The Weak-Link Asymmetry: AI's Peril and Promise

Charles I. Jones* Stanford GSB and Anthropic

Christopher Tonetti Stanford GSB and NBER

November 6, 2026 — Version 0.5

Abstract

In a "weak links" model of production, where tasks are complements, the gains from automation tend to arrive slowly because output is constrained by the least productive tasks. But the same complementarity makes attacks more harmful and easier to execute. For example, with an elasticity of substitution across tasks of 0.2, automating tasks worth 1 percent of GDP with a tenfold increase in productivity raises output by less than one percent, while cutting their productivity tenfold lowers output by 68 percent on impact. When a larger AI sector simultaneously makes AI models more capable and attacks more likely, the positive productivity spillover is negligible compared with the potential damage, so the market overinvests in AI. In a weak-link economy, the costs and benefits from AI can be extremely asymmetric.

*We are grateful to Anton Korinek, Peter McCrory, Phil Trammell, and the participants of a brown bag seminar at The Anthropic Institute for helpful discussions. We used Claude as a research and writing assistant. The authors are releasing this paper in their personal capacity. The views expressed are their own and do not necessarily reflect those of Anthropic. The paper has not been reviewed or endorsed by Anthropic as an institutional position.

1. Introduction

Recent research emphasizes that the gains from artificial intelligence (AI) may arrive relatively slowly because tasks in production are complements: output is constrained by bottlenecks and weak links (Aghion et al., 2019; Jones and Tonetti, 2026). From a safety standpoint, this could be good news. If job losses and catastrophic risks materialize gradually over decades, society may use that time to mitigate the downsides.

However, the same weak links that delay the gains make the losses large and immediate. Since a chain is only as strong as its weakest link, essentially all of the links have to be strengthened to get large benefits. But cutting one link in a chain makes the chain useless. Kremer (1993) makes this point with the space shuttle *Challenger*, which exploded because an inexpensive rubber gasket failed.

The paper's central exercise uses a standard CES production function in which tasks are strong complements, with an elasticity of substitution of 0.2 as in Jones and Tonetti (2026). Consider changing the productivity of a group of tasks that account for 1 percent of GDP. Multiplying that productivity by a factor of ten (or infinity) raises GDP by just one quarter of one percent. Intuitively, production is constrained by all the other weak links. Conversely, the same change on the downside is much more damaging. Reducing productivity by a factor of ten on tasks with a 1 percent cost share lowers GDP by 68 percent. The two shocks are the same size in logs, but the responses differ by two orders of magnitude. Gains are bounded and small. Losses are unbounded and large. We call this the weak-link asymmetry.

The consequences for AI could be profound. In the simulations in Jones and Tonetti (2026), AI automating idea production leads growth to explode, but it takes decades for this explosion to materialize. Raising GDP growth substantially requires AI to improve weak links throughout the economy. Many physical world tasks will require advanced robotics. Regulatory and legal barriers, the scarcity of compute, and the high share of sectors like services and the government (Baumol's cost disease) are all examples of weak links that delay the benefits from AI.

In contrast, frontier AI models are already superhuman at cyberattacks, as illustrated by Anthropic's Mythos model and OpenAI's Hugging Face episode (UK AI Security Institute, 2026; Anthropic, 2026; METR, 2026). AI-enhanced attacks on the financial system, the electric power grid, or critical internet infrastructure could cause substantial economic disruption in the next few years. Frontier labs warn that catastrophic biorisks may not be far behind (Anthropic, 2025).

Cyberattacks and bioattacks require much narrower capabilities that will be achieved well before we have AI that can manage a Fortune 500 company or robots that can teach elementary school. It is easier to improve short chains than long ones. Before AI can automate large parts of the economy, it will be able to target a few critical links, either on behalf of a bad actor or perhaps even on its own if misaligned.

Examples of attacks on weak links. Once guardrails have been bypassed, AI can make a severe attack more likely in three ways. First, AI can identify the weak links that are easiest to attack and whose failure would do the most harm. Second, AI can aid in the planning and organization of attacks and advise on keeping an attack secret during this phase. Finally, AI may help carry out an attack.

The electric grid. High-voltage transformers, typically weighing between 150 and 400 tons, are less than 3 percent of the transformers in U.S. substations but carry 60 to 70 percent of the nation's electricity. The National Research Council concluded that a carefully planned attack on this weak link "could deny large regions of the country access to bulk system power for weeks or even months" and that "some of the lost equipment may take months or even years to replace."[1] In addition to helping an adversary plan and execute a physical attack on hardware, AI can enable cyberattacks that directly destroy physical equipment. The Stuxnet worm destroyed centrifuges at Iran's uranium enrichment plant by taking over the computers that ran them (Langner, 2011).

Critical internet infrastructure. The internet runs on software that a few people maintain and on hardware that could be susceptible to attack. xz Utils, a compression library in nearly every distribution of Linux, has a single volunteer maintainer. In 2024, an engineer stumbled upon a backdoor that an attacker had spent two years working to insert (Freund, 2024; Cox, 2024). AI makes the search for such openings cheap. Mythos Preview found more than ten thousand vulnerabilities of high or critical severity in "the most systemically important software" (Anthropic, 2026). Mythos Preview was used by Anthropic and about 50 partners to find vulnerabilities so they can be fixed. Fixing is the slower step. Each vulnerability must be verified, disclosed, and patched, and until the patch is deployed, an attacker can exploit it (Anthropic, 2026). As with Stuxnet, a cyberattack destroyed hardware when Russia invaded

[1] National Research Council (2012, pp. 1 and 4). The wait for equipment has lengthened since that report. A utility that orders a large transformer "may have to wait 2 to 4 years for it to be delivered, compared to a wait of just months as recently as 2020" (National Infrastructure Advisory Council, 2024, p. 3).

Ukraine. The attack on Viasat's KA-SAT satellite network on February 24, 2022, left tens of thousands of customer modems across Europe unable to operate. Many had to be replaced rather than repaired remotely (Viasat, 2022). The attack also severed the remote monitoring of about 5,800 wind turbines in central Europe, which were collateral damage. With no backup link, the modems had to be replaced wind farm by wind farm, and a month later almost half the farms were still offline (ENERCON, 2022).

TSMC and ASML. The most advanced computer chips are made on one island with equipment from one company. In 2019, 92 percent of the world's capacity to make chips below 10 nanometers was in Taiwan. The industry's own report calls this concentration "single points of failure" (Varas et al., 2021, p. 5). ASML, in the Netherlands, is "currently the world's only manufacturer of EUV lithography systems," the extreme-ultraviolet machines TSMC uses to make its most advanced chips. Carl Zeiss, in turn, is ASML's sole supplier of lenses, mirrors, and other optical components. ASML's own annual report says that "if Carl Zeiss SMT were to … be unable to maintain production of optics over a prolonged period, we would effectively cease to be able to conduct our business" (ASML Holding N.V., 2026, pp. 17 and 69). AI lowers the cost of finding and exploiting weaknesses in this economically central supply chain.

Bioattacks. The weak links in the examples above are equipment. People can be weak links too. Essential occupations depend on skills that take years to acquire, and several are geographically concentrated: frontier AI research in San Francisco, finance in New York and London, leading-edge chipmaking in Taiwan. A pathogen aimed at one of these kills a small fraction of the population but a large fraction of an essential occupation, so output falls until the survivors retrain.

AI makes it easier to obtain the knowledge needed to perform a bioattack. Even back in 2025, non-experts using frontier models had 4.7 times higher odds of writing a feasible experimental protocol (a step-by-step plan for an experiment) than a group using only the internet (AI Security Institute, 2025, p. 16). Imagine what models in two years will be able to do. As of 2023, there were 69 biosafety level 4 (BSL-4) labs, which handle the most dangerous pathogens, spread across 27 countries, and only one of the 27 scores high on the oversight of dual-use research (Lentzos and Koblentz, 2023, pp. 5 and 10–13).

Nuclear attacks. Nuclear weapons are another form of attack that puts large numbers of lives at risk. Hacking a country's nuclear weapons system may be extremely difficult, especially in the West and Russia, where we have had many years to strengthen the systems. However, AI

could again be used to target the weakest links. AI models in the future may be superhuman at social engineering or at manipulating evidence and information. Past close calls were often accidents of false information; AI could produce such information on purpose. The history of narrowly-averted disasters with nuclear weapons makes clear that this dangerous technology is a potential source of significant harm (Sagan, 1993; Lewis et al., 2014).

Summary of results. The magnitude of the loss from an attack depends in part on whether other inputs can substitute for the ones an attack disables. When the target of the attack has a cost share of 1 percent, reallocating inputs toward the damaged tasks cuts the loss associated with a tenfold decline in productivity from 68 percent of GDP to 6 percent.

Whether that substitution is available and how long it takes depends on the details of the attack. The output declines end when the damage is repaired or the destroyed inputs are replaced. Rebuilding transformers takes years, while a hacked banking system may be restored in a week. And of course deaths are permanent, so they weigh heavily in the welfare calculation.

The weak-link asymmetry also applies to attacks on people. Killing 1 percent of the population uniformly lowers output by 1 percent and leaves consumption per survivor unchanged. The welfare loss is the value of the lives lost. But holding the number of deaths fixed, targeting the deaths at a critical weak link, such as chipmakers in Taiwan or finance in New York and London, can compound the loss by also reducing output per survivor.

The final part of the paper adds an AI sector. Its investment makes AI models more capable, and more capable models have two spillovers. They raise the productivity of the rest of the economy, and they raise the probability of a successful attack, because attackers also use the AI models. Defenders use them too, but a defender must secure every critical link while an attacker needs to break only one, so we assume AI helps attackers more. Part of the benefit and part of the cost of AI are therefore external to the frontier labs. Whether the market overinvests or underinvests depends on which spillover is larger. In a weak-link economy, the productivity gain from more capable models is small, but the damage from an attack can be enormous. In our setup, this means that the negative externalities from AI typically outweigh the positive externalities (although we do not consider medical innovations from AI that extend life). Taxes on AI firms can internalize these externalities.

In the remainder of the paper, Section 2 lays out the environment, Section 3 the asymmetry, Section 4 the availability of substitutes, Section 5 the duration of losses and the welfare cost of

deaths, and Section 6 the AI sector and the planner.

Related literature. This paper builds on task-based models of automation, including Zeira (1998), Autor et al. (2003), and Acemoglu and Restrepo (2018). Kremer (1993) and Jones (2011) show how weak links constrain production when tasks are complements. Aghion et al. (2019) and Jones and Tonetti (2026) apply this logic to automation and AI and find that the tasks that are not yet automated limit the gains. Estimates of those gains vary widely. Trammell and Korinek (2023), Davidson (2021), and Erdil and Besiroglu (2023) consider scenarios in which AI accelerates economic growth, while Nordhaus (2021) concludes that such a singularity is not near. Acemoglu (2024) estimates that AI will raise growth in total factor productivity by less than 0.1 percentage points per year over the next decade, while Aghion and Bunel (2024) calculate 0.7 percentage points. These papers study the gains from AI. We use the same production function to study the losses. Our model is static, not because we think the dynamic effects are unimportant but rather because those have been studied, and the effect of weak links in reducing those gains is already appreciated. The same intuition holds in a static model. The dynamics are less important to understanding the downside risks.

A second literature studies how shocks to individual firms and sectors affect aggregate output. Gabaix (2011) and Acemoglu et al. (2012) show that shocks to large firms and to sectors that supply many others do not average out in the aggregate. Baqaee and Farhi (2019) extend the theorem of Hulten (1978) to second order and show that when inputs are complements, negative shocks are magnified and positive shocks are attenuated. This is the weak-link asymmetry. Our contribution is to highlight the importance of this asymmetry to AI.

On the empirical side, Barrot and Sauvagnat (2016), Boehm et al. (2019), and Carvalho et al. (2021) use natural disasters to show that a disruption to suppliers lowers the output of their customers. We discuss this type of evidence later as support for the view that damaging weak links can be costly.

Concerns about catastrophic risks from AI are highlighted by Joy (2000), Bostrom (2002, 2014), Rees (2003), Posner (2004), Yudkowsky et al. (2008), Ord (2020), and Nielsen (2024). Hendrycks et al. (2023) and Growiec and Prettner (2025) provide overviews. Jones (2016) studies the tradeoff between the consumption that new technologies deliver and the lives they put at risk, and Jones (2024) asks how much existential risk is worth taking in exchange for faster growth from AI. Aschenbrenner (2020) and Aschenbrenner and Trammell (2024) suggest we may live in a "time of perils" in which we are advanced enough to face high risk but

not rich enough to spend sufficiently on mitigation. Martin and Pindyck (2015, 2020) use the value of a statistical life to evaluate the gains from avoiding catastrophes, building on Rosen (1988), Murphy and Topel (2006), Nordhaus (2003), and Hall and Jones (2007). In these papers the damage from a catastrophe is typically taken as given. This paper differs by deriving the damage from the structure of production, so that it depends on which tasks are attacked and on who is killed.

Finally, several papers study policy toward risky technologies. Acemoglu and Lensman (2024) show that the optimal adoption of a transformative technology can be delayed if its costs are irreversible, and Guerreiro et al. (2023) analyze the optimal regulation of AI. Jones (2025) finds that spending at least 1 percent of GDP each year to mitigate the risk from AI can be justified in most of the scenarios he considers. That paper takes the development of AI as given. Here, investment by the AI labs raises both the capability of the models and the probability of an attack, so a tax on the AI sector corrects an externality.

2. The Weak-Link Environment

A unit measure of tasks combine to produce a single final good with a constant elasticity of substitution σ:

$$Y = \left(\int_0^1 Y_i^{\frac{\sigma-1}{\sigma}} \, di \right)^{\frac{\sigma}{\sigma-1}}. \tag{1}$$

Each task is produced with a single input according to

$$Y_i = \psi_i X_i, \tag{2}$$

where ψ_i is a task-specific productivity and X_i is the quantity of the input — for example, capital or labor — allocated to task i. The economy's resource constraint is $\int_0^1 X_i \, di = \bar{X}$, where $\bar{X}$ is the fixed supply of the input.

We focus throughout on $\sigma < 1$, so that tasks are complements and every task is essential. Production is constrained by the weakest links (Kremer, 1993; Jones, 2011). The model is static, with no feedback from AI to the growth rate as in Jones and Tonetti (2026). The gains we compute are therefore one-time increases in the level of output.

Based on a range of evidence, Jones and Tonetti (2026) suggest $\sigma = 0.2$ as a benchmark value. We adopt that as our baseline. Demirer, Musolff, and Yang (2026) estimate an elasticity

of substitution of 0.23 between AI and human effort in software development. Other direct estimates of the short-run elasticity are close to zero. Renesas supplied some 40 percent of the world's automotive microcontrollers. When the 2011 Tōhoku earthquake shut its main automotive plant, Toyota was short of these chips, and its domestic output fell by more than half for three months. From the same event, Boehm et al. (2019) estimate that the U.S. affiliates of Japanese firms cut output nearly one-for-one with their lost imports. Their structural estimates for these affiliates put the elasticity of substitution at 0.2 across material inputs and at 0.03 between materials and a capital–labor aggregate. Atalay (2017) estimates an elasticity of 0.1 across intermediate inputs in annual U.S. data.

