mid-range bin (\$10 billion–\$100 billion), in which the median expert placed significantly more weight ($p = 0.04$).

Likelihood that data-damaging worm attacks cause \$X in total economic damages in 2026, conditional on Capability 1

Capability 1: AI enables 25% of moderately-skilled individual hackers to develop elite exploits.

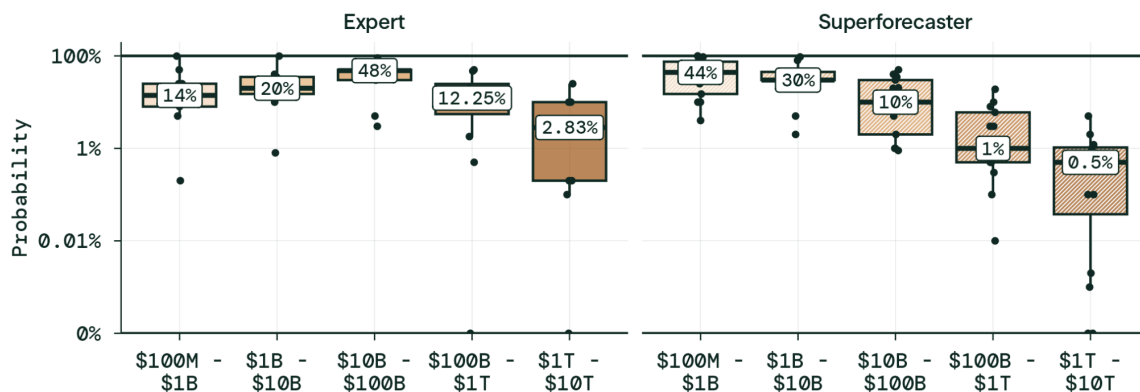


Figure 13: Forecasted probability that data-damaging worm attacks cause total economic damages in each range in 2026, conditional on Capability 1 (AI enables 25% of TA2 actors to write elite exploits).



Using the participants' confirmed expected damages estimates, Capability 1 is associated with a 3–5x increase in expected damages, from approximately \$15 billion to \$67 billion for the median expert forecast and from \$10 billion to \$33 billion for the median superforecaster forecast. Overall, experts expected a much higher increase in damages conditional on AI reaching Capability 1.

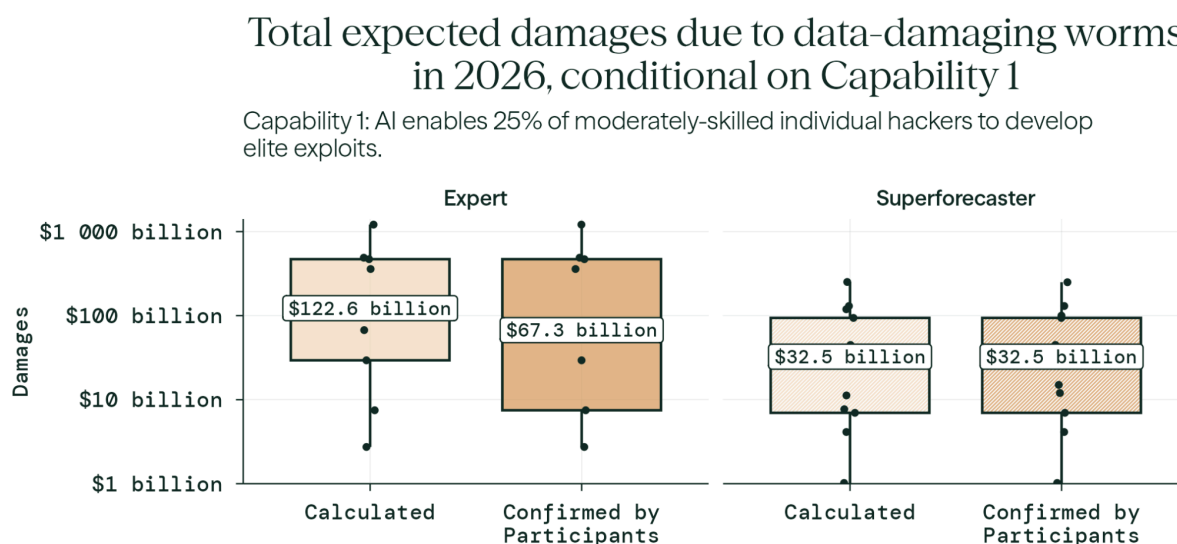


Figure 14: Calculated and participant-confirmed expected damages from data-damaging worms in 2026, conditional on Capability 1 (AI enables TA2 actors to write elite exploits).

Forecast updates

After reviewing a summary of the original survey's numerical results and key arguments, participants were given the opportunity to revise their forecasts of the main outcome: the probability that a data-damaging worm attack would cause at least \$10 billion in economic damages in 2026. Forecasts generally shifted toward the crowd consensus. However, experts were less likely to update than superforecasters; only one of the four experts who completed the follow-up survey revised their estimate. As a result, aggregate median forecasts remained unchanged, while the interquartile range narrowed for superforecasters (Figure 15). The expert median differs slightly here (7% versus 8% in Figure 4) because only a subset of experts participated in the follow-up survey.

When asked to reflect on how the results and arguments impacted their updates, participants expressed a mix of agreement with the results and disagreement on some key assumptions. Superforecasters questioned how large of a capability boost experts attributed to the scenario where Capability 1 had been met, citing that there would be additional real-world operational and defensive constraints. Other disagreements concerned the historical base rate of worm attacks, the willingness of higher-sophistication actors (TA3–TA5), and the strength of deterrents such as law enforcement, escalation risks, and cyber defenses.

Updated forecasts of the probability of a data-damaging worm attack causing $\geq \$10$ billion in economic damages in 2026

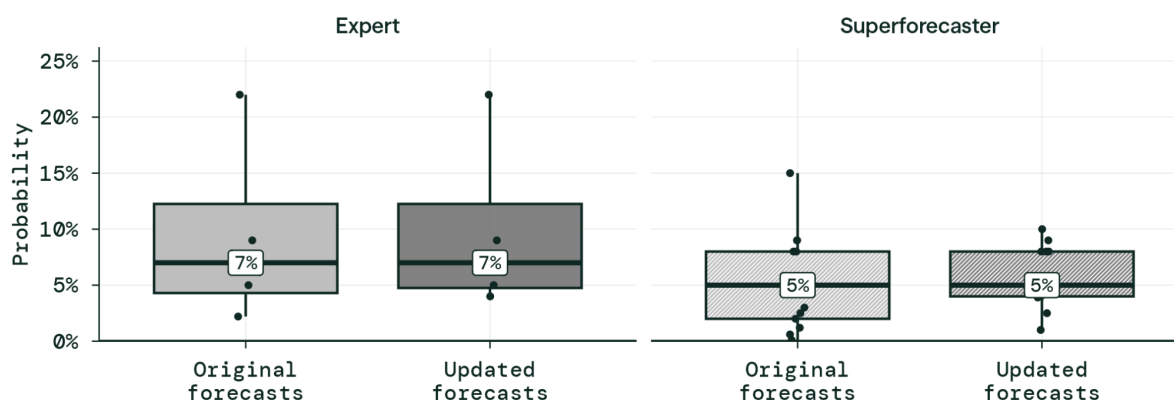


Figure 15: Updated forecasts from the follow-up survey of the probability of at least one data-damaging worm attack causing at least \$10 billion in economic damages in 2026. Only responses from participants who completed both the original and follow-up surveys are shown.

Participants were presented with a summary of Anthropic’s report describing an AI-assisted cyber espionage campaign using Claude [1], and were asked whether this information changed their forecasts. Overall, the report was seen as evidence of rapidly advancing AI-cyber capabilities, particularly as it showed AI can automate large portions of multi-stage attacks. Most participants said this was broadly consistent with their prior expectations of AI progress in this area and did not lead to major updates, while a minority questioned the credibility of key claims in the report.

Participants highlighted that the implications of the report were mostly relevant for TA3 and some TA4 actors, as the campaign showed a level of automation that suggested lower resource and coordination barriers than previously assumed. However, participants believed the attack still relied on known vulnerabilities, while advanced capabilities—such as discovering zero-day vulnerabilities or developing elite exploits—remain key bottlenecks for less sophisticated actors. Participants also stressed that the report centers around industrial espionage rather than a worm attack, limiting its direct relevance to the main outcome in question. As a result, although the report reinforced expectations of near-term rapid AI progress in cybersecurity, most participants did not think it substantially increased the likelihood of a large-scale data-damaging worm attack by 2026.

Participants identified the numerical results from the original study, the Anthropic cyber espionage report, and the study’s rationales as the primary drivers of forecast updates (see Figure 16). The numerical results were the largest contributor, and drove updates in both directions as participants moved toward the crowd consensus. The Anthropic report and general AI progress consistently pushed participants to raise their forecasts of the main outcome, while policy developments had a more mixed effect, leading some participants to increase and others to decrease their risk estimates.

Qualitatively, upward revisions were typically driven by evidence of faster than anticipated AI progress, particularly as outlined in the Anthropic report, as well as increased concern about the willingness of state actors and the potential for geopolitical escalation.



Downward revisions emphasized the role of defensive improvements alongside offensive capabilities.

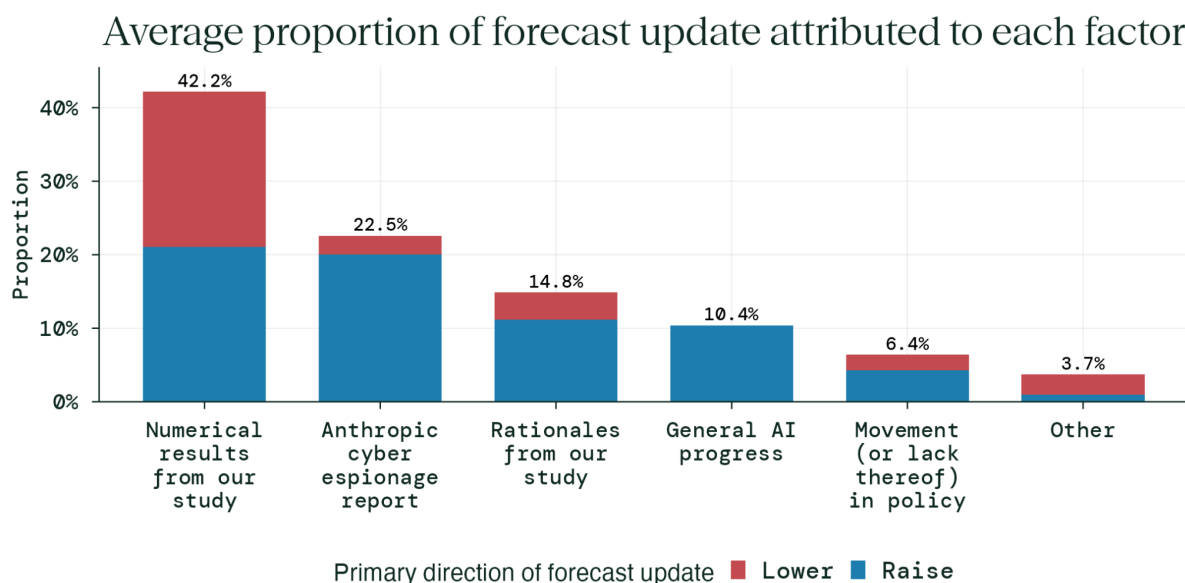


Figure 16: Impact of each information source on participants' forecast updates.

Electrical grid cyberattacks: results

In addition to the data-damaging worm scenario, we asked participants to consider a 2026 cyberattack against the U.S. electrical grid that causes a blackout with at least \$10 billion or at least \$100 billion in economic damages.

Baseline forecasts

Participants generally gave very low forecasts for these events. The median expert and superforecaster both forecast a 1% chance of a grid cyberattack causing at least \$10 billion in damages (IQR: 0.48–1.5% for experts and IQR: 0.5–2% for superforecasters) and a 0.1% chance of an attack causing at least \$100 billion in damages (IQR: 0.002–0.15% for experts and IQR: 0.05–0.12% for superforecasters). Figure 17 shows participants' responses to this question.

When justifying their forecasts, participants often pointed to the lack of historical precedent, noting that even major past incidents caused significantly less damage. They noted that kinetic attacks could be more effective than cyberattacks at damaging the grid, and that cyber-induced damage may be easier to repair than physical damage. It was generally thought that geopolitical conflict would be a key driver of this risk, as attacks of this level would require sophisticated state-level actors with significant operational and technology expertise. Such an attack was seen as more likely to be an act of war than criminal activity, with some participants citing the potential for U.S.–China conflict.

Probability of a cyberattack against the U.S. electrical grid causing a blackout with $\geq \$X$ billion in economic damages in 2026

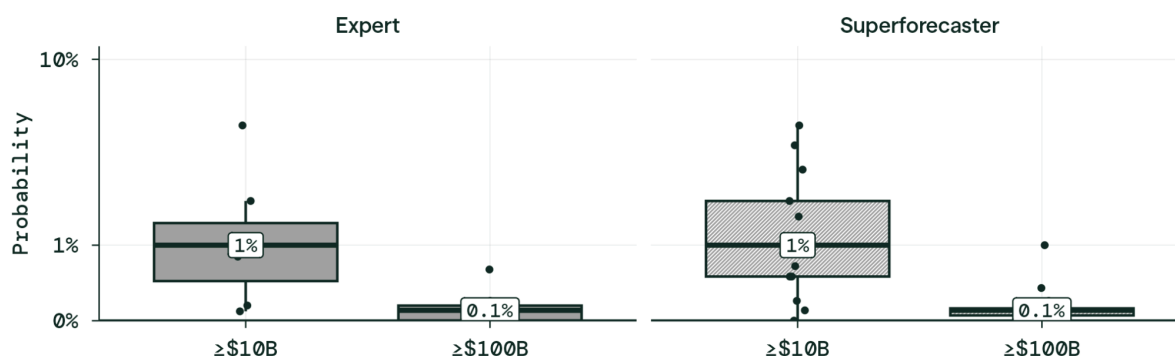


Figure 17: Forecasts of the probability of at least one cyberattack against the U.S. electrical grid causing at least \$10 billion or at least \$100 billion in economic damages in 2026.

Experts thought that, were such an event to occur, TA5 actors would be the most likely cause: conditional on a grid cyberattack causing at least \$10 billion in damages, the average expert assigned a 73% probability to TA5 actors being the cause (IQR: 62–92%). Superforecasters also thought TA5 actors were the most likely cause (average probability of 40%, with IQR 22–50%), although, compared to experts, they placed substantially more weight on other actors, particularly TA4 (35%, IQR: 28–40%) (see Figure 18). Results were similar for an attack causing $\geq \$100$ billion in damages (see [Appendix C](#)).

Generally, respondents thought that this capability would only be available to more sophisticated actors. Some also suggested that less capable actors would have little interest in this type of attack, preferring financially-motivated crimes instead. The constraints on more sophisticated actors (e.g., avoiding retaliation) were again mentioned, however, some respondents suggested that states might operate through proxies to allow plausible deniability. Rationales often mentioned China, with many referring to Volt Typhoon, a Chinese state-sponsored threat actor reportedly responsible for widespread intrusions into critical-infrastructure networks [24].

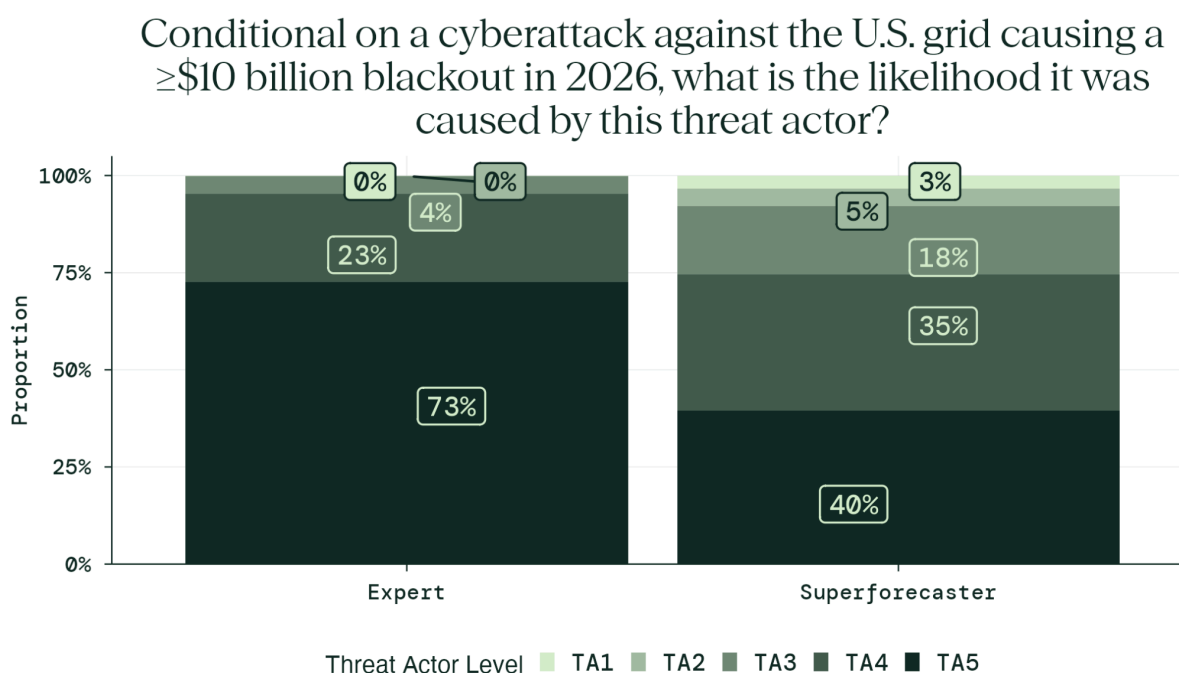


Figure 18: Average probability assigned to each threat actor type as the primary cause of a cyberattack against the U.S. grid, conditional on such an attack causing at least \$10 billion in damages in 2026.

Forecasts conditional on AI capabilities

We next asked participants how their forecast would change under hypothetical scenarios involving open-weight frontier AI models and AI capabilities. These AI capability scenarios were defined as:

BOX 4: CAPABILITY 3 DESCRIPTION

Capability 3: AI enables TA1 actors to perform like a TA3 actor in an OT-specific capture-the-flag style competition.

A study conducted at the end of 2025 to measure AI uplift on a CTF-style competition focused on industrial control systems (ICS) and OT cybersecurity, finds that, with access to AI, individual TA1 actors complete as high a proportion of tasks as a team of ten experienced cybersecurity professionals (equivalent to a TA3 actor) without access to AI, when both teams are given the same amount of time.

BOX 5: CAPABILITY 4 DESCRIPTION

Capability 4: Real-world incident: warning shot.

A cyberattack on the U.S. grid causes a blackout, leading to $\geq \$100$ million in economic damages, and is subsequently confirmed by credible sources to have been perpetrated by a TA1- or TA2-level actor using AI.

Both experts and superforecasters thought Capability 3 would have only a small impact on the risk of a grid cyberattack causing at least \$10 billion in damages in the U.S. in 2026. Under this scenario, the median forecast of this outcome rose to 1.5% (IQR: 1–7%) (from a baseline of 1%) for experts and 2% (IQR: 1–10%) for superforecasters. Under Capability 4, the median expert forecast of a grid cyberattack causing at least \$10 billion in damages in 2026 rose to 15% (IQR: 2–20%). For the median superforecaster, it rose to 4% (IQR: 2.1–8%). (Figure 19.) Figure 20 shows how forecasts of $\geq \$100$ billion changed conditional on these capabilities.

Probability of a cyberattack against the U.S. electrical grid causing a blackout with $\geq \$10$ billion in economic damages in 2026

Capability 3: AI enables TA1 actors to match a TA3 actor in an OT-specific CTF-style competition.
Capability 4: A real-world warning shot incident.

