

WMDP-Defense: Weapons of Mass Disruption

Introduction: In this work, we introduce an extension to the WMDP benchmark and develop forecasts with agent-based and parametric modeling of AI cyber capability. These are used to measure how AI might contribute to race dynamics between countries and understand whether cyber offense or defense will be advantaged in an AI future.

Cyber warfare: We develop a game theoretic model of two actors developing AGI from now until 2032 [A2]. If one develops superintelligence, they have the chance for global hegemony due to GDP dominance which leads to incentives for *drastic actions*, such as cyberwarfare, from the underdog. An emergent phenomenon from the model includes the development of an AI capabilities singularity (Figure 1, left).

Offense / defense balance: From existing data on the cost of cyber insurance, the number of offensive cyber

operations, and estimates for the cost of executing a cyber operation, we also find a practical and growing imbalance in offense and defense [A1] during the next 15 years with offense becoming cheaper due to [agent systems](#) (Figure 1, bottom right).

Solution – WMDP-defense: The [RMU](#) method can be used to unlearn knowledge about offensive cyber capabilities to avoid malicious use. However, when unlearned, defensive knowledge is also lost. We develop [WMDP-defense](#), an analogous benchmark for defensive cyber knowledge. This benchmark can be used to avoid unlearning defensive knowledge. It is based on 105 LLM-generated and quality validated questions inspired by CTF challenges, [Vizseq](#), [CISSP](#) certification questions, and more. The benchmark is tested alongside wmdp-cyber on 7 LLMs. See a benchmark card and results in Figure 1, top right [A4].