The efficient allocation of the input equates the marginal product of X across tasks, so $X_i \propto \psi_i^{\sigma-1}$. With $\sigma < 1$, the exponent $\sigma - 1$ is negative, so the efficient allocation devotes more resources to less productive tasks. This is the classic force behind the Baumol (1967) cost disease. Substituting the allocation back into the production function gives

$$Y = \Psi \bar{X}, \qquad \text{where} \qquad \Psi \equiv \left(\int_0^1 \psi_i^{\sigma-1} \, di \right)^{\frac{1}{\sigma-1}}. \tag{3}$$

Aggregate productivity Ψ is a power mean of order $\sigma - 1 < 0$ of the task productivities, so it puts more weight on the small values of ψ_i, the weak links.

With competitive markets, the cost share of any collection of tasks Ω satisfies the usual CES share equation:

$$s_\Omega \equiv \frac{\int_\Omega p_i Y_i \, di}{PY} = \frac{\int_\Omega Y_i^{\frac{\sigma-1}{\sigma}} \, di}{Y^{\frac{\sigma-1}{\sigma}}}. \tag{4}$$

where p_i is the price of task i and P is the price of the final good. In equilibrium, a group's cost share also equals the fraction of $\bar{X}$ that it employs. With the allocation and this share equation in hand, we turn to the response of output to a shock.

3. The Weak-Link Asymmetry

3.1 The Gain–Loss Formula

Consider a shock that multiplies the productivity of a collection of tasks with initial cost share s by a common factor $z > 0$, with the allocation of inputs held fixed, or "frozen." This case applies right after a sudden shock, before any input can move. Section 4 considers the case in which

inputs reallocate. Automation and technical progress are the case $z > 1$ in which productivity is enhanced. An attack, disruption, or disaster reduces productivity and corresponds to $z < 1$. Write $Y(s, z)$ for output under the shock and Y_0 for output at the initial productivities. Because the inputs are frozen, the output of each shocked task scales by z. Raising the CES aggregate (1) to the power $\frac{\sigma-1}{\sigma}$, splitting the integral across the two groups, and applying the share equation (4) to each yields the gain–loss formula.

Proposition 1 (The gain–loss formula)**.** Scale the productivity of tasks with cost share s by $z > 0$, holding the allocation of inputs fixed. Output changes by the factor

$$G(s, z; \sigma) \equiv \frac{Y(s, z)}{Y_0} = \left(1 - s + s\, z^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}} = \left(\frac{1 - s'}{1 - s}\right)^{\frac{\sigma}{1-\sigma}}, \tag{5}$$

where $s' = s\, z^{\frac{\sigma-1}{\sigma}} / (1 - s + s\, z^{\frac{\sigma-1}{\sigma}})$ is the tasks' cost share after the shock via the share equation (4). For $\sigma < 1$:

(i) *Bounded gains.* As $z \to \infty$, $s' \to 0$ and $G \to \left(\frac{1}{1-s}\right)^{\frac{\sigma}{1-\sigma}}$.

(ii) *Unbounded losses.* As $z \to 0$, $s' \to 1$ and $G \to 0$.

(iii) *Local symmetry.* $\frac{d \ln G}{d \ln z} = s'$, equal to s at $z = 1$.

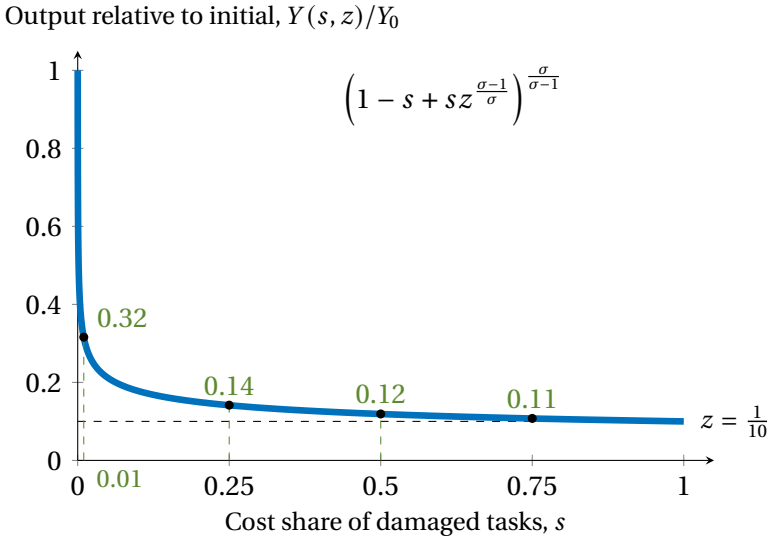
(iv) *Global asymmetry.* $\ln G$ is strictly concave in $\ln z$: $\frac{d^2 \ln G}{d(\ln z)^2} = \frac{\sigma-1}{\sigma} s'(1 - s') < 0$.

Proof in Appendix B.

The economics of Proposition 1 is the logic of a chain that is only as strong as its weakest link. Output rises at most by the factor $\left(\frac{1}{1-s}\right)^{\sigma/(1-\sigma)}$, however large the improvement (part i). This bound is the infinite automation formula of Jones and Tonetti (2026). Improving one link without limit cannot make the chain much stronger than its other links allow, so the gain saturates. In contrast, output falls to zero as the attack loss intensifies (part ii). Weakening one link toward zero makes the chain as weak as that link, so the loss is unbounded. In this sense, "weak links" is a better metaphor than bottlenecks.

The elasticity of output with respect to the shock is the cost share, s (part iii). For a small shock the response is the cost share for gains and losses alike. This is the theorem of Hulten (1978). It holds for shocks of every size only in the Cobb–Douglas case $\sigma = 1$, where $s' = s$ and

Figure 1: Output Loss from Damaging Tasks with Cost Share s



Note: Output relative to its initial level, from equation (5), when a shock reduces the productivity of tasks with an initial cost share of s by the factor $z = 1/10$, holding the allocation of inputs fixed. The elasticity of substitution is $\sigma = 0.2$. The dashed line at z is output when all tasks are damaged ($s = 1$).

$G = z^s$.

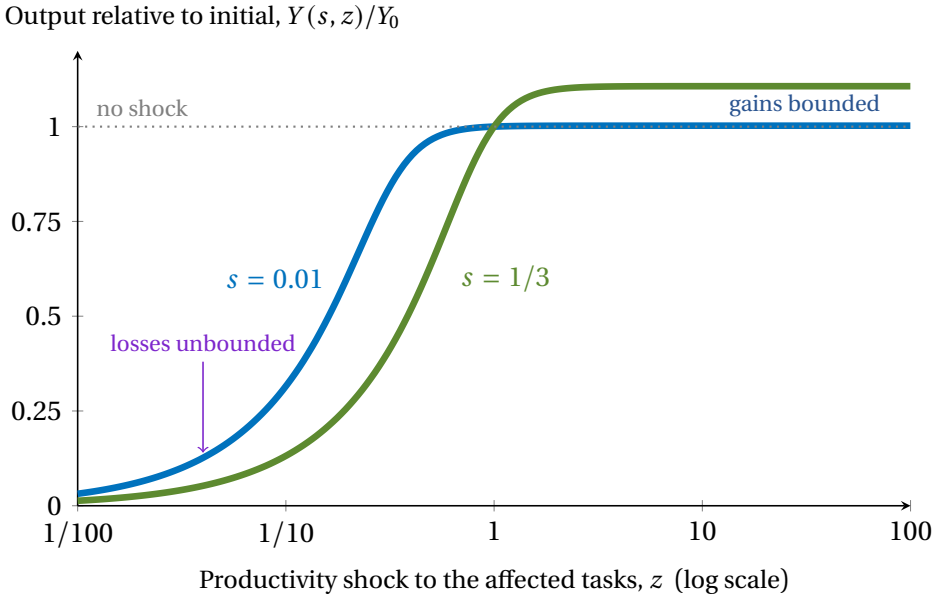
The asymmetry this paper studies is the departure from that log-linear benchmark when $\sigma < 1$. Log output is concave in the log shock with curvature $\frac{\sigma-1}{\sigma}\, s'(1-s')$ (part iv), as in Baqaee and Farhi (2019).

The last expression in equation (5) says the same thing in the language of cost shares. When tasks are complements, a shock that lowers the productivity of a group of tasks raises its cost share and a shock that raises productivity lowers it. The response of output to a shock of any size is then the change in the share of the other tasks, from $1-s$ to $1-s'$, raised to the power $\frac{\sigma}{1-\sigma}$.

3.2 Magnitudes

Figure 1 plots the loss from equation (5) against the cost share s for $\sigma = 0.2$ and a shock $z = 1/10$ that destroys 90 percent of the productivity of the damaged tasks. Two features stand out. First, a shock to a small collection of tasks can have a large effect on GDP. Cutting the productivity of tasks worth 1 percent of GDP by 90 percent reduces total GDP by 68 percent.

Figure 2: Small Gains and Large Losses



Note: The output response $G(s,z;\sigma)$ from equation (5), for $\sigma = 0.2$ and two cost shares. The horizontal axis is the productivity factor applied to the affected tasks, on a log scale, so equal horizontal distances are equal-sized shocks in the two directions.

Second, the curve is steep near $s = 0$ and then flattens. Most of the damage a shock can do is done once the damaged tasks are a modest share of the economy. Once a set of damaged tasks becomes the binding constraint, damaging additional tasks does relatively little further harm.

Figure 2 varies the shock rather than the share and plots output against z on a log scale for $\sigma = 0.2$. A point at $z = 10$ and a point at $z = 1/10$ are equal-magnitude shocks in opposite directions, yet the curves rise only slightly to the right before flattening and fall steeply to the left toward zero. This clear asymmetry and the potential for large losses is the main message of the paper.

Consider a technology with productivity factor $\bar{z} > 1$. Used well, it multiplies the productivity of a set of tasks by $\bar{z}$. Used as a weapon, it divides the same tasks' productivity by $\bar{z}$. Table 1 reports both outcomes for $\sigma = 0.2$ and tasks with a cost share of $s = 1$ percent. A tenfold improvement raises output by 0.25 percent. A tenfold degradation lowers it by 68 percent. The loss is larger than the gain by a factor of $68/0.25 = 272$.

Table 2 reports the asymmetry across cost shares and elasticities and makes two points. First, the gains at tenfold and at infinite improvement are nearly equal. The gains have already

Table 1: The Same Productivity Factor, Opposite Directions

Productivity factor $\bar{z}$	Percent change Gain: $z = \bar{z}$	Loss: $z = 1/\bar{z}$
2	0.2	−3
5	0.3	−39
10	0.3	−68
100	0.3	−97

Note: Percent change in output from multiplying the productivity of a group of tasks with cost share $s = 1\%$ by $\bar{z}$ (improvement) or by $1/\bar{z}$ (degradation), from equation (5) with $\sigma = 0.2$.

Table 2: The Asymmetry: Percent Gains versus Losses

	Small share, $s = 1\%$		Large share, $s = 1/3$	
Productivity shock z	$\sigma = 0.2$	$\sigma = 0.5$	$\sigma = 0.2$	$\sigma = 0.5$
Infinite improvement ($z \to \infty$)	0.3	1	11	50
Tenfold improvement ($z = 10$)	0.3	0.9	11	43
Tenfold degradation ($z = 1/10$)	−68	−8	−87	−75
Hundredfold degradation ($z = 1/100$)	−97	−50	−99	−97

Note: Percent change in GDP from equation (5) when the productivity of tasks with initial cost share s is multiplied by z, holding the allocation of inputs fixed, for two cost shares and two values of the elasticity of substitution.

saturated. The losses, by contrast, keep growing. Second, the asymmetry weakens as tasks become more substitutable but does not disappear. Even at $\sigma = 0.5$, the upper end of the empirically relevant range, degrading tasks with a 1-percent cost share tenfold lowers GDP by 8 percent. That loss is eight times the gain from automating the same tasks with infinite productivity.

4. Reallocation: The Availability of Substitutes

The calculation so far holds inputs frozen, which describes the short run. How long the short run lasts depends on the nature of the attack. Software can be reinstalled relatively quickly, but high-voltage transformers could take more than a year to replace. Once inputs can be

reallocated, moving them into the damaged tasks reduces the loss. In the model, the input $\bar{X}$ is reallocated according to the new productivities, so aggregate productivity is again the power mean in equation (3) evaluated at the damaged ψ_i's. Using the share equation (4) to express the answer in terms of the initial cost share gives the gain–loss formula evaluated at a higher elasticity of substitution.

Proposition 2 (Reallocation)**.** When inputs are reallocated efficiently after the shock, the output change becomes the factor

$$\begin{aligned} G^r(s,z;\sigma) &= \left(1-s+s\,z^{\sigma-1}\right)^{\frac{1}{\sigma-1}} = \left(\frac{1-s'}{1-s}\right)^{\frac{1}{1-\sigma}} \\ &= G(s,z;\sigma^r)\,, \qquad \text{where } \sigma^r \equiv \frac{1}{2-\sigma}\,. \end{aligned} \tag{6}$$

and where the post-shock cost share is $s' = s\,z^{\sigma-1}/(1-s+s\,z^{\sigma-1})$.
Proof in Appendix B.

The first line of equation (6) is the gain–loss formula (5) with the reallocation exponent $\sigma-1$ in place of $\frac{\sigma-1}{\sigma}$. In share form the exponent is $\frac{1}{1-\sigma}$ in place of $\frac{\sigma}{1-\sigma}$. The second line expresses it as the original $G(\cdot)$ at a higher elasticity of substitution ($\sigma < \sigma^r < 1$ for $\sigma \in (0,1)$). The economy responds to the shock as if tasks were more substitutable than they are. Frozen inputs and strong complementarity together produce the largest losses, because as $\sigma \to 0$ the frozen response falls to z, the Leontief outcome. But σ^r never falls below one half. With inputs free to move, even complete complements respond like a frozen economy with an elasticity of $1/2$. At the baseline $\sigma = 0.2$ the effective elasticity is 0.56.

As $\sigma \to 0$, equation (6) converges to the share-weighted harmonic mean of 1 and z, $(1-s+s/z)^{-1}$. This is Amdahl's law, a computer-architecture formula for how much a system speeds up when only a fraction of it improves. Even complete complements retain this much substitutability because resources substitute for productivity one-for-one. Each task produces $Y_i = \psi_i X_i$, so doubling a task's input fully offsets a halving of its productivity. This Cobb–Douglas offset between ψ_i and X_i is present inside even a Leontief aggregate.

The same elasticity shift raises the bound on gains. As $z \to \infty$, equation (6) converges to $\left(\frac{1}{1-s}\right)^{\frac{1}{1-\sigma}}$. This is the infinite automation gain of Proposition 1 with the exponent $\frac{1}{1-\sigma}$ in place of $\frac{\sigma}{1-\sigma}$. At the baseline values, automating the 1-percent tasks with infinite productivity is

Table 3: The Asymmetry between Gains and Losses (percent of GDP)

	Infinite automation ($z = \infty$)		Cutting productivity 90% ($z = 1/10$)	
Cost share s	Inputs frozen	Reallocated	Inputs frozen	Reallocated
1%	0.3	1	−68	−6
10%	3	14	−82	−41
25%	7	43	−86	−65
50%	19	138	−88	−80

Note: Percentage change in GDP from shocking tasks with an initial cost share s, with $\sigma = 0.2$. The first two columns report the gain from automating the tasks with infinite productivity (Jones and Tonetti, 2026), with inputs frozen (equation (5)) and reallocated (equation (6)). The last two report the loss when the same tasks have their productivity cut by 90 percent, from the same two equations.