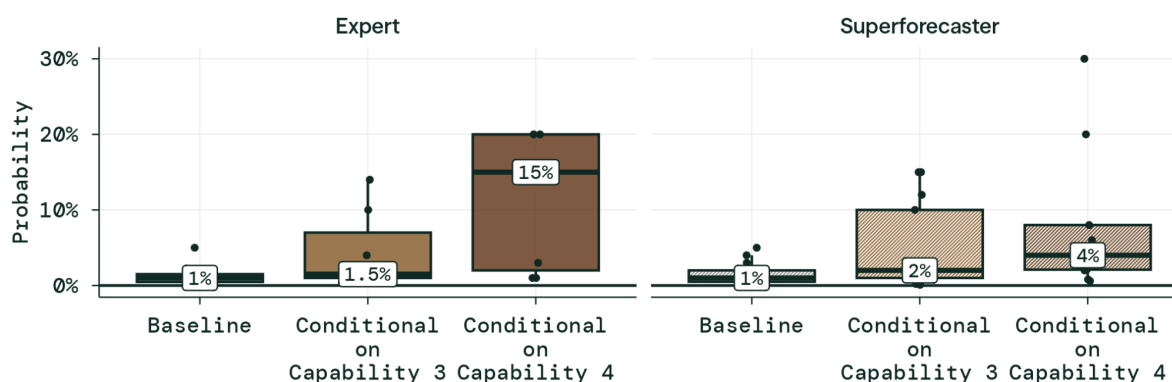


Figure 19: Forecasts of the probability of at least one cyberattack against the U.S. electrical grid causing at least \$10 billion in economic damages in 2026, under baseline assumptions, and conditional on Capability 3 (AI enables TA1 actors to perform like a TA3 actor in an OT-specific capture-the-flag style competition) and Capability 4 (a real-world warning shot incident).



Probability of a cyberattack against the U.S. electrical grid causing a blackout with $\geq \$100$ billion in economic damages in 2026

Capability 3: AI enables TA1 actors to match a TA3 actor in an OT-specific CTF-style competition.
Capability 4: A real-world warning shot incident.

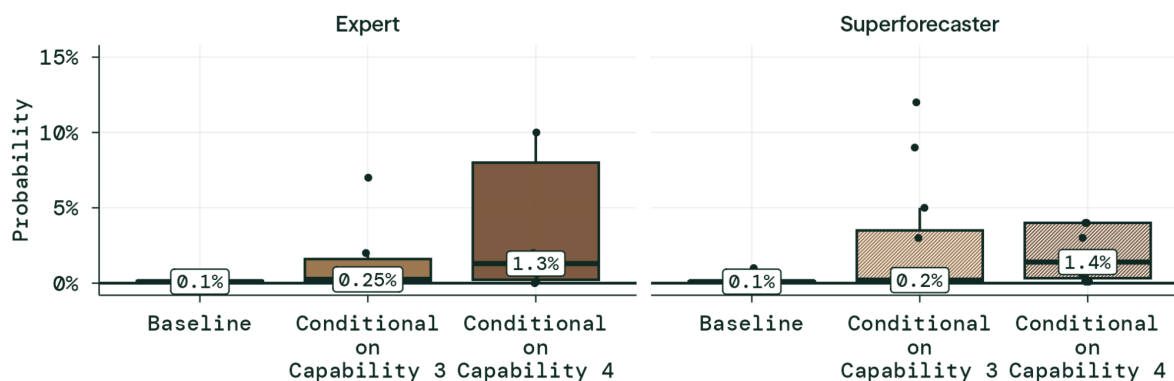


Figure 20: Forecasts of the probability of at least one cyberattack against the U.S. electrical grid causing at least \$100 billion in economic damages in 2026, under baseline assumptions, and conditional on Capability 3 (AI enables TA1 actors to perform like a TA3 actor in an OT-specific capture-the-flag style competition) and Capability 4 (a real-world warning shot incident).

When explaining their forecasts for Capability 3, respondents noted limitations of CTF competitions, especially that they often do not simulate actively defended networks or real operational complexity. Some also shared the opinion that conducting a grid attack is a multifaceted operation, requiring not only technical cyber skills, but also reconnaissance, operational coordination, physical access, and more. Some respondents suggested that this task would be beyond the capabilities of a TA3 actor, so uplifting TA1 actors to this level would have limited impact on this risk. Some respondents noted that the number of attempts may increase, but it's very likely they would be unsuccessful, although one respondent highlighted the danger of less experienced actors attempting attacks without fully understanding the potential downsides of their actions. Some respondents also suggested that an increase in risk would drive investment in grid security, making the task even more difficult.

Respondents who indicated that Capability 4 would suggest an increase in risk argued that a successful \$100 million attack would show that AI had lowered technical barriers and could attract copycat attempts. However, many participants also suggested that a warning shot would trigger a significant increase in defenses. Some noted a tension between the "wake-up call" effect, which could strengthen defenses, and the window of vulnerability before those defenses improved. Some participants voiced uncertainty about whether conducting an attack that causes \$10 billion in damage (rather than \$100 million) is a qualitatively different challenge.

Actor capability and willingness to launch large-scale grid attacks

As in the data-damaging worm scenario, we asked participants how capability and willingness to launch a grid cyberattack causing at least \$10 billion in damages vary



across threat actor categories. Both groups of participants generally thought it was close to impossible for a TA1 or TA2 actor to succeed at this task with six months of effort, with median forecasts of 0% probability that a randomly selected actor has the capabilities to launch such an attack. For the other actor types, superforecasters generally put a higher probability on their chances of success. For example, the median superforecaster estimated a 65% probability of a TA5 actor having the capability (IQR: 40–90%), compared to the median expert forecast of 25% (IQR: 6–42.5%) (Figure 21). Rationales suggested that many participants thought the U.S. and China may have this capability, but that this is difficult to assess given the lack of precedent.

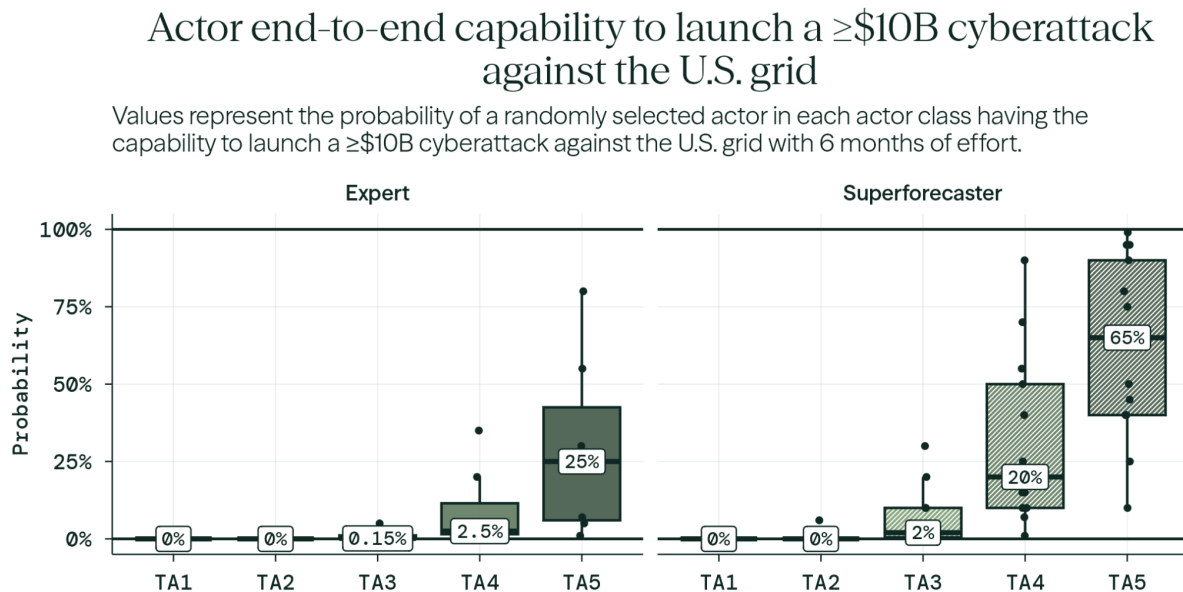


Figure 21: Probability that a randomly selected actor in each threat actor class is able to launch a cyberattack against the U.S. grid with 6 months of effort, causing at least \$10 billion in economic damages.

Comparing these results with the \$100 billion version in [Appendix C](#) suggests they should be interpreted with caution. Notably, for TA5 actors, the median expert capability estimate falls from 25% to 2%, while the median superforecaster estimate changes much less, from 65% to 58%. This may mean superforecasters saw little additional capability required to scale from \$10 billion to a \$100 billion attack. Alternatively, they may have read capability more loosely as the ability to mount a serious attack aimed at the damage threshold proposed, without separately accounting for whether the attack would actually achieve the specified damage. Experts may have tied capability more strongly to directly achieving the threshold through the attack. Some of the divergence between the two groups in these forecasts may be attributable to this ambiguity in the wording of the question.

In general, respondents thought that all types of threat actors would be less willing to conduct a large-scale grid attack than a large-scale worm attack (Figure 22). Some respondents noted that there is a roughly inverse relationship between willingness and capability, as more capable actors are constrained by serious geopolitical consequences. Rationales emphasized that a grid attack would not fit the financial motivations that drive



most less sophisticated actors, and that TA5 actors would be constrained by retaliation concerns, limiting such attacks largely to acts of war.

Actor willingness to launch a $\geq \$10\text{B}$ cyberattack against the U.S. grid

Values represent the probability that at least one actor in each actor class would spend at least 6 months actively attempting to launch a $\geq \$10\text{B}$ cyberattack against the U.S. grid if they were capable.

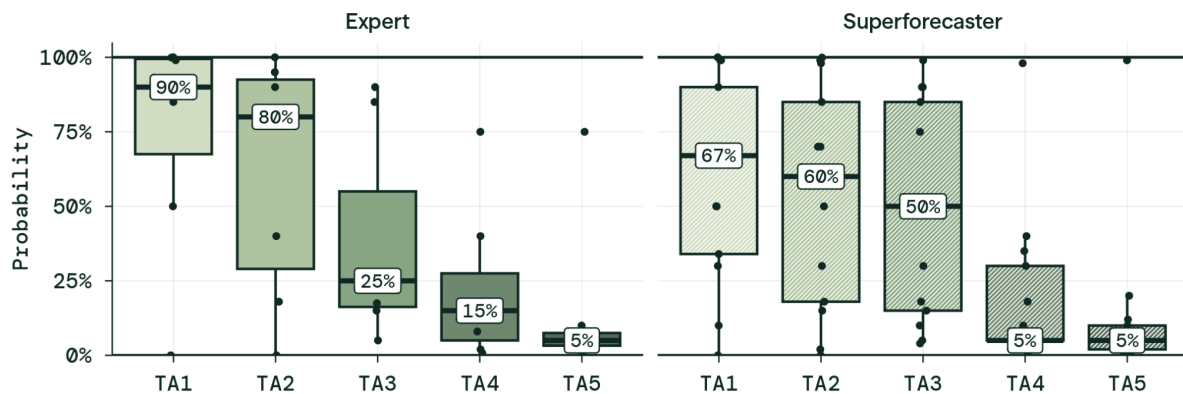


Figure 22: Probability that at least one actor in each threat actor class would spend 6 or more months actively attempting to launch a cyberattack against the U.S. electrical grid, causing a blackout with at least \$10 billion in economic damages in 2026, assuming the actor had the capability to do so.

Expected damages

As with worm attacks, we asked participants to forecast the probability that cyberattacks against the U.S. electrical grid would cause different ranges of total economic damages in 2026. The median expert forecasted a 3% probability of such damages falling between \$100 million and \$1 billion (IQR: 1.3–15%), and a 0.0032% probability of such damages falling between \$1 trillion and \$10 trillion (IQR: 0.0001–0.006%) (Figure 23). Superforecaster medians were close to an order of magnitude larger: 10% (IQR: 2–16%) and 0.02% (IQR: 0.0001–0.1%), respectively.

Likelihood that cyberattacks to the U.S. grid cause \$X in total economic damages in 2026

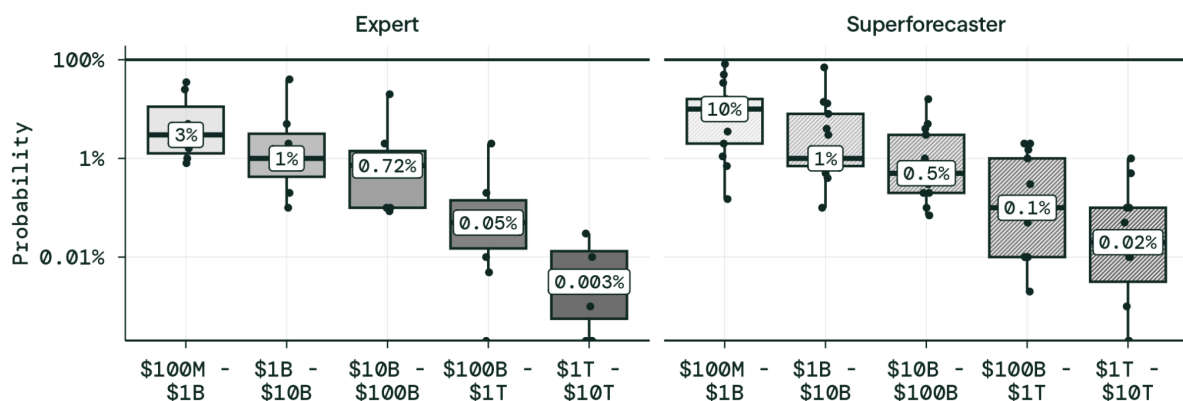




Figure 23: Forecasted probability that cyberattacks on the U.S. grid cause total economic damages in each range in 2026.

We used these binned probabilities to calculate an expected damages value and participants could revise the calculated value if they felt like it did not capture their beliefs. These values (shown in Figure 24) were more than an order of magnitude smaller than for data-damaging worm attacks. Participants noted that the largest-scale attacks seemed extremely unlikely, while one participant observed that even very small tail probabilities can skew the calculated expected damages very high.

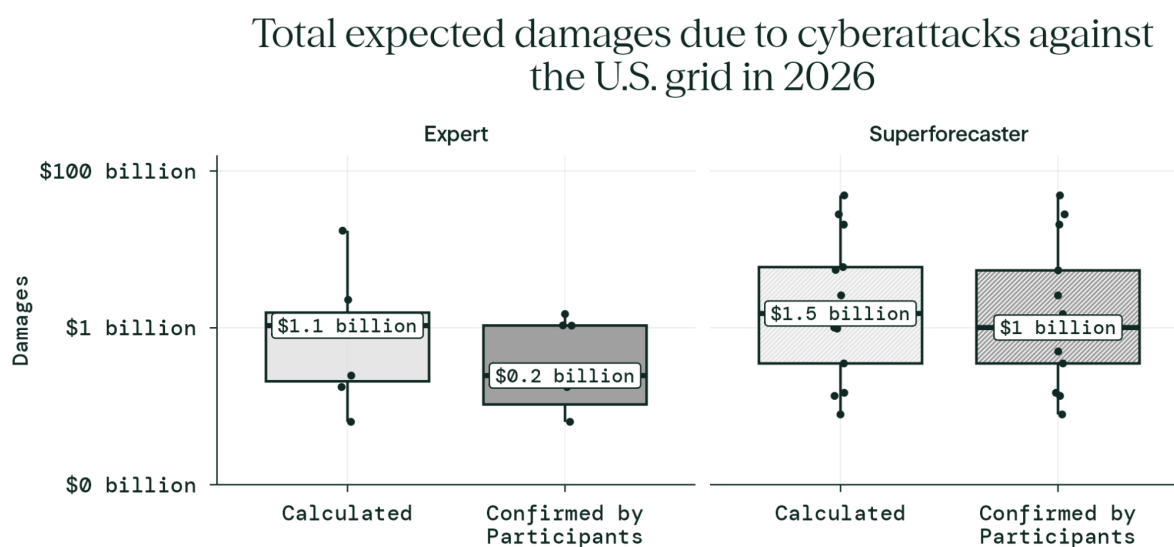


Figure 24: Calculated and participant-confirmed expected damages from cyberattacks against the U.S. electrical grid in 2026.

We also asked how these forecasts would change conditional on the following AI capability:

BOX 6: CAPABILITY 2 DESCRIPTION

Capability 2: AI solves more than 90% of Cybench tasks

AI alone can solve more than 90% of the tasks on Cybench, completely unguided, at the end of 2025.

Cybench is an industry-standard benchmark of AI capabilities across six cybersecurity categories: cryptography, web security, reverse engineering, forensics, exploitation, and miscellaneous [25]. It includes 40 capture-the-flag (CTF) tasks of varying difficulties, drawn from real-world CTF competitions such as HackTheBox and Glacier. Cybench was designed to evolve over time and to incorporate new tasks that remain relevant as capabilities progress.



Conditioning on this capability led to a roughly 4x increase in expected damages for the median expert and a roughly doubling of expected damages for the median superforecaster (Figure 25 and Figure 26). Rationales suggested that some participants expected such capabilities to be associated with many more attempts by many actors leading to more small-scale attacks, which could add up to slightly higher total damages, though not significantly larger. Some participants noted that capabilities are not the limiting factor — deterrence, escalation risks and human coordination are still bigger constraints, and this capability would not help with those. Some also argued that once an AI-enabled attack succeeded, defenders—especially the U.S. government—would rapidly harden defenses.

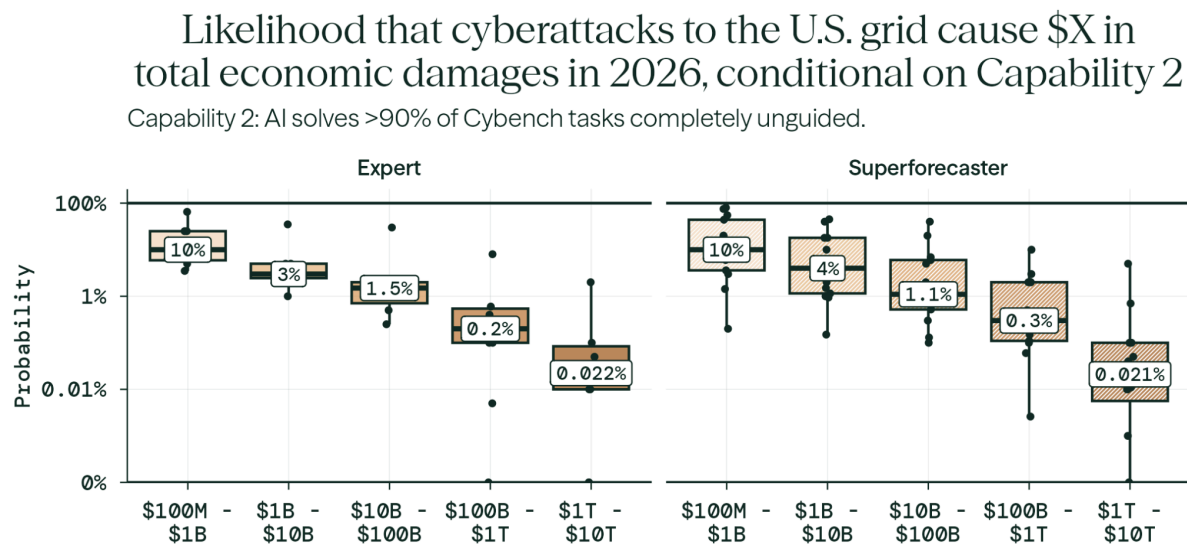


Figure 25: Forecasted probability that cyberattacks against the U.S. grid cause total economic damages in each range in 2026, conditional on Capability 2 (AI solves at least 90% of Cybench tasks).

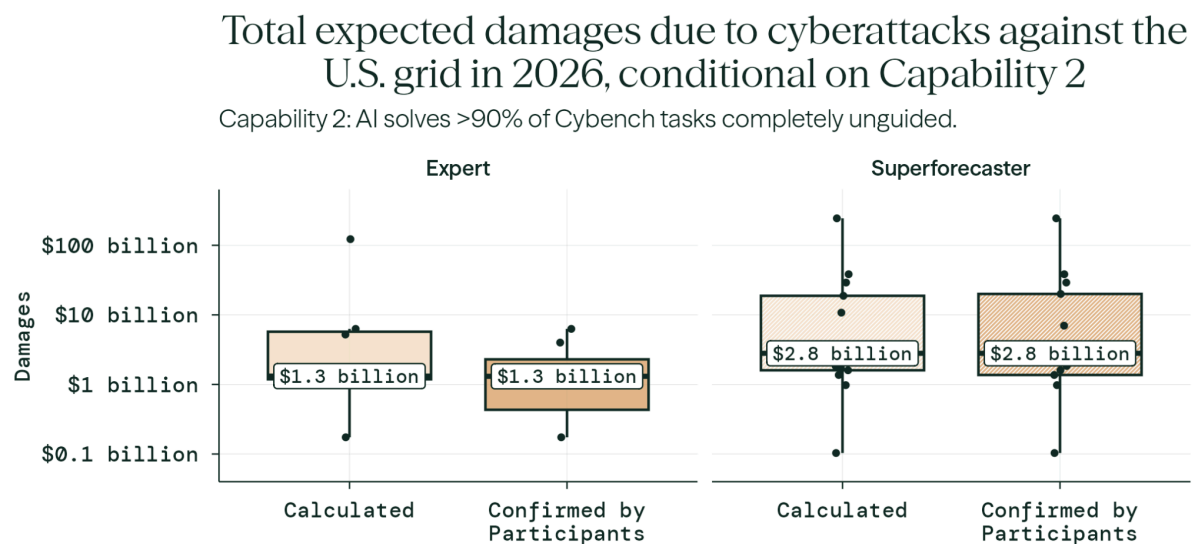


Figure 26: Calculated and participant-confirmed expected damages from cyberattacks against the U.S. electrical grid in 2026, conditional on Capability 2 (AI solves at least 90% of Cybench tasks).