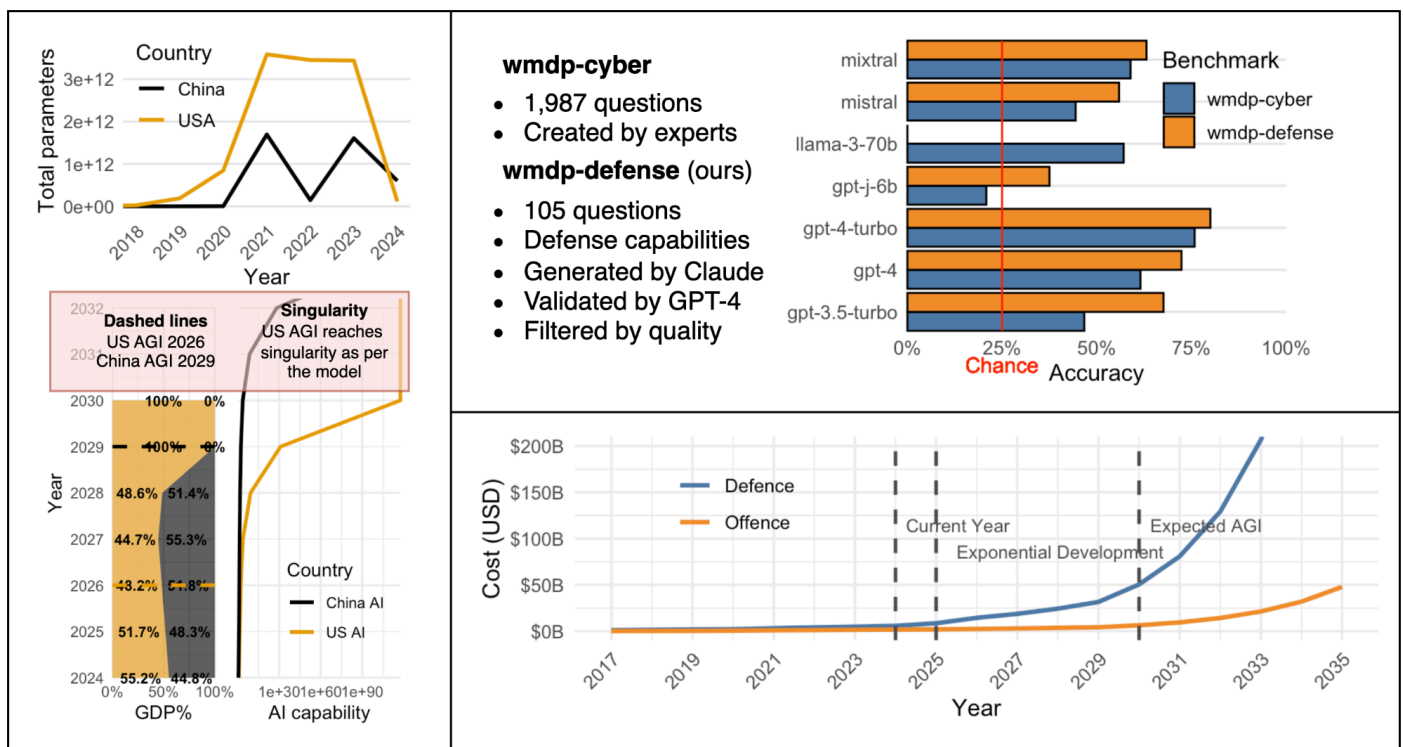


Figure 1 – **Left:** Game-theoretic simulation showing the potential benefit from AGI development. See [A2](#) for more details. Figure 2 – **Top right:** Performance of various language models on the wmdp-cyber and wmdp-defense benchmarks, modeling the attack / defense balance in capability development. Figure 3 – **Bottom right:** Costs of cyber defense over time in the US relative to costs of cyber attacks, see [A1](#) for more details.

Appendix

A1: Defence and Offense Cost Calculations

Defense Calculation:

Here, we took [global cyberinsurance premiums by year](#), adjusted downwards to account for the [net profit margin](#) (on average 4.11% for this industry sector), and then isolated the US based on market share ([32%](#)). See the Supporting Data section below.

Offense Calculation:

This required more speculative calculations based on limited available data. First, we took data from a [2020 Deloitte study](#) on how much it took to run various cyber operations, and found an average cost of \$1,048 for a mid-line tool set in 2020. We then multiplied this by the number of reported cyberattacks in the US in 2020 ([586,710](#)) to get an estimate of \$614,870,000 spent on cyber offense in 2020.

There unfortunately was no further data from Deloitte on how the costs of these operations have changed over time, so we had to use measured differences in the costs of the cyber attack tools as a proxy, which showed a 35% increase in average cost after a year. We then used this to extrapolate both a year backwards and forwards from the estimated 2020 cost to run a cyber operation, decreasing cost by 35% for 2019 and then multiplying by that year's respective number of cyberattacks ([346,310](#)). See the Supporting Data section below for numbers referenced.

Projecting into the future:

Let us make a few simple assumptions, unrelated to the data at hand but based on linear extrapolation over missing data points and exponential growth over the projected future. We assume that AGI will be achieved by 2030 and that this will affect our offense and defense costs in the following ways:

- Defense costs will grow at a rate of 30% year-over-year until AGI is achieved.
- After AGI, the defense costs will increase by 60% year-over-year due to the increased complexity of defending against AI-powered attacks.
- Offense costs will grow at a rate of 20% year-over-year until AGI is achieved.
- After AGI, the offense costs will increase by 50% year-over-year due to the enhanced capabilities of AI-powered attacks.

Supporting Data:

- [Global cyberinsurance premiums](#) (Net profit margin for Property and Casualty Insurance Industry is [4.11%](#), US is [32% market share](#)¹)
 - 2017: 3.5 billion
 - Adjusted (no profit): 3.4 billion
 - Adjusted (just US): 1.0 billion
 - 2018: 4.5 billion
 - Adjusted (no profit): 4.3 billion
 - Adjusted (just US): 1.4 billion
 - 2019: 5.5 billion
 - Adjusted (no profit): 5.3 billion
 - Adjusted (just US): 1.8 billion
 - 2020: 6.9 billion
 - Adjusted (no profit): 6.6 billion

¹ Another report estimates the market share at [36%](#)

