

Catastrophic Human Misuse of Advanced AI

A Multi-Threshold Assessment of Intent, Capability, Defense, and Recovery

Rodney Malpert, J.D., M.A.

Cornell University

RodMalpert.com

Circulation edition | 29 September 2026

The reference scenario through 2039 assigns approximately 30% to at least one major systemic disaster, 10% to war-scale devastation or worse, and 2% to recovery-impairing civilizational catastrophe materially enabled by deliberate human misuse of advanced AI. The estimates are subjective and scenario-dependent.

Prepared through human and AI collaboration. Specific contributions and source limitations are disclosed at the end.

Contents

Abstract	3
1 The question and the limits of the estimate	4
2 Three severity thresholds	5
3 Intent, access, execution, defense, and recovery	7
4 Biology and the strongest challenge to a low upper-tier estimate	8
5 Cyber, critical infrastructure, and restoration	11
6 Military escalation, finance, and other pathways	12
7 Quantitative synthesis and dependence	14
8 Attribution, the frozen counterfactual, and companion conclusions	15
9 Timing, protection, and the remaining exposure	17
10 Evidence that should change the estimate	17
11 Defensive priorities	18
12 Limitations and conclusion	19
About the Human Author	20
Contributions and disclosure	20
Appendix A Numerical specification and audit	21
References and source status	23

Abstract

Deliberate human misuse of advanced AI can cause severe harm before any autonomous system becomes civilization's dominant strategic actor. This paper assesses three nested outcomes during a reference exposure scenario from 29 September 2026 through 31 December 2039: at least one major systemic disaster, war-scale devastation or worse, and recovery-impairing civilizational catastrophe. The central judgments are approximately 30%, 10%, and 2%, respectively. Broad subjective plausibility ranges are 12%-55%, 3%-25%, and 0.4%-8%. They are not empirically calibrated confidence intervals or outputs of a fitted uncertainty distribution.

The analysis separates serious human intent, dangerous access, practical execution, defensive failure, severity, and recovery. Cyber and infrastructure dominate the lower threshold because digital access can be close to action. Biology dominates the upper tail because a successful event can propagate beyond its initiating location. States have extensive resources but strong ordinary reasons to avoid indiscriminate destruction; small actors have fewer resources and an important extreme tail of motives. AI can change both capability and willingness to act by making dangerous operations appear more controllable.

The 2% upper-tier estimate depends chiefly on effective response, geographic diversity, and surviving industrial recovery capacity. It does not rest on permanent containment of dangerous information. Repeated attempts, common dependencies, or practical biological capability diffusing before defense would challenge that judgment sharply. A transparent cumulative-hazard specification reproduces the central ladder, but its parameters are authored assessments rather than measured rates. The companion AI Risk paper addresses a different conditional superintelligence transition; neither its percentages nor this paper's can be added without matching scope and modeling overlap.

1 The question and the limits of the estimate

1.1 What is being estimated

The primary question is the probability that at least one sufficiently severe event occurs in which deliberate harmful human use of advanced AI makes a material causal contribution. A human principal might be a state, an organization, an individual, or an insider. The principal supplies the strategically relevant harmful objective; an AI may do much of the planning and execution. The eventual scale need not be intended. An attempted bounded operation that escapes its intended limits still belongs here when the harmful objective originated with a human.

Material contribution requires more than AI being present. AI must appreciably increase the chance or severity of the qualifying harm by changing access, planning, execution, speed, scale, defensive failure, or escalation. Ordinary wars, natural pandemics, financial crashes, and infrastructure failures do not enter merely because contemporary institutions use AI. Conversely, an operation need not require superintelligence to enter the estimate. Misuse of capabilities already available in September 2026 is within scope.

The headline is an event-attribution probability, not the percentage of all disasters caused by AI and not the net increase in all-cause catastrophe risk from future AI progress. Useful AI can prevent other disasters while materially enabling a particular harmful event. Those facts can coexist. A secondary counterfactual comparison helps discipline attribution but does not replace the primary question.

1.2 Calendar horizon and the human-control window

The reporting horizon is explicit: 29 September 2026 through the end of 2039. The reference scenario allows substantial improvement and diffusion of dangerous capability while defensive institutions absorb useful AI unevenly. It is a scenario-conditioned assessment, not an estimate averaged over all possible dates of superintelligence arrival. It does not condition on superintelligence ever arriving, and it does not assume that civilization reaches the ASI threshold without earlier harm.