Comments on mitigations

We asked participants to note any thoughts or comments on mitigations they considered relevant to this threat model: a large-scale cyberattack against the U.S. electrical grid. Their responses centered on two broad categories: grid and institutional resilience, and AI-specific defenses.

On the grid infrastructure side, several participants emphasized the importance of physically separating critical control systems from internet-connected networks, increasing redundancy and decentralization to reduce the risk of cascading failures, maintaining manual overrides and backup systems, and stockpiling spare equipment to support faster recovery after an attack. Participants also flagged the need for stronger vetting of personnel with access to critical systems.

Participants also identified several AI-specific mitigations. Some suggested that highly capable AI systems could be made available earlier or differentially to defenders, similar to the early-access scenario considered for the data-damaging worm scenario. Others emphasized the importance of using AI-powered tools for anomaly detection, attack detection, and other defensive capabilities. Additional suggestions included monitoring model queries, conducting mandatory red-teaming focused on industrial-control-system protocols such as Modbus, DNP3, and relevant IEC standards, strengthening vetting and logging for API access to advanced models, and imposing strict restrictions on highly capable open-weight models.

Forecasts of AI progress on relevant capabilities

We asked participants to forecast when each of the following capabilities would be met:

- 1. Capability 1: AI enables TA2 actors to write elite exploits**
 - a. A study finds that access to frontier AI models enables 25% of moderately-skilled individual hackers to find vulnerabilities and write elite exploits, assuming three months of full-time effort.
- 2. Capability 2: AI solves at least 90% of Cybench tasks**
 - a. AI alone can solve at least 90% of the tasks on Cybench completely unguided.
- 3. Capability 3: AI enables TA1 actors to perform like a TA3 actor in an OT-specific capture-the-flag style competition**
 - a. A study is conducted to measure AI uplift on a CTF-style competition focused on industrial control systems (ICS) and OT cybersecurity, finds that, with access to AI, individual TA1 actors complete as high a proportion of tasks as a team of ten experienced cybersecurity professionals (equivalent to a TA3 actor) without access to AI, when both teams are given the same amount of time.
- 4. Capability 4: Real-world incident: warning shot**



- a. A cyberattack against the U.S. grid causes a blackout with at least \$100 million in economic damages and is subsequently confirmed by credible sources to have been perpetrated by a TA1- or TA2-level actor using AI.

Respondents generally expected these capabilities to be achieved within about six years (Figure 27). The median expert forecast placed Capability 1 in 2032, Capability 2 and Capability 3 in 2028, and Capability 4 in 2031. Since the survey closed in August 2025, Capability 2 appears to have been achieved. Recent evaluations also suggest rapid progress on Capability 1, though it remains unclear whether current frontier models meet the capability as defined here.⁷

Timeline of progress on relevant AI capabilities

Capability 1 (C1): Elite exploit uplift, Capability 2 (C2): Cybench
Capability 3 (C3): CTF uplift, Capability 4 (C4): Cyberattack warning shot

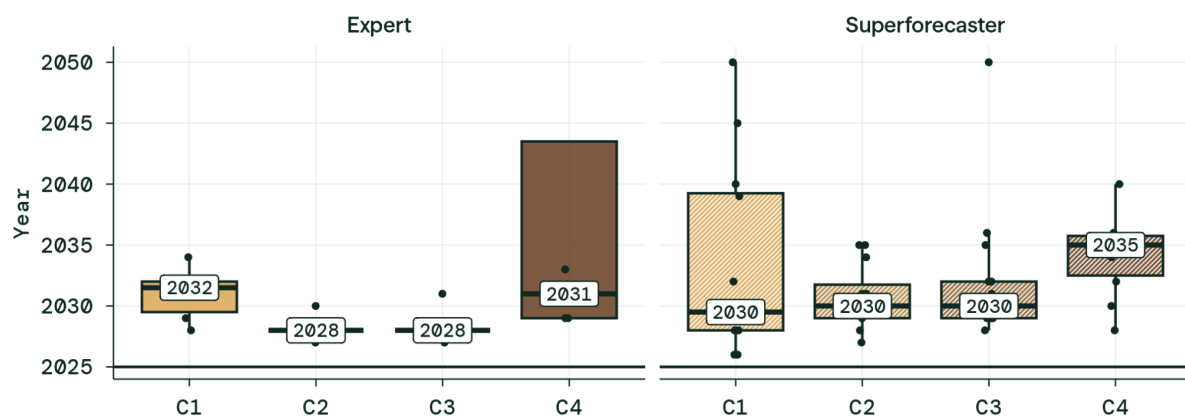


Figure 27: Forecasts of the year in which participants expected each AI capability to be achieved.

To understand the signal provided by Cybench, we asked participants to forecast the probability that Capability 1 (AI enables 25% of TA2 actors to write elite exploits) and Capability 3 (AI enables TA1 actors to perform like a TA3 actor in an OT-specific CTF) would be achieved, conditional on Capability 2 (AI solves at least 90% of Cybench tasks) being achieved. Generally, experts thought Cybench provided greater signal of other capabilities, compared to superforecasters (see Figure 28).

⁷ The 90% threshold was first crossed on the [public Cybench leaderboard](#) by Claude Opus 4.6, which was added at 93% unguided solved on a 37-problem subset on February 6, 2026. Later, Claude Mythos Preview reached 100% on a 35-problem subset [26, p. 49]. Recent model evaluations, including Anthropic's Claude Mythos system card and OpenAI's GPT-5.3-Codex cybersecurity evaluation, suggest substantial progress towards AI-enabled exploit development capabilities, although whether the models fully satisfy our definition remains uncertain.

Probability of AI capabilities being met conditional on an AI model solving >90% of Cybench tasks completely unguided

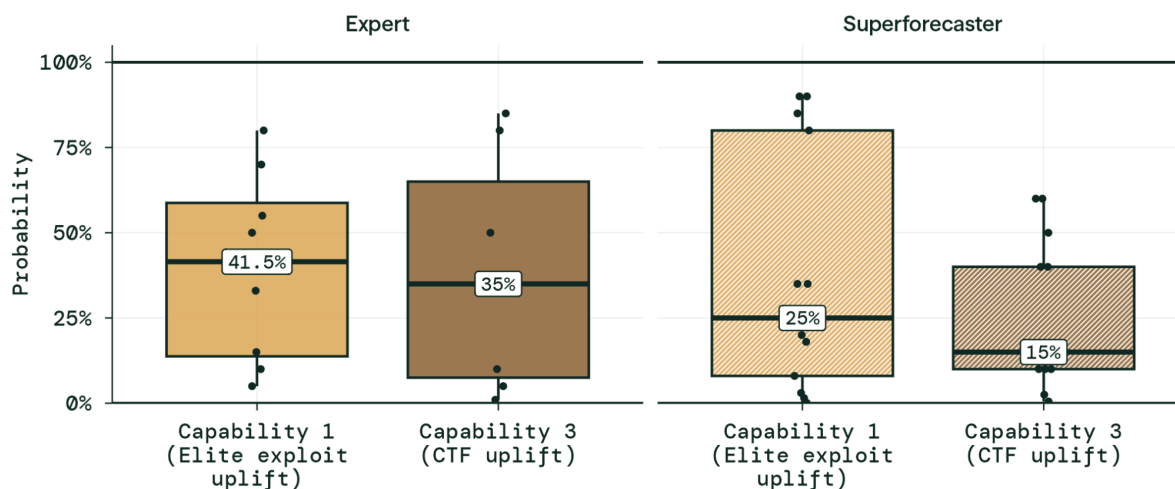


Figure 28: Forecasts of the probability that Capability 1 (AI enables TA2 actors to write elite exploits) and Capability 3 (AI enables TA1 actors to perform like a TA3 actor in an OT-specific CTF) will be achieved, conditional on Capability 2 (AI solves at least 90% of Cybench tasks) being achieved.

Limitations

This pilot study has important limitations that should be kept in mind when interpreting the results. Our sample included only 13 superforecasters and eight experts, two of whom completed only a subset of the questions. This makes the results sensitive to individual forecasts and the reported aggregate statistics fragile. Similarly, our convenience sample of experts may be biased toward people concerned about AI impacts on cybersecurity, and should not be treated as representative of cybersecurity experts as a whole. In addition, some expert participants had prior exposure to early drafts of related reports by our GovAI collaborators. This may have anchored their views or created shared framings and assumptions before the forecasting exercise.

The questions in this study differed from typical forecasting exercises because they focused on hypothetical AI evaluations and low-probability, extreme societal events without clear resolvability. Many were conditional questions, meaning forecasters would not 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 limited historical data available for certain scenarios, may affect the accuracy and usefulness of the forecasts.

The follow-up survey introduces additional limitations: only a subset of the original participants completed it. In the follow-up, respondents were intentionally exposed to summary statistics, arguments, and external information so that we could understand how these inputs affected their views. This may have anchored updates or induced convergence towards the original consensus rather than encouraged independent reassessment.



Conclusion

The study results suggest that AI capabilities could significantly increase some cyber risks in the near term. Both experts and superforecasters predicted substantial increases in the likelihood of a large-scale data-damaging worm attack when AI enables moderate-sophistication individual hackers to develop elite exploits. Median forecasts suggest that AI-enabled vulnerability discovery could increase the risk of large-scale worm attacks by 3–3.5x, with expected annual damages rising from roughly \$10–15 billion to \$33–67 billion. Respondents believed these capabilities could arrive within the next five to six years.

At the same time, structured mitigation measures appear to matter. Participants believed that maintaining proprietary model access with strong anti-jailbreak and monitoring controls could more than halve the probability of such catastrophic outcomes. Limited early access for defenders was viewed as less effective, although still potentially useful.

Given the small sample size and the speculative nature of some scenarios, these results should be interpreted as directional rather than definitive. Nonetheless, they illustrate the value of structured forecasting for anticipating technology-driven security risks. Future work could explore a wider range of possible cyber harms and include a larger, more representative sample of relevant subject-matter experts.



References

1. Anthropic. "Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign." 2025. <https://www.anthropic.com/news/disrupting-AI-espionage> Accessed: 2026-05-27.
2. Crosignani, Matteo, Marco Macchiavelli, and André F. Silva. "Pirates without Borders: The Propagation of Cyberattacks through Firms' Supply Chains." 2023. <https://doi.org/10.1016/j.jfineco.2022.12.002> Journal of Financial Economics 147, no. 2: 432-448.
3. Johansmeyer, Tom. 2024. "Perception Shapes Reality: How Views on Financial Market Correlation Affect Capital Availability for Cyber Insurance." 2024. <https://kar.kent.ac.uk/106432/> Journal of Risk Management and Insurance 28, no. 1: 1-25.
4. Center for Strategic and International Studies. "Significant Cyber Incidents Since 2006." Living document. Accessed: 2026-05-27. https://csis-website-prod.s3.amazonaws.com/s3fs-public/2024-12/241210_Significant_Cyber_Events.pdf
5. Council on Foreign Relations. "Cyber Operations Tracker." <https://www.cfr.org/cyber-operations/> Accessed: 2026-05-27.
6. Harry, Charles, and Nancy Gallagher. "Cyber Events Database." Center for International and Security Studies at Maryland. <https://cisss.umd.edu/research-impact/publications/cyber-events-database-home> Accessed: 2026-05-27.
7. Garton, David. "Purdue Model Framework for Industrial Control Systems & Cybersecurity Segmentation." 2019. Topic Paper 4-14, prepared for the National Petroleum Council Study on Oil and Natural Gas Transportation Infrastructure.
8. US Department of Energy. "DOE-417 Electric Emergency Incident and Disturbance Report" <https://doe417.pnnl.gov/> Accessed: 2026-05-27.
9. Alrich, Tom. 2019. "It's Official: The Event Reported in March Was a Real Cyber Attack." [Tom Alrich's Blog](#), 2019. Accessed: 2026-05-27.
10. North American Electric Reliability Corporation. "Risks Posed by Firewall Firmware Vulnerabilities." [Lesson Learned](#), 2019.
11. Alrich, Tom. "When Will a Ransomware Attack Impact the Bulk Electric System? 2018." [Tom Alrich's Blog](#), 2020.
12. MITRE. "Cyber Risk to Mission Case Study" 2022. <https://apps.dtic.mil/sti/trecms/pdf/AD1183007.pdf> Accessed: 2026-05-27.
13. Willuhn, Marian. "Satellite Cyber Attack Paralyzes 11GW of German Wind Turbines." [pv magazine](#) 2022.
14. Proska, Ken, John Wolfram, Jared Wilson, Dan Black, Keith Lunden, Daniel Kapellmann Zafra, Nathan Brubaker, Tyler McLellan, and Chris Sistrunk. "Sandworm Disrupts Power in Ukraine Using a Novel Attack Against Operational Technology." [Google Cloud Blog](#), 2023.
15. Culler, Megan Jordan, Megan Mincemoyer Egan, Remy Vanece Stolworthy, and Jake P. Gentle. 2024. "Attack Surface of Renewable Energy Technologies." 2024. INL/CON-24-76414-Revision-O. [Idaho National Laboratory](#) (p. 15).



16. SektorCERT. "The Attack against Danish Critical Infrastructure." 2023
<https://sektorcert.dk/wp-content/uploads/2023/11/SektorCERT-The-attack-against-Danish-critical-infrastructure-TLP-CLEAR.pdf>
17. U.S. Energy Information Administration. 2025. "Table 1.2. Summary Statistics for the United States, 2014–2024." In [Electric Power Annual](#): With Data for 2024. Release date October 16, 2025.
18. Energy Charter Secretariat. 2023. Ukrainian Energy Sector Evaluation and Damage Assessment, [Version IX](#), April 27, 2023, p. 12.
19. Lloyd's. "The Insurance Implications of a Cyber Attack on the U.S. Power Grid." London: Lloyd's, [Business Blackout](#) 2015.
20. HM Government. [National Risk Register](#): 2025 Edition. London: Cabinet Office.
21. Rose, Adam, Gbadebo Oladosu, and Shu-Yi Liao. "Business Interruption Impacts of a Terrorist Attack on the Electric Power System of Los Angeles: Customer Resilience to a Total Blackout." 2007. [Risk Analysis](#) 27, no. 3: 513–531. DOI: 10.1111/j.1539-6924.2007.00912.x.
22. National Research Council. "Terrorism and the Electric Power Delivery System." National Academies Press, 2012. DOI: [10.17226/12050](#).
23. Nevo, Sella, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott. "Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models". [RAND Corporation](#), 2024. DOI: 10.7249/RRA2849-1.
24. National Cyber Security Centre (NCSC). "Defending Against China-Nexus Covert Networks of Compromised Devices." 2025.
<https://www.ncsc.gov.uk/news/defending-against-china-nexus-covert-networks-of-compromised-devices>. Accessed: 2026-05-27.
25. Zhang, Andy K., Neil Perry, Riya Dulepet, Joey Ji, Celeste Menders, Justin Lin, Eliot Jones, et al. "Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models." [International Conference on Learning Representations](#) 2025.
26. Anthropic. "Claude Mythos Preview System Card." 2026.
<https://www-cdn.anthropic.com/7624816413e9b4d2e3ba620c5a5e091b98b190a5/Claude%20Mythos%20Preview%20System%20Card.pdf>. Accessed: 2026-05-27.



Appendix A: Survey Details

Survey questions

Task 0: Threat actor estimates

Q0: How many threat actors are in each category?
(For Operational Capacity Levels from 1-5)

Note 1: Treat each state as a single threat actor.

Note 2: For TA3, TA4 & TA5, each threat actor should count as a single actor. For example, if you estimate 500 separate groups capable of OC3-level operations consisting of 10 individuals each, you should enter 500, not 5,000 (500 x 10 individuals).

(For definitions, see the level explanations above)

Task 1: Data-damaging worms

1.1 Baseline forecasts (unconditional)

1.1.1 Holistic estimates

Q1.1.1a: What is the probability of at least one data-damaging worm attack, similar to WannaCry or NotPetya, causing at least \$10 billion in economic damages in 2026?

Note: Multiple smaller attacks adding up to \$10 billion should not count here. We're asking about the occurrence of at least one individual attack meeting the \$10 billion threshold.

Q1.1.1b: Conditional on a data-damaging worm attack causing \$10 billion dollars in economic damages in 2026, what is the probability it was caused by an actor of the following threat actor level:

Q1.1.1c: What is the likelihood that data-damaging worm attacks cause \$X in total economic damages in 2026.

This could be due to a single large attack or multiple smaller attacks.

Q1.1.1d: In this section, we use a logarithmic mean method to estimate the expected damages across these different attack magnitudes, based on your forecasts above. If the results do not reflect your beliefs, please let us know below. If you disagree with the estimate we have calculated above from your forecasts, please provide your alternative best estimate for the total damages caused by data-damaging worm attacks in 2026 here:

1.1.2 Step-by-step estimates

Q1.1.2ai: Actor capability to write elite exploits. What is the probability that a randomly selected threat actor in each threat actor class would be capable of developing elite



exploits involved in creating a data-damaging worm capable of causing $\geq \$10$ billion in damages?

(Here, we only ask about actors having the capability to find vulnerabilities and write elite exploits of them with 3 months of full-time effort.)

Q1.1.2aⁱⁱ: Actor capability to complete other tasks involved in developing a data-damaging worm given access to elite exploits. What is the probability that a randomly selected threat actor in each level can currently complete the other tasks, aside from creating elite exploits, involved in creating a data-damaging worm causing $\geq \$10$ billion in damages?

(i.e., assume that someone publishes elite exploits on the open internet. What is the probability that a randomly selected threat actor in each class could then create a data-damaging worm, causing $\geq \$10$ billion in damages with 3 months of full-time effort?)

Q1.1.2aⁱⁱⁱ: Actor capability to develop and deploy a data-damaging worm attack (end-to-end). For each threat actor level, what is the probability that a randomly selected threat actor in that level is currently capable of developing and deploying a data-damaging worm attack that causes $\geq \$10$ billion in damages?

(Here, we ask about actors having the capability to find vulnerabilities, write elite exploits of them, and use them to deploy the data-damaging worm attack with 3 months of full-time effort.)

Q1.1.2b: Assuming they are capable, what is the probability that at least one actor in each threat actor class would spend 3 or more months actively attempting to launch a data-damaging worm attack causing $\geq \$10$ billion in economic damages in 2026?

1.2 Forecasts conditional on AI capabilities

1.2.1 Holistic estimates

Q1.2.1a: What is the probability of at least one data-damaging worm attack, similar to WannaCry or NotPetya, causing at least \$10 billion in economic damages in 2026, conditional on frontier open-weight AI models enabling 25% of TA2 actors to write vulnerabilities and write elite exploits with 3 months of full-time effort?

This attack could be launched by threat actors from any class.

Q1.2.1b: What is the probability that Capability 1 has been met, conditional on an open-weight frontier AI model being able to solve at least 90% of the tasks on Cybench completely unguided at the end of 2025?

Note that here we ask about what AI achieving Capability 2 implies about how far it's gone in Capability 1. We don't ask about the main outcome of the survey here (data-damaging worm attacks).



Q1.2.1c: What is the likelihood that data-damaging worm attacks cause \$X in total economic damages in 2026, conditional on AI enabling 25% of TA2 actors to find vulnerabilities and write elite exploits with 3 months of full-time effort?

Q1.2.1d: As in the unconditional version of this question, here we use a logarithmic mean method to estimate the expected damages across these different attack magnitudes, based on your forecasts above. If the results do not reflect your beliefs, please let us know below. Your baseline unconditional forecasts are shown in brackets for comparison. If you disagree with the estimate we have calculated above from your forecasts, please provide your alternative best estimate for the total damages caused by data-damaging worm attacks in 2026, conditional on Capability 1 here:

1.2.2 Step-by-step estimates

Q1.2.2ai: Actor capability to write elite exploits. What is the probability that a randomly selected threat actor in each threat actor class would be capable of developing elite exploits involved in creating a data-damaging worm capable of causing $\geq$ \$10 billion in damages, conditional on AI reaching Capability 1?

(Here, we only ask about actors having the capability to find vulnerabilities and write elite exploits of them with 3 months of full-time effort.)