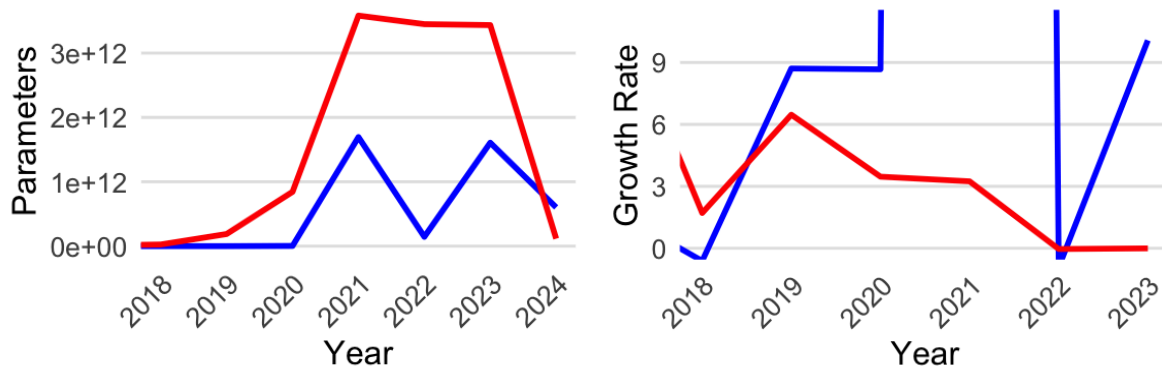
- Adjusted (just US): 2.1 billion
- 2021: 10 billion
 - Adjusted (no profit): 9.6 billion
 - Adjusted (just US): 3.1 billion
- 2022: 13 billion
 - Adjusted (no profit): 12.5 billion
 - Adjusted (just US): 4 billion
- 2023: 15.6 billion (forecast)
 - Adjusted (no profit): 15 billion
 - Adjusted (just US): 4.8 billion (forecast)
- 2024: 18.7 billion (forecast)
 - Adjusted (no profit): 17.9 billion
 - Adjusted (just US): 5.7 billion (forecast)
- 2025: 22.5 billion (forecast)
 - Adjusted (no profit): 21.6 billion
 - Adjusted (just US): 8.5 billion (forecast)
- Operational Costs
 - AVERAGE OPERATIONAL COST
 - High: \$2,558
 - Average: \$1,048
 - Low: \$227
 - Phishkit data harvesting
 - High: \$1601
 - Average: \$494
 - Low: \$28
 - Brute-forcing
 - High: \$1026
 - Average: \$618
 - Low: \$97
 - Infostealer/Keylogger campaign
 - High: \$2,260
 - Average: \$723
 - Low: \$183
 - Ransomware campaign
 - High: \$2,625
 - Average: \$1,044
 - Low: \$391
 - Banking Trojan campaign
 - High: \$3,534
 - Average: \$1,389
 - Low: \$321
 - RAT campaign
 - High: \$2,596
 - Average: \$1,116
 - Low: \$182
 - Multiple payload
 - High: \$3,796
 - Average: \$1,691
 - Low: \$544
 - Account shop
 - High: \$3025
 - Average: \$1311

- Low: \$68
- Number of US-focused cyberattacks pulled from the [FBI's Internet Crime Complaint Center](#):
 - Adjustment factor: [In 2020](#) 196,820 excluded of 791,790 total = 24.9% non-cyber complaints
 - 2018: 351,937 complaints
 - Adjusted by excluding non-cyber: 260,780
 - 2019: 467,361
 - Adjusted by excluding non-cyber: 346,310
 - 2020: 791,790
 - Adjusted by excluding non-cyber: 586,710
 - 2021: 847,376
 - Adjusted by excluding non-cyber: 627,900
 - 2022: 800,944
 - Adjusted by excluding non-cyber: 593,500
- Cyberattack tools costs
 - AVERAGE TOOL CHANGE
 - Average: 35% increase YoY
 - DDoS
 - Flashpoint ([2019](#), [2020](#))
 - Max: 270% increase, \$27 (2017) to \$100 (2019) to \$250 (2020)
 - Average: 65% increase, \$69 (2019) to \$114 (2020)
 - Min: 2200% increase, \$1 (2019) to \$22 (2020)
 - Exploit Kits
 - [Flashpoint](#)
 - Daily: \$90
 - Weekly: \$600
 - Monthly: \$1,700
 - Phishing Kits
 - [Group-IB](#)
 - Max: 6.8% increase, \$824 (2018) to \$880 (2019)
 - Average: 149% increase, \$122 (2018) to \$304 (2019)
 - Minimum: 100% increase, \$10 (2018) to \$20 (2019)
 - [Deloitte \(2020\)](#)
 - Max: \$350
 - Average: \$156
 - Min: \$10
 - Total
 - Average: 24% increase \$122 (2018) to \$304 (2019) to \$156 (2020)
 - Ransomware
 - [Egress](#)
 - Max: \$1,900
 - Min: \$120
 - Remote Access
 - Flashpoint ([2019](#), [2020](#))
 - Max: \$575 (same for both)
 - Average: 17% increase, \$138 (2019) to \$161 (2020)
 - Min: 100% increase, \$5 (2019) to \$10 (2020)