The human-misuse hazard can cease earlier along a particular history if durable protection prevents qualifying misuse, or if humans cease to be the strategically relevant principals. Those are different reasons for a declining hazard. Protection is favorable; supersession by an autonomous regime may be profoundly unfavorable. Neither the arrival of a stronger AI nor the passage of a year establishes protection.

The end of 2039 is a reporting cutoff, not a prediction that misuse becomes impossible then. Later risk remains unestimated unless a continuation scenario is supplied. Earlier protection and exposure continuing through 2045 are shown as separate sensitivities. They change the assumed trajectory, rather than mechanically multiplying a constant annual hazard.

The companion RSI paper describes a conditional March 2028-to-March 2031 takeoff scenario. Its chronology assumes that a civilization-ending misuse or loss-of-control event does not terminate the narrative. It supplies neither the probability nor the date of durable protection. A reference misuse window extending into the late 2030s is an analytical choice; it is not a conclusion deduced from that three-year story. [12]

1.3 Causal ownership and cases outside both headlines

Highly delegated human operations remain human misuse. A case enters the autonomous-risk analysis when an independently formed or substituted AI objective is necessary to produce the qualifying harm. An AI carrying out a human harmful objective recklessly does not become an autonomous takeover simply because it chooses the implementation. Mixed cases require an explicit event record naming the human objective, the AI's contribution, and the escalation mechanism. They should not be silently counted twice or omitted twice.

Ordinary abuse and shocks below Tier 1 are excluded. So are accidents without a deliberate harmful human objective, except where they are consequences of a qualifying deliberate operation. This distinction leaves important unsafe deployment and negligent-use risks outside the present estimate.

Deliberate human use of AI for durable authoritarian domination also needs a separate outcome model. It may be devastating without producing the mortality or recovery failures used here. The companion paper contains a severe-dystopia category, but its numerical scope excludes deliberate human misuse. The existence of that category does not establish that human-directed authoritarian lock-in is quantitatively covered there. Together these two papers therefore remain an incomplete assessment of all AI harms. [11]

2 Three severity thresholds

Recovery capacity is the central distinction. A catastrophe can kill many people and still leave the means of reconstruction. Another event can have lower initial mortality yet destroy the institutions, expertise, infrastructure, and coordination on which recovery depends. Death counts inform severity; they do not replace the causal assessment.

Threshold	Central estimate	Working criterion
Tier 1 or worse	About 30%	A major systemic disaster requiring extraordinary emergency mobilization, with prolonged loss of critical functions and substantial human or economic harm.
Tier 2 or worse	About 10%	War-scale devastation requiring years of reconstruction; substantial recovery capacity survives.
Tier 3	About 2%	A geographically extensive catastrophe destroys or fragments enough industrial and institutional capacity that a credible route to restoring modern civilization is absent.

2.1 Major systemic disaster

Tier 1 requires a national or international emergency beyond ordinary large-scale fraud or a temporary service interruption. A major country, or several countries, loses normal operation of multiple critical systems for weeks or months and suffers substantial mortality or deep economic damage. Extraordinary mobilization and possibly external assistance are required. Core governmental, technical, and productive recovery capacity remains.

An EMP-like national infrastructure emergency is a severity analogy, not a claim that AI itself creates an electromagnetic pulse. A severe multi-sector cyber crisis or a very large deliberately caused outbreak can also qualify. A dramatic benchmark result, one compromised company, or a high financial loss alone does not establish Tier 1.

2.2 War-scale devastation with recovery capacity

Tier 2 involves destruction comparable in reconstruction burden to that suffered by a heavily affected major country in the Second World War, or comparably grave international devastation. Recovery can take years and transform economic and political life, but surviving expertise, outside assistance, production, and coordinating institutions offer a credible route back.