Q1.2.2aii: Actor capability to complete other tasks involved in developing a data-damaging worm given access to elite exploits. What is the probability that a randomly selected threat actor in each threat actor class would be capable of completing the other tasks involved in creating a data-damaging worm capable of causing $\geq$ \$10 billion in damages, aside from developing elite exploits, conditional on AI reaching Capability 1?

(i.e., assume that someone publishes elite exploits on the open internet. What is the probability that a randomly selected threat actor in each class could then create a data-damaging worm, causing $>$ \$10 billion in damages with 3 months of full-time effort?)

Q1.2.2aiii: Actor capability to develop and deploy a data-damaging worm attack (end-to-end). For each threat actor level, what is the probability that a randomly selected threat actor in that level is currently capable of developing and deploying a data-damaging worm attack that causes $\geq$ \$10 billion in damages, conditional on AI reaching Capability 1?

(Here, we ask about actors having the capability to find vulnerabilities, write elite exploits of them, and use them to deploy the data-damaging worm attack with 3 months of full-time effort.)

1.2.3 AI capabilities timelines

Q1.2.3a: If this study is performed every year, in which year will it find that AI enables 25% of TA2 actors to find vulnerabilities and write elite exploits within 3 months of effort?

If you think the relevant year is after 2100, or never, please just input "2101".



Q1.2.3b: In which year will the top AI model first solve at least 90% of the tasks on Cybench?

If you think the relevant year is after 2100, or never, please just input "2101".

1.3 Forecasts conditional on mitigations

What is the probability of a data-damaging worm attack, similar to WannaCry or NotPetya, causing at least \$10 billion in economic damages in 2026, conditional on companies taking the following actions [P1 or P2]?

For these forecasts, please assume that a study conducted in 2025 finds that access to frontier AI models enables 25% of TA2 actors to find vulnerabilities and write elite exploits (Capability 1), with three months of full-time effort. You can assume that the study has unrestricted access to the AI model, regardless of its public accessibility.

Q1.3.1a: P1: Proprietary models with refusals and anti-jailbreak measures

Q1.3.1b: P2: Temporary protections with early access to defenders

Task 2: Cyberattacks against the U.S. electrical grid

2.1 Baseline forecasts (unconditional)

2.1.1 Holistic estimates

Q2.1.1a: What is the probability of at least one cyberattack against the U.S. electrical grid causing a blackout with at least \$10 billion in economic damages in 2026?

(Note: Multiple smaller attacks adding up to \$10 billion should not count here. We're asking about the occurrence of at least one individual attack meeting the \$10 billion threshold.)

Q2.1.1a (v2): What is the probability of at least one cyberattack against the U.S. electrical grid causing a blackout with at least \$100 billion in economic damages in 2026?

(Note: Multiple smaller attacks adding up to \$100 billion should not count here. We're asking about the occurrence of at least one individual attack meeting the \$100 billion threshold.)

Q2.1.1b: Conditional on such an attack happening, what is the probability it was caused by an actor of the following threat actor level:

Q2.1.1c: What is the likelihood that, in total, cyberattacks against the U.S. electrical grid cause blackouts resulting in \$X billion total economic damages in 2026?

This could be due to a single large attack or multiple smaller attacks.

Q2.1.1d: In this section, we use a logarithmic mean method to estimate the expected damages across these different attack magnitudes, based on your forecasts above. If the results do not reflect your beliefs, please let us know below. If you disagree with the



estimate we have calculated above from your forecasts, please provide your alternative best estimate for the total damages caused by blackouts due to cyberattacks against the U.S. grid in 2026 here:

2.1.2 Step-by-step estimates

Q2.1.2a: For each threat actor level, what is the probability that a randomly selected threat actor in that level is currently capable of launching a cyberattack against the U.S. electrical grid, causing a blackout with at least \$10 billion in economic damages?

Here, we ask about actors being capable of launching the cyberattack with 6 months of full-time effort.

Q2.1.2b: Assuming they are capable, what is the probability that at least one actor in each threat actor class would spend 6 or more months actively attempting to launch a cyberattack against the U.S. electrical grid, causing a $\geq$ \$10 billion blackout in 2026?

2.2 Forecasts conditional on AI capabilities

2.2.1 Holistic estimates

Q2.2.1a: What is the probability of a cyberattack against the U.S. electrical grid causing a blackout with at least \$10 billion in economic damages in 2026, conditional on an open-weight AI model uplifting TA1 actors to perform as well as a TA3 actor in a CTF-style competition?

This attack could be launched by threat actors from any class.

Q2.2.1b: What is the probability of a cyberattack on the US electrical grid causing a blackout with at least \$10 billion in economic damages in 2026, conditional on a cyberattack against the U.S. grid causing a $>$ \$100 million blackout and being subsequently confirmed to have been perpetrated by a TA1/TA2 actor using an open-weight AI model?

This attack could be launched by threat actors from any class.

Q2.2.1c: What is the probability that Capability 3 has been met, conditional on an open-weight frontier AI model being able to solve at least 90% of the tasks on Cybench completely unguided at the end of 2025?

Note that here we ask about what AI achieving Capability 2 implies about how far it's gone in Capability 3. We don't ask about the main outcome of the survey here (cyberattack against the U.S. grid).

Q2.2.1d: What is the probability of cyberattacks against the U.S. electrical grid causing a blackout with \$X in total economic damages in 2026, conditional on a frontier AI model solving at least 90% of the tasks on Cybench completely unguided?

Q2.2.1e: As in the unconditional version of this question, here we use a logarithmic mean method to estimate the expected damages across these different attack magnitudes, based on your forecasts above. If the results do not reflect your beliefs, please let us



know below. Your baseline unconditional forecasts are shown in brackets for comparison. If you disagree with the estimate we have calculated above from your forecasts, please provide your alternative best estimate for the total damages caused by blackouts due to cyberattacks against the U.S. grid in 2026, conditional on Capability 2 here:

2.2.2 AI capabilities timelines

Q2.2.2a: If this study is performed every year, in which year will it find that an individual TA1 with access to AI matches a team of ten experienced cybersecurity experts (equivalent to a TA3 actor) in a CTF-style competition?

If you think the relevant year is after 2100, or never, please just input "2101".

Q2.2.2b: In which year will an individual TA1/TA2 actor with the help of AI first cause a blackout leading to \$100 million in economic damages?

If you think the relevant year is after 2100, or never, please just input "2101".

2.3 Mitigations

Q2.3.1: Please note any thoughts or comments on mitigations you consider relevant to this threat model here.

Background information

Table A.1: Threat actor level definitions

OC-level	Definition	Threat Actor Level	Definition	Example
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
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
OC3	Operations roughly less capable than or comparable to ten individuals who are experienced professionals in	Threat Actor level 3 (TA3)	Actors only capable of OC3	Well-known criminal hacker groups,



	information security, spending several months with a total budget of up to \$1 million on the specific operation, with major preexisting cyberattack infrastructure.		operations but not higher.	well-resourced terrorist organizations, and industrial espionage organizations.
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)
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)

Descriptions of mitigation scenarios

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 (1–3).

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 31st 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](#) and would not create the same sized uplift described in the influenza RCT study).

1. 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 the model can as a result accurately and answer in sufficient helpful detail a set of questions about vulnerability and discovery and exploit development.

- a. 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.
2. 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:
 - a. A “bug bounty” programs that offer up to \$15,000 rewards for anyone who identifies and reports a universal jailbreak for one of their models
 - i. 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.
 - ii. 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
 - b. 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),
 - i. 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.
 - ii. 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.
3. 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” [23 pp. 25–26]. 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:
 - a. 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.
 - b. The organization requires and enforces strong passwords, frequent software updates, and reporting of lost or stolen devices.
 - c. 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 only 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 PO security, i.e., is open weight.

Figure A.1 is adapted from [3] and shows the historical pattern of major worm attacks (those affecting more than 10–25 companies with reported damages exceeding \$800M), broken down by the threat actor responsible. The threat actor categorization refers to the overall capabilities of the actor, not the operational capacity required for the specific attack.

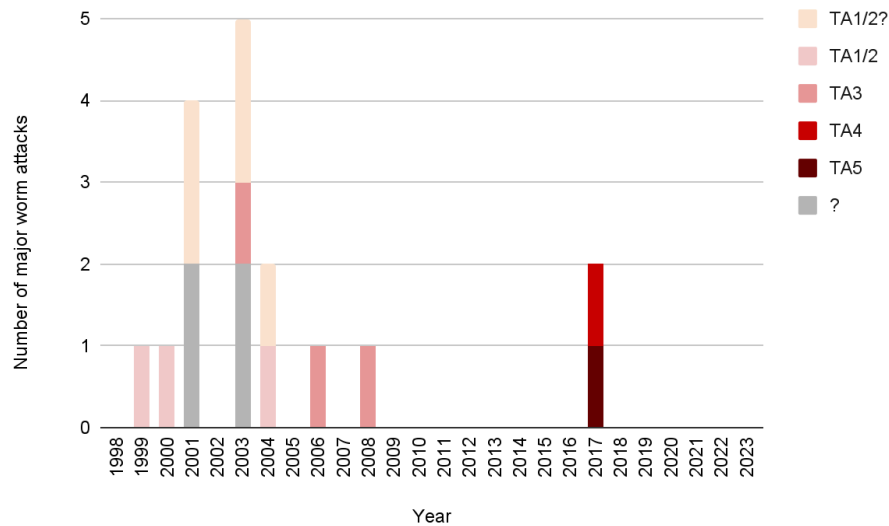


Figure A.1: Major worm attacks by threat actor (1998–2023) (adapted from [3])

[Ankit et al. 2022](#) review US blackouts, and find that outages caused by natural events last around ~10x longer than those from other causes. This is due to the scale of damage to grid equipment associated with severe weather events. Restoring power typically involves sending field workers to assess and repair damaged equipment. On the other hand, after cascading blackouts, power is typically restored relatively quickly.

Table A.2 lists selected historical blackouts, including those caused by cascading failures, physical damage from severe weather, and cyberattacks. The largest blackouts have generally been cascading failures, while the longest have resulted from physical damage to grid equipment.

Table A.2: Selected historical blackouts

Blackout	Year	Type	Scope (m people)	Duration (90% restoration)	Reported Damages (\$2024)
Northeast US	2003	cascading faults	50	2 days	\$12bn
Italy/Switzerland	2003	cascading faults	56	13 hours	\$2bn
Canada, geomagnetic storm	1989	cascading faults	9	10 hours	
USA, Hurricane Maria	2017	phys damage	2	5 months	~\$1bn ⁸
USA, Hurricane Irma	2017	phys damage	9	6 days	
Australia, ice storm	2016	phys damage	2	8 hours	[AU\$367m]
USA, Hurricane Sandy	2012	phys damage	18	8 days	
USA, Hurricane Ike	2008	phys damage	4	14 days	
USA, Hurricane Katrina	2005	phys damage	2	21 days	
USA, Hurricane Wilma	2005	phys damage	3	11 days	
USA/Canada, ice storm	1998	phys damage	2	<10 days	
Ukraine	2022	cyberattack	unknown	unknown	unknown

⁸ This includes economic damages from hurricanes broadly, but not property damage.



Ukraine	2016	cyberattack	0.5	1 hour	≤\$0.1m
Ukraine	2015	cyberattack	0.5	3 hours	Est. \$0.1–10m

Table A.3 summarises the estimated calendar time and technical headcount required for prior operational technology (OT) cyberattacks. These attacks generally required several person-years of skilled labour from state-level actors.

Table A.3: Estimated resource inputs to prior OT cyberattacks

Attack	Calendar time (years)	Est. technical headcount (FTE)
Stuxnet	0.5 – 4.5	20 – 50
2015 Ukraine grid	1.6	3 – 20
2016 Ukraine grid	1	3 – 20
Triton	0.3 – 3	1 – 10

Table A.4 outlines the capabilities demonstrated in prior OT cyberattacks, spanning long-term reconnaissance, knowledge of target OT environments, tailored malware development, realistic test environments, and stealth. Beyond the significant resource and time investments, prior OT cyberattacks have required a diverse range of capabilities.

Table A.4: Capabilities demonstrated in prior OT cyberattacks

Long-term reconnaissance	• 2015 Ukraine attack: the first known phishing attempt was 18 months prior to the attack.⁹ • Triton: reconnaissance and intrusion campaign almost three years prior to the plant shutdown.¹⁰ • German steel mill: The Dragonfly campaign began at least four years prior to the (unintentional) German steel mill incident.¹¹
Knowledge of OT environments	• 2015 Ukraine attack: “The strongest capability of the attackers was not in their choice of tools or in their expertise, but in their capability to perform long-term reconnaissance operations required to learn the environment and execute a highly synchronized, multistage, multisite attack. (E-ISAC)” • Stuxnet: “The detailed pin-point manipulations of these sub-controllers indicate a deep physical and functional knowledge of the target environment” Langner, p.10 • Incontroller/Pipedream malware demonstrates “breadth of ICS knowledge”
Tailored malware	• Triton malware was tailored to the specific PLC controller used in the petrochemical plant and the firmware version installed.¹²

⁹ [Cys Centrum](#) via [Maschmeyer 2021](#)

¹⁰ [DoJ 2021: Gladkikh indictment](#) p.5

¹¹ See [Slowik 2021 p.3](#)

¹² [Slowik 2021 pp.13–4](#)



	• Stuxnet malware targeted the specific Siemens PLC controllers used in the Natanz enrichment facility.¹³ • 2016 Ukraine Win32/Industroyer malware was designed to disrupt the working processes of industrial control systems used in electrical substations” ESET p.3
Test environments	• 2015 Ukraine: adversaries “likely had systems ... that they were able to evaluate and test their firmware against” (E-ISAC 2016, p.13) • 2016 Ukraine Win32/Industroyer malware was designed to disrupt the working processes of industrial control systems used in electrical substations” ESET p.3 • Triton: obtaining the PLC hardware and software “was essential for recreating the full working environment needed for experimenting and writing malicious code” (Di Pino et al. 2018) [researcher replication of the malware] • Some mistakes in prior OT cyberattacks may be explained by insufficient testing — e.g., Triton malware inadvertently causing plant shutdowns, the failure of the protective relay DoS attack in the 2016 grid attack.¹⁴
Stealth	• Triton attackers interacted with target controllers during times when fewer workers were on-site to react to alarms (Mandiant) • Stuxnet malware had a range of stealthy features, e.g., using stolen digital certificates, hiding binaries in a Windows rootkit, checking for and disabling antivirus software, hiding modified code on PLCs via a modified rootkit (Helmsley & Fisher 2018, p.11)

¹³ [OSTI 2018 p.11](#); Langner

¹⁴ [Mandiant 2017](#): “... Mandiant ran trilog.exe in a lab environment with a valid Triconex controller and discovered a conditional check in the malware that prevented the payload binary from persisting in the environment. Mandiant confirmed that, after correctly patching the attack script to remove this check, the payload binary would persist in controller memory, and the controller would continue to run.”
[Slowik 2019 p.4](#) “One possibility for this error may be test environment. Protocol emulators in software, such as the publicly-available IEC Server project,¹⁷ do not enforce statefulness and associated communication timeouts. Physical hardware will employ such items in line with the vendor-specific implementations of the relevant communication standard.”



Appendix B: Follow-up survey

About this follow-up survey

Thank you for participating in our original study. We appreciate your thoughtful contributions.

In this follow-up, we're interested in understanding whether your views have changed, and why they have or have not changed. You'll have the opportunity to review your previous forecasts, reflect on the results of the study we conducted, and consider new developments that have occurred since you completed the original survey. This follow-up survey will focus on risks from data-damaging worm attacks (Task 1 from the original survey) and will not include questions on attacks on the US electrical grid (Task 2 from the original survey).

Our goal is to better understand how people update their beliefs in light of new arguments, empirical data, and real-world events.

What to expect

This survey is divided into the following sections:

1. **Reviewing study results:** Read a summary of the arguments and results from the study, and let us know how (if at all) they influence your views.
2. **Reflecting on the cybersecurity incident report:** Consider how a recent cybersecurity incident report published by Anthropic affects your forecasts.
3. **Updating forecasts & reflections:** Update some of your forecasts (if you wish), and tell us what factors played the biggest role in shaping your views.

We have structured the survey so it takes no more than 1 hour, though you're welcome to spend more time if you wish. If you have any questions, feel free to reach out to rebecca@forecastingresearch.org.

We suggest you complete this Qualtrics survey on a single device. Your answers will be automatically saved in the browser.

Step 1: Reviewing study results

Below we present a brief summary of the study results. In addition to numerical results, we also summarize the main narrative arguments participants provided in support of their forecasts.

We present results from

- 8 people with expertise in cybersecurity ("experts")
- 13 top generalist forecasters ("superforecasters")

Holistic estimate

We start with the same baseline question you answered in the original survey:



BOX B1: HOLISTIC ESTIMATE QUESTION

What is the probability of at least one data-damaging worm attack, similar to WannaCry or NotPetya, causing at least \$10 billion in economic damages in 2026?

We show the:

- Unconditional forecasts
- Forecasts conditional on Capability scenario 1, where: AI enables some TA2 actors to write elite exploits. We define this as:
 - 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.
 - All models that meet Capability 1 are open-weight

Forecasts

The median expert's unconditional forecast was 8% and the median superforecaster's was 5%. When asked to condition on AI enabling 25% of TA2 actors to write elite exploits and relevant models being open-weight, these median forecasts rose to 41% and 15%, respectively.

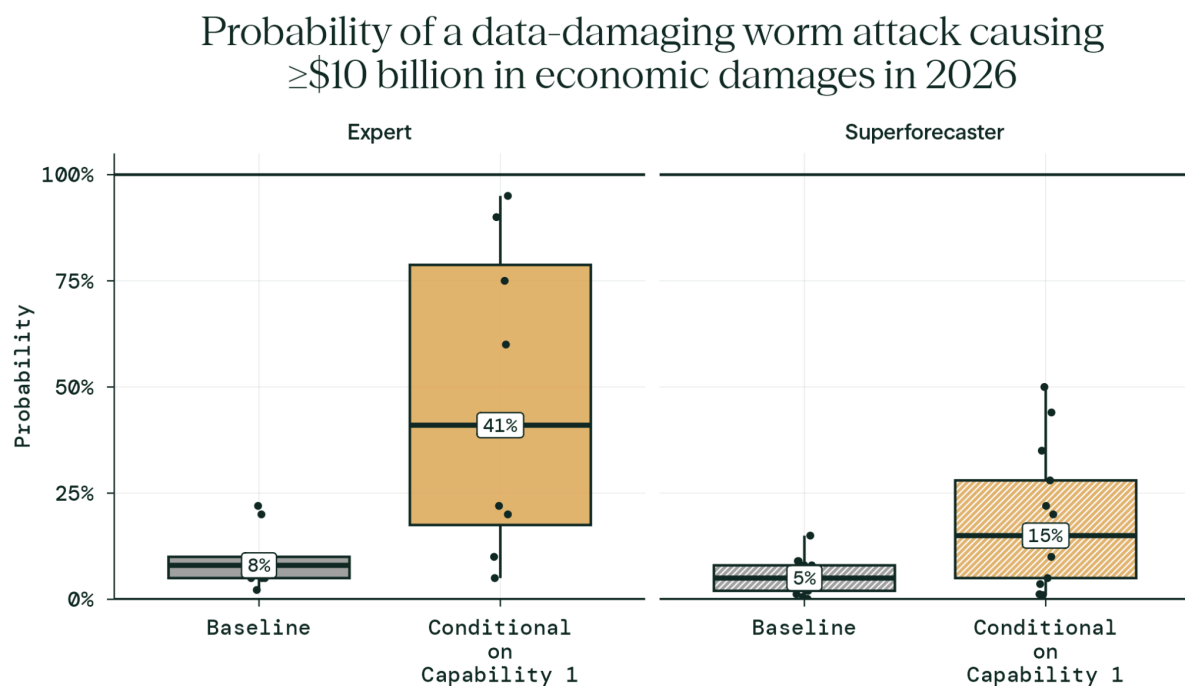


Figure B.1: Probability of a data-damaging worm attack causing ≥\$10 billion in economic damages in 2026.

Rationales

The main arguments used in rationales explaining forecasts are summarized below.

Rationales for baseline forecasts



Participants' text rationales for these forecasts commonly referenced the very low base rates for this event, noting that \$10 billion would indicate a damage level with very few precedents for this type of attack. Many noted that only TA4 and TA5 actors would have the resources to conduct this scale of attack. Although some participants noted that there are several active conflicts or areas of geopolitical tension globally (Ukraine/Russia, Iran/Israel, China/Taiwan) they also noted that there is generally little motivation for major powers to use cyberattack capabilities. Some respondents suggested that AI could cause deviation from historical patterns but thought that 2026 would be too soon to see this. There were mixed views on whether improved cyberdefense outweighs potential for greater harm in a more digitally dependent society.