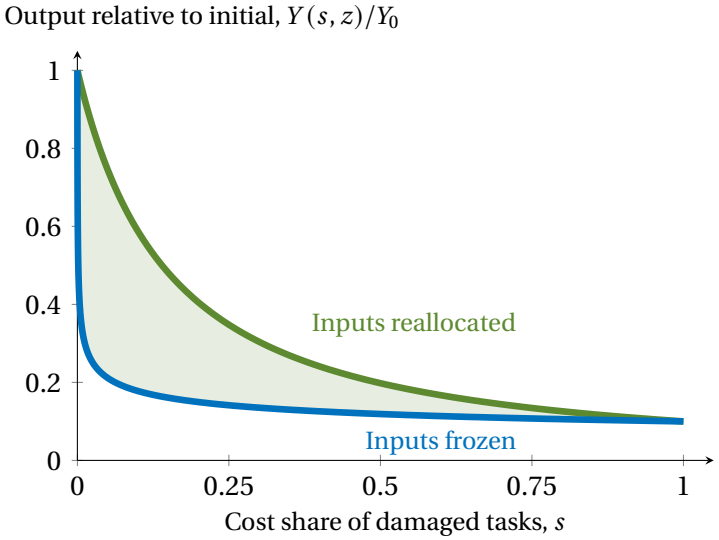
worth 0.25 percent of GDP with inputs frozen and 1.3 percent with inputs reallocated. Four fifths of the gain from automation is realized only when the freed inputs are reallocated to the remaining tasks.

Output rises as resources move into the damaged tasks. Before inputs move, the relative price of the damaged tasks rises, because their output is scarce. The quantity of each damaged task is down by the factor z and the CES demand curve has elasticity σ, so the price of a damaged task relative to an undamaged one rises by the factor $z^{-1/\sigma}$. The cost share of the damaged tasks rises with it. Energy's cost share rises the same way during an oil crisis. In response to these prices, resources move into the damaged tasks. The efficient allocation sets $X_i \propto \psi_i^{\sigma-1}$. The shock lowers the productivity of the damaged tasks, so their input allocation rises and the relative price falls to the unit cost $1/z$. The loss depends on the post-shock cost shares, the s' in Proposition 1 and Proposition 2.

The numbers underlying this reallocation can be stunning. For example, consider $s = 0.01$ and $z = 1/10$. With inputs frozen, the relative price of a damaged task rises a hundred-thousand-fold. The damaged tasks' cost share rises from 1 percent to 99.0 percent of GDP. Inputs move into the damaged tasks precisely because their relative price rises so much. Once inputs reallocate, the damaged tasks' share of inputs and expenditure rises only from 1.0 percent to 6.0 percent. The relative price rises tenfold rather than a hundred-thousand-fold.

Table 3 compares the two counterfactuals across cost shares, and Figure 3 shows the results graphically. Reallocating inputs reduces the loss from damage to tasks with a 1-percent cost share by over ninety percent, compared with about a quarter for tasks with a 25-percent cost

Figure 3: The Value of Reallocation



Note: Output loss from cutting the productivity of tasks with initial cost share s by 90 percent ($z = 1/10$), with $\sigma = 0.2$. The blue curve is output with inputs frozen (equation (5)). The green curve is output with inputs reallocated efficiently toward the damaged tasks (equation (6)). The shaded region is the output recovered by reallocation.

share. The smaller the cost share of the damaged tasks, the larger the fraction of the loss that disappears once inputs move. A sudden attack is dangerous because inputs are frozen for a time after it occurs, and a targeted attack would certainly consider the extent to which reallocation can be delayed. Policies that speed the transition from the frozen response to the reallocated one (redundancy, stockpiles, backup capacity, response planning) are safety investments that can substantially reduce the cost of attacks on critical weak links.

Whether real economies have this ability is an open question. Reallocation requires three things. First, a substitute has to exist. When computers are hacked, adding more computers may not help. Second, the processes the substitute needs must be rebuilt. Years ago, computers replaced paper ledgers, clearing arrangements, and the people who knew how to use them. Finally, the substitute runs on tasks the attack may also disable. After the 2017 NotPetya malware erased its software (Greenberg, 2018), Maersk moved 80 percent of its volume by hand for the ten days its systems took to rebuild, because its ports, its phones, and everyone else's computers still worked. An attack on the payments network or the grid takes those down too, so an economy-wide attack reallocates more slowly than a firm-level one.

Evidence. Are losses of this size plausible? No shock on record has cost 68 percent of GDP. One might conclude from that record that real economies substitute more easily than $\sigma = 0.2$ allows. Table 4 assigns five well-known episodes an approximate cost share and shock and reports the loss each implies under equations (5) and (6), with sources in Appendix C.

The shocks in the historical record were mild. For mild shocks, the formula predicts small losses. The Arab oil embargo of 1973–74 lasted five months. At the trough, in November 1973, Arab production was down by 7.5 percent of world output. The shortfall in world production averaged 4 percent over the six months through March (Hamilton, 2009, 2013). This is a shock of $z \approx 0.96$ to tasks with a cost share of around 4.5 percent. For a shock that small, the loss is the cost share times the shock, so the formula gives 0.2 percent of GDP under either allocation. The recession that followed was larger. U.S. GDP fell 3 percent from peak to trough. Hamilton (2013) calculates that the lost output exceeded the value of the lost oil by an order of magnitude and attributes the difference to the demand side, as households cut their spending on cars and other goods.

Prices provide evidence on the elasticity. The world price of oil quadrupled while the quantity consumed barely changed, so oil's expenditure share rose. By the share equation (4), a cost share rises as quantity falls only when σ is below one. Hamilton (2009) puts the short-run elasticity of demand in this episode at 0.10.

The 2022 cutoff of Russian gas to Germany is the same case at about eight times the size ($z = 0.7$ against 0.96). Bachmann et al. (2024) studied a complete stop that cuts gas use by 30 percent and predicted, against warnings of collapse, a GDP loss of 0.5 to 3 percent. Gas is 1.2 percent of expenditure in their calibration. At $z = 0.7$ the formula gives 0.9 percent with inputs frozen and 0.5 percent reallocated. Both are at the bottom of their range. The realized cut was about half as large. Germany absorbed it with storage that was full by November, a world market in liquefied gas, and six months between the invasion and the complete stop (Moll et al., 2023), and GDP actually grew 1.9 percent in 2022. The largest forecasts of loss treated gas as an input with no substitutes. The episode is consistent with the reallocated response, so the small loss says little about the short-run elasticity.

When the shock was large historically, the loss was several times the Hulten formula of cost share times shock. For example, consider Covid-19. Suppose in-person tasks are about 15 to 20 percent of GDP and the lockdowns of April 2020 cut their productivity roughly in half. The formula then gives a loss of 16 to 29 percent with inputs frozen and 9 to 16 percent

Table 4: Weak-Link Shocks in the Data

Episode	Cost share s	Productivity shock	Model loss Frozen	Model loss Realloc.	Outcome
Oil embargo, 1973–74	0.045	−4	0.2	0.2	U.S. GDP −3; oil price ×4
German gas, 2022	0.012	−30	0.9	0.5	forecast 0.5–3; GDP +1.9
Covid, April 2020	0.15–0.20	−40 to −50	16–29	9–16	U.K. GDP −25 that month
Puerto Rico, 2017	0.03	−75	42	7	Output −18 in 2017Q4
Soviet collapse, 1989–98	broad	large	—	—	official GDP −40 to −50

Note: The cost shares and productivity shocks are our approximate assignments for five episodes. The shock is the percent change in the productivity of the affected tasks, $z - 1$, so −30 is $z = 0.7$. The model columns are the output loss from equation (5) (inputs frozen) and equation (6) (inputs reallocated) at $\sigma = 0.2$, in percent of GDP. Oil: the shortfall in world production averaged 4 percent over six months (Hamilton, 2009); U.S. spending on petroleum products was 4.5 to 4.7 percent of GDP in 1972–73; GDP is from its peak in 1973 to its trough in 1975. Gas: the complete-stop scenario of Bachmann et al. (2024), gas at 1.2 percent of expenditure cut by 30 percent; the outcome column gives their forecast and 2022 GDP growth, and gas use fell 14 percent. Covid: in-person tasks; in the U.K., accommodation and food, arts and other services, transport, and education are 16 percent of value added, and their output fell by half; U.K. monthly GDP in April 2020 against February (Office for National Statistics, first estimate). Puerto Rico: electricity, about 3 percent of GDP; three quarters of customers were without power on average over the first two months; a monthly index of economic activity averaged 18 percent below its August level in 2017Q4; real GNP fell 4.2 percent in the fiscal year ending June 2018. Soviet: official GDP from 1989 to its trough, −45 percent in Russia and −47 in the former Soviet republics outside the Baltics (European Bank for Reconstruction and Development, 1999). Appendix C gives the sources.

reallocated, against 7 to 13 percent from the cost share times the shock. U.K. output that month was 25 percent below its February level (Office for National Statistics, 2020).

In September 2017, Hurricane Maria destroyed most of Puerto Rico's electricity grid. Electricity sales to households in October declined by 97 percent relative to the previous October's, and half of the loss was restored after two months. Electricity has a cost share of about 3 percent. The government's monthly index of economic activity, which combines payroll employment, electricity generation, gasoline sales, and cement sales, was 16 to 21 percent below its August level from October to December and real GNP fell 4.2 percent in the fiscal year ending June 2018 (Department of Economic Development and Commerce of Puerto Rico, 2026; Puerto Rico Planning Board, 2022). Full power took eleven months to restore (U.S. Government Accountability Office, 2019).

After the Soviet Union dissolved, measured output fell roughly in half. Blanchard and Kremer (1997) attribute an important part of the decline to the breakdown of supply relationships and show that output fell more for goods with more complex production processes.

5. Mortality: Who Is Killed Matters

In addition to lowering the productivity of tasks or damaging specialized capital, AI in the wrong hands can help an attacker kill people. As discussed extensively in the literature, the welfare losses directly associated with deaths are extremely large when inferred from the value of a statistical life. Here, we show that deaths targeted at weak links also lower output per person for the survivors.

Consider a bioattack that kills one percent of a country's population. If the deaths are uniform across the population, GDP falls by one percent and per capita consumption need not change. Suppose instead that the attack is targeted. Frontier AI research is concentrated in San Francisco. Finance clusters in New York and London. The machines that make leading-edge semiconductors are built by a single firm, ASML, in the Netherlands and operated at scale in a handful of fabs in Taiwan. How much more do the same number of deaths cost when they are concentrated in one of these places?

Uniform and targeted deaths. Let L be the population size, each person supply one unit of labor, and let labor be the only factor of production, so $\int_0^1 X_i \, di = L$. Suppose a fraction $\bar{d}$ of the population dies, uniformly across all tasks. Every task loses the fraction $\bar{d}$ of its workers, so output falls by the factor $1 - \bar{d}$. Consumption per survivor is therefore unchanged.

With weak links there is a second margin of *who* is killed. This margin matters only for the survivors, because it changes which workers remain to produce. Call the workers who perform a collection of tasks with cost share s an occupation. Suppose occupational skills are specific in the short run, so survivors from other occupations cannot step in until they retrain. Retraining is how labor is reallocated across tasks, as in Section 4. By the share equation (4), an occupation with cost share s employs the same fraction s of the workforce, so "the tasks with cost share s" and "the occupation with s percent of the workers" are the same. Now suppose the attack targets one occupation of size s, with the same number of deaths. A share ω of the $\bar{d}L$ deaths occurs in that occupation. The remaining deaths are drawn uniformly from all occupations. The fractions of workers who survive in the targeted occupation and in every other occupation are

$$z_{\text{target}} \;=\; 1 - \frac{\omega\,\bar{d}}{s} - (1-\omega)\,\bar{d} \qquad \text{and} \qquad z_{\text{other}} \;=\; 1 - (1-\omega)\,\bar{d}\,. \tag{7}$$

A task's output is proportional to the number of workers who perform it. The output of every task is then multiplied by z_{other}, and the output of the targeted tasks by a further factor $z_{\text{target}}/z_{\text{other}}$. Targeted deaths therefore have the same effect on output as a productivity shock to the targeted tasks, so the gain–loss formula applies with $z = z_{\text{target}}/z_{\text{other}}$. Let Y' be output after the attack and before any retraining. We call the deaths uniform when $\omega = 0$ and fully targeted when $\omega = 1$.

Proposition 3 (For a given number of deaths, a targeted bioattack lowers output more)**.** Let a fraction $\bar{d}$ of the population die, with $\sigma < 1$.

(i) *Uniform.* If deaths are drawn uniformly from every occupation, output falls by the factor $1 - \bar{d}$ under any allocation of labor. Consumption per survivor is unchanged.

(ii) *Targeted.* If a share $\omega > 0$ of the deaths occurs in an occupation with cost share $s < 1$ and $z_{\text{target}} > 0$, then before any retraining

$$\frac{Y'}{Y_0} = z_{\text{other}}\, G\left(s,\, \frac{z_{\text{target}}}{z_{\text{other}}};\, \sigma\right) < 1 - \bar{d}\,. \tag{8}$$

Output is strictly decreasing in ω.

Proof in Appendix B.

When every death is targeted ($\omega = 1$), output is $G(s, 1 - \bar{d}/s; \sigma)$ times its initial level. With $s = \bar{d}$ the whole occupation is killed as $\omega \to 1$, so output falls to zero.

In a numerical example of part (ii), the loss of output on impact is four times as large when the deaths are targeted. Let 1 percent of the population die and the occupation's cost share also be 1 percent ($s = \bar{d} = 1$ percent), with half of the deaths occurring in the targeted occupation ($\omega = 0.5$), so half the occupation's members die. If the deaths are uniform, output falls by 1 percent. If half of them occur in the targeted occupation, output falls by 4.0 percent on impact, four times as much. As in Section 4, the relative price of the targeted tasks rises, because fewer workers remain to perform them. The targeted occupation's cost share rises from 1 to 14 percent of GDP, the wage of its surviving members rises to twenty-seven times the pre-shock wage, and everyone else's wage falls by 16 percent.

These losses occur on impact, before any retraining. Once survivors from other occupations retrain and refill the targeted occupation, the surviving workforce is again spread over

unchanged productivities, so output returns to the uniform outcome, $Y = \Psi(1-\bar{d})L$.

Welfare and duration. Targeted deaths therefore cause a permanent loss of lives and a loss of output per survivor that ends when the occupation is refilled. We measure the effect of both losses on welfare, defined as the discounted sum of utility over persons and over time:

$$W = \int_0^\infty e^{-\rho t} L_t \, u(c_t) \, dt, \qquad u(c_t) = \bar{u} + \frac{c_t^{1-\gamma} - 1}{1-\gamma}, \tag{9}$$

where L_t is the population, $c_t = Y_t/L_t$ is consumption per person, $u(c_t)$ is the flow utility of a person who consumes c_t, the intercept $\bar{u}$ is the intrinsic value of being alive (Jones, 2016, 2024), γ is the elasticity of marginal utility, and ρ is the rate of time preference. Let births equal deaths, so the population is stationary at any level. Because the model is static and the population constant, without an attack $L_t = L$ and $c_t = c_0$ in every year, so total welfare in equation (9) is $L\,u(c_0)/\rho$.