For pandemic calibration, an illustrative mortality band of roughly 50-100 million helps locate the intended severity. Describing this loosely as COVID multiplied by ten is hazardous. WHO estimated approximately 14.9 million excess deaths in 2020-2021 alone, including indirect effects; multiplying that measure by ten already exceeds 100 million. Reported deaths and excess mortality are different baselines. The stated 50-100 million band, together with disruption and reconstruction burden, is the actual illustration here. It is not a mandatory numerical cutoff. [8]

2.3 Recovery-impairing civilizational catastrophe

Tier 3 requires more than a global recession, even an exceptionally deep one. It requires widespread loss or fragmentation of the people, specialized knowledge, productive systems, trusted records, infrastructure, and coordination necessary to restore modern industrial civilization. Broad geographic correlation matters because intact external centers can otherwise supply knowledge, medical assistance, capital, machinery, and personnel.

The classification asks whether there remains a credible industrial recovery pathway over approximately the following decade, rather than inferring permanent collapse from a terrible first year. That diagnostic horizon is not a claim that recovery after ten years is impossible. Where restoration remains genuinely uncertain, classification should remain uncertain too. A rigorous forecast would carry that ambiguity rather than decide it by rhetoric.

Assessment should record mortality and morbidity; duration of critical-function loss; survival of technical expertise and institutions; production, supply chains, and trusted records; and geographic breadth. No single dimension establishes Tier 3. Nor does restoration of a few functions establish recovery if industrial replenishment and institutions remain broken.

2.4 Nested probabilities and the maximum severity reached

Every Tier 3 history also qualifies for Tier 2 or worse and Tier 1 or worse. The central ladder therefore implies 70% with no qualifying event, 20% with a maximum at Tier 1, 8% with a maximum at Tier 2, and 2% reaching Tier 3. The 10% and 2% are contained within 30%; they are not added to it.

The ratios 10/30 and 2/10 are conditionals about the maximum severity observed over the whole reporting period. They do not show that an individual Tier 1 attack has a one-third chance of escalating, or that each Tier 2 event has a one-fifth chance of destroying recovery.

Several separate events can occur in one history. Per-attack escalation would require a different event-level model.

3 Intent, access, execution, defense, and recovery

3.1 A causal chain with changing bottlenecks

Every pathway follows the same scaffold: a serious harmful attempt; access to useful capability and resources; practical execution; failure of preventive or responsive defense; and sufficient severity and recovery impairment. Each stage is conditional on the preceding history, the actor, the time, and the surrounding defensive environment. Multiplication of conditional probabilities is valid without assuming independence; treating casually guessed marginal probabilities as independent is not.

A capability gain has its largest effect when it removes a binding constraint. Better written advice matters less if the actor still lacks practical execution. More reliable execution can matter enormously if other requirements have already become ordinary. Similarly, an attack that succeeds operationally can remain far below Tier 3 when external recovery bases survive. The appropriate description is a changing set of constraints, not one permanent bottleneck.

Intent is also dynamic. A system that makes an operation appear selective, deniable, geographically bounded, or reversible can make a strategic actor more willing to attempt it. The actor can be wrong about those properties. AI can increase catastrophe risk by creating false confidence in controllability without creating a new desire to destroy civilization.

3.2 Actor classes are not interchangeable

States possess extensive technical, financial, intelligence, and organizational resources. Under ordinary conditions they also have strong incentives to avoid indiscriminate destruction: retaliation, reciprocal vulnerability, economic blowback, alliance costs, regime survival, and harm to their own populations. State risk concentrates in wars, acute crises, perceived first-mover opportunities, and situations where leaders believe an operation can be contained.

Organized non-state actors often depend on supporters, sanctuary, financing, and organizational survival. Most are not indiscriminate destruction maximizers. Apocalyptic groups, exterminatory factions, terminal-defeat situations, and internal teams with goals different from leadership belong in the tail. AI can reduce the need for a large organization and make those internal divergences more consequential.

Individuals and small cells have negligible per-person catastrophic intent but a very large possible population. Feasibility can also create intent at the margin by making a formerly imaginary operation seem possible. There is no sound inference from very few historical expert attackers to very few future AI-assisted attempts. The relevant population consists of people with motive, access, persistence, opportunity, and enough practical support; its size must be investigated rather than dismissed.

3.3 Access has several channels

Centralized services can refuse harmful requests, restrict tools, monitor use, and investigate accounts. Their safeguards are consequential. Dangerous access can nevertheless occur through downloadable weights, unauthorized credentials, compromised customer or vendor infrastructure, specialized tools, or poorly controlled deployments. Those channels have different costs and defensive opportunities.