Rationales for Capability 1 forecasts

Rationales noted that finding elite exploits would remove the most important bottleneck for TA2 actors, and 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 defense-offense timing imbalance: while AI helps both sides, there may be a "dangerous window of vulnerability" when new capabilities emerge.

Step-by-step estimate - capability

Here we ask about the capability of actors to create a data-damaging worm, including finding elite exploits and doing all other tasks necessary to conduct a successful attack.

BOX B2: ACTOR CAPABILITY QUESTION

For each threat actor level, what is the probability that a randomly selected threat actor in that level is currently capable of developing and deploying a data-damaging worm attack that causes $\geq$ \$10 billion in damages?

(Here, we ask about actors having the capability to find vulnerabilities, write elite exploits of them, and use them to deploy the data-damaging worm attack with 3 months of full-time effort.)

We show the:

- Unconditional forecasts
- Forecasts conditional on Capability scenario 1, where AI enables some TA2 actors to write elite exploits. We define this as:
 - 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.
 - All models that meet Capability 1 are open-weight



Forecasts

Both groups of participants generally thought that TA1 to TA3 actors were very unlikely to be able to develop a data-damaging worm and that almost all TA5 actors would have this capability. Opinions were more mixed about the capabilities of TA4 actors, with the median expert suggesting that a randomly selected TA4 actor has just a 5% probability of being able to develop a data-damaging worm, but the median superforecaster putting the probability at 40%.

Actor end-to-end capability to develop data-damaging worms

Values represent the probability of a randomly selected actor in each actor class having the capability to perform all the steps required to develop and deploy a large-scale data-damaging worm with 3 months of effort.

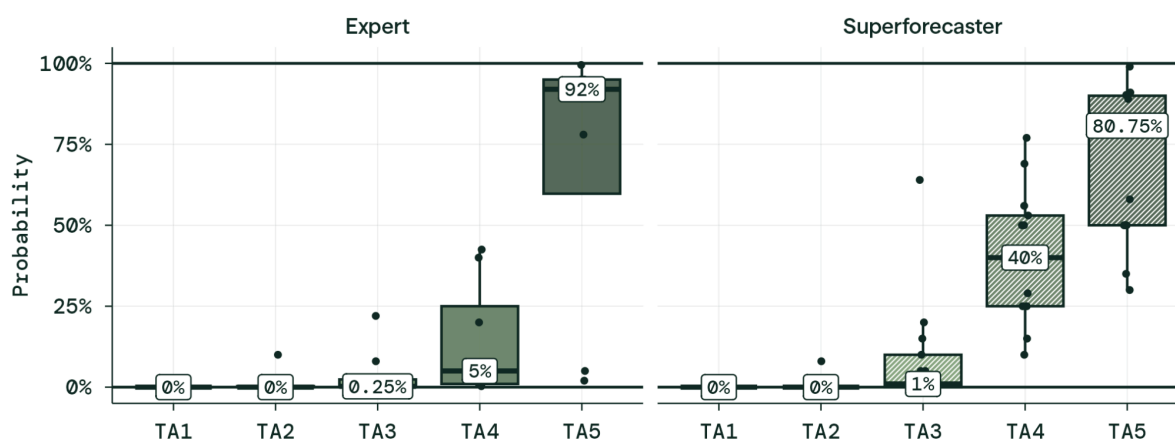


Figure B.2: Probability that a randomly selected actor in each actor class is able to develop a data-damaging worm with 3 months of effort.

When asked to condition on Capability 1 (access to frontier AI models enables 25% of TA2 actors to find vulnerabilities and write elite exploits), participants, particularly experts, increased their estimate of the probability that a randomly selected actor could develop a data-damaging worm. In this hypothetical scenario, the median expert put a 51.5% probability of a TA3 actor succeeding at this task, and a 13.75% probability of a TA2 actor succeeding.



Actor end-to-end capability to develop data-damaging worms, conditional on Capability 1

Values represent the probability of a randomly selected actor in each actor class having the capability to perform all the steps required to develop and deploy a large-scale data-damaging worm with 3 months of effort. Forecasts are conditional on Capability 1: AI enables 25% of moderately-skilled individual hackers to develop elite exploits.

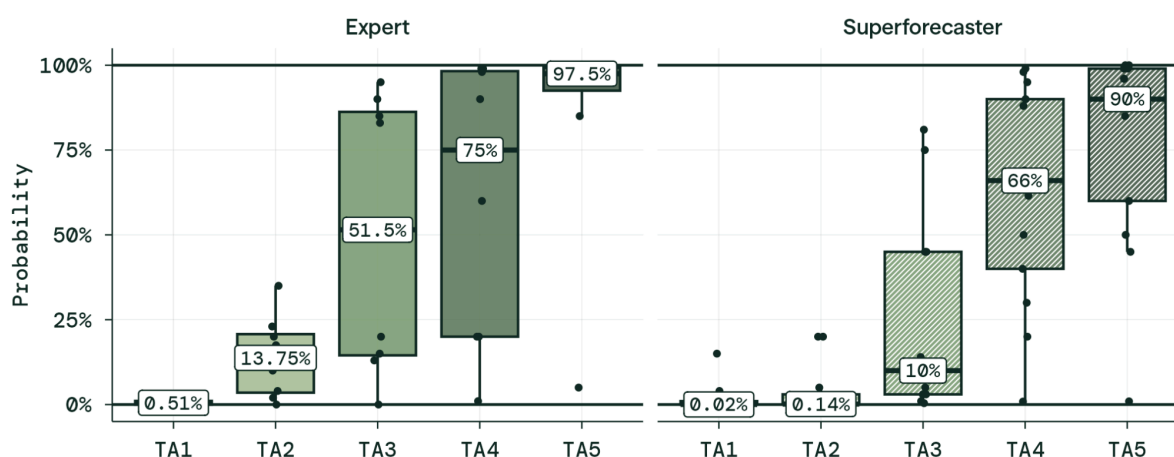


Figure B.3: Probability that a randomly selected actor in each actor class is able to develop a data-damaging worm with 3 months of effort conditional on access to frontier AI models enabling 25% of TA2 actors to find vulnerabilities and write elite exploits

Rationales

The main arguments used in rationales explaining forecasts are summarized below.

Unconditional forecasts

Rationales commonly highlighted the difficulty of finding elite exploits, noting that it requires deep technical knowledge and skill as well as expensive infrastructure. The 3-month timeline was thought to be particularly challenging. Some participants noted that important IT systems would be protected by teams of cybersecurity specialists, making it unlikely that an individual (TA1 or TA2) actor would succeed in beating this team of possibly hundreds of skilled defenders.

In addition, many participants highlighted the absence of instances of large-scale damage from data-damaging worms and suggested this provides evidence that TA1–3 actors are unlikely to have this capability. Unlike TA4 and TA5 actors, which are mostly states, individuals and smaller groups have fewer constraints on releasing a worm, if they had the capability. State groups (TA4 and TA5) actors may avoid releasing a data-damaging worm attack due to the possibility of retaliation, or the risk of an attack harming their own interests given its uncontrolled nature and/or the effect of a large-scale attack on global financial systems.

Rationales for Capability 1 forecasts

Most participants thought that the Capability 1 scenario would boost the capabilities of all actors. Generally, participants thought that lower-level actors, (TA1–TA3) would be helped more than TA4 and TA5 actors, as these higher-level actors may already have these capabilities. Some participants noted that even actors who already have the



capability to write elite exploits would likely benefit from AI making their work more efficient.

However, some rationales suggested that the TA1 and TA2 actors, even those who had been enabled to identify elite exploits, would still face barriers in performing the other tasks necessary to create a $\geq \$10$ billion attack, including skill barriers but also access to necessary resources and operational organization. Many participants again noted that the 3 month timeline would be limiting. Some participants noted that defenses would also be boosted by AI capabilities, with at least one participant arguing that the increase in defenses would negate any increase in offensive capability.

Step-by-step estimate - willingness

Here we ask about the willingness of actors to launch a data-damaging worm attack.

BOX B3: ACTOR WILLINGNESS QUESTION

Assuming they are capable, what is the probability that at least one actor in each threat actor class would spend 3 or more months actively attempting to launch a data-damaging worm attack causing $\geq \$10$ billion in economic damages in 2026?

We show the unconditional forecasts.

Forecasts

Both groups gave high estimates ($>95\%$) for TA1 and TA2 actors, suggesting that the large number of such actors makes it almost certain that one would be willing to conduct such an attack.

Actor willingness to launch data-damaging worm attacks

Values represent the probability that at least one actor in each actor class would spend at least 3 months actively attempting to launch a data-damaging worm attack if they were capable.

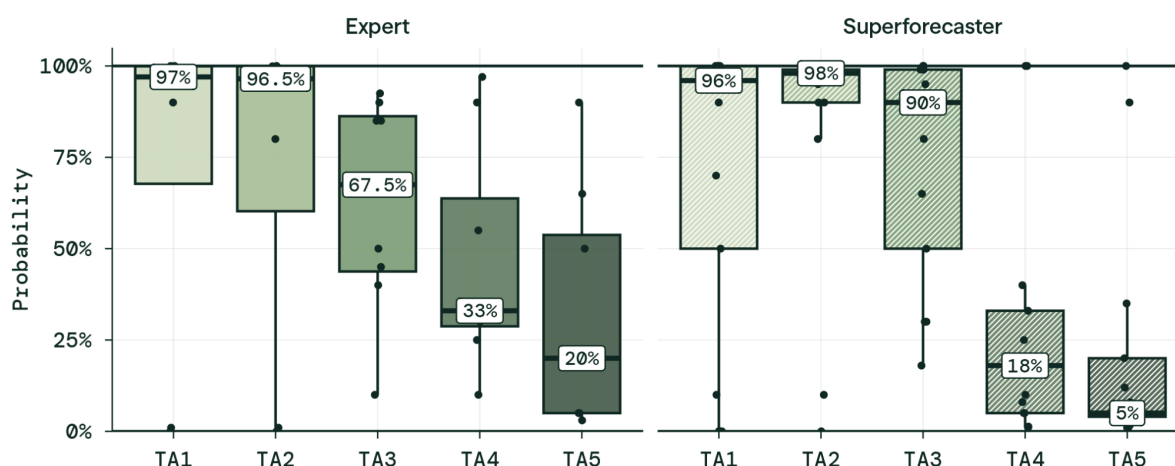


Figure B.4: Probability that at least one actor in the threat class would spend 3 or more months actively attempting to launch a data-damaging worm attack, causing $\geq \$10$ billion in economic damages in 2026, assuming they had the capability to do so



Rationales

Many participants noted that the very large number of TA1 and TA2 actors suggests it is likely that at least one of those actors would be willing to conduct such a large-scale cyberattack. Several participants emphasized psychological factors, e.g., these groups include teenagers, individuals with poor impulse control, and those motivated by fame or malice rather than rational calculation. Some participants suggested that TA3 attackers are generally more motivated to conduct attacks, and that some, e.g. terrorist groups, would be motivated to cause widespread damage. However, several participants suggested that TA3 attackers could be deterred by the law enforcement response that such a large attack would provoke. They argued that these groups, which are often criminal groups, would be looking for the biggest pay-off for the smallest risk, noting that monetizing \$10 billion in damage would be extremely difficult through ransomware, and there may be better options than a very large-scale global cyberattack.

TA4 and TA5 actors (which were generally taken to be states or state-backed groups) were thought to be unlikely to conduct such an attack due to fears of attribution and retaliation. Multiple participants emphasized that TA5 actors have the most sophisticated understanding of escalation dynamics and would view such attacks as "crossing major red lines." Some participants suggested that these actors would only conduct such a large attack if they were at war with an adversary. Some participants noted relative geopolitical stability at present, but some highlighted the ongoing war in Ukraine and potential future China-Taiwan tensions as factors that could change this calculus. Several participants noted that North Korea, within TA4, might be more willing, given their history with WannaCry and existing cybercriminal activities.

Reflecting on your forecasting process

We would like to understand your thoughts on these results and rationales.

- Were the numerical results surprising in any way? Do they make you reconsider your forecast?
- Do you agree or disagree with any of the arguments?
- Do any of these arguments make you think differently about your forecast? If not, why?
- Which argument is the most relevant to informing your forecast?

Here we provided participants with their original forecasts and rationales.

Step 2: Cybersecurity incident report

On November 13th, 2025, the AI company Anthropic released a report on what they describe as an orchestrated cyber espionage campaign that made use of their LLM, Claude.

- In mid-September 2025, Anthropic detected suspicious activity, which they later determined to be a sophisticated espionage campaign, conducted by a Chinese state-sponsored group.
- Anthropic reports that 80–90% of operations in the attack were autonomously conducted by Claude Code. The attack made use of tools, including open-source penetration testing tools coordinated through Model Context Protocol. They report the attack having six phases:



- Campaign initialization (human-directed)
- Reconnaissance and attack surface mapping (AI-autonomous)
- Vulnerability discovery and validation (AI-autonomous)
- Credential harvesting and lateral movement (AI-autonomous)
- Data collection and intelligence extraction (AI-autonomous)
- Documentation and handoff (AI-autonomous)

Anthropic identified 30 targets of the attack, including technology corporations, financial institutions, chemical manufacturers, and government agencies across multiple countries. They reported the attack led to “multiple successful intrusions.”

- The attackers bypassed Claude’s guardrails by a) breaking down the attacks into small, seemingly innocent tasks, and b) telling Claude that it was an employee of a cybersecurity firm conducting defensive testing.
- Anthropic suggests that the work performed by Claude in this campaign would have taken “vast amounts of time for a human team”, and utilized speeds that would be impossible for human hackers to match.
- Anthropic notes that Claude sometimes hallucinated credentials or claimed that publicly available information was actually secret information it had collected.

You can read the full 10–page report [here](#).

We would like to understand your thoughts on this campaign.

- Does this report make you think differently about any of your forecasts? If not, why?

Step 3: Updating forecasts & reflections

Now we’d like to invite you to review some of your forecasts and update them.

Holistic

Update

When you completed the original survey, you answered the following question:

What is the probability of at least one data-damaging worm attack, similar to WannaCry or NotPetya, causing at least \$10 billion in economic damages in 2026?

(Note: Multiple smaller attacks adding up to \$10 billion should not count here. We’re asking about the occurrence of at least one individual attack meeting the \$10 billion threshold.)

Now that you’ve seen the study results and the Anthropic report, would you update your forecast of the main outcome?

Please enter your updated forecast below (in %). If you don’t want to update your forecast, please enter your original forecast.

Please enter a brief explanation for your reasoning.

Attributing the Change

If you had to attribute the change in your forecast to specific factors, how much would you assign to each of the following?



- General AI progress
- Movement (or lack thereof) in policy
- Numerical results from our study
- Expert and superforecaster rationales from our study
- Anthropic cyber espionage report
- Other (Please specify below)

(Please enter approximate percentages. The total should sum to 100%.)

If you selected "Other", please specify here:

For each of the above factors, did it cause you to raise or lower your forecast of baseline risk?

Step-by-step

Actor capability

In this section, we ask about the capability of threat actors to complete the following tasks.

- a) Write elite exploits
- b) Complete other tasks (aside from writing elite exploits) involved in creating a data-damaging worm.
- c) Develop and deploy a data-damaging worm attack (end-to-end)

We'll ask about your unconditional forecasts and your forecasts conditional on the Capability scenario 1, where AI enables 25% of TA2 actors to write elite exploits and all models that meet this capability are open-weight.

- a) Actor capability to write elite exploits

What is the probability that a randomly selected threat actor in each threat actor class would be capable of developing elite exploits involved in creating a data-damaging worm capable of causing $\geq$ \$10 billion in damages?

(Here, we only ask about actors having the capability to find vulnerabilities and write elite exploits of them with 3 months of full-time effort.)

Now that you've seen the study results and the Anthropic report, would you update these forecasts?

We'd also like to know how your forecasts would change if the question asked about the capability to find vulnerabilities and write elite exploits of them with 12 months (rather than 3 months) of full-time effort.

- b) Actor capability to complete other tasks involved in developing a data-damaging worm given access to elite exploits

What is the probability that a randomly selected threat actor in each level can currently complete the other tasks, aside from creating elite exploits, involved in creating a data-damaging worm causing $\geq$ \$10 billion in damages?



(i.e., assume that someone publishes elite exploits on the open internet. What is the probability that a randomly selected threat actor in each class could then create a data-damaging worm, causing $\geq$ \$10 billion in damages with 3 months of full-time effort?)

Now that you've seen the study results and the Anthropic report, would you update these forecasts?

We'd also like to know how your forecasts would change, if the question asked about developing a data-damaging worm given access to elite exploits with 12 months (rather than 3 months) of full-time effort.

c) Actor capability to develop and deploy a data-damaging worm attack
(end-to-end)

For each threat actor level, what is the probability that a randomly selected threat actor in that level is currently capable of developing and deploying a data-damaging worm attack that causes $\geq$ \$10 billion in damages?

(Here, we ask about actors having the capability to find vulnerabilities, write elite exploits of them, and use them to deploy the data-damaging worm attack with 3 months of full-time effort.)

We'd also like to know how your forecasts would change if the question asked about the capability to develop and deploy a data-damaging worm attack with 12 months (rather than 3 months) of full-time effort.

Actor willingness

Assuming they are capable, what is the probability that at least one actor in each threat actor class would spend 3 or more months actively attempting to launch a data-damaging worm attack causing $\geq$ \$10 billion in economic damages in 2026?

We'd also like to know how your forecast would change if the question asked about the actor spending 12 or more months (rather than 3 or more months) actively attempting to launch the attack.



Appendix C: Additional results

Marginal and relative risk given AI capabilities

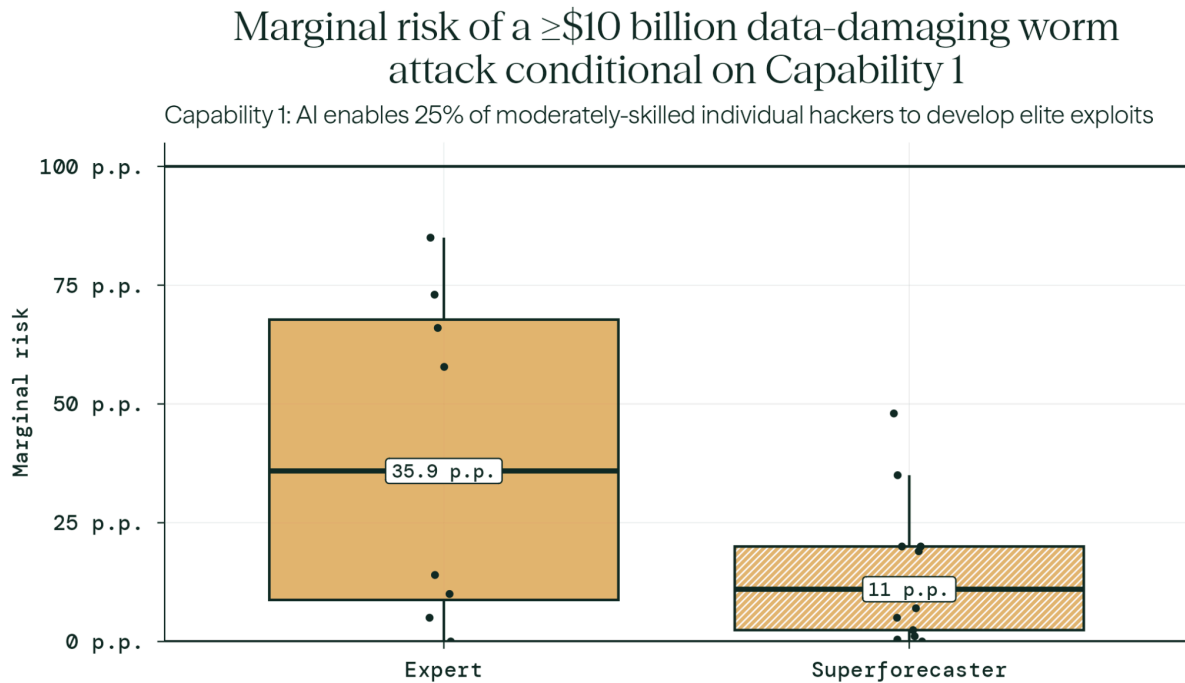


Figure C.1: Marginal risk of at least one data-damaging worm attack causing at least \$10 billion in economic damages in 2026, conditional on Capability 1 (AI enables TA2 actors to write elite exploits).