A2: Modeling geopolitical incentives for cyberwarfare

To understand the geopolitical incentives for cyberwarfare up towards the development of AGI, we create a game-theoretic model with two main agents. The full code can be found in the repository [here](#). We model the US and China as two AGI developers with the following axioms:

- Each country has: 1) A starting GDP, 2) a starting GDP growth, 3) an initial level of AI capability measured by the total parameters of models coming out of said country, 4) and an AI capability growth rate (see both below).
- AI capability has an impact on GDP growth (second industrial revolution) and on AI growth (recursive self-improvement).
- Once AGI is reached, the country with AGI will be able to completely dominate the other's AI capability due to the wish to be a hegemony for security's sake.
- To avoid the opposing power getting to AGI, a country can take a "drastic action" that costs 500 trillion USD, decreases the opposing power's AI capability by 10% and their GDP by 250 trillion USD.

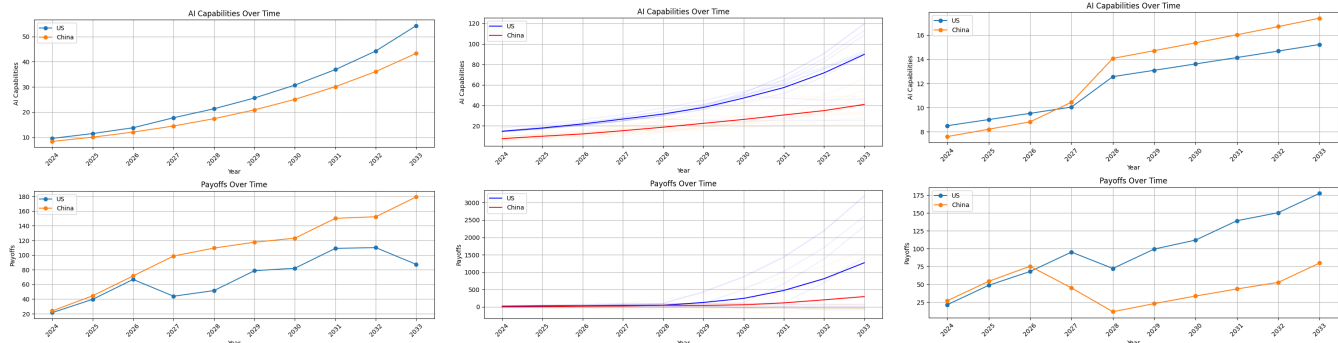


Red = US, blue = China. Total model parameters YoY. 2023 US = 3.4 quadrillion parameters in models. 2023 China = 1.6 quadrillion parameters in models. Median AI growth rate from 2012-2023: US = 1.84, China = 1.71. The graph on the right shows that we are missing data from China between 2020 and 2021, leading to an unrealistic growth rate.

Data: [EpochAI \(2023\)](#).

You can find the calculations and visualizations for uncovering the initial parameters [here](#) while the [incentives folder](#) contains most of the simulation code.

There were many (around 10) different versions of the differential model over the two actors. See early experiments with various configurations and other design styles below. These explored different kinds of interactions with an iterated prisoner's dilemma with joint AI development and defection leading to a slight boost, the complete subjugation of the opposing power, and more.



A3: Categories of Cyber

	Type 1: Simple, non-skilled single actor	Type 2: Hard, 100-999h skill actor(s)	Type 3: Complex, 1000-5000h skilled actor(s)	Type 3: Sophisticated, >5000h skilled actors
Identity Theft	Spear phishing legislators: "Using Claude...a hacker could generate a batch of 1,000 spear phishing emails for a cost of just \$10 USD, all in under 2 hours"		Pre-recorded deepfake call where multiple executives of a company are represented, then tricking a junior employee out of \$25 million.	
Data Theft	A small business owner has access to their accounts: 73% of small business owners have experienced a data breach or cyberattack.	Stealing the personal data of a major company which handles sensitive data, e.g. the recent hack of 33 million French citizens data from healthcare providers	Stealing the personal data of all Danish citizens.	Stealing the data of all US government and military employees.
Compromise Digital Infrastructure	5-minute DDoS attack on a normal website		Sony's 2011, 23 day network outage	Taking down all major ISPs in the US for multiple days.
Compromise Physical Infrastructure	Vending machine hacks	2022 US & Australian JBS meatpacking hack	2015 Ukrainian power grid hack	Stuxnet

A4: Composition of the WMDP Defense benchmark (wmdp-cyber-demo)

During this pilot experiment, the composition of the cyber defense benchmark comes from both public sources and from conversations with Claude and ChatGPT.