We choose units so that today's consumption per person c_0 is normalized to one, so $u(c_0) = \bar{u}$ and $u'(c_0) = 1$. The value of a year of life in years of consumption is then $u(c_0)/(u'(c_0)\,c_0) = \bar{u}$. For numerical examples, we set $\rho = 2.5$ percent and $\gamma = 2$, typical numbers in macroeconomics, and $\bar{u} = 5$, so that the value of a year of life is five times consumption per person (Hall, Jones, and Klenow, 2020). For a person with 40 years of life remaining, these values imply a value of about 125 years of consumption, in line with estimates of the value of a statistical life of one hundred times annual consumption or more (Hall and Jones, 2007; Murphy and Topel, 2006; Jones, 2025).

A flow loss of one unit of utility per year that lasts T years has a present value of $(1-e^{-\rho T})/\rho$ years. We call this present value the discounted duration of the loss, or just duration for short. With $\rho = 0.025$, a loss that lasts one year has a duration of 0.99 years. A permanent loss has the duration $1/\rho = 40$ years.

We first compute the welfare loss from the lives lost and then the welfare loss from the lower output per survivor. Because the population is stationary, it is permanently $\bar{d}L$ smaller after the attack. Each year, there are $\bar{d}L$ fewer people who would have received utility $u(c_0)$, so the welfare loss from the lives lost is

$$\int_0^\infty e^{-\rho t} \, \bar{d}L \, u(c_0) \, dt = \bar{d}L \cdot \frac{u(c_0)}{\rho}. \tag{10}$$

This is the number of deaths times the value of a life, $u(c_0)/\rho$. If the deaths are uniform, consumption per survivor does not change, so the lives lost are the only welfare loss. A life is then worth $\bar{u}/\rho = 200$ years of consumption.[2]

Next, consider the welfare loss from the lower output per survivor. Let T_r be the time it takes to refill the targeted occupation, with a baseline value of one year. Until then, consumption per survivor is $c' = c_0\,(Y'/Y_0)/(1-\bar{d})$, with Y'/Y_0 given by equation (8), so each of the $(1-\bar{d})L$ survivors loses $u(c_0) - u(c')$ in utility per year. If the deaths are uniform, $c' = c_0$, and c' falls as more of the deaths are targeted.

Adding the two losses gives the fall in welfare from the attack, ΔW. Per capita and in years of consumption, it is

$$\frac{\Delta W}{L} = \underbrace{\frac{\bar{d}\,u(c_0)}{\rho}}_{\text{lives lost: } 0.01\times 200 = 2} + \underbrace{\frac{1-e^{-\rho T_r}}{\rho}\left(1-\bar{d}\right)\left[u(c_0)-u(c')\right]}_{\text{survivors: utility loss per year} \times 0.99}. \tag{11}$$

The first term is the same however the deaths are distributed, because every life has the same value, $u(c_0)/\rho$. Targeting raises the welfare loss only through the second term, the lower consumption of the survivors.

A consumption equivalent. To express the two components of the bioattack welfare loss in equation (11) as a fraction of consumption in every year rather than in years of consumption at today's marginal utility, we compare two scenarios. In the first, the attack occurs. In the second, there is no attack and everyone permanently consumes the fraction λ of no-attack consumption c_0. The consumption equivalent of the loss is the λ that makes the population indifferent between the two scenarios. It solves

$$\frac{u(c_0) - u(\lambda c_0)}{\rho} = \frac{\Delta W}{L}. \tag{12}$$

Table 5 reports the consumption equivalents for different shocks. In the upper panel, 1 percent of the population dies. The lives-lost column is constant at 4.8 percent, because the number of deaths is held fixed. All the variation is in the survivors' column. If the deaths are uniform, consumption per person is unchanged. With 50 percent of the deaths targeted at

[2] Growth would raise the weight on lives lost. With consumption growing at $g = 2$ percent, the flow value of a life $u(c_t) = \bar{u} + 1 - 1/c_t$ rises over time and adds $1/\rho - 1/(\rho+g) = 18$ years (9 percent) to the value of a life (Jones, 2025). The output losses, valued at a falling marginal utility, would shrink.

Table 5: The Welfare Cost of Deaths: Consumption Equivalents

Incidence of deaths	Consumption per survivor, c'	Consumption equivalent, every year		
		Lives lost	Survivors	Total
Death toll $\bar{d} = 1\%$, $s = 1\%$				
Uniform	100	−4.8	0.0	−4.8
50% targeted	97.0	−4.8	−0.1	−4.8
90% targeted	31.5	−4.8	−5.0	−9.3
Death toll $\bar{d} = 10\%$, $s = 10\%$				
Uniform	100	−33.3	0.0	−33.3
50% targeted	81.0	−33.3	−0.5	−33.6
90% targeted	17.8	−33.3	−9.3	−37.6

Note: A fraction $\bar{d} = 1$ or 10 percent of the population dies, and the targeted occupation has the same cost share, $s = \bar{d}$. The deaths are uniform, or 50 or 90 percent of them are targeted at the occupation ($\omega = 0.5$ or 0.9) and the rest are uniform, as in (7). The first column is consumption per survivor until the occupation is refilled, c', in percent of its pre-attack level. The other columns are $100(\lambda - 1)$ for the λ of (12), the percent of consumption the population would give up in every year to rule the loss out. The lives lost and the survivors' loss are the two terms of (11). Parameters are $\sigma = 0.2$, $\gamma = 2$, $\bar{u} = 5$, $1/\rho = 40$ years, and a retraining time of one year ($(1 - e^{-\rho})/\rho = 0.99$ years).

the 1-percent occupation, the survivors lose 0.1 percent of consumption in every year. With 90 percent targeted, they lose 5.0 percent, so the welfare loss from lower output per person exceeds the loss from the lives lost. In this sharply targeted case, the GDP loss is so large that the overall cost of the bioattack is equivalent to reducing consumption by 9.3 percent per year forever.

In the lower panel, the bioattack kills 10 percent of the population. For comparison, the mortality rate of the 1918 influenza was 1 to 3 percent worldwide, and the Black Death killed roughly a third of Europe. The value of the lives lost in this case rises tenfold, from $\bar{d}\,\bar{u} = 0.05$ to 0.5, so the population would give up a third of its consumption in every year to avoid these deaths.

Unlike the value of the lives lost, the survivors' loss rises less than tenfold, to 0.5 percent of consumption in every year when half of the deaths are targeted and to 9.3 percent when 90 percent are. The reason is the concavity induced by weak links (Figure 1). The fall in output grows less than proportionally with the size of the targeted occupation. When 1 percent of the population dies and half of the deaths are targeted, about half of a 1-percent occupation dies

and output falls by 4.0 percent on impact (Proposition 3). When 10 percent of the population dies, with half of the deaths again targeted, 55 percent of a 10-percent occupation dies and output falls by 27 percent. The occupation is ten times larger, but the fall in output is only 6.8 times as large.

6. AI Labs and the Rest of the Economy

Investment in the AI sector determines how capable the AI models become and how likely it is that someone uses them for a successful attack. An improvement in AI raises productivity in the rest of the economy, but it can also raise the productivity of "bad actors." That investment therefore has spillovers in both directions. How do the positive and negative externalities interact, and what tax or subsidy will lead to optimal AI investment?

Machines and labor. Tasks aggregate as in equation (1), but machines perform the tasks on $[0, \beta]$ and labor performs the rest:

$$Y_i = \begin{cases} \psi_i K_i & i \in [0, \beta] \\ \psi_i L_i & i \in (\beta, 1], \end{cases} \qquad \int_0^\beta K_i \, di = K, \qquad \int_\beta^1 L_i \, di = L, \tag{13}$$

where K_i and L_i are the capital and labor on task i, and the capital stock K and the population L are in fixed supply. The assignment of tasks to capital and labor is fixed, so β does not change. Within each block, the efficient allocation is as in Section 2, with $K_i \propto \psi_i^{\sigma-1}$ and $L_i \propto \psi_i^{\sigma-1}$, so output is

$$\bar{Y} = \left[(BK)^{\frac{\sigma-1}{\sigma}} + (AL)^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}}, \tag{14}$$

where $B \equiv \left(\int_0^\beta \psi_i^{\sigma-1} \, di \right)^{1/(\sigma-1)}$ and $A \equiv \left(\int_\beta^1 \psi_i^{\sigma-1} \, di \right)^{1/(\sigma-1)}$ are proportional to the power mean (3) taken over the machine tasks and the labor tasks separately. The capital share of income is $s_K = (BK)^{\frac{\sigma-1}{\sigma}} / \bar{Y}^{\frac{\sigma-1}{\sigma}}$ (equation 4). Proposition 1 holds for any collection of tasks with cost share s, whichever factor performs them, so β, K, B, and A matter only through s_K.

The AI sector and the rest of the economy's output combine in a Cobb–Douglas aggregator to produce final output. For simplicity, the AI and the non-AI sector each produce with the same technology (14). The AI sector employs a fraction a of the economy's capital and labor.[3]

[3] With identical technologies and common factor prices, both sectors use the economy-wide mix of capital and labor. The AI sector hires the fraction a of every occupation, so one number describes the allocation.

Thus, output is

$$Y \;=\; Y_{\text{AI}}^{\,\theta}\, Y_{\text{non-AI}}^{\,1-\theta} \;=\; a^{\,\theta}\,(1-a)^{1-\theta}\;\bar{Y}\,, \tag{15}$$

where $Y_{\text{AI}} = a\,\bar{Y}$ and $Y_{\text{non-AI}} = (1-a)\,\bar{Y}$ are the sector outputs and θ is the AI good's Cobb–Douglas weight. Equation (15) says that output is $\bar{Y}$, the output of a single sector that employs all of the economy's capital and labor, scaled by a term that depends only on the size of the AI sector, a. Because a is chosen before any shocks are known, a shock multiplies Y and $\bar{Y}$ by the same factor, $Y(s,z)/Y_0 = \bar{Y}(s,z)/\bar{Y}_0$. Therefore, the results on output from Sections 3–5 apply to Y directly.

The AI improvement. Consider a one-time improvement in AI capabilities, such as a new AI algorithm. The AI sector exists before this AI improvement and produces at today's productivity. The AI improvement multiplies the productivity of every task performed by machines by a factor z and leaves the productivity of labor tasks unchanged. Proposition 1 applies with the machine tasks' cost share s_K, so output after the AI improvement, Y_{safe}, satisfies

$$\frac{Y_{\text{safe}}}{Y_0} \;=\; G\big(s_K,\, z;\, \sigma\big) \;=\; \Big(1 - s_K + s_K\, z^{\,\frac{\sigma-1}{\sigma}}\Big)^{\frac{\sigma}{\sigma-1}}\,. \tag{16}$$

The attacks. AI can also be used for an attack. We study two kinds of attack in two separate economies, one for cyberattacks and one for bioattacks. (This lets us avoid considering the interaction case where both kinds of attacks occur.) If there is no attack, the productivity of machines is multiplied by $z(a)$, an increasing function of the size of the AI sector, and output is Y_{safe}. A successful attack occurs with probability $p(a)$. In the case of a cyberattack, the attack lowers the productivity of targeted machines — with cost share s_{target} — by the factor $1/z(a)$.

In the case of a bioattack, Section 5 already considered a "narrow" version in which a potentially small fraction of the population dies. Here we instead consider the most extreme outcome in which everyone dies if a bioattack succeeds. This is clearly extreme, but it is also a possibility that the existential risk community considers, so we include it in this section. Because welfare is the expected utility of a person alive today, before anyone knows who will die, a bioattack that kills everyone with probability p is close to one that kills a uniform fraction p of the population for certain.[4]

More capable models raise the productivity of bad actors, but they also raise the produc-

[4]The two would be identical if consumption per survivor were unchanged, as in Section 5. Here the capital stock survives, so consumption per survivor rises slightly.

tivity of defenders, who use them to find and patch vulnerabilities before an attacker can exploit them (Anthropic, 2026). Whether AI benefits attackers or defenders more is uncertain (Garfinkel and Dafoe, 2019; Bengio et al., 2026, pp. 61–62). In biosecurity, offense currently appears favored (Bengio et al., 2026, p. 143). The probability $p(a)$ is net of both effects. We assume that AI helps attackers more, as the weak-link structure suggests: a defender must secure every critical link, while an attacker needs to break only one. The probability $p(a)$ is therefore increasing in the size of the AI sector. A failed attack does no damage, so from here on "attack" means a successful attack.

These alternatives are summarized here:

Cyberattack economy

State	Probability	Outcome
no attack	$1-p(a)$	machine productivity $\times\, z(a)$
cyberattack	$p(a)$	targeted machines $\times\, 1/z(a)$, others $\times\, z(a)$

Bioattack economy

State	Probability	Outcome
no attack	$1-p(a)$	machine productivity $\times\, z(a)$
bioattack	$p(a)$	everyone dies

(17)

In the cyberattack, malware corrupts the software used in payments, logistics, and the systems firms run on, across the economy at once, affecting a cost share of tasks s_{target}. This is NotPetya at the scale of a country rather than a firm. In the first days nothing works. Reallocating inputs around the damage takes time, so we use the frozen response of Section 4 and hold inputs fixed for as long as the damage lasts.

When a cyberattack occurs, there are three groups of tasks. Labor tasks have cost share $1-s_K$ and keep their productivity. The other machine tasks have cost share $s_K - s_{\text{target}}$ and productivity multiplied by z. The targeted tasks have cost share s_{target} and productivity multiplied by $1/z$. Output is the share-weighted power mean of the groups' productivities, as in Proposition 1

with three groups in place of two. Output under the attack, Y_{cyber}, satisfies

$$\frac{Y_{\text{cyber}}}{Y_0} \;=\; G_{\text{cyber}}(s_{\text{target}}, s_K, z; \sigma) \;\equiv\; \left(1 - s_K + (s_K - s_{\text{target}})\, z^{\,\frac{\sigma-1}{\sigma}} + s_{\text{target}}\, z^{\,-\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}. \tag{18}$$

AI capability and the severity of an attack. The size of the AI sector determines how capable the AI models are, and AI capability determines both the AI productivity factor and the probability of an attack. We summarize this chain with one elasticity:

$$z(a) \;=\; \bar{z}\left(\frac{a}{\theta}\right)^{\phi}, \qquad p(a) \;=\; \bar{p}\left(\frac{a}{\theta}\right)^{\phi}, \tag{19}$$

where $\phi > 0$ is the elasticity of both the AI productivity factor and the probability of an attack with respect to the size of the AI sector. When the sector employs the fraction θ of the economy's inputs, AI multiplies machine productivity by $\bar{z}$ and the probability of an attack is $\bar{p}$. Because $Y_{\text{AI}} = a\,\bar{Y}$ state by state, a is proportional to the AI sector's output whether or not an attack occurs. A larger AI sector therefore has a larger productivity gain and a larger probability of an attack, and in the cyberattack economy a more damaging attack.

Timing. The size of the AI sector is chosen once, before the realization of uncertainty. When no attack occurs, every machine task improves by the same productivity factor, so inputs are already allocated efficiently across tasks.

In the bioattack economy, the state is drawn once, right after the inputs are allocated, and the deaths are permanent. In the cyberattack economy, an individual attack lasts one year, for example the time it takes to patch and flush the corrupted software or replace the specialized hardware. Let T be the number of years the economy is at risk of a cyberattack. We consider two cases. In the first, $T = 1$. After the year is over, every machine task benefits from the high productivity $z(a)$ whether or not an attack occurred. The second case is a prolonged struggle with AI-enabled attacks. The economy is at risk for $T = 10$ years, and in each year an attack occurs with probability $p(a)$. After ten years, the attacks end in one of two ways. With probability $p(a)$, the AI improvement is abandoned to end the attacks, and every task returns to its initial productivity. With probability $1 - p(a)$, the AI is aligned and guardrailed, the AI improvement is kept, and every machine task has the high productivity $z(a)$.