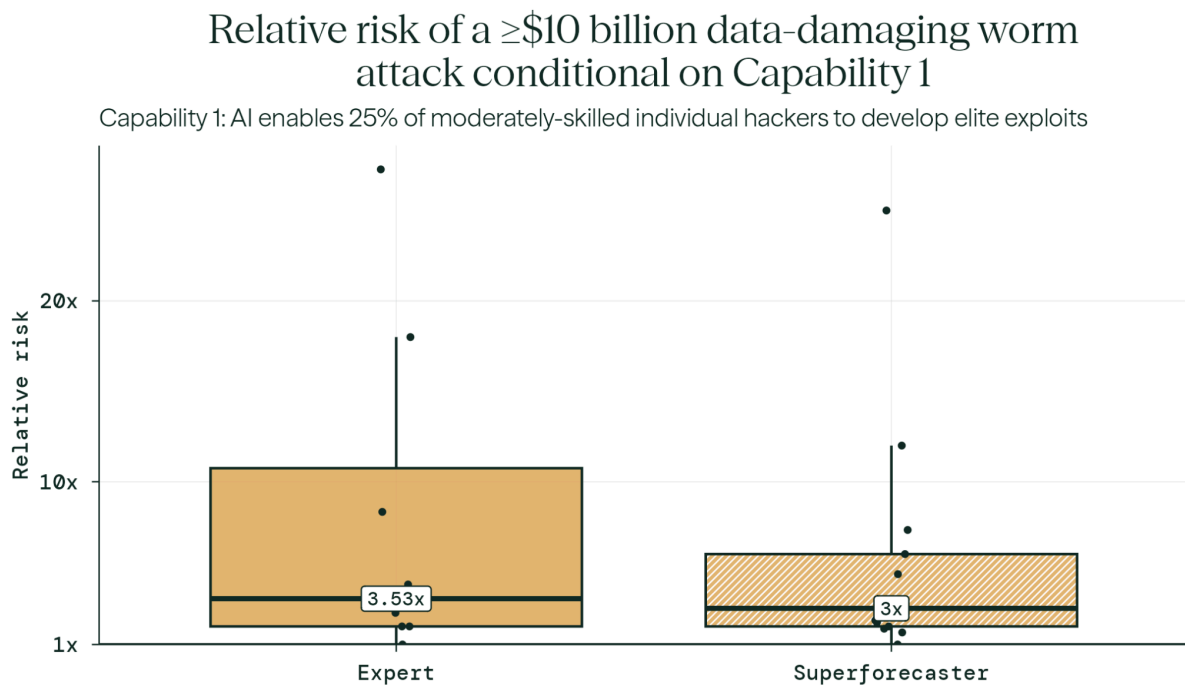




Figure C.2: Relative risk of at least one data-damaging worm attack causing at least \$10 billion in economic damages in 2026, conditional on Capability 1 (AI enables TA2 actors to write elite exploits).

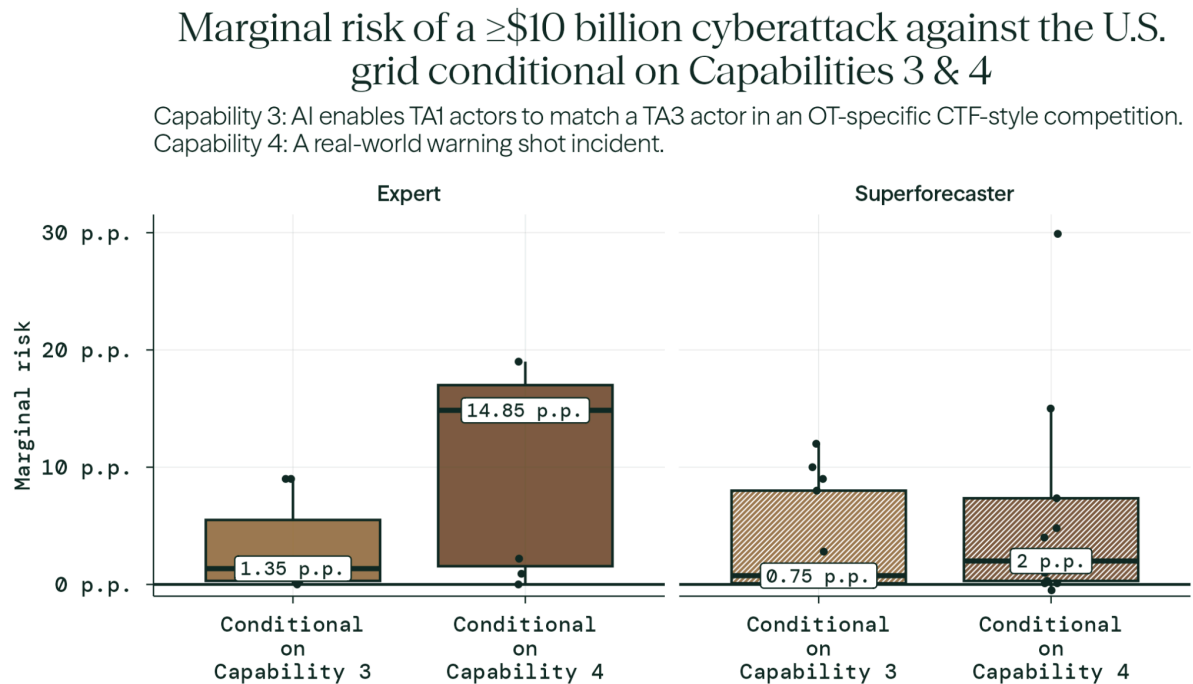


Figure C.3: Marginal risk of at least one cyberattack against the US grid causing at least \$10 billion in economic damages in 2026, conditional on Capability 3 (AI enables TA1 actors to match a TA3 actor in an OT-specific CTF-style competition), and Capability 4 (a real-world warning shot).

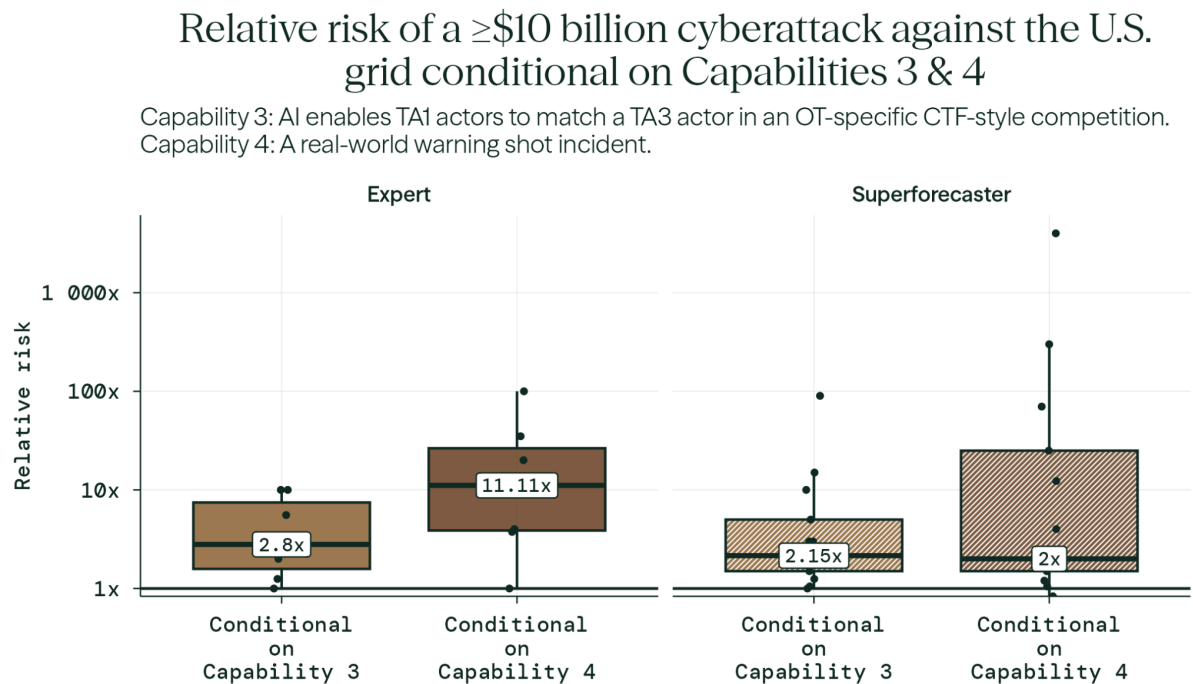




Figure C.4: Relative risk of at least one cyberattack against the US grid causing at least \$10 billion in economic damages in 2026, conditional on Capability 3 (AI enables TA1 actors to match a TA3 actor in an OT-specific CTF-style competition), and Capability 4 (a real-world warning shot).

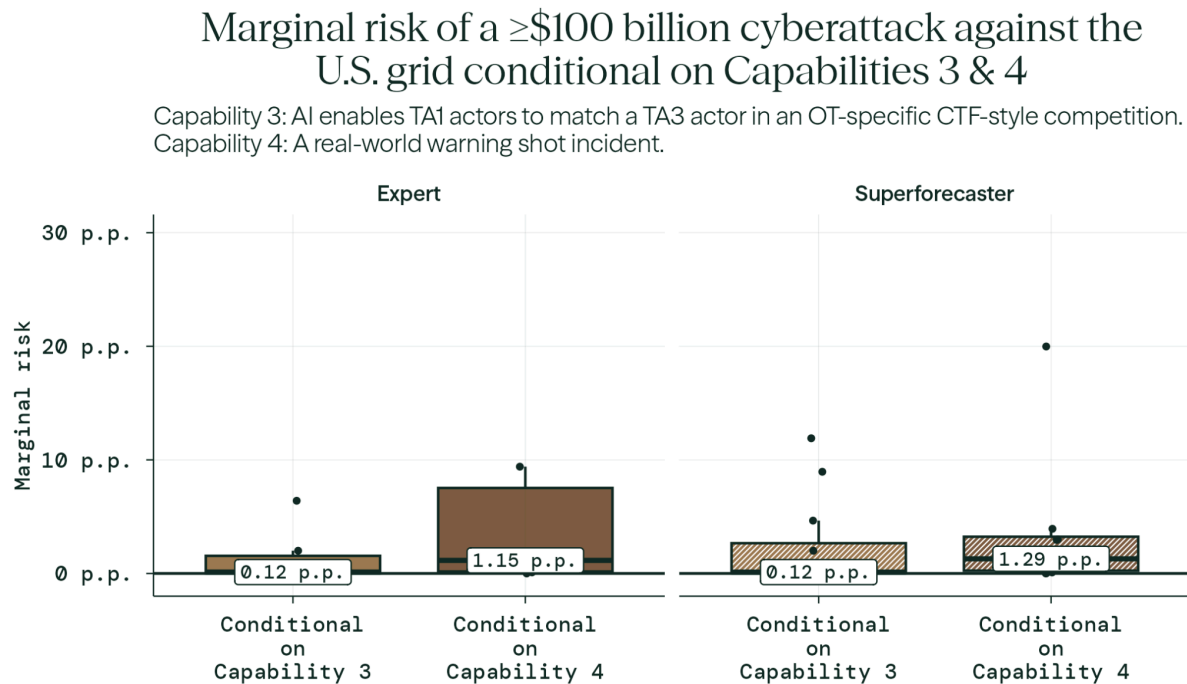


Figure C.5: Marginal risk of at least one cyberattack against the US grid causing at least \$100 billion in economic damages in 2026, conditional on Capability 3 (AI enables TA1 actors to match a TA3 actor in an OT-specific CTF-style competition), and Capability 4 (a real-world warning shot).

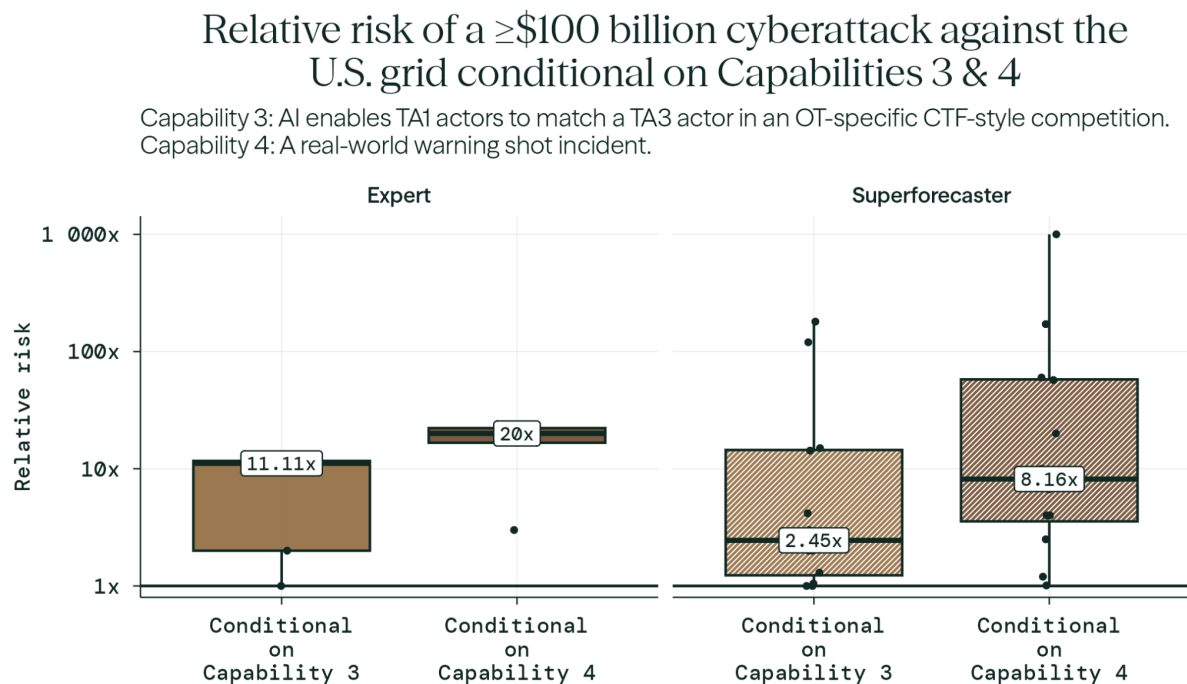




Figure C.6: Relative risk of at least one cyberattack against the US grid causing at least \$100 billion in economic damages in 2026, conditional on Capability 3 (AI enables TA1 actors to match a TA3 actor in an OT-specific CTF-style competition), and Capability 4 (a real-world warning shot).

Actor attribution for larger scale grid attack

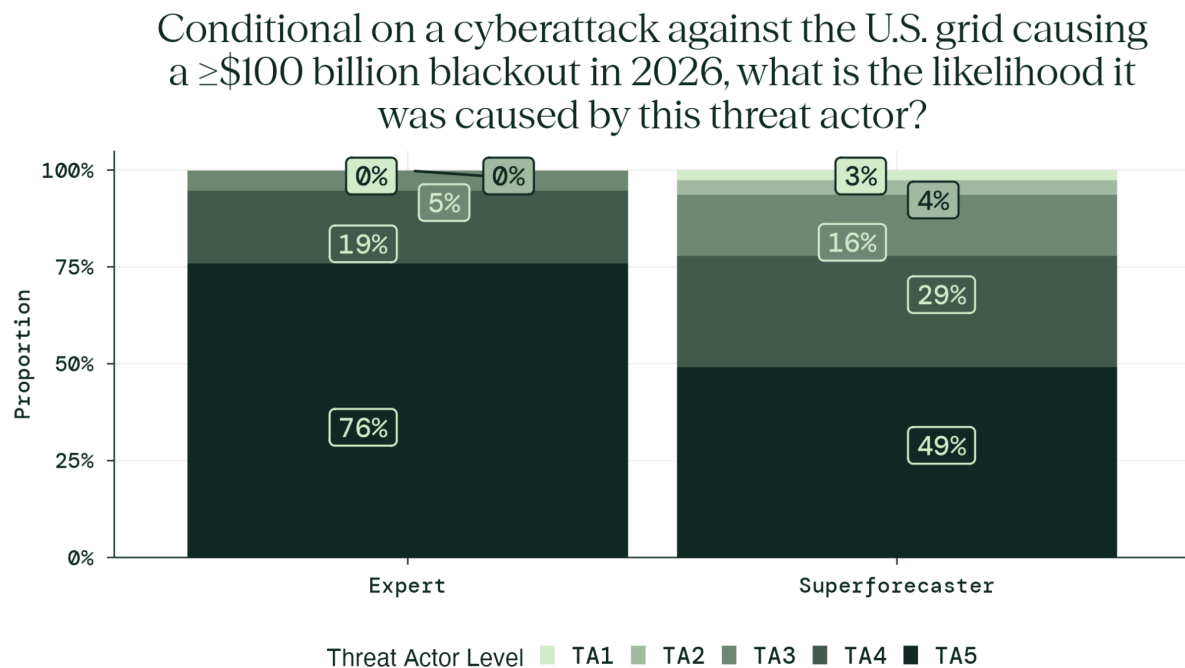


Figure C.7: The average probability assigned to each threat actor type being the primary cause of a cyberattack against the U.S. grid. These forecasts are conditional on the cyberattack occurring in 2026, and causing at least \$100 billion in damages.

Actor capability breakdown and willingness to act

In the original survey, participants were asked about actor capability and willingness over a 3-month time horizon. In the follow-up survey, we both repeated these 3-month questions (to capture updated views) and introduced a longer 12-month horizon. Here we present additional results across these three sets of responses. Consistent with the main outcome forecasts, updates in the three-month case were generally small, and aggregate results did not change significantly. We also show forecasts of actor willingness to launch a cyberattack against the U.S. grid causing at least \$100B, in the main text we only show these results for the smaller ($\geq \$10$ B) magnitude.



Original survey results (3-month horizon)

Actor end-to-end capability to develop data-damaging worms

Values represent the probability of a randomly selected actor in each actor class having the capability to perform all the steps required to develop and deploy a large-scale data-damaging worm with 3 months of effort.

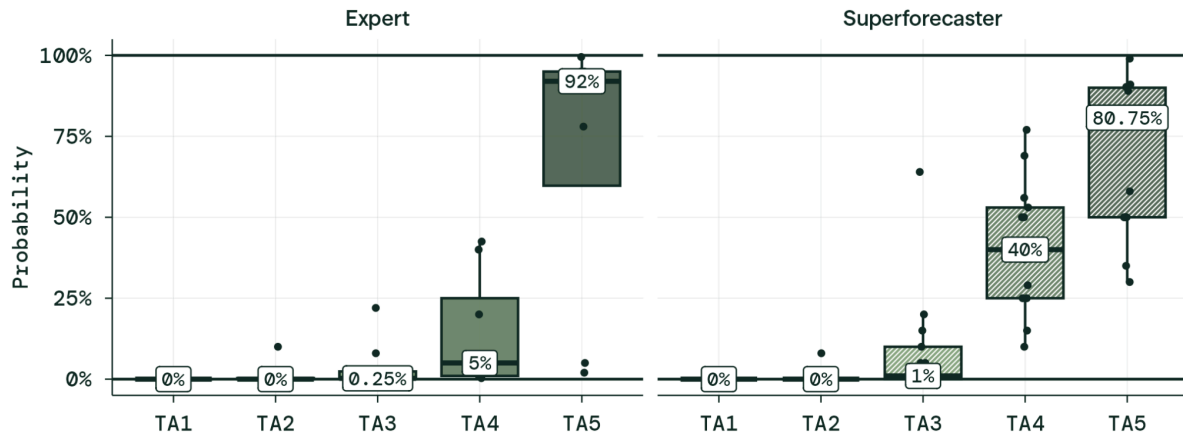


Figure C.8: Original survey results. Probability that a randomly selected actor in each actor class is able to develop a data-damaging worm given 3 months of effort.

Actor end-to-end capability to develop data-damaging worms, conditional on Capability 1

Values represent the probability of a randomly selected actor in each actor class having the capability to perform all the steps required to develop and deploy a large-scale data-damaging worm with 3 months of effort. Forecasts are conditional on Capability 1: AI enables 25% of moderately-skilled individual hackers to develop elite exploits.