Purchased inference-time computation is a resource dimension across access channels, not a bypass of safeguards by itself. In some cyber evaluations more computation substantially improves the performance of a fixed model. That does not demonstrate the same effect in biology or show that any weak model can be made frontier-level by spending more. Capability depends on the model, the environment, the tools, and the amount of computation together. [2]

Stolen service credentials likewise do not automatically remove the provider's model safeguards. They can supply unauthorized access, conceal identity, or shift costs while monitoring and refusal remain. Theft of model weights is a materially different event. Conflating them overstates what the evidence about credential theft establishes.

3.4 Defense is a system of partially dependent layers

Useful layers include provider safeguards; secure distribution and access; physical and ecosystem controls; and detection, containment, and recovery. A shared identity system, cloud dependency, software component, or flawed defender can cause several apparent layers to fail together. A diagram with multiple boxes is not proof of independent protection.

Defenders can use the same improvements in reasoning, coding, scientific assistance, and coordination as attackers. They may have privileged telemetry, institutional authority, trusted update channels, and existing capacity. They can also face procurement delays, legacy infrastructure, fragmented incentives, and a need to justify action. Offensive access and defensive absorption are different processes; neither side automatically wins.

4 Biology and the strongest challenge to a low upper-tier estimate

4.1 What the evidence establishes

The International AI Safety Report 2026 describes increasingly strong biological assistance but mixed evidence about practical uplift. It cites o3 outperforming 94% of domain experts on a virology-protocol troubleshooting benchmark. This is a benchmark ranking, not a 94% probability that a novice can execute a dangerous operation. Practical work remains different from answering questions, and realistic evaluation is constrained. The report also describes evidence of assistance on acquisition-proxy tasks, so treating all practical uplift as unshown would be too reassuring. [1]

The appropriate inference is that the knowledge and troubleshooting barriers have weakened substantially, while the degree of substitution for tacit skill, experimental judgment, and complete physical execution remains uncertain. Future laboratory assistance and automation

can weaken those remaining constraints. The central estimate must therefore be conditional on their trajectory; it cannot treat today's physical friction as permanent.

4.2 Motives and perceived selectivity

An indiscriminate global biological event generally harms a state's own leadership, economy, military, and population. That limits ordinary strategic intent. Apparent selectivity or controllability can change the decision even when those properties are unreliable. Organized actors face related constraints; small apocalyptic, suicidal, or exterminatory actors may care much less about blowback.

The non-state tail is the most serious challenge to a reassuring calculation. Once AI materially substitutes for scarce expertise, the number of actors capable of persistent attempts can rise. Their motives need not be sophisticated, stable, or rational in the usual strategic sense. Lowering a skill barrier can alter the joint distribution of motive and capability even when the frequency of extreme motives remains unchanged.

4.3 Physical controls are valuable but incomplete

Physical resources, laboratory implementation, and an actual consequential outbreak remain necessary. Screening and access controls can raise cost and expose suspicious activity, but neither information controls nor one procurement chokepoint covers the whole biological threat space.

IBBIS's June 2026 synthesis-map report describes screening commitments concentrated among a subset of providers and gaps involving intermediaries and different jurisdictions. This supports a fragmented-coverage assessment. It does not justify treating a provider-count percentage as the fraction of worldwide synthesis orders screened. Large providers and small intermediaries have very different volumes, and public commitments do not establish independently audited practice. No global coverage percentage is assumed here. [6]

The May 2025 U.S. executive order directed revision or replacement of the earlier synthesis-screening framework. A directive is not evidence that comprehensive implementation already exists. The current legislative status of individual bills is not needed for this estimate, and this paper makes no unverified claim about their passage or enactment. Broader, audited screening remains a substantive defense, without being treated as universal or sufficient. [7]

4.4 Propagation, response, and industrial survival

Biology has a distinctive scaling mechanism: a successful event can propagate without the attacker individually reaching every location. That makes it the leading upper-tail pathway in this assessment. Its actual reach and severity remain dependent on the biological mechanism, the environment, detection, behavior change, medical response, and the continuity of essential systems.

A catastrophic outbreak also mobilizes defense. Public-health response, changes in movement and contact, medical countermeasures, emergency production, and surviving regions can preserve an industrial recovery route. AI may accelerate those responses. It may also improve offensive iteration or misinformation. The deciding variables are delays, reliability,

deployment capacity, and correlated failure, not a general claim that AI favors either offense or defense.