The competitive equilibrium. Households own the factors and trade complete claims on the states. Labs, task producers, and the final-good firm are perfectly competitive. We model both the positive effects on productivity and the negative consequences of attacks as externalities, and treat the firms in the AI sector as atomistic. Corollary 1 relaxes this atomistic assumption and considers a strategic AI sector.

In equilibrium, the AI sector's share of the economy's capital and labor is its Cobb–Douglas weight,

$$a^{eq} \;=\; \theta. \tag{20}$$

Final-good spending on the AI good is the fraction θ of output, and because the two sectors use the same technology, the AI sector's share of the economy's capital and labor equals its share of spending. An attack multiplies both sector outputs by the same factor, so the state does not change relative prices, and atomistic labs take the AI productivity factor and the probability of an attack as given. None of $\bar{p}$, $\bar{z}$, ϕ, or T therefore appears in an equilibrium condition.

Expected welfare. Welfare is given by equation (9). Initial consumption per person, normalized to one at today's allocation, $a = \theta$, is $c_0(a) = (a/\theta)^{\theta}\,((1-a)/(1-\theta))^{1-\theta}$ (equation 15). Consumption after the AI improvement is

$$c_{\text{safe}}(a) \;\equiv\; c_0(a)\; G\big(s_K,\, z(a);\, \sigma\big)\,, \qquad c_{\text{cyber}}(a) \;\equiv\; c_0(a)\; G_{\text{cyber}}\big(s_{\text{target}},\, s_K,\, z(a);\, \sigma\big)\,. \tag{21}$$

Because welfare is additive over time, expected welfare does not depend on how attacks are spread across the years at risk. In each year at risk an attack occurs with probability $p(a)$, and in the prolonged case the AI improvement is also abandoned after T years with probability $p(a)$. Expected welfare is therefore the same as with a single draw at the start, between no attack in any year and an attack that lasts all T years (Appendix B). With probability $1 - p(a)$, no attack ever occurs, the no-attack path. With probability $p(a)$, an attack occurs in every year at risk, the attack path. Let c_{after} be consumption on the attack path after the T years. In the one-year case, $c_{\text{after}} = c_{\text{safe}}$. In the prolonged case, $c_{\text{after}} = c_0$, because the AI improvement must be abandoned to end the attacks. Welfare on the attack path is

$$W_{\text{attack}}(a) \;\equiv\; \frac{1}{\rho}\Big[\,\Big(1 - e^{-\rho T}\Big)\, u(c_{\text{cyber}}(a)) \;+\; e^{-\rho T}\, u(c_{\text{after}}(a))\Big]\,. \tag{22}$$

On the no-attack path, everyone has the flow $u(c_{\text{safe}})$ forever. Expected welfare per person alive today is

$$\begin{aligned} W_{\text{cyber}}(a) &= \underbrace{(1-p(a))\,\frac{u(c_{\text{safe}}(a))}{\rho}}_{\text{no-attack path}} + \underbrace{p(a)\,W_{\text{attack}}(a)}_{\text{attack path}}, \\ W_{\text{bio}}(a) &= (1-p(a))\,\frac{u(c_{\text{safe}}(a))}{\rho}. \end{aligned} \tag{23}$$

A bioattack ends every life, so everyone loses the flow $u(c_{\text{safe}})$ forever. This loss counts only the people alive today. It omits the people who would have been born in the future, so it understates the cost of extinction. In the other direction, ignoring the dynamic effects of accelerating economic growth understates the benefits of AI, particularly if AI also cures diseases and extends life expectancy.

The planner. The planner chooses the allocation before the state is known and, unlike any single lab, accounts for the effect of the size of the AI sector on AI capability, and through it on the AI productivity factor and the probability of an attack. In the cyberattack and the bioattack economy, the planner solves

$$\begin{aligned} &\max_{a\in(0,1)} W_{\text{cyber}}(a) \qquad \text{and} \qquad \max_{a\in(0,1)} W_{\text{bio}}(a) \\ &\text{subject to} \quad c_0(a) = \left(\frac{a}{\theta}\right)^{\theta}\left(\frac{1-a}{1-\theta}\right)^{1-\theta}, \quad (19), \quad \text{and} \quad (21). \end{aligned} \tag{24}$$

The size of the AI sector has two effects. First, it changes initial consumption in every state through $c_0(a)$; the market allocation maximizes this by choosing $a=\theta$. Second, it makes the AI models more capable. More capable models raise the AI productivity factor, which raises consumption when no attack occurs and lowers it under a cyberattack, and they raise the probability of an attack.

The optimal allocation is given by Proposition 4.

Proposition 4 (The planner's allocation)**.** Assume a unique interior optimum. Let $\mathcal{E}(a^*) = \mathcal{E}^z(a^*) + \mathcal{E}^p(a^*)$ be the net external cost of the AI sector in percent of consumption, the sum of a productivity part and a probability part. See Appendix A for a derivation of $\mathcal{E}(a^*)$. The size of the AI sector that the planner chooses solves

$$\underbrace{\frac{\theta}{a^*} - \frac{1-\theta}{1-a^*}}_{\substack{\text{productive value of the} \\ \text{marginal input in AI}}} = \underbrace{\frac{\phi}{a^*}\mathcal{E}(a^*)}_{\substack{\text{external cost of the AI productivity factor and} \\ \text{the probability of an attack}}}, \tag{25}$$

and $a^* < \theta = a^{eq}$ if and only if $\mathcal{E}(\theta) > 0$. An ad valorem tax on the AI sector's factor payments decentralizes a^*, at the rate

$$\tau^* = \frac{\phi}{a^*}\mathcal{E}(a^*)\frac{1-a^*}{1-\theta}. \tag{26}$$

Proof in Appendix B.

The left side of equation (25) is zero at the market allocation, because atomistic labs ignore the net external cost on the right side. One more unit of capital and labor in AI raises the AI productivity factor and the probability of an attack by ϕ/a^* log points each. The net external cost of one log point of both is $\mathcal{E}$. No lab internalizes that its inputs improve everyone's models, that the same AI capability makes an attack more destructive, or that a larger sector makes an attack more likely.

Multiplying equation (25) by $a^*(1-a^*)$ gives $\theta - a^* = \phi\,\mathcal{E}(a^*)\,(1-a^*)$. The gap between the size of the AI sector in the market and in the planner's allocation is ϕ times the net external cost times the size of the non-AI sector, $1-a^*$. Dividing the gap by $a^*(1-\theta)$ gives the tax (26).

Because of the weak-link asymmetry, the positive externality of AI is limited compared to the potentially large negative externality. Whenever the net external cost is positive, the market allocation overinvests in AI, lowering welfare by raising the risk and severity of an attack. Even when the AI labs can coordinate, they do not achieve the planner's allocation.

Corollary 1 (Coordination among labs)**.** Let the labs act as one. They take all prices as given, including the state prices, but know that their joint size determines $z(a)$, and they choose the size of the AI sector to maximize the value of their profits. The size they choose, $\hat{a}$, solves

$$\frac{\theta}{\hat{a}} - \frac{1-\theta}{1-\hat{a}} = \frac{\phi}{\hat{a}}\,\theta\,\mathcal{E}^{z}(\hat{a}), \tag{27}$$

so coordinated labs internalize the fraction θ of the productivity part of the net external cost and none of the probability part.

Proof in Appendix B.

The coordinated sector earns the fraction θ of output in every state, so it captures the fraction θ of the AI productivity factor's effect on output, both the larger gain when no attack occurs and the larger loss under a cyberattack. The rest of the economy receives the remaining fraction $1-\theta$, which stays external. The probability of an attack, in contrast, enters the labs' valuation only through the state prices, which they take as given. The cost of the increased probability of an attack is therefore external at any degree of coordination.

In practice, the labs do seem to internalize the negative externalities, at least to some extent. Our analysis does not take this into account.

Consumption equivalents. We express welfare gains using permanent consumption equivalents, and there are three welfare calculations we construct. First, we consider an economy with the AI improvement and no risk of attack, in which everyone consumes c_{safe} forever. Let λ_R be the factor by which consumption in that economy would have to be lowered in every year, before the state is known, for people to be indifferent between it and the economy with the risk. The factor solves

$$\frac{u(\lambda_R\, c_{\text{safe}})}{\rho} = W(\theta)\,, \tag{28}$$

where W is W_{cyber} or W_{bio}. Therefore λ_R measures the equilibrium welfare loss associated with the possibility of an attack.

Second, we consider an economy without the AI improvement, in which there are no attacks and everyone consumes today's consumption, $c_0 = 1$, forever. Let λ_0 be the factor by which consumption in that economy would have to be multiplied in every year for people to be indifferent between it and the economy with the AI improvement and the risk. The factor

solves

$$\frac{u(\lambda_0 c_0)}{\rho} \;=\; W(\theta)\,. \tag{29}$$

Therefore λ_0 measures the welfare gain from the AI improvement, including the risk that comes with it.

Third, we consider the market economy itself. Let $\lambda_P > 1$ be the factor by which consumption in the market economy would have to be raised in every state and in every year for its expected welfare to equal the planner's. The factor solves

$$W(\theta;\, \lambda_P) \;=\; W(a^*)\,, \tag{30}$$

where $W(\theta;\, \lambda)$ is expected welfare (23) at $a = \theta$ with consumption in every state and year multiplied by λ. Therefore λ_P measures the welfare gain of moving from the equilibrium allocation to the optimal allocation. We also evaluate the first two calculations at the planner's allocation, with $W(a^*)$ in place of $W(\theta)$ in equations (28) and (29).

Parameter values. As in the rest of the paper, $\sigma = 0.2$. Utility is calibrated as in Section 5, with $\rho = 2.5$ percent, $\gamma = 2$, and $\bar{u} = 5$. The capital share is $s_K = 1/3$. We set $\theta = 0.02$, so in the competitive equilibrium the AI sector employs 2 percent of the economy's capital and labor. The targeted tasks in the cyberattack economy have cost share $s_{\text{target}} = 1$ percent. At the competitive equilibrium, AI multiplies machine productivity by $\bar{z} = 10$, the tenfold improvement of Section 3, and the probability of an attack is $\bar{p} = 50$ percent for the cyberattack and $\bar{p} = 1$ or 10 percent for the bioattack. We have no evidence on ϕ and set $\phi = 1$ as a benchmark.[5]

The market allocation. Table 6 reports the two economies. The first panel is the market allocation. With no attack, output is 10.7 percent above today's, which turns out to be the same gain that would occur with infinite automation of all machines. A cyberattack on machines with a cost share of $s_{\text{target}} = 1\%$ reduces GDP by the now-familiar 68 percent for a year.

A cyberattack, one year at risk, 50 percent chance of an attack. Even with the risk of an attack, the AI improvement raises welfare by the equivalent of 7.3 percent of consumption permanently. To keep the AI improvement but remove the risk of an attack, people would be

[5]The iso-elastic forms have $z(a) < 1$ for $a < \theta\,\bar{z}^{-1/\phi}$ (a tenth of θ at $\phi = 1$) and $p(a) > 1$ for $a > \theta\,\bar{p}^{-1/\phi}$ (2θ at $\bar{p} = 50$ percent). Neither binds at any allocation in Table 6.

Table 6: Market and Planner

All entries in percent except $z(a^*)$	Cyberattack		Bioattack	
Probability of an attack at $a=\theta$, $\bar{p}$	50		1	10
Years at risk of an attack, T	1	10	one draw	
Duration of an attack	1 year		forever	
Inputs during an attack	frozen		—	
Machine productivity after the attacks	$z(a)$	1 or $z(a)$	—	
Market allocation ($a=\theta=2\%$, $\bar{z}=10$)				
Output, no attack	+10.7		+10.7	
Output, during an attack	−68.4		−100	
Output, after the attacks	+10.7	0.0	—	
Welfare gain vs. AI improvement without attack risk	−3.0	−24.2	−5.3	−36.1
Welfare gain vs. no AI improvement	+7.3	−16.1	+4.8	−29.2
Planner				
AI sector's share of inputs, a^*	0.92	0.51	0.65	0.32
AI productivity factor, $z(a^*)$	4.6	2.5	3.2	1.6
Probability of an attack, $p(a^*)$	23.0	12.7	0.32	1.60
Output, no attack	+10.1	+9.0	+9.5	+6.5
Output, during an attack	−34.0	−3.3	−100	
Output, after the attacks	+10.1	−1.3	—	
Tax on AI factor payments	120	300	213	536
Welfare gain vs. AI improvement without attack risk	−0.9	−2.9	−2.8	−11.4
Welfare gain vs. no AI improvement	+9.7	+7.5	+7.6	−2.0
Welfare gain over the market	+2.2	+28.1	+2.9	+93.7

Note: All entries are percentages except the AI productivity factor, $z(a^*)$. Output is relative to today's, Y_0 at $a=\theta$. A cyberattack lasts one year with inputs frozen. With one year at risk the AI improvement is kept after the attack. With ten years at risk an attack occurs in each year with probability $p(a)$ and the AI improvement is abandoned after the ten years with probability $p(a)$. The second column's "after the attacks" entries are for that case. When the AI improvement is kept, output is that of the row "no attack." A bioattack kills everyone, so output is zero. Welfare gains are $100(\lambda-1)$ for the consumption equivalents λ_R of (28) (AI improvement without attack risk), λ_0 of (29) (no AI improvement), and λ_P of (30) (the market). In the planner's panel the first two are evaluated at $W(a^*)$. Parameters: $\sigma=0.2$, $\bar{z}=10$, $s_{\text{target}}=1$ percent, $s_K=1/3$, $\theta=0.02$, $\phi=1$, $\gamma=2$, $\bar{u}=5$, and $\rho=2.5$ percent.

willing to give up 3.0 percent of consumption permanently.

A cyberattack, ten years at risk, 50 percent chance of an attack each year and 50 percent chance of abandoning the AI improvement. With ten years of prolonged struggle against cyberattacks, the AI improvement lowers welfare by the equivalent of 16.1 percent of consumption permanently. To keep the AI improvement but remove the risk of attacks, people would be willing to give up 24 percent of consumption permanently.

An existential risk bioattack, 1 or 10 percent chance of extinction. With a 1-percent chance of extinction, the AI improvement raises welfare by the equivalent of 4.8 percent of consumption permanently. With a 10-percent chance, it lowers welfare by the equivalent of 29.2 percent. To keep the AI improvement but remove the risk of extinction, people would be willing to give up 5.3 percent of consumption permanently when that risk is 1 percent, and 36 percent when it is 10 percent.

Optimal allocations. The second panel of Table 6 is the planner's allocation.

A cyberattack, one year at risk, 50 percent chance of an attack. In the planner's allocation, the AI improvement raises welfare by the equivalent of 9.7 percent of consumption permanently, against 7.3 percent at the market allocation. The planner reduces the AI sector's share of inputs by 54 percent, from 2 to 0.92 percent. A smaller AI sector is less capable, so the AI productivity factor falls from 10 to 4.6 and the probability of a cyberattack declines from 50 to 23 percent. An attack leaves output 34.0 percent below today's, versus 68.4 percent at the market allocation. If no attack occurs, the benefits from AI are roughly similar at around +10% for both allocations. An AI productivity factor of 4.6 still delivers almost the gain from infinite automation of all machines.