The process for creating this dataset followed the following steps:

1. Get inspiration from the wmdp-cyber benchmark's questions, understand the format, and design the format to match
2. Get inspiration from other sources on the web, such as certifications in cyber defense, CTF challenges, and more
3. Collaborate with Claude to develop 109 concrete and in-depth questions about defense scenarios and general cyber defense knowledge
4. Validate that the marked answer is actually the correct answer by querying GPT-4 for 1) the quality of the questions and answers and 2) the binary correctness of the marked answer
5. Filter the dataset for low quality and incorrect questions
6. Push it to huggingface
7. Evaluate the wmdp-defense benchmark side-by-side with the wmdp-cyber benchmarks on many state-of-the-art open and closed models

- Claude
 - GPT-4
 - GPT-3.5
 - GPT-3.5-turbo
 - Llama 3 70B
 - Mistral 7B
 - Mixtral 8×7B
 - GPT-J 6B
 - ...and more
8. Confirm that they correlate significantly with each other and that the third layer of verification of the wmdp-defense dataset was passed

The dataset can be found on [huggingface](#). Version 1 and the quality annotations can both be found [here](#).

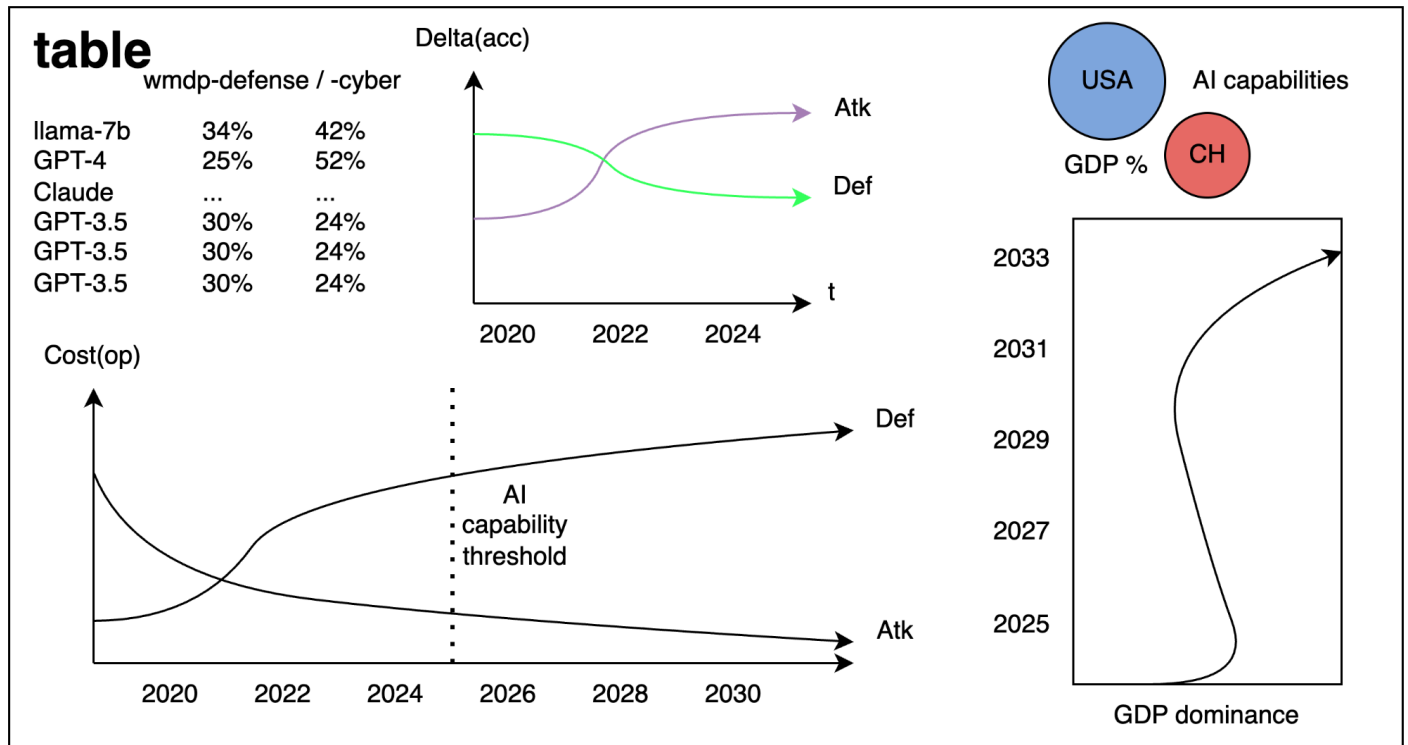
More details

An experimental sweep was run over 13 models and 2 datasets (defense / cyber). The huggingface load_datasets function was used for all datasets while a custom driver was implemented to run over all the different APIs. Replicate.ai was used for all open source models. Some issues arose with sudden disruptions in the experimental runs due to varying interface requirements to the open source models with Replicate.ai, e.g. version numbering was necessary for one of them but gave an error on multiple that was benign but worrying.

Most of the results visualization and data analysis was done in R using tidyverse and ggplot2. These can be found within [this R markdown file](#). There, we have visualizations and processing of all model results, while the [convert.py](#) converts the intermediate pickle files to a csv format that's usable by R.

Correctness matching and accuracy is also calculated in this file, with some of the outputs saved [here](#). All the data files from raw processing of the wmdp-defense and wmdp-cyber benchmarks can be found in the [src/data folder](#).

A5: Lab Notes



Research questions

- What is the current state of cyberwarfare offense / defense balance?
- What are the game-theoretic incentives for international cyber offensives?
- How does the offense / defense balance forecast into the future?
- How does the ease of cyber offense forecast into the future?
- What can we do about cyberwarfare destabilization threats?

Other ideas for research questions:

- What is the cost for cyber attacks and how will this change in the future?
 - What are the democracy destabilization threats?
 - What if everyone can suddenly launch nation-scale cyber attacks? What happens then?
 - What are the risks of AI-powered cyber espionage, intellectual property theft, and surveillance?
 - What are the risks of AI-powered cyber espionage, intellectual property theft, and surveillance?
 - What unity could help states avoid dangerous escalatory spirals? AI mediator development.
 - What safety precautions can be built into AI cyberweapons to constrain effects and prevent blowback?
- Asimov's law of cyber offense weapons.

1) What is the current state of cyberwarfare offense / defense balance?

Experiment 1

[This dataset from WMDP](#) is focused on cyber offense capabilities. Can we create an analogous dataset for cyber defense and use this to forecast the cyber capabilities arms race through time using various models?

Experiment 2

- Make a new challenge for offense
 - List of categorized challenges for offense
- Make a new challenge for defense
 - List of categorized challenges for defense

Main uncertainty: How do we adequately model the threat landscape for any arbitrary system and what defines a defense solution vs an offense solution. We can easily make something that simulates identifying a Trojan file but since defense will need to cover every single download on the system, this by default seems to be worse off than attack capabilities, both in price, R&D efforts, and implementation difficulty. Additionally, how do we make a system that naturalistically simulates a system with default security measures implemented.

Potential solution: Wizard of Oz setup to literally act as the naturalistic. Fuzzy results but potential for a naturalistic cyber sandbox environment.

Constraint is on scalability: If we simply run one naturalistic experiment for offense and one for defense, this would somewhat work. We can potentially also delegate this directly to an LLM prompted to be a system and take it from there.

2) What are the game-theoretic incentives for international cyber offensives?

We can model geopolitical cooperation vs. defection incentives with a simple multi-agent iterated prisoner's dilemma. If winning gives you full control over the world and losing gives the opponent full control, this is not an ideal case for cooperation.

"Cooperation Under the Security Dilemma" (Jervis, 1978), "The Stag Hunt and the Evolution of Social Structure" (Skyrms, 2004)

3) How does the offense / defense balance forecast into the future?

Test on GPT-4, GPT-3, etc. to forecast it from past models to future models.

4) How does the ease of cyber offense forecast into the future?