Biomedical success alone is insufficient if electricity, logistics, manufacturing, trusted communication, or governance fails. Conversely, exceptional mortality alone does not establish that those systems cannot be rebuilt. The upper tier requires widespread failure of the response and recovery mechanisms, including outside sources of assistance.

4.5 Central judgments and what they require

The illustrative biological marginals are approximately 8% for Tier 1 or worse, 5% for Tier 2 or worse, and 1.2% for Tier 3 during the reference window. These are judgments, not outputs derived from the 94% benchmark. They make biology less likely than cyber to initiate a qualifying lower-tier emergency but much more important in the upper tail.

An 8% lower-tier judgment requires a fairly small aggregate probability that practical execution and response failure produce a qualifying outbreak before the window ends. It cannot be defended merely by saying extreme intent is rare. It relies on real execution constraints, intervention, and detection retaining substantial force while useful capability diffuses. If those constraints fall early, the lower-tier biological probability should rise sharply.

The 1.2% upper-tier judgment additionally requires a steep severity filter: many biological disasters, including some extraordinarily lethal ones, leave a recovery route. That is plausible but contestable. Historical pandemic recovery does not establish recovery from every future engineered event, and geographic diversity can be less protective when the mechanism is genuinely global.

4.6 Repeated attempts and the concentration of risk

If N attempts each have probability q of causing a qualifying event, and the attempts are independent conditional on a fixed setting, the probability of at least one event is $1 - (1 - q)^N$. The approximation is useful for exposing how a large actor pool can defeat an apparently reassuring per-attempt probability. The independence and fixed-defense assumptions are strong; the example is not an estimate of future attacker numbers or practical success.

Illustrative attempts	Illustrative chance per attempt	Chance of at least one qualifying event
100	0.1%	About 10%
1,000	0.1%	About 63%
10,000	0.01%	About 63%

Real attempts share technical barriers, can learn from one another, and trigger defensive responses. A global breakthrough can raise their common success probability; an effective intervention can lower it. Conditional on a shared capability-and-defense regime, many opportunities remain consequential even when attempts are not independent unconditionally. Perfect correlation and complete independence are both poor default assumptions.

There is a second stress test. Within the reference biological history estimates, $1.2/8 = 15\%$ is the ratio of upper-tier to lower-tier risk. If the biological Tier 1-or-worse probability rose to 50% while that history-level ratio remained 15%, biological Tier 3 alone would be 7.5%. The aggregate 2% headline would then fail. Holding the ratio fixed is illustrative, not a forecast: practical capability and defensive weakness might change both quantities together.

This is why the central estimate is fragile to broad practical biological uplift. It is not a claim that a civilization-breaking attack remains technically remote for every conceivable system, and it should not survive decisive evidence that many determined actors can reliably cross the physical stage before response improves.

5 Cyber, critical infrastructure, and restoration

5.1 Capability is advancing and permission matters

The UK AI Security Institute's multi-step evaluations found average progress on a 32-step corporate-network range rising from 1.7 steps for an August 2024 model to 9.8 for a February 2026 model at a fixed computation budget. Increasing the budget yielded gains of up to 59% within the tested range. The industrial-control range was harder. Neither range included active defenders that could block or slow the agent. The result establishes improving raw capability under those conditions, not the probability of a national disaster. [2]

A July 2026 joint UK-US assessment found Kimi K3 behind the leading closed frontier on preliminary cyber tests, with safeguards that did not prevent offensive assistance during evaluation. The report distinguishes selected capability testing from protected public deployments and notes limitations of its simulated network range. A fixed one-year lag between all open and closed capabilities is not assumed. Threshold crossing matters more than an average lag. [3]

OpenAI rates GPT-6 Astra at its Critical cybersecurity threshold, conditional on appropriate tools and access, while describing strengthened deployment protections. Anthropic's September 2026 report describes deeper orchestration and misuse of customer credentials; it explicitly says the cited campaigns did not compromise Anthropic's own systems. Unauthorized service access and theft of frontier weights are different risks. Provider reports are useful directional evidence, with selection and attribution limitations. [4,5]

5.2 From intrusion to a systemic emergency

Cyber operations can move rapidly from access to digital action, reducing the skill and team requirements for disruption. Individuals, criminal organizations, and states can all participate, although their motives and targets differ. Theft or extortion that never causes prolonged critical-function loss remains below the threshold.