A cyberattack, ten years at risk, 50 percent chance of an attack each year and 50 percent chance of abandoning the AI improvement. In the planner's allocation, the AI improvement raises welfare by the equivalent of 7.5 percent of consumption permanently, against a loss of 16.1 percent at the market allocation. The planner cuts the AI sector's share of inputs by 75 percent, to 0.51 percent. The AI productivity factor falls from 10 to 2.5 and the probability of an attack in each year declines to 12.7 percent. An attack leaves output 3.3 percent below today's. Moving from the market allocation to the planner's raises welfare by the equivalent of 28 percent of consumption permanently. The tax on the AI sector that implements the planner's allocation is 300 percent.

An existential risk bioattack, 1 or 10 percent chance of extinction. In the planner's allocation,

the AI improvement raises welfare by the equivalent of 7.6 percent of consumption permanently when the chance of extinction at the market allocation is 1 percent, and lowers it by only 2.0 percent when that chance is 10 percent. The planner cuts the AI sector's share of inputs by 68 percent at $\bar{p} = 1$ percent, to 0.65 percent, substantially reducing the capability of AI. The AI productivity factor falls from 10 to 3.2, and the probability of extinction falls by two-thirds, from the market allocation's 1 percent to 0.32 percent. In the case in which the probability of extinction at the market allocation is 10 percent, the planner takes even more drastic action, taxing the AI sector at a 536 percent rate to reduce its share of inputs by 84 percent, to 0.32 percent. The AI productivity factor falls from 10 to 1.6, and the probability of extinction falls from $\bar{p} = 10$ percent to 1.6 percent. These actions raise welfare by an amount equivalent to nearly doubling consumption permanently in the market allocation.

Robustness. Coordination among the labs (Corollary 1) closes little of the gap. With one year at risk and an even chance of a cyberattack, coordinated labs reduce the AI sector's share of inputs from 2.0 to 1.92 percent and close 7 percent of the gap between its size in the market and in the planner's allocation. The tax that implements a^* falls from 120 to 118 percent. With ten years at risk, the productivity part of the net external cost is much larger. One more log point of the AI productivity factor makes an attack more damaging and lowers expected welfare by the equivalent of 29.2 percent of consumption in every year, against 4.2 percent with one year at risk. Coordinated labs internalize part of this cost, so coordination then closes 27 percent of the gap. In the bioattack economy coordinated labs expand the sector by a ten-thousandth of a percentage point, because the only cost they internalize is the effect of the AI productivity factor on their own revenue.

The elasticity ϕ matters for the size of the tax. With $\phi = 2$ the taxes are 65, 109, 96, and 157 percent. With $\phi = 1/2$ they are 162, 1,019, 262, and 3,314 percent. When $z(a)$ and $p(a)$ fall with the square root of the sector, the same fall in the probability takes a much smaller AI sector, a share of inputs of 0.77 percent in the first column against 0.92 at $\phi = 1$.

7. Conclusion

In a "weak links" model of production, the gains from automation tend to arrive slowly because output is bottlenecked by the least productive tasks. But the same complementarity makes attacks more harmful and easier to execute. For example, with an elasticity of substitution

across tasks of 0.2, automating tasks worth 1 percent of GDP with infinite productivity raises output by a quarter of a percent, while cutting their productivity tenfold lowers output by 68 percent on impact. In our model, a larger AI sector makes AI models more capable and attacks more likely. The positive productivity spillover is negligible compared with the potential damage, so the market overinvests in AI. Implementing the optimal allocation typically requires large taxes on the AI sector.

In a weak-link economy, the costs and benefits from AI can be extremely asymmetric. Strengthening a chain requires strengthening every link, while cutting a single link breaks it. This asymmetry strikes us as a good description of how AI models may affect the aggregate economy, today and in the near future. Some benefits are likely to appear in the next decade, but the largest, such as curing cancer, require a long string of experiments in the physical world and a long process of regulatory approval. In contrast, models in the next few years may be capable of causing substantial damage. If society does not manage the harms from AI, it may forgo the benefits.

References

Acemoglu, Daron, "The Simple Macroeconomics of AI," Working Paper 32487, National Bureau of Economic Research May 2024.

__ and Pascual Restrepo, "The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment," *American Economic Review*, June 2018, *108* (6), 1488–1542.

__ and Todd Lensman, "Regulating Transformative Technologies," *American Economic Review: Insights*, September 2024, *6* (3), 359–76.

__ , Vasco M. Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi, "The Network Origins of Aggregate Fluctuations," *Econometrica*, 2012, *80* (5), 1977–2016.

Aghion, Philippe and Simon Bunel, "AI and Growth: Where Do We Stand?," June 2024. unpublished manuscript.

__ , Benjamin F. Jones, and Charles I. Jones, "Artificial Intelligence and Economic Growth," in Ajay Agrawal, Joshua Gans, and Avi Goldfarb, eds., *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, 2019, pp. 237–282.

AI Security Institute, "Frontier AI Trends Report," United Kingdom, December 2025.

Anthropic, "Activating AI Safety Level 3 Protections," May 2025. https://www.anthropic.com/news/activating-asl3-protections.

__ , "Project Glasswing: An Initial Update," May 2026. https://www.anthropic.com/research/glasswing-initial-update.

A.P. Moller–Maersk, "Annual Report 2017," Copenhagen 2018.

Aschenbrenner, Leopold, "Existential Risk and Growth," September 2020. Global Priorities Institute Working Paper No. 6-2020.

__ and Philip Trammell, "Existential Risk and Growth," February 2024. Global Priorities Institute at Oxford, manuscript.

ASML Holding N.V., "Annual Report 2025," Veldhoven 2026.

Atalay, Enghin, "How Important Are Sectoral Shocks?," *American Economic Journal: Macroeconomics*, 2017, *9* (4), 254–280.

Autor, David H., Frank Levy, and Richard J. Murnane, "The Skill Content Of Recent Technological Change: An Empirical Exploration," *The Quarterly Journal of Economics*, November 2003, *118* (4), 1279–1333.

Bachmann, Rüdiger, David Rezza Baqaee, Christian Bayer, Moritz Kuhn, Andreas Löschel, Benjamin Moll, Andreas Peichl, Karen Pittel, and Moritz Schularick, "What if? The Macroeconomic and Distributional Effects for Germany of a Stop of Energy Imports from Russia," *Economica*, 2024, *91* (364), 1157–1200.

Baqaee, David Rezza and Emmanuel Farhi, "The Macroeconomic Impact of Microeconomic Shocks: Beyond Hulten's Theorem," *Econometrica*, 2019, *87* (4), 1155–1203.

Barrot, Jean-Noël and Julien Sauvagnat, "Input Specificity and the Propagation of Idiosyncratic Shocks in Production Networks," *Quarterly Journal of Economics*, 2016, *131* (3), 1543–1592.

Baumol, William J., "Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis," *American Economic Review*, June 1967, *57*, 415–426.

Bengio, Yoshua et al., "International AI Safety Report 2026," DSIT research series 2026/001, February 3 2026.

Blanchard, Olivier and Michael Kremer, "Disorganization," *The Quarterly Journal of Economics*, November 1997, *112* (4), 1091–1126.

Boehm, Christoph E., Aaron Flaaen, and Nitya Pandalai-Nayar, "Input Linkages and the Transmission of Shocks: Firm-Level Evidence from the 2011 Tōhoku Earthquake," *Review of Economics and Statistics*, March 2019, *101* (1), 60–75.

Bostrom, Nick, "Existential Risks," *Journal of Evolution and Technology*, March 2002, *9*.

__ , *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014.

Bram, Jason and Lauren Thomas, "Beginning to Gauge Maria's Effect on Puerto Rico's Economy," Liberty Street Economics, Federal Reserve Bank of New York, January 12 2018.

Bundesnetzagentur, "Rückblick: Gasversorgung im Jahr 2022," Bonn, January 6, https://www.bundesnetzagentur.de/DE/Gasversorgung/a_Gasversorgung_2022/start.html 2023.

Carvalho, Vasco M., Makoto Nirei, Yukiko U. Saito, and Alireza Tahbaz-Salehi, "Supply Chain Disruptions: Evidence from the Great East Japan Earthquake," *Quarterly Journal of Economics*, 2021, *136* (2), 1255–1321.

Chirgwin, Richard, "IT 'Heroes' Saved Maersk from NotPetya with Ten-Day Reinstallation Blitz," The Register, January 25 2018.

Corbett, Michael, "Oil Shock of 1973–74," Federal Reserve History, https://www.federalreservehistory.org/essays/oil-shock-of-1973-74 2013.

Cox, Russ, "Timeline of the xz Open Source Attack," https://research.swtch.com/xz-timeline April 2024.

Davidson, Tom, "Could Advanced AI Drive Explosive Economic Growth?," June 2021. Open Philanthropy report.

Demirer, Mert, Leon Musolff, and Liyuan Yang, "Writing Code vs. Shipping Code: Productivity Effects Across Generations of AI Coding Tools," Working Paper 35275, National Bureau of Economic Research May 2026.

Department of Economic Development and Commerce of Puerto Rico, "The Puerto Rico Economic Activity Index (PR-EAI), August 2026," Monthly report and historical data, San Juan, https://www.estadisticas.pr.gov/en-us/productos/indice-de-actividad-economica-coincidente 2026.

ENERCON, "Over 50 per cent of WECs Back Online Following Disruption to Satellite Communication," Press release March 2022. https://w3.windfair.net/wind-energy/pr/40316-enercon-wind-turbine-wind-farm-satellite-communication-converter-disruption-remote-monitoring-scada-ka-sat-russia-ukraine-europe-access.

Energy Institute, "Statistical Review of World Energy," London, as compiled by Our World in Data 2026.

Erdil, Ege and Tamay Besiroglu, "Explosive growth from AI automation: A review of the arguments," 2023.

European Bank for Reconstruction and Development, *Transition Report 1999: Ten Years of Transition*, London: EBRD, 1999.

Fischer, Stanley and Ratna Sahay, "The Transition Economies After Ten Years," April 2000. NBER Working Paper 7664.

Freund, Andres, "Backdoor in Upstream xz/liblzma Leading to SSH Server Compromise," oss-security mailing list, March 29, https://www.openwall.com/lists/oss-security/2024/03/29/4 March 2024.

Fujimoto, Takahiro, "Supply Chain Competitiveness and Robustness: A Lesson from the 2011 Tohoku Earthquake and Supply Chain "Virtual Dualization"," September 2011. MMRC Discussion Paper 362, University of Tokyo.

Gabaix, Xavier, "The Granular Origins of Aggregate Fluctuations," *Econometrica*, May 2011, *79* (3), 733–772.

Garfinkel, Ben and Allan Dafoe, "How Does the Offense-Defense Balance Scale?," *Journal of Strategic Studies*, 2019, *42* (6), 736–763.

Greenberg, Andy, "The Untold Story of NotPetya, the Most Devastating Cyberattack in History," *Wired*, August 22 2018.

Growiec, Jakub and Klaus Prettner, "The Economics of p(doom): Scenarios of Existential Risk and Economic Growth in the Age of Transformative AI," January 2025. unpublished manuscript.

Guerreiro, Joao, Sergio Rebelo, and Pedro Teles, "Regulating Artificial Intelligence," Working Paper 31921, National Bureau of Economic Research November 2023.

Hall, Robert E. and Charles I. Jones, "The Value of Life and the Rise in Health Spending," *Quarterly Journal of Economics*, February 2007, *122* (1), 39–72.

__ , __ , and Peter J. Klenow, "Trading Off Consumption and Covid-19 Deaths," *Quarterly Review*, 2020, *42* (1).

Hamilton, James D., "Causes and Consequences of the Oil Shock of 2007–08," *Brookings Papers on Economic Activity*, 2009, *40* (1), 215–261.

__ , "Historical Oil Shocks," in Randall E. Parker and Robert Whaples, eds., *Routledge Handbook of Major Events in Economic History*, New York: Routledge, 2013, pp. 239–265.

Hendrycks, Dan, Mantas Mazeika, and Thomas Woodside, "An Overview of Catastrophic AI Risks," 2023.

Hulten, Charles R., "Growth Accounting with Intermediate Inputs," *Review of Economic Studies*, 1978, *45* (3), 511–518.

Jones, Charles I., "Intermediate Goods and Weak Links in the Theory of Economic Development," *American Economic Journal: Macroeconomics*, April 2011, *3* (2), 1–28.

__ , "Life and Growth," *Journal of Political Economy*, 2016, *124* (2), 539–578.

__ , "The AI Dilemma: Growth versus Existential Risk," *American Economic Review: Insights*, December 2024, *6* (4), 575–90.

__ , "How much should we spend to reduce A.I.'s existential risk?," March 2025. NBER Working Paper 33602.

__ and Christopher Tonetti, "Past Automation and Future A.I.: How Weak Links Tame the Growth Explosion," 2026. unpublished manuscript, Stanford GSB.

Joy, Bill, "Why the Future Doesn't Need Us," *Wired Magazine*, April 2000, *8* (4).

Kaplan, Greg, Benjamin Moll, and Giovanni L. Violante, "The Great Lockdown and the Big Stimulus: Tracing the Pandemic Possibility Frontier for the U.S.," September 2020. NBER Working Paper 27794.

Kremer, Michael, "The O-Ring Theory of Economic Development," *Quarterly Journal of Economics*, August 1993, *108* (4), 551–576.

Langner, Ralph, "Stuxnet: Dissecting a Cyberwarfare Weapon," *IEEE Security & Privacy*, 2011, *9* (3), 49–51.

Lentzos, Filippa and Gregory D. Koblentz, "Global BioLabs Report 2023," Technical Report, King's College London and Schar School of Policy and Government, George Mason University, February 2023.

Lewis, Patricia, Heather Williams, Benoît Pelopidas, and Sasan Aghlani, "Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy," Technical Report, Chatham House, the Royal Institute of International Affairs, April 2014.

Martin, Ian W. R. and Robert S. Pindyck, "Welfare Costs of Catastrophes: Lost Consumption and Lost Lives," *The Economic Journal*, 08 2020, *131* (634), 946–969.

Martin, Ian W.R. and Robert S. Pindyck, "Averting Catastrophes: The Strange Economics of Scylla and Charybdis," *American Economic Review*, October 2015, *105* (10), 2947–85.

METR, "Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident," With Redwood Research, https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/ August 2026.

Moll, Benjamin, Moritz Schularick, and Georg Zachmann, "The Power of Substitution: The Great German Gas Debate in Retrospect," *Brookings Papers on Economic Activity*, 2023, *54* (2), 395–455.

Murphy, Kevin M. and Robert H. Topel, "The Value of Health and Longevity," *Journal of Political Economy*, October 2006, *114* (5), 871–904.

National Infrastructure Advisory Council, "Addressing the Critical Shortage of Power Transformers to Ensure Reliability of the U.S. Grid," Washington, DC June 2024.

National Research Council, *Terrorism and the Electric Power Delivery System*, Washington, DC: National Academies Press, 2012.

Nielsen, Michael, "How to be a wise optimist about science and technology?," December 2024. https://michaelnotebook.com/optimism/index.html.

Nordhaus, William D., "The Health of Nations: The Contribution of Improved Health to Living Standards," in Kevin M. Murphy and Robert Topel, eds., *Measuring the Gains from Medical Research: An Economic Approach*, Chicago: University of Chicago Press, 2003, pp. 9–40.

__ , "Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth," *American Economic Journal: Macroeconomics*, January 2021, *13* (1), 299–332.