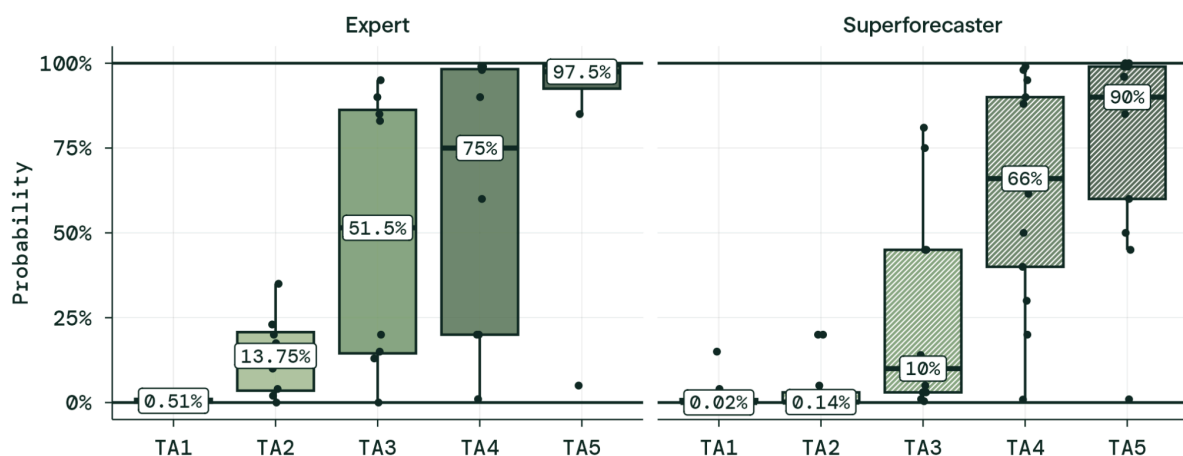


Figure C.9: Original survey results. Probability that a randomly selected actor in each actor class is able to develop a data-damaging worm conditional on Capability 1 given 3 months of effort.



Actor capability to write elite exploits

Values represent the probability of a randomly selected actor in each actor class having the capability to write elite exploits with 3 months of effort.

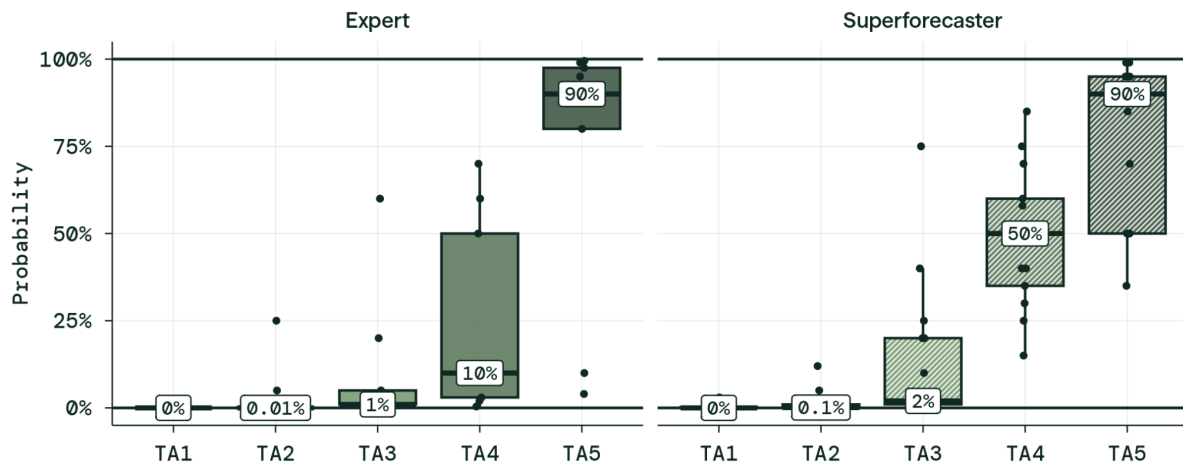


Figure C.10: Original survey results. Probability that a randomly selected actor in each actor class is able to write elite exploits given 3 months of effort.

Actor capability to write elite exploits, conditional on Capability 1

Values represent the probability of a randomly selected actor in each actor class having the capability to write elite exploits with 3 months of effort. Forecasts are conditional on Capability 1: AI enables 25% of moderately-skilled individual hackers to develop elite exploits.

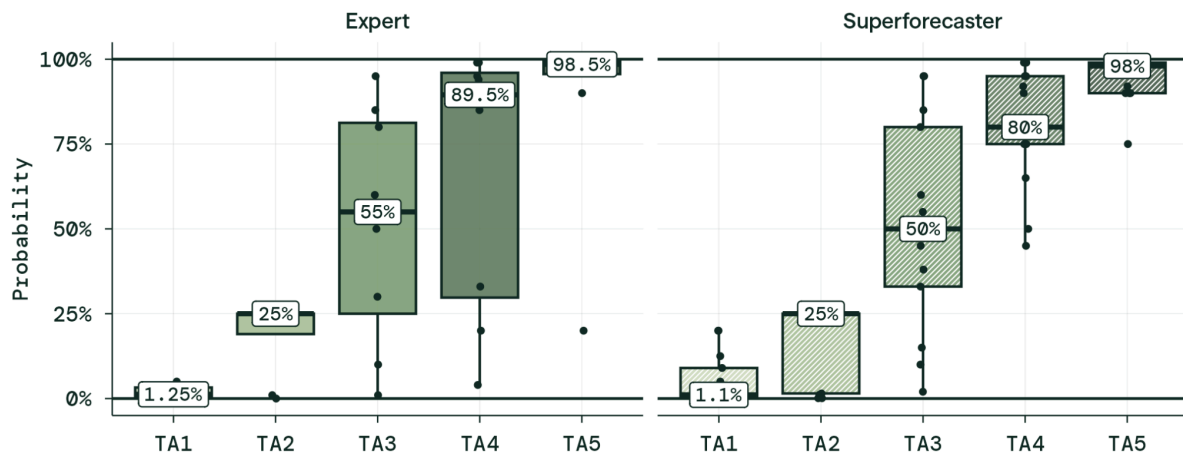


Figure C.11: Original survey results. Probability that a randomly selected actor in each actor class is able to write elite exploits conditional on Capability 1 given 3 months of effort.



Actor capability to do other steps (other than write elite exploits) of developing a data-damaging worm

Values represent the probability of a randomly selected actor in each actor class having the capability to perform the other steps (other than write elite exploits) required to develop and deploy a large-scale data-damaging worm with 3 months of effort.

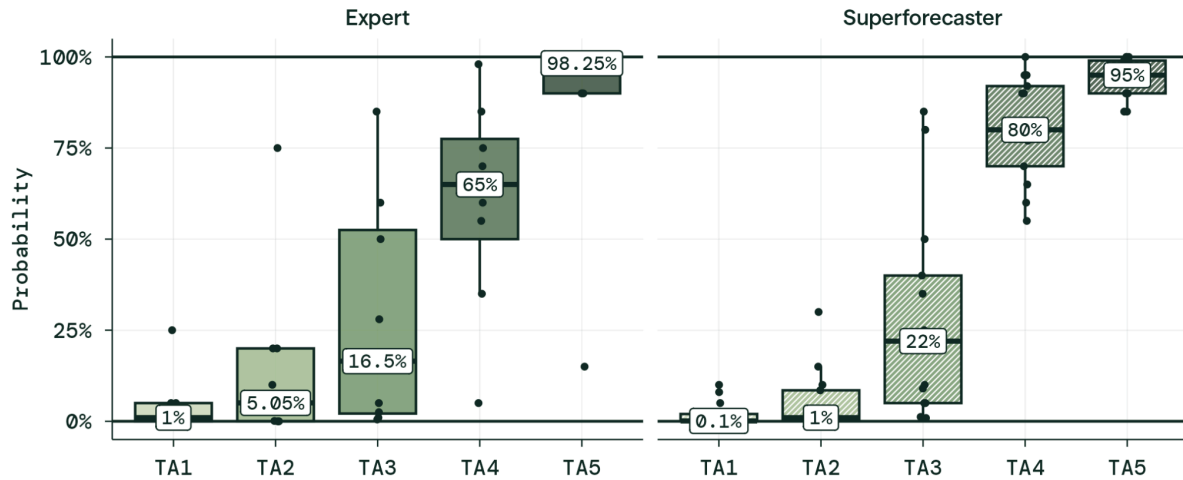


Figure C.12: Original survey results. Probability that a randomly selected actor in each actor class is able to do the other steps required to develop and deploy a data-damaging worm given 3 months of effort.

Actor capability to do other steps (other than write elite exploits) of developing a data-damaging worm, conditional on Capability 1

Values represent the probability of a randomly selected actor in each actor class having the capability to perform the other steps (other than write elite exploits) required to develop and deploy a large-scale data-damaging worm with 3 months of effort. Forecasts are conditional on Capability 1: AI enables 25% of moderately-skilled individual hackers to develop elite exploits.

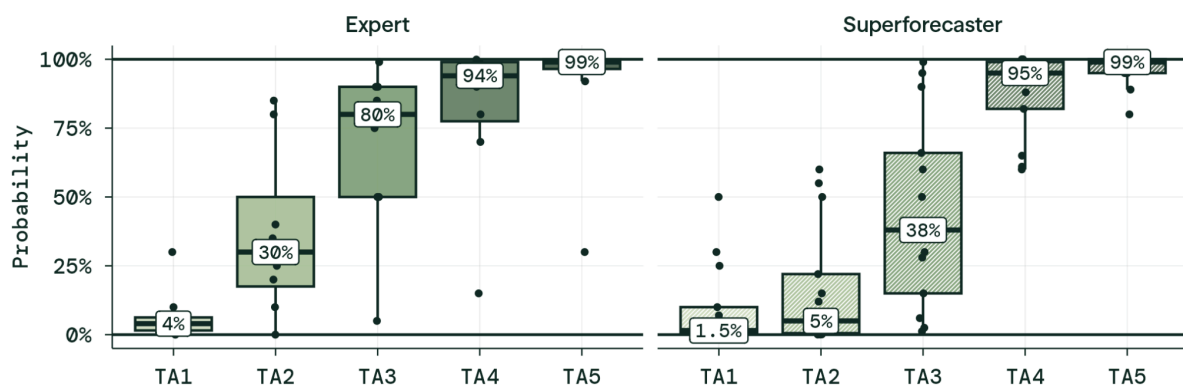


Figure C.13: Original survey results. Probability that a randomly selected actor in each actor class is able to do the other steps required to develop and deploy a data-damaging worm conditional on Capability 1 given 3 months of effort.



Actor end-to-end capability to launch a $\geq \$100\text{B}$ cyberattack against the U.S. grid

Values represent the probability of a randomly selected actor in each actor class having the capability to launch a $\geq \$100\text{B}$ cyberattack against the U.S. grid with 6 months of effort.

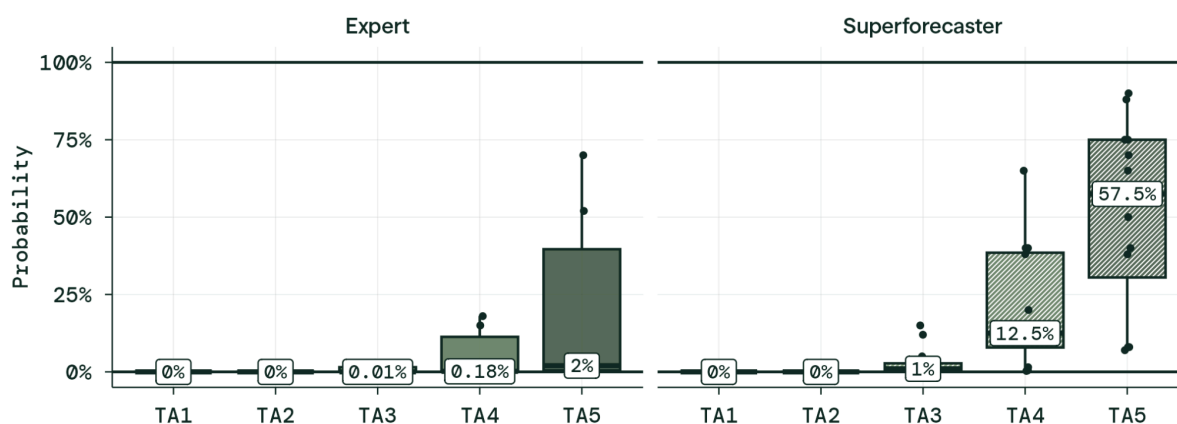


Figure C.14: Probability that a randomly selected actor in each actor class is able to launch a cyberattack against the U.S. grid with 6 months of effort, leading to at least \$100 billion in economic damages.

Follow-up survey (3-month horizon, updated responses)

Follow-up study: Actor end-to-end capability to develop data-damaging worms

Values represent the probability of a randomly selected actor in each actor class having the capability to perform all the steps required to develop and deploy a large-scale data-damaging worm with 3 months of effort.

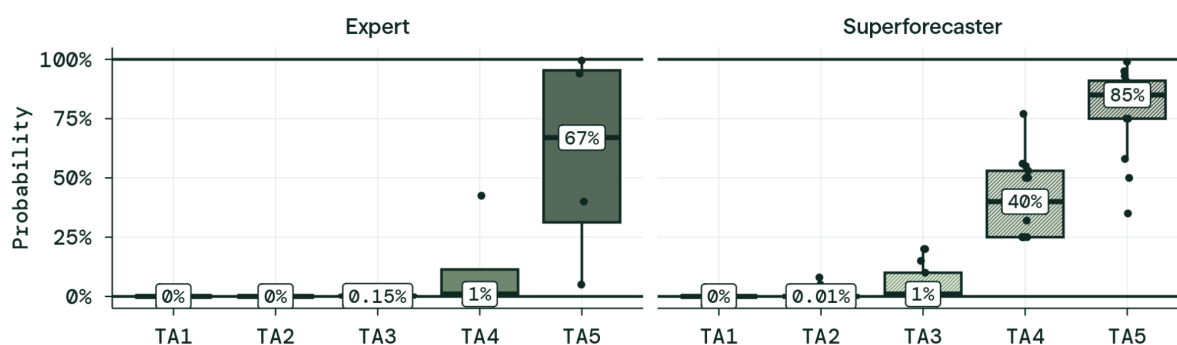


Figure C.15: Follow-up survey results. Probability that a randomly selected actor in each actor class is able to develop a data-damaging worm given 3 months of effort.

Follow-up study: Actor end-to-end capability to develop data-damaging worms, conditional on Capability 1

Values represent the probability of a randomly selected actor in each actor class having the capability to perform all the steps required to develop and deploy a large-scale data-damaging worm with 3 months of effort. Forecasts are conditional on Capability 1: AI enables 25% of moderately-skilled individual hackers to develop elite exploits.

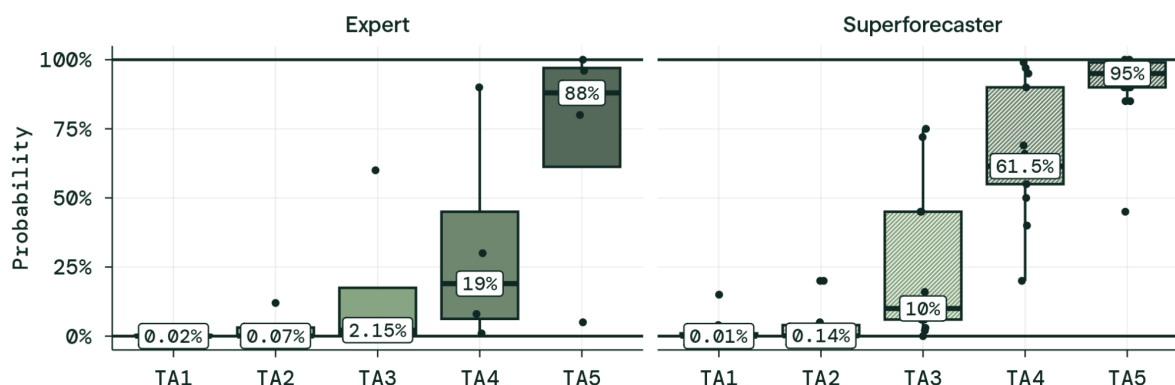


Figure C.16: Follow-up survey results. Probability that a randomly selected actor in each actor class is able to develop a data-damaging worm conditional on Capability 1 given 3 months of effort.

Follow-up study: Actor capability to write elite exploits

Values represent the probability of a randomly selected actor in each actor class having the capability to write elite exploits with 3 months of effort.

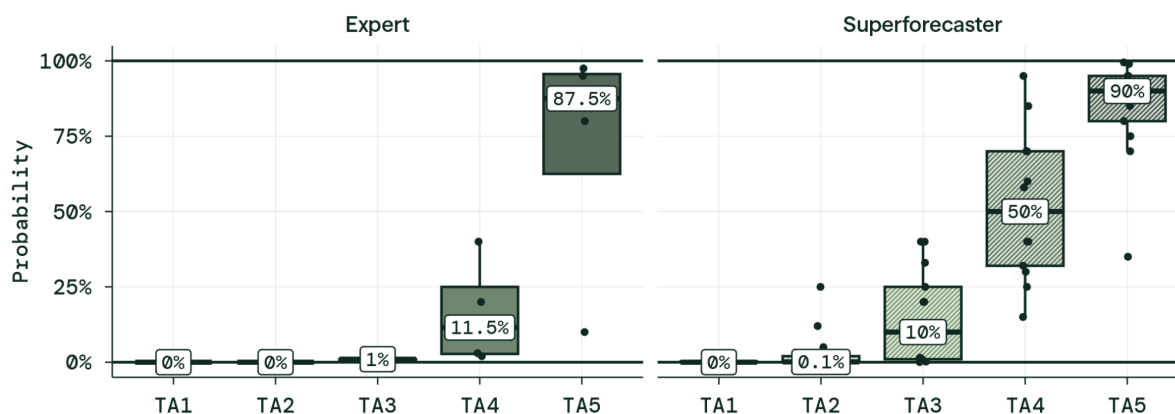


Figure C.17: Follow-up survey results. Probability that a randomly selected actor in each actor class is able to write elite exploits given 3 months of effort.

Follow-up study: Actor capability to write elite exploits, conditional on Capability 1

Values represent the probability of a randomly selected actor in each actor class having the capability to write elite exploits with 3 months of effort. Forecasts are conditional on Capability 1: AI enables 25% of moderately-skilled individual hackers to develop elite exploits.

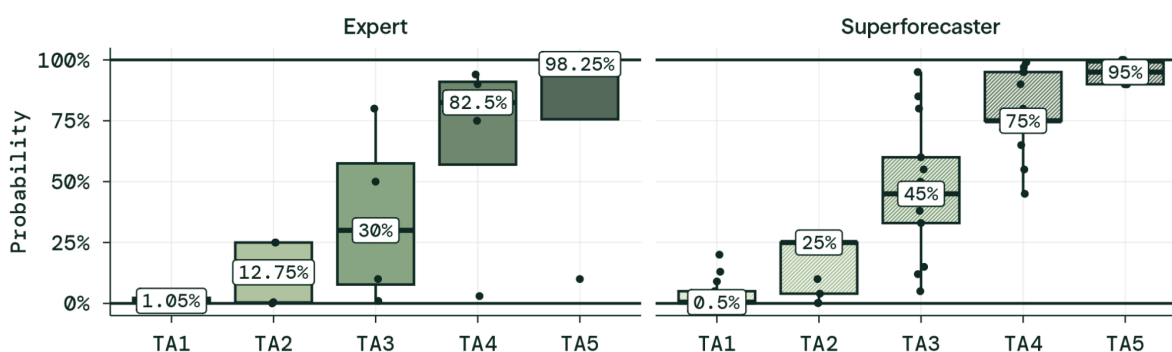


Figure C.18: Follow-up survey results. Probability that a randomly selected actor in each actor class is able to write elite exploits conditional on Capability 1 given 3 months of effort.

Follow-up study: Actor capability to do other steps (other than write elite exploits) of developing a data-damaging worm

Values represent the probability of a randomly selected actor in each actor class having the capability to perform the other steps (other than write elite exploits) required to develop and deploy a large-scale data-damaging worm with 3 months of effort.

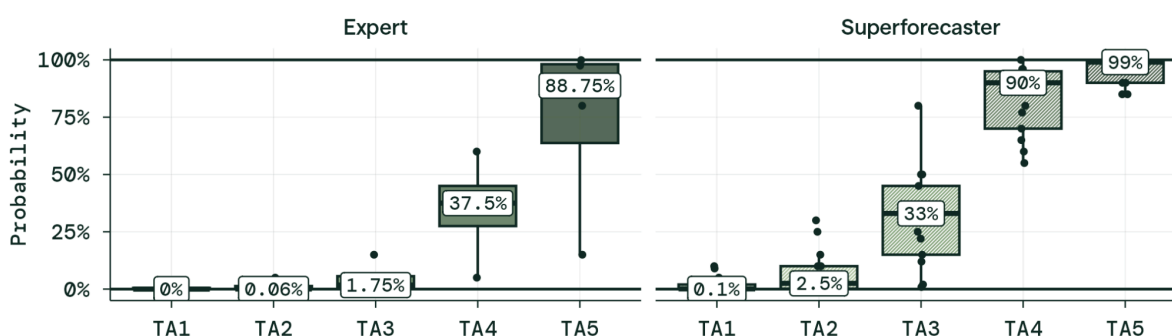


Figure C.19: Follow-up survey results. Probability that a randomly selected actor in each actor class is able to do the other steps required to develop and deploy a data-damaging worm given 3 months of effort.