Infrastructure dependencies make a systemic emergency plausible: energy, communications, water, transport, health services, and digital coordination support one another. CISA's dependency framework is evidence for analyzing these relationships, not proof that every intrusion triggers a cascade. Operators differ in technology, geography, manual fallback, staffing, and recovery capacity. [9]

Tier 1 can follow from a concentrated national disruption without eliminating external repair capacity. Tier 2 needs much greater physical or sustained economic devastation. Tier 3 requires broad correlated failure that persists while defenders isolate networks, restore trusted systems, replace equipment, mobilize personnel, and draw on outside support.

5.3 The restoration contest is a separate capability

An attacker able to gain access does not thereby have the ability to prevent recovery. That latter contest involves persistence, shared dependencies, physical damage, disruption of replacement supplies, loss of trusted backups, and weaknesses in restoration institutions. A test that ends at intrusion or data theft does not observe it.

Common dependencies are the principal upward mechanism. A shared cloud, identity service, software component, communications network, or automated defender can synchronize failures. Diversity, segmentation, offline control, independent recovery records, workforce continuity, and practiced restoration are downward mechanisms. New defensive AI can strengthen them, but a single widely deployed defender may also introduce a common failure mode.

5.4 Central judgments

The illustrative cyber and infrastructure marginals are approximately 15% for Tier 1 or worse, 2.5% for Tier 2 or worse, and 0.35% for Tier 3. The lower-tier weight reflects improving delegation and the closeness of digital access to execution. The steep upper-tier reduction reflects the further requirements for extensive physical harm and failure of restoration.

These weights should rise substantially after credible evidence of sustained disruption across heterogeneous critical organizations against active defenders, especially if restoration repeatedly fails. More impressive intrusion scores alone chiefly update the earlier stages. Successful restoration under realistic adversarial pressure would supply more relevant reassurance than a general claim that the systems are secure.

6 Military escalation, finance, and other pathways

6.1 Military and strategic escalation

The military pathway concerns deliberate harmful human operations in which AI materially changes initiation, targeting, operational success, attribution, warning, or escalation. It does not count the entire background probability of great-power war. Nor does every military use of AI constitute misuse. An AI-assisted coercive or offensive action can enter even if the ultimate catastrophe was unintended.

AI can change incentives by increasing confidence in surveillance, targeting, defense, or operational control. It can also compress decision time or make actors fear that delay gives an adversary an overwhelming advantage. These mechanisms coexist with potential improvements in information and verification; the net effect is not uniformly adverse.

SIPRI's February 2026 survey describes signs of Chinese and U.S. AI integration around nuclear-related systems and limited transparency in several national postures. It records expressed support for human control over nuclear use without treating that declaration as

proof of implementation. The uncertainty is strategic and institutional as well as technical. [10]

The illustrative military marginals are approximately 5%, 2.5%, and 0.5% across the three thresholds. State resources make upper-tier outcomes possible, while retaliation and regime survival ordinarily discourage their deliberate pursuit. Severe crises, misperception, and false confidence in bounded action create the tail. The most relevant safeguards preserve decision time, independent warning channels, survivable deterrence, and meaningful human deliberation.

6.2 Finance, identity, and trusted records

Financial systems can suffer large AI-enabled fraud, cyber disruption, manipulation, or identity compromise. A qualifying crisis requires substantial loss of critical functions and economic harm, rather than merely a large theft or volatile asset prices. Trusted records, payments, and settlement are also recovery substrates for every other pathway.

Reconciliation, independent records, liquidity support, emergency legal authority, and degraded-mode operation can support recovery. Such defenses may fail when infrastructure and governance fail simultaneously. The CPMI-IOSCO cyber-resilience guidance treats prevention, response, and recovery as system-level responsibilities; the existence of guidance does not demonstrate adequate deployment. [13]

The illustrative finance-as-initiator marginals are approximately 3%, 0.5%, and 0.1%. Finance receives a small standalone upper-tier weight because its gravest consequences usually require a wider physical or institutional breakdown. When it amplifies another pathway, it is not counted as a separate independent catastrophe.