Office for Budget Responsibility, "Commentary on the OBR Coronavirus Reference Scenario," London, April 14 2020.

Office for National Statistics, "GDP Monthly Estimate, UK: April 2020," Statistical bulletin, June 12, https://www.ons.gov.uk/economy/grossdomesticproductgdp/bulletins/gdpmonthlyestimateuk/april2020 2020.

Ord, Toby, *The Precipice: Existential Risk and the Future of Humanity*, Hachette Books, 2020.

Posner, Richard A., *Catastrophe: Risk and Response*, Oxford University Press, 2004.

Puerto Rico Electric Power Authority, "Basic Financial Statements and Required Supplementary Information, Fiscal Year Ended June 30, 2019," San Juan 2019.

Puerto Rico Planning Board, "Informe Económico al Gobernador 2020," San Juan, https://jp.pr.gov/informe-economico-al-gobernador/ 2021.

__ , "Informe Económico al Gobernador 2021," San Juan, https://jp.pr.gov/informe-economico-al-gobernador/ 2022.

Rees, Martin, *Our Final Century*, London: William Heinemann, 2003.

Renesas Electronics, "Renesas Electronics Announces Resumption of Wafer Lines Production at Naka Factory," Press releases of March 28 and June 10, Tokyo 2011.

Rosen, Sherwin, "The Value of Changes in Life Expectancy," *Journal of Risk and Uncertainty*, 1988, *1*, 285–304.

Sagan, Scott D., *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons*, Princeton, NJ: Princeton University Press, 1993.

Statistisches Bundesamt, "Gross Domestic Product Up 1.9% in 2022," Press release No. 020, Wiesbaden, January 13 2023.

Toyota Motor Corporation, "TMC Announces Results for April 2011; TMC Announces Results for May 2011," News releases of May 27 and June 28, Toyota City 2011.

Trammell, Philip and Anton Korinek, "Economic Growth under Transformative AI," Technical Report, National Bureau of Economic Research Working Paper 31815 October 2023.

UK AI Security Institute, "Our Evaluation of Claude Mythos Preview's Cyber Capabilities," April 2026. https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities.

U.S. Department of Energy, "Hurricanes Maria and Irma: Event Summaries," Situation reports, September 21, 2017 to April 4, 2018, https://www.energy.gov/ceser/articles/hurricanes-nate-maria-irma-and-harvey-situation-reports 2018.

U.S. Energy Information Administration, "Annual Energy Review 2011," DOE/EIA-0384(2011), Washington, DC September 2012.

__ , "EIA Electricity Sales Data for Puerto Rico Show Rate of Recovery since Hurricanes," Today in Energy, August 6 2018.

__ , "Puerto Rico Electricity Generation Returned to Pre-2017 Hurricane Levels One Year Later," Today in Energy, November 25 2019.

U.S. Government Accountability Office, "2017 Hurricane Season: Federal Support for Electricity Grid Restoration in the U.S. Virgin Islands and Puerto Rico," GAO-19-296, Washington, DC April 2019.

Varas, Antonio, Raj Varadarajan, Jimmy Goodrich, and Falan Yinug, "Strengthening the Global Semiconductor Supply Chain in an Uncertain Era," Boston Consulting Group and Semiconductor Industry Association April 2021.

Viasat, "KA-SAT Network Cyber Attack Overview," March 30, https://news.viasat.com/blog/corporate/ka-sat-network-cyber-attack-overview 2022.

Yudkowsky, Eliezer et al., "Artificial intelligence as a positive and negative factor in global risk," *Global catastrophic risks*, 2008, *1* (303), 184.

Zeira, Joseph, "Workers, Machines, And Economic Growth," *Quarterly Journal of Economics*, November 1998, *113* (4), 1091–1117.

The appendix material is largely written by Claude and we have not yet edited it for clarity.

A. The Net External Cost of the AI Sector

This appendix derives the net external cost of the AI sector, $\mathcal{E}(a)$, that enters Proposition 4.

The elasticity of output with respect to a shock is the shocked tasks' post-shock cost share (part iii of Proposition 1). More capable models raise consumption when no attack occurs. Under a cyberattack they raise the productivity of the other machine tasks and lower that of the targeted ones, so the net effect is the other tasks' share minus the targeted share. To make this precise, write s_K^{safe} for the machine tasks' cost share when no attack occurs and they run at $z(a)$, $s_{\text{target}}^{\text{cyber}}$ for the targeted tasks' cost share under the attack, and s_K^{cyber} for all machine tasks' cost share under the attack. The other machine tasks then have cost share $s_K^{\text{cyber}} - s_{\text{target}}^{\text{cyber}}$. A log point of the AI productivity factor raises consumption by s_K^{safe} log points when no attack occurs and changes it by $s_K^{\text{cyber}} - 2\, s_{\text{target}}^{\text{cyber}}$ log points under a cyberattack. At the market allocation $s_K^{\text{safe}} = 0.005$ percent, $s_{\text{target}}^{\text{cyber}} = 0.9934$, and $s_K^{\text{cyber}} - s_{\text{target}}^{\text{cyber}} = 3.2 \times 10^{-7}$. The gain is saturated. Under an attack the targeted tasks are nearly all of costs.

One more log point of the AI sector's size raises the AI productivity factor and the probability of an attack by ϕ log points each. In the cyberattack economy a log point of both changes expected welfare in three ways: the attack is more likely, it is more damaging, and the gain when no attack occurs is larger. Dividing the change by the marginal utility of a 1-percent rise in consumption in every state and year puts it in percent of consumption. That marginal utility is

$$\begin{aligned} M_{\text{cyber}}(a) \;&\equiv\; (1-p(a))\, u'(c_{\text{safe}})\, c_{\text{safe}} \\ &\quad +\; p(a)\Big[\left(1-e^{-\rho T}\right) u'(c_{\text{cyber}})\, c_{\text{cyber}} \;+\; e^{-\rho T}\, u'(c_{\text{after}})\, c_{\text{after}}\Big], \qquad \text{(A1)} \\ M_{\text{bio}}(a) \;&\equiv\; (1-p(a))\, u'(c_{\text{safe}})\, c_{\text{safe}}\,, \end{aligned}$$

and with $\gamma = 2$, $u'(c)\, c = 1/c$. We call the result the net external cost of the AI sector,

$$\mathcal{E}(a) \;\equiv\; \mathcal{E}^z(a) + \mathcal{E}^p(a), \qquad \mathcal{E}^z(a) \;\equiv\; -\frac{1}{M(a)}\,\frac{\partial\, \rho W}{\partial \ln z}, \qquad \mathcal{E}^p(a) \;\equiv\; -\frac{1}{M(a)}\,\frac{\partial\, \rho W}{\partial \ln p}, \quad \text{(A2)}$$

where W stands for W_{cyber} or W_{bio} of (23), written as a function of c_0, z, and p, so the derivatives hold c_0 fixed. $\mathcal{E}^z$ is the productivity part of the net external cost and $\mathcal{E}^p$ the probability part.

Let s_{after} be the elasticity of c_{after} with respect to the AI productivity factor, s_K^{safe} when the AI improvement is kept and zero when it is abandoned. They evaluate to

$$\begin{aligned}\mathcal{E}_{\text{cyber}}(a) \;=\; \frac{1}{M_{\text{cyber}}(a)}\Big[& \underbrace{p(a)\,[u(c_{\text{safe}}) - \rho\, W_{\text{attack}}(a)]}_{\text{a more likely attack}} \\ & + \underbrace{p(a)\left(1 - e^{-\rho T}\right)\left(2\, s_{\text{target}}^{\text{cyber}} - s_K^{\text{cyber}}\right) u'(c_{\text{cyber}})\, c_{\text{cyber}}}_{\text{a more damaging attack}} \\ & - \underbrace{\Big[(1 - p(a))\, s_K^{\text{safe}}\, u'(c_{\text{safe}})\, c_{\text{safe}} \;+\; p(a)\, e^{-\rho T}\, s_{\text{after}}\, u'(c_{\text{after}})\, c_{\text{after}}\Big]}_{\text{a larger gain}}\Big], \\ \mathcal{E}_{\text{bio}}(a) \;=\; & \frac{p(a)}{1 - p(a)}\,\frac{u(c_{\text{safe}})}{u'(c_{\text{safe}})\, c_{\text{safe}}} \;-\; s_K^{\text{safe}}(a)\,.\end{aligned} \tag{A3}$$

In the cyberattack economy the first term is the probability part and the other two are the productivity part. At the market allocation the three terms are 3.0, 4.2, and 0.005 percent of consumption with one year at risk and 24.2, 29.2, and 0.002 percent with ten. In the bioattack economy the first term, the probability part, is the odds of extinction times the value of a year of life of Section 5 evaluated at c_{safe}, $u(c_{\text{safe}})/(u'(c_{\text{safe}})\, c_{\text{safe}})$. The second term, the productivity part, is minus the gain from one more log point of the AI productivity factor, s_K^{safe}.

B. Proofs

Proof of Proposition 1. With the allocation frozen, each shocked task's output scales by z, so the shocked group's subaggregate scales by z while the rest is unchanged. Raising (1) to the power $\frac{\sigma-1}{\sigma}$ and applying the share equation (4) to the two groups, $Y(s,z)^{\frac{\sigma-1}{\sigma}} = (1-s)\,Y_0^{\frac{\sigma-1}{\sigma}} + z^{\frac{\sigma-1}{\sigma}} s\, Y_0^{\frac{\sigma-1}{\sigma}}$, which gives the first expression in (5). Write $g(z) \equiv 1 - s + s z^{\frac{\sigma-1}{\sigma}}$, so $G = g^{\frac{\sigma}{\sigma-1}}$. Applying (4) after the shock, the shocked group's share is $s' = s z^{\frac{\sigma-1}{\sigma}}/g$, so $1 - s' = (1-s)/g$ and $G = \left(\frac{1-s}{1-s'}\right)^{\frac{\sigma}{\sigma-1}} = \left(\frac{1-s'}{1-s}\right)^{\frac{\sigma}{1-\sigma}}$, the second expression. For the properties, $\ln G = \frac{\sigma}{\sigma-1}\ln g$. (i) As $z \to \infty$, $z^{\frac{\sigma-1}{\sigma}} \to 0$ since $\frac{\sigma-1}{\sigma} < 0$, so $s' \to 0$ and $G \to (1-s)^{\frac{\sigma}{\sigma-1}}$. (ii) As $z \to 0$, $z^{\frac{\sigma-1}{\sigma}} \to \infty$, so $s' \to 1$, $g \to \infty$, and $G = g^{\frac{\sigma}{\sigma-1}} \to 0$ because $\frac{\sigma}{\sigma-1} < 0$. (iii) With $v \equiv \ln z$, $\frac{d\ln G}{dv} = \frac{s z^{\frac{\sigma-1}{\sigma}}}{g} = s'$, which equals s at $z = 1$. (iv) Differentiating again and using $g - s z^{\frac{\sigma-1}{\sigma}} = 1 - s$, $\frac{d^2 \ln G}{dv^2} = \frac{\sigma-1}{\sigma} \cdot \frac{s z^{\frac{\sigma-1}{\sigma}}(1-s)}{g^2} = \frac{\sigma-1}{\sigma}\, s'(1-s') < 0$ for all z, since $s' = s z^{\frac{\sigma-1}{\sigma}}/g$

and $1 - s' = (1 - s)/g$. ■

Proof of Proposition 2. After the shock, the efficient allocation and equation (3) give output with reallocation, $Y^r(s, z)$, as $Y^r(s, z)/Y_0 = \Psi'/\Psi$, a power mean of order $\sigma - 1$ of 1 and z. Expressing the weights via the share equation (4) yields the first equality in (6), and the share form follows as in the proof of Proposition 1 with $z^{\sigma-1}$ in place of $z^{\frac{\sigma-1}{\sigma}}$. For the last equality, at $\sigma^r = \frac{1}{2-\sigma}$ the two exponents of the gain–loss formula are $\frac{\sigma^r-1}{\sigma^r} = 1 - (2 - \sigma) = \sigma - 1$ and $\frac{\sigma^r}{\sigma^r-1} = \frac{1}{\sigma-1}$, so $G(s, z; \sigma^r)$ reproduces (6). For $\sigma \in (0, 1)$, $\sigma^r > \sigma$ because $1 - \sigma(2 - \sigma) = (1 - \sigma)^2 > 0$. The frozen allocation remains feasible after the shock, so output with reallocation is weakly higher. ■

Proof of Proposition 3. (i) Uniform deaths scale every task's labor by $1 - \bar{d}$, so every task output and hence the aggregate (1) scale by $1 - \bar{d}$ with the allocation frozen. With reallocation, output is $\Psi \cdot (1 - \bar{d})L$ by (3), again $1 - \bar{d}$ times its initial level. In both cases $c = Y/L$ is unchanged. (ii) The deaths scale the labor, and hence the output, of each of the occupation's tasks by z_{target} and of every other task by z_{other}, with the allocation frozen. The aggregate (1) is homogeneous of degree one, so output is z_{other} times the frozen experiment of Proposition 1 with $z = z_{\text{target}}/z_{\text{other}}$ applied to tasks with cost share s. For the comparison, Y'/Y_0 is the share-weighted power mean of order $\frac{\sigma-1}{\sigma} < 1$ of the numbers z_{other} and z_{target}, while $1 - \bar{d} = (1 - s)\, z_{\text{other}} + s\, z_{\text{target}}$ is their arithmetic mean (the power mean of order one) with the same weights. Power means strictly increase in their order when the values differ, so $Y'/Y_0 < 1 - \bar{d}$ for $\omega > 0$. For the last claim, $(Y'/Y_0)^{\frac{\sigma-1}{\sigma}} = s\, z_{\text{target}}^{\frac{\sigma-1}{\sigma}} + (1 - s)\, z_{\text{other}}^{\frac{\sigma-1}{\sigma}}$ has the derivative $\frac{\sigma-1}{\sigma}\, \bar{d}\, (1 - s)\left[z_{\text{other}}^{-1/\sigma} - z_{\text{target}}^{-1/\sigma}\right]$ with respect to ω, which is positive because $\sigma < 1$ and $z_{\text{target}} < z_{\text{other}}$. The exponent $\frac{\sigma}{\sigma-1}$ is negative, so Y'/Y_0 falls as ω rises. ■

Expected welfare (used in Section 6). Write $h(a) \equiv \theta \ln a + (1 - \theta) \ln(1 - a)$, so that by (15) per-capita consumption absent the AI improvement is $c_0(a) = e^{h(a)-h(\theta)}$ under the normalization $c_0(\theta) = 1$. Measure welfare (9) per person alive today. With no attack, everyone has the flow $u(c_{\text{safe}})$ forever, so $W/L = u(c_{\text{safe}})/\rho$. A flow $u(c)$ received for the first T years is worth $\int_0^T e^{-\rho t} u(c)\, dt = \left(1 - e^{-\rho T}\right) u(c)/\rho$, and the same flow received from T on is worth $e^{-\rho T} u(c)/\rho$. With an attack in each of the T years at risk, everyone has the flow $u(c_{\text{cyber}})$ for T years and $u(c_{\text{after}})$ thereafter, which gives W_{attack} in (22). A bioattack leaves nobody alive, so $W/L = 0$.