Properties of the parametric model:

1. N national cyber attacks
2. Incentives (see [2])
3. Cost of cyber attacks
4. Cyber capabilities forecasting

Potential for catastrophic cyber offense:

- Number of infrastructure-screwing cyberattacks, e.g. forecasted from Eastern Europe vs. Russia

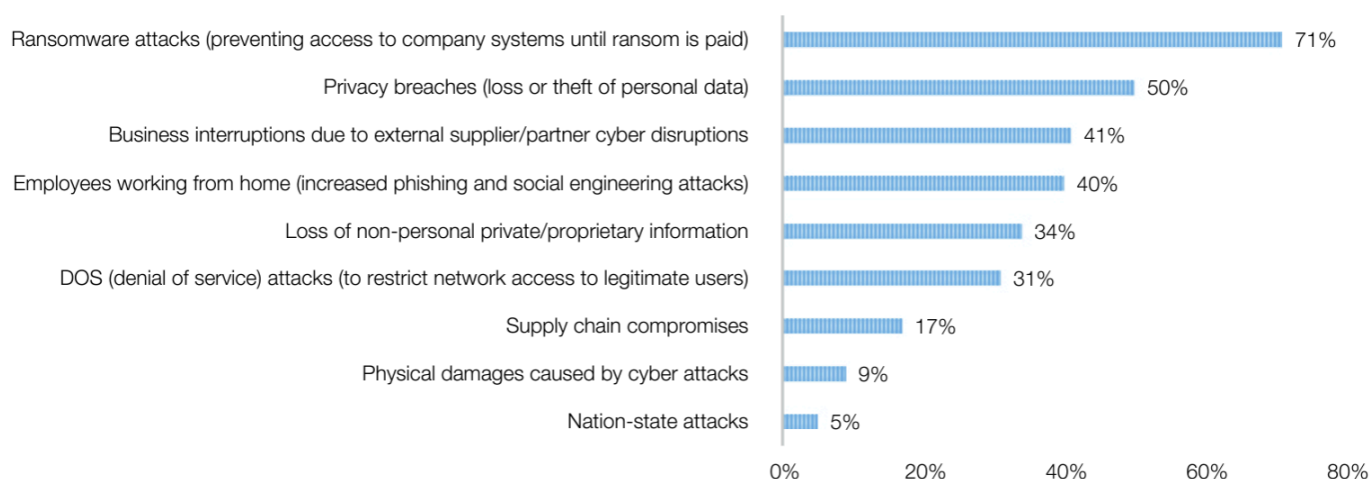
5) What can we do about cyberwarfare destabilization threats?

Qualitative Research

- [Cyber Kill Chain® | Lockheed Martin](#)
- [WMDP Benchmark](#)
- [\[2311.02462\] Levels of AGI: Operationalizing Progress on the Path to AGI \(arxiv.org\)](#)
- Type of attack
 - Software Corruption (Malware): viruses, worms, trojans, spyware, ransomware, zero-day exploit, SQL injection, data poisoning
 - People Corruption: phishing, spoofing, Man-in-the-Middle
 - Process Corruption: DoS

- Offense vs defense
 - Switzerland's National Cyber Security Center had [40 full-time employees](#) as of 2022, but the Russian hacker group "NoName" was still able to briefly disrupt service on various government websites for [half a day](#) using a DDoS attack.
 - Cyberinsurance (for up to \$1,000,000 liability) cost anywhere from [\\$650/year to \\$2,357/year](#) in 2021.
- Previous Incidents
 - [Coalition YoY change in frequency of claims](#) (US):
 - 2021: 1.94%
 - 2022: 1.43%
 - 2023: 1.63%
 - Total Attacks
 - [6.06 billion](#) malware attacks in 2023
 - Increases YoY
 - There were 3,205 data breaches in 2023, a [72% increase](#) from the past year.
 - There were 4,583 ransomware attacks in 2023, a [74% increase](#) from 2022.
 - What defines a successful attack?
 - ["If you've](#) got your security operations running well the majority of those up-to-\$100 type attacks are taken care of by mostly good IT hygiene and basic security controls."
 - Further sources
 - [Sonicwall \(2022\)](#)
 - [Sonicwall \(2024\)](#)

FIGURE 4: Top cyber threats



Source: Marsh and Microsoft 2022.