6.3 Other physical harms and capability transfer

Chemical, radiological, and AI-enabled conventional or autonomous-system harms form a residual pathway with illustrative marginals of 3%, 0.4%, and 0.05%. Material access and physical scale generally constrain these routes more strongly than digital execution. A remote possibility of an extreme combination is retained rather than assuming that every residual route is harmless.

Model theft, credential compromise, illicit capability transfer, and failure of access governance are enabling mechanisms. They can activate several pathways and must not become an extra row of catastrophe probability added to each one. Similarly, a biological emergency with infrastructure and financial consequences is a compound history, not three independent draws from separate domains.

7 Quantitative synthesis and dependence

7.1 What the numbers represent

The estimates synthesize empirical capability and incident evidence with authored judgments about future intent, access, execution, defense, severity, and recovery. No observed sample identifies the probability of civilization-scale AI misuse. The numerical specification is a transparent predictive scenario, not a statistically fitted posterior. Calling it Bayesian in spirit does not create likelihoods that have not been estimated.

Illustrative pathway marginal	Tier 1 or worse	Tier 2 or worse	Tier 3
Cyber and critical infrastructure	15%	2.5%	0.35%
Biology	8%	5%	1.2%
Military escalation	5%	2.5%	0.5%
Finance as initiator	3%	0.5%	0.1%
Other physical harms	3%	0.4%	0.05%
Combined central judgment	About 30%	About 10%	About 2%

These pathway figures are overlapping marginal judgments, not mutually exclusive causal shares. A history can contain more than one qualifying pathway, and a compound event can have more than one material cause. Their sums are 34%, 10.9%, and 2.2%. Independence would give unions of approximately 30.1%, 10.5%, and 2.2%. Those calculations are diagnostics, not the definition of the final estimates.

The central totals lie between the largest marginal and the sum of marginals at every tier, as probability coherence requires. They are close to the independence diagnostic but do not establish independence. A causal-initiation decomposition would instead allocate each qualifying compound history to one defined initiating category, with shares adding to the combined probability. This paper does not pretend that its overlapping marginals are that decomposition.

7.2 Dependence can change both frequency and severity

At fixed marginals, overlap can reduce the union relative to an independence calculation. But correlation is not a general downward correction to catastrophe risk. A common shock can increase the marginals themselves, and several failures together can turn individually recoverable disruptions into Tier 3. Pairwise positive association alone does not determine a many-event union.

A complete assessment would condition on shared regimes: rapid diffusion with weak response; effective defense with heterogeneous systems; strategic crisis; or another common environment. It would then model attack opportunities and response within each regime and average over the regime uncertainty. The central assessment compresses those possibilities. The absence of an estimated regime distribution is an important structural limitation.

7.3 A coherent aggregate hazard specification

Appendix A gives a simple specification that reproduces 30%-10%-2% and makes the horizon explicit. For a threshold k , the cumulative hazard H_k produces $P_k = 1 - \exp(-H_k)$. The central totals imply $H_1 = 0.35667$, $H_2 = 0.10536$, and $H_3 = 0.02020$. These parameters summarize the assessed opportunity, capability, and response history; they are not measured annual attack rates.

A nested marked-event construction uses nonnegative hazards for Tier 1-only events, Tier 2-only events, and Tier 3 events. Higher-severity events count automatically toward lower thresholds. It gives coherent probabilities for the maximum observed severity even when several events occur. This construction verifies bookkeeping; it neither proves the rates nor reproduces the full strategic response to every event.

7.4 Uncertainty and scenarios

Scenario judgment	Tier 1 or worse	Tier 2 or worse	Tier 3
Faster defensive absorption and stronger recovery	20%	5%	0.5%
Reference scenario through 2039	30%	10%	2%
Rapid dangerous diffusion and weaker response	50%	25%	8%

These are coherent alternative judgments about several inputs together. They are not quantiles of a sampled distribution and not estimates of a particular policy's effectiveness. The broad plausibility ranges of 12%-55%, 3%-25%, and 0.4%-8% summarize additional subjective uncertainty. Their endpoints are not simultaneous scenarios, and outcomes beyond them remain possible.