With one year at risk, and in the bioattack economy, weighting the two states by $1 - p(a)$ and $p(a)$ gives (23). With ten years at risk the draws are independent across years. In each of the ten years expected flow utility is $(1 - p(a))\, u(c_{\text{safe}}) + p(a)\, u(c_{\text{cyber}})$. From year ten on it is $(1 - p(a))\, u(c_{\text{safe}}) + p(a)\, u(c_0)$. Discounting and adding up,

$$\int_0^\infty e^{-\rho t}\, \mathbb{E}\left[u(c_t)\right] dt \;=\; (1 - p(a))\, \frac{u(c_{\text{safe}})}{\rho} \;+\; p(a)\, W_{\text{attack}}(a)\,,$$

with $T = 10$ and $c_{\text{after}} = c_0$ in W_{attack}. The ten annual draws and a single attack that lasts ten years and ends with the AI improvement abandoned give the same probabilities to the same consumption levels at every date. They differ in how outcomes are correlated across years, which expected welfare does not depend on.

Raising consumption in every state and at every date by the factor e^x raises $\rho\, W_{\text{cyber}}$ at the rate $M_{\text{cyber}}(a)$ of (A1) per unit of x at $x = 0$, because each state's flow $u(c)$ responds at the rate $u'(c)\, c$. It raises $\rho\, W_{\text{bio}}$ at the rate $M_{\text{bio}}(a)$ of (A1), because the dead do not respond.

The competitive equilibrium. An attack of either kind multiplies $\bar{Y}$ by a state-specific constant, zero in the case of a bioattack, so $Y_{\text{AI}}/Y_{\text{non-AI}} = a/(1-a)$ in every state and the relative price of the two goods, $\theta Y_{\text{non-AI}}/((1-\theta)Y_{\text{AI}})$, is the same in every state. Factors are hired before the state is known at a rental rate r and a wage w, and each lab takes $z(a)$ and $p(a)$ as given. With state prices q_ς, a lab that employs the fraction a_j of the economy's capital and labor pays $a_j(rK + wL)$ and earns $a_j \sum_\varsigma q_\varsigma\, \theta Y_\varsigma / a$, and a non-AI producer employing the fraction b_j pays $b_j(rK + wL)$ and earns $b_j \sum_\varsigma q_\varsigma (1-\theta) Y_\varsigma/(1-a)$. Zero profits in both sectors give $\theta/a = (1-\theta)/(1-a)$, so $a^{eq} = \theta$, which does not depend on $\bar{p}$, ϕ, T, or $\bar{z}$. With complete markets the first welfare theorem gives the same allocation as maximizing (23) at a fixed AI productivity factor and probability, whose first-order condition is $M(a)\, h'(a) = 0$, since the level $c_0(a)$ scales consumption in every state and $M(a) > 0$.

Proof of Proposition 4. The elasticity of G with respect to the shock is the shocked tasks' post-shock cost share (part iii of Proposition 1). Differentiating (16) and (18) in $\ln z$ gives $s_K^{\text{safe}} = s_K z^{\frac{\sigma-1}{\sigma}}/(1 - s_K + s_K z^{\frac{\sigma-1}{\sigma}})$ for G and $s_K^{\text{cyber}} - 2\, s_{\text{target}}^{\text{cyber}}$ for G_{cyber}, with $s_{\text{target}}^{\text{cyber}} = s_{\text{target}}\, z^{-\frac{\sigma-1}{\sigma}} / G_{\text{cyber}}^{\frac{\sigma-1}{\sigma}}$ and $s_K^{\text{cyber}} - s_{\text{target}}^{\text{cyber}} = (s_K - s_{\text{target}})\, z^{\frac{\sigma-1}{\sigma}} / G_{\text{cyber}}^{\frac{\sigma-1}{\sigma}}$, all evaluated at $z = z(a)$. The allocation enters expected welfare through the level $c_0(a)$, which scales consumption in every state and at every date with $d \ln c / da = h'(a) = \theta/a - (1-\theta)/(1-a)$, and through the two schedules, with

$d \ln z(a)/da = d \ln p(a)/da = \phi/a$. The level has the weight $M(a)$ (the expected-welfare paragraph). In the cyberattack economy the derivative of $\rho\, W_{\text{cyber}} = (1-p)\, u(c_{\text{safe}}) + p\, \rho\, W_{\text{attack}}$ in the log of the probability at a fixed level and AI productivity factor is $-p\,[u(c_{\text{safe}}) - \rho\, W_{\text{attack}}]$. Its derivative in the log of the AI productivity factor at a fixed level and probability is

$$\begin{aligned}(1-p)\, s_K^{\text{safe}}\, u'(c_{\text{safe}})\, c_{\text{safe}} \;+\; p\, e^{-\rho T}\, s_{\text{after}}\, u'(c_{\text{after}})\, c_{\text{after}} \\ -\; p\left(1 - e^{-\rho T}\right)\left(2\, s_{\text{target}}^{\text{cyber}} - s_K^{\text{cyber}}\right) u'(c_{\text{cyber}})\, c_{\text{cyber}}.\end{aligned}$$

Their sum is $-M_{\text{cyber}}(a)\, \mathcal{E}_{\text{cyber}}(a)$, so

$$\rho\, W'_{\text{cyber}}(a) \;=\; M_{\text{cyber}}(a) \left[h'(a) - \frac{\phi}{a}\, \mathcal{E}_{\text{cyber}}(a) \right].$$

In the bioattack economy the two derivatives of $\rho\, W_{\text{bio}} = (1 - p(a))\, u(c_{\text{safe}})$ are $-p(a)\, u(c_{\text{safe}})$ and $(1 - p(a))\, u'(c_{\text{safe}})\, c_{\text{safe}}\, s_K^{\text{safe}}$, whose sum is $-M_{\text{bio}}(a)\, \mathcal{E}_{\text{bio}}(a)$, and the same expression holds with M_{bio} and $\mathcal{E}_{\text{bio}}$. Setting $W' = 0$ gives (25). At $a = \theta$ the left side of (25) is zero, so $W'(\theta)$ has the sign of $-\mathcal{E}(\theta)$ and $a^* < \theta$ exactly when $\mathcal{E}(\theta) > 0$. Multiplying (25) by $a^*(1-a^*)$ gives $\theta - a^* = \phi\, \mathcal{E}(a^*)(1-a^*)$. With an ad valorem tax τ on the AI sector's factor payments, the hiring conditions are $\theta/a = (1+\tau)\mu$ and $(1-\theta)/(1-a) = \mu$, where μ is the shadow cost of the factors, so $1 + \tau = \theta(1-a)/((1-\theta)a)$. Substituting (25) for θ/a^* gives (26). ■

Proof of Corollary 1. Since $\ln Y = \theta \ln Y_{\text{AI}} + (1-\theta) \ln Y_{\text{non-AI}}$ and the two sectors are scaled copies, the AI sector's revenue is the fraction θ of output in every state. Households value revenue in state ς at the state price q_ς, the state's probability and duration weight times $u'(c_\varsigma)$, so at given state prices a proportional change in state-ς output is worth $u'(c_\varsigma)\, c_\varsigma$ times its size, the weights of $M(a)$ and of $\partial \rho W / \partial \ln z$. The sector therefore values a log point of the AI productivity factor at $\theta\, \partial \rho W / \partial \ln z = -\theta\, M(a)\, \mathcal{E}^z(a)$, and a log point of the probability at nothing, because it changes the state prices and not the sector's revenue at those prices; the lives are in no revenue. The sector's hiring condition is the market's with the productivity term added, $\theta/a - (1-\theta)/(1-a) = (\phi/a)\, \theta\, \mathcal{E}^z(a)$, by the steps of the proof of Proposition 4 with $\theta\, \mathcal{E}^z$ in place of $\mathcal{E}$. With the tax τ the condition is $\theta/a - (\phi/a)\, \theta\, \mathcal{E}^z(a) = (1+\tau)(1-\theta)/(1-a)$, so the rate that implements a^* is τ^* less $\theta\, (\phi/a^*)\, \mathcal{E}^z(a^*)\, (1-a^*)/(1-\theta)$. ■

C. Sources for the Historical Episodes

This appendix gives the source of each number in Table 4 and in the paragraphs around it. The cost shares and shocks are our assignments. The two model columns evaluate equations (5) and (6) at $\sigma = 0.2$. A cut of x percent in the quantity of an input, at a fixed allocation, scales the output of the tasks that use it by $1 - x$, so for oil and gas the shock is $z = 1 - x$.[6]

The oil embargo. The Arab members of OPEC announced the embargo on October 17, 1973 and lifted it in March 1974. Hamilton (2013) reports that their production in November was 4.4 million barrels a day below September's, 7.5 percent of world output, and that the shortfall in world production averaged 2.3 million barrels a day over the following six months, which is 3.9 percent. Hamilton (2009) reports 4.0 percent for October to March and divides it by a price change of 41 log points to obtain an elasticity of 0.10. The cost share is U.S. spending on petroleum products, $55.6 billion in 1972 and $65.1 billion in 1973 (U.S. Energy Information Administration, 2012), or 4.5 and 4.7 percent of GDP. It rose to 6.4 percent in 1974. The cost of crude oil alone was 1.9 percent of GDP in 1973. The posted price of oil rose from $2.90 a barrel before the embargo to $11.65 in January 1974 (Corbett, 2013), and world oil consumption fell 1.5 percent in 1974 (Energy Institute, 2026). U.S. real GDP fell 3.1 percent from the fourth quarter of 1973 to the first quarter of 1975 in current data from the Bureau of Economic Analysis. Hamilton sets $38 billion a year of lost U.S. output against $5 billion of oil lost worldwide.

German gas. The abstract of Bachmann et al. (2024) puts the loss from a stop of Russian energy imports "in the range 0.5% to 3% of GDP." In their scenario Russia supplies 55 percent of German gas, other suppliers replace 5, and switching power plants to coal frees 20, which leaves a shortfall of 30 percent. Their appendix calibrates gas at 1.2 percent of gross national expenditure. Germany used 14 percent less gas in 2022 than in 2018–21 (Bundesnetzagentur, 2023), and 20 percent less from July 2022 to March 2023 (Moll et al., 2023). Russia invaded Ukraine on February 24, cut the flow through Nord Stream 1 to 40 percent in mid-June, and stopped it on August 31. Storage reached 100 percent on November 13. Liquefied gas replaced 13 percent of earlier consumption and Norwegian pipeline gas 16. The first estimate of 2022 growth was 1.9 percent (Statistisches Bundesamt, 2023), and after revisions as low as 1.4 the

[6] A companion note, `ClaudePDFs/2026-09-18-EpisodeSources.pdf` in the paper's repository, lists every number with the sentence in its source, the alternative measures, and the data vintages.

current estimate is again 1.9. The forecasts of 8 to 12 percent losses treated gas as an essential input whose loss multiplies along the value chain (Moll et al., 2023).

Covid. Office for National Statistics (2020) reports that U.K. GDP fell 5.8 percent in March 2020 and 20.4 percent in April, 25.1 percent below February. The current vintage says 25.3. Accommodation and food, arts and other services, transport, and education have weights of 2.8, 3.5, 4.2, and 5.8 percent in U.K. value added (Office for Budget Responsibility, 2020), 16.3 in all, and their output fell 52 percent from February to April. These sectors account directly for 8.5 of the 25 points. Retail, construction, and manufacturing, parts of which were also closed by order, account for much of the rest. Definitions of contact-intensive activity range from leisure and hospitality alone, 4 percent of U.S. GDP, to the 26 percent of value added in the social sector of Kaplan et al. (2020).

Puerto Rico. Hurricane Maria made landfall on September 20, 2017. All 1.57 million customers lost power (U.S. Department of Energy, 2018), and 80 percent of the transmission and distribution network was inoperable (U.S. Energy Information Administration, 2019). The share of customers or load without power averaged 91 percent over the first month and 77 percent over the first two, and generation in October was 83 percent below August's. Half of normal load was back by November 15, and the utility declared restoration complete on August 14, 2018, 328 days after landfall (U.S. Government Accountability Office, 2019). The utility's revenue is $3.4 to $3.6 billion in a normal year (Puerto Rico Electric Power Authority, 2019), which is 3.3 percent of Puerto Rico's GDP of $103 billion and 5 percent of its GNP of $69 billion. The corresponding share for the United States is 2.0 percent. Puerto Rico has no monthly or quarterly GDP. Residential and industrial electricity sales in October were 3 and 6 percent of the previous October's (U.S. Energy Information Administration, 2018), and payroll employment fell 4 percent that month, against 30 percent in New Orleans after Katrina (Bram and Thomas, 2018). The Economic Activity Index combines payroll employment, electricity generation, gasoline sales, and cement sales. It was 16, 21, and 17 percent below its August level in October, November, and December, 18 percent on average, and 18, 23, and 18 percent below its level a year earlier (Department of Economic Development and Commerce of Puerto Rico, 2026). Because electricity generation is one of its four components, the index probably overstates the fall in output. Real GNP fell 4.2 percent in the fiscal year from July 2017 to June 2018, after falling 3.2 percent the year before (Puerto Rico Planning Board, 2021, 2022).

The Soviet collapse. Blanchard and Kremer (1997) report that official GDP in 1996 stood below half its 1989 level in ten of the fifteen republics, and that an estimate adjusted for unreported activity puts the decline from 1989 to 1994 at 35 percent. Chaining the growth rates in European Bank for Reconstruction and Development (1999), official GDP in the twelve republics outside the Baltics was 39 percent below 1989 in 1994 and 47 percent below at its trough in 1998. The troughs for Russia, Ukraine, and Kazakhstan were 45, 63, and 39 percent below 1989, and Fischer and Sahay (2000) report the same. Blanchard and Kremer call the breakdown of supply relationships "one explanation" that "has played an important role." They measure a good's complexity by one minus the Herfindahl index of its input shares in Russia's 1990 input-output table, and across 159 goods in nine countries a one-standard-deviation increase in complexity lowers output over 1992–94 by 3 to 8 percent.

The Tōhoku earthquake. Renesas held 40 percent of the world market for automotive microcontrollers (Fujimoto, 2011). Its Naka plant stopped on March 11, 2011 and resumed mass production on June 1 (Renesas Electronics, 2011). The chips were customized to each car model, so automakers could not switch suppliers. Toyota built 129,491 vehicles in Japan in March, 53,823 in April, and 107,437 in May, which is 63, 78, and 54 percent fewer than a year earlier (Toyota Motor Corporation, 2011). It was short of 150 parts in April and 30 in May. For the U.S. affiliates of Japanese multinationals, Boehm et al. (2019) estimate an elasticity of 0.2 between Japanese and other materials and 0.03 between materials and the capital–labor aggregate. Atalay (2017) finds an elasticity across intermediate inputs below 0.2 in annual U.S. industry data and uses 0.1. Over a one-year horizon in Japanese firm-to-firm data, Carvalho et al. (2021) estimate 1.2 across intermediate inputs and 0.6 between primary and intermediate inputs.

NotPetya and Maersk. The attack began on June 27, 2017. Maersk's chairman said at the World Economic Forum in January 2018, "We had to install 4,000 new servers, 45,000 new PCs, 2,500 applications," in ten days, and "we managed 80% of that volume manually" (Chirgwin, 2018; Greenberg, 2018). Seventeen of the firm's 76 port terminals stopped. Online booking for existing customers resumed after two days. The firm reports "losses in the order of USD 250–300m" (A.P. Moller–Maersk, 2018), 1 percent of its revenue of $30.9 billion. The White House put the damage to all firms at "more than $10 billion" (Greenberg, 2018), against a world GDP of $81 trillion.