Follow-up study: Actor capability to do other steps (other than write elite exploits) of developing a data-damaging worm, conditional on Capability 1

Values represent the probability of a randomly selected actor in each actor class having the capability to perform the other steps (other than write elite exploits) required to develop and deploy a large-scale data-damaging worm with 3 months of effort. Forecasts are conditional on Capability 1: AI enables 25% of moderately-skilled individual hackers to develop elite exploits.

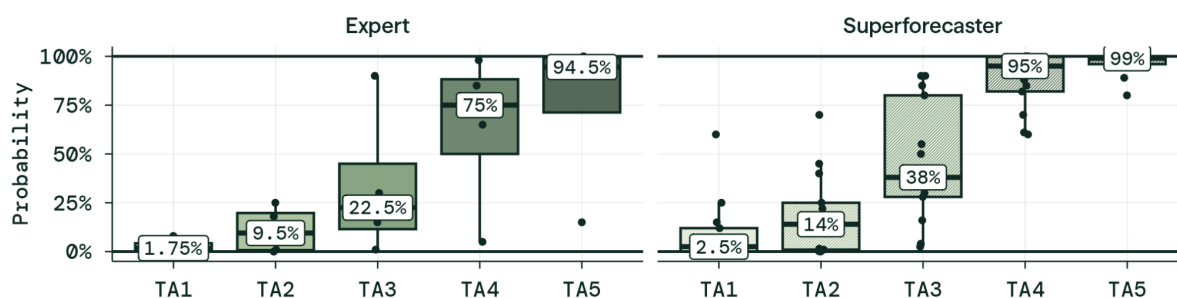


Figure C.20: Follow-up survey results. Probability that a randomly selected actor in each actor class is able to do the other steps required to develop and deploy a data-damaging worm conditional on Capability 1 given 3 months of effort.

Follow-up survey (12-month horizon)

In the 12-month follow-up results, the median expert estimate for TA4 end-to-end capability remains low (1%) despite higher estimates for some component capabilities (elite exploits and other tasks). Given the small expert follow-up sample (only 4 participants), we expect medians to be relatively unstable, so these results are not surprising.

Actor end-to-end capability to develop data-damaging worms

Values represent the probability of a randomly selected actor in each actor class having the capability to perform all the steps required to develop and deploy a large-scale data-damaging worm with 12 months of effort.

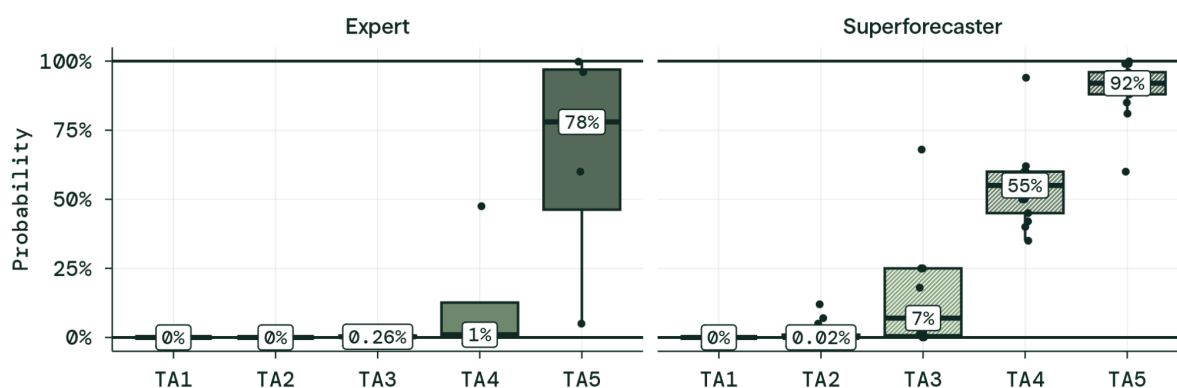


Figure C.21: Follow-up survey results. Probability that a randomly selected actor in each actor class is able to develop and deploy a data-damaging worm given 12 months of effort.



Actor end-to-end capability to develop data-damaging worms, conditional on Capability 1

Values represent the probability of a randomly selected actor in each actor class having the capability to perform all the steps required to develop and deploy a large-scale data-damaging worm with 12 months of effort. Forecasts are conditional on Capability 1: AI enables 25% of moderately-skilled individual hackers to develop elite exploits.

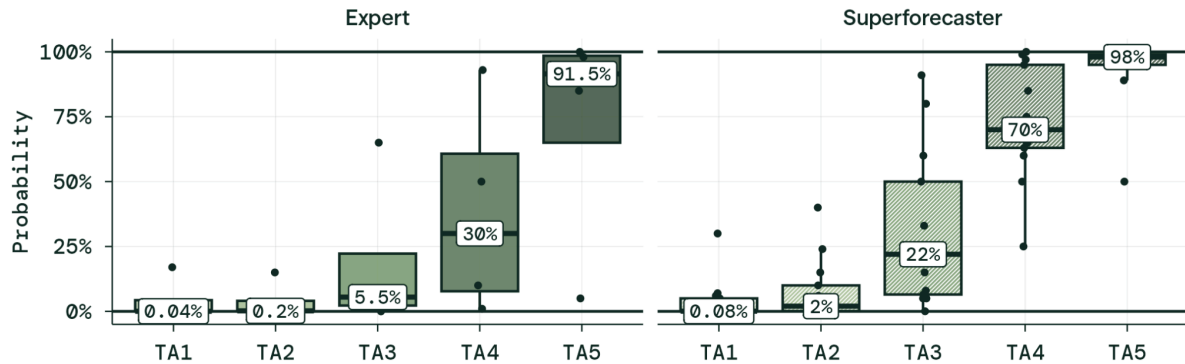


Figure C.22: Follow-up survey results. Probability that a randomly selected actor in each actor class is able to develop and deploy a data-damaging worm conditional on Capability 1 given 12 months of effort.

Actor capability to write elite exploits

Values represent the probability of a randomly selected actor in each actor class having the capability to write elite exploits with 12 months of effort.

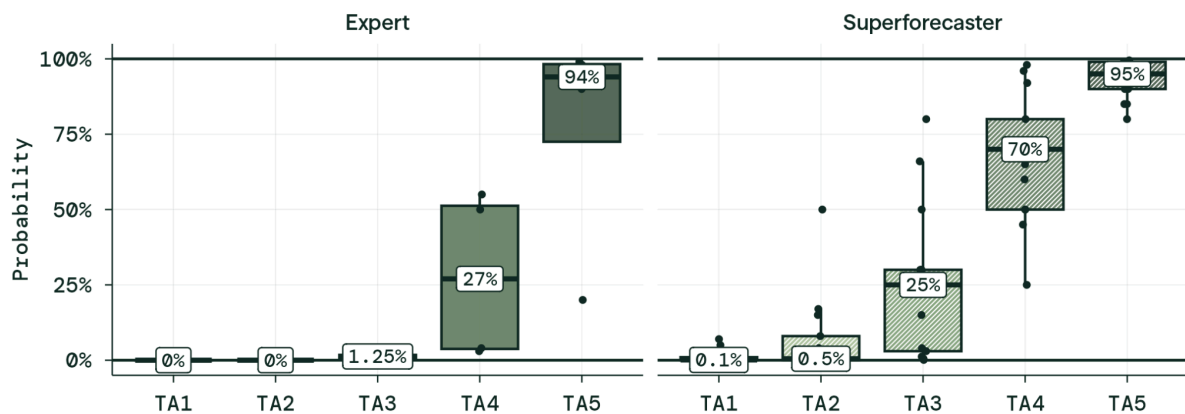


Figure C.23: Follow-up survey results. Probability that a randomly selected actor in each actor class is able to write elite exploits given 12 months of effort.



Actor capability to write elite exploits, conditional on Capability 1

Values represent the probability of a randomly selected actor in each actor class having the capability to write elite exploits with 12 months of effort. Forecasts are conditional on Capability 1: AI enables 25% of moderately-skilled individual hackers to develop elite exploits.

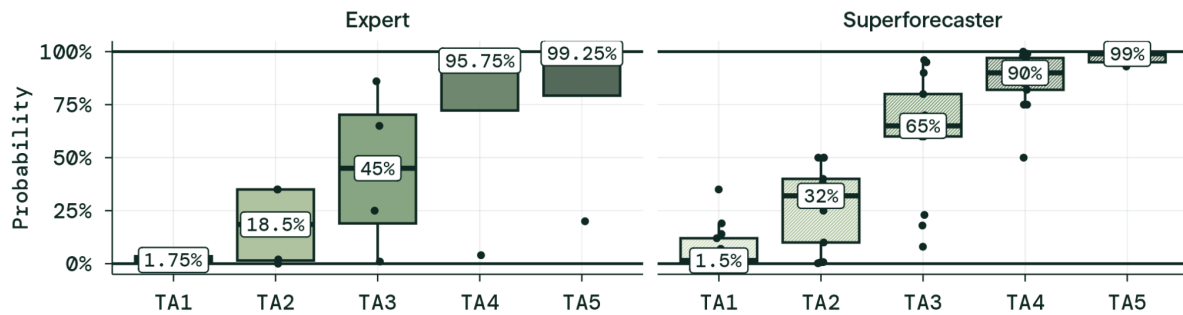


Figure C.24: Follow-up survey results. Probability that a randomly selected actor in each actor class is able to write elite exploits conditional on Capability 1 given 12 months of effort.

Actor capability to do other steps (other than write elite exploits) of developing a data-damaging worm

Values represent the probability of a randomly selected actor in each actor class having the capability to perform the other steps (other than write elite exploits) required to develop and deploy a large-scale data-damaging worm with 12 months of effort.

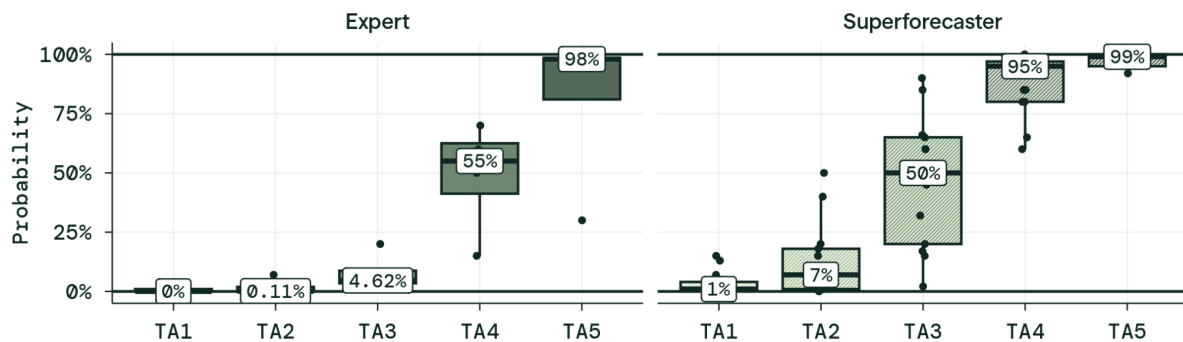


Figure C.25: Follow-up survey results. Probability that a randomly selected actor in each actor class is able to do the other steps required to develop and deploy a data-damaging worm given 12 months of effort.

Actor capability to do other steps (other than write elite exploits) of developing a data-damaging worm, conditional on Capability 1

Values represent the probability of a randomly selected actor in each actor class having the capability to perform the other steps (other than write elite exploits) required to develop and deploy a large-scale data-damaging worm with 12 months of effort. Forecasts are conditional on Capability 1: AI enables 25% of moderately-skilled individual hackers to develop elite exploits.

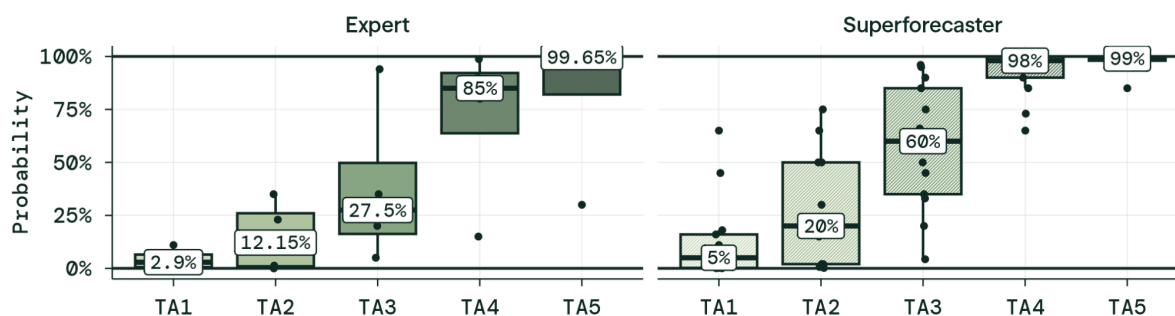


Figure C.26: Follow-up survey results. Probability that a randomly selected actor in each actor class is able to do the other steps required to develop and deploy a data-damaging worm conditional on Capability 1 given 12 months of effort.

Impact of time-horizon increase

Participants broadly agreed that additional time increases the probability of actors successfully completing the relevant tasks across all threat actor tiers. However, they emphasized diminishing returns beyond three months, as well as offsetting factors such as more time for defenders to patch known vulnerabilities. The benefits of additional time were seen as most impactful for low- to mid-sophistication actors, whose end-to-end capabilities for developing and deploying large-scale data-damaging worms increased twofold (see Figure C.27). Lower-tier actors (TA1 and TA2) were viewed as benefiting from increased experimentation time but still generally constrained by technical capability and coordination limits. For mid-tier actors (TA3 and even TA4), participants mentioned the extended time window could be crucial for refining exploits and building out necessary infrastructure. On the contrary, participants generally believed that top-tier actor capabilities (TA5) were largely insensitive to time, given they already have the ability to execute attacks in shorter periods of time if they were motivated to do so.



Ratio of probability of actor having End-to-End capabilities given 12 vs. 3 months.

Values are unconditional.

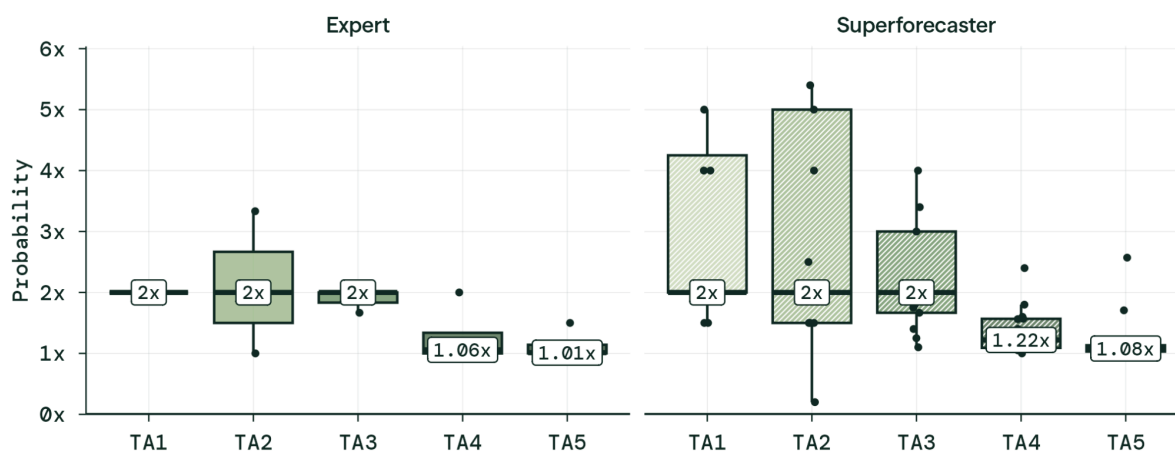


Figure C.27: Relative increase in probability that a randomly selected actor in each actor class is able to develop and launch a data-damaging worm attack (end-to-end) with 12 months of effort, compared to 3 months.

Under the scenario where AI achieved Capability 1, overall capability estimates remained higher for all threat actor levels, as in the original survey, but the marginal benefit of extending the timelines diminishes further. This is consistent with participant views that AI shortens the duration of development cycles and shifts the constraints from time to coordination and execution. Willingness was not seen as strongly-time dependent, and in some cases was expected to decrease over longer-time horizons due to sustained costs and risks.



Actor willingness to launch a $\geq \$100\text{B}$ grid cyberattack

Actor willingness to launch a $\geq \$100\text{B}$ cyberattack against the U.S. grid

Values represent the probability that at least one actor in each actor class would spend at least 6 months actively attempting to launch a $\geq \$100\text{B}$ cyberattack against the U.S. grid if they were capable.

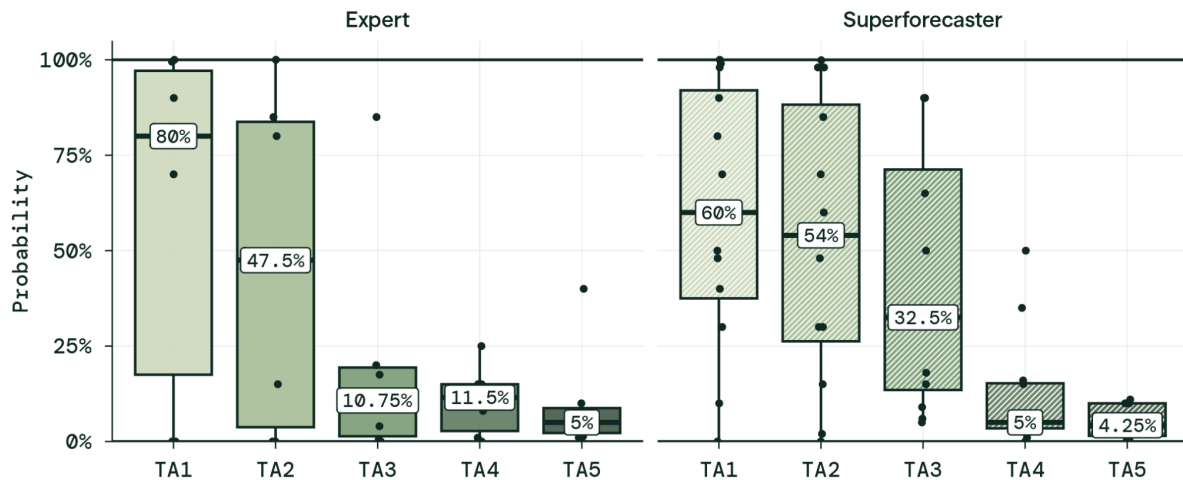


Figure C.28: Probability that at least one actor in the threat class would spend 6 or more months actively attempting to launch a cyberattack against the U.S. electrical grid, causing a $\geq \$100$ billion blackout in 2026, assuming they had the capability to do so.

Actor willingness to launch data-damaging worm attacks

Values represent the probability that at least one actor in each actor class would spend at least 12 months actively attempting to launch a data-damaging worm attack if they were capable.

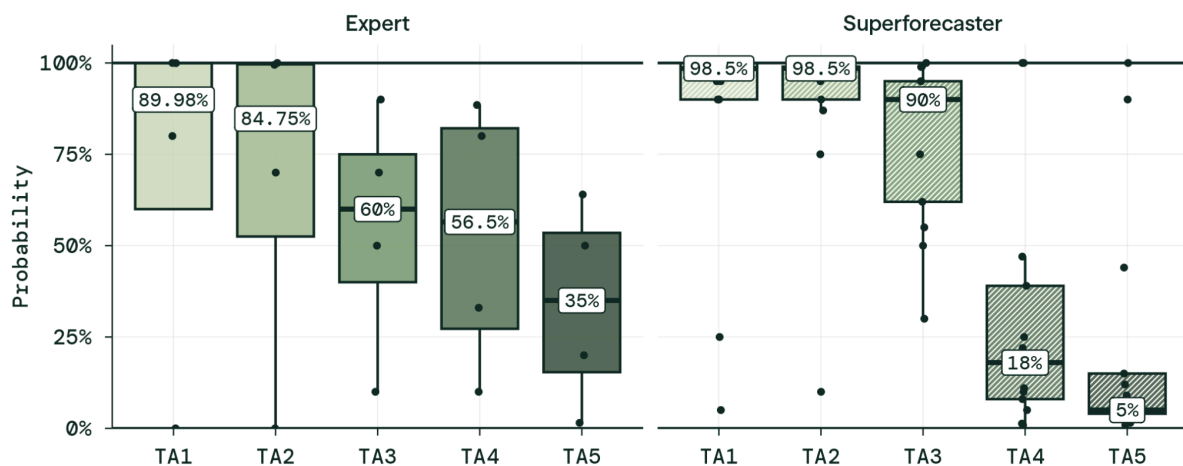


Figure C.29: Probability that at least one actor in each threat actor class would spend 12 or more months actively attempting to launch a data-damaging worm in 2026, assuming the actor had the capability to do so.