The upper tier is particularly uncertain because end-to-end practical capability and the correlation of recovery failures are poorly identified. The present 2% judgment deserves less confidence than its decimal notation might suggest. Its purpose is to expose a low-single-digit central assessment and the mechanisms that would invalidate it, not to certify a small risk.

8 Attribution, the frozen counterfactual, and companion conclusions

8.1 The secondary counterfactual

The proposed comparison freezes AI capability and accessible deployment near January 2026 levels while attempting to match non-AI conditions. It should freeze both offensive and defensive AI assistance, not preserve all future defensive improvements while removing only offense. The exercise is difficult because institutions, actors, and crisis behavior adapt to the technology; their endogenous differences cannot simply be assumed away.

For a matched severity event, the net risk increment is the probability under the future-AI scenario minus the probability under the frozen-AI scenario. That difference can be smaller than the probability of an event materially enabled by AI, and can even be negative if prevention dominates elsewhere. Neither quantity is the other by definition.

This paper does not estimate a defensible numerical increment because it lacks a matched model for the frozen world. It uses the counterfactual to test causal claims, especially military ones. If AI changes neither the likelihood nor the severity of a conflict under a credible paired-world assessment, the conflict cannot be attributed to future AI progress. Existing harmful AI capability can nevertheless remain within the primary misuse estimand.

8.2 Why the two risk papers cannot be added

The companion AI Risk paper assigns approximately 25% to destructive extinction and 27% to enduring severe dystopia, giving about 52% combined; any modeled severe harm is about 57%. It conditions on civilization-relevant superintelligence arriving under a rapid-transition scenario, beginning before an in-scope severe outcome. It uses roughly 25% global mortality or comparably deep breakdown as its extreme-event severity anchor. [11]

This paper begins now, does not require ASI, and reports national as well as global disasters through 2039. Its Tier 3 concerns impaired industrial recovery and is neither extinction nor the companion's broader dystopia category. Even the event labels that sound similar have different definitions. The probability of their union requires a common capability history, horizon, severity definition, and treatment of mixed cases.

For genuinely matched events A and M, $P(A \text{ or } M) = P(A) + P(M) - P(A \text{ and } M)$. Using $P(A) + P(\text{not } A) P(M \text{ given not } A)$ is equivalent. The 30% headline here is not $P(M \text{ given not } A)$. Early misuse can weaken defense, change public support for control, accelerate authoritarian responses, or prevent the later transition entirely. Favorable AI protection can reduce misuse. Those dependencies are part of the joint question, not small optional corrections.

8.3 The moral-transformative route is a different issue

The companion paper examines welfare-centered ethics and the possibility of a system favoring transformation or replacement. It distinguishes conscious continuation of existing persons from preservation of records, coercive transformation, and destruction followed by replacement. That route is not imported as another human-misuse probability merely because a human once endorsed the ethical theory.

If a human deliberately commands a destructive operation, the principal boundary may place the case here; if an AI independently adopts the decisive harmful objective, it belongs in the autonomous analysis. Consent, continuity, and the actual causal objective determine the classification. A genuinely continuous voluntary enhancement is not automatically catastrophe in either paper.

9 Timing, protection, and the remaining exposure

Protection or exposure scenario	Tier 1 or worse	Tier 2 or worse	Tier 3
Effective protection by about 2032	About 14%	About 4%	About 0.8%
Reference exposure through 2039	About 30%	About 10%	About 2%
Exposure continues through 2045	About 40%	About 15%	About 3.5%

The early-protection row changes the assumed defensive trajectory through the reporting horizon. The longer-exposure row adds a continuation scenario. Neither is a linear rescaling of calendar years. Capability, access, adaptation, and restoration evolve together, and lower-threshold events may occur before dangerous capability becomes widely diffused.

Protection would require evidence that relevant actors cannot readily obtain and execute a qualifying operation and that failures are detected, contained, or restored before crossing the threshold. One well-aligned frontier system is insufficient when other systems, deployments, jurisdictions, and physical resources remain accessible.

The longer window also need not be uniformly worse at every annual point. AI-assisted defenses can improve faster than attackers; public alarm can produce reforms; a dangerous early event can reveal vulnerabilities. An early event can instead disable defense or prompt risky emergency centralization. A proper history-dependent model would allow both responses rather than presuming a benign closure date.

10 Evidence that should change the estimate

Update rules should distinguish capability, dangerous access, practical execution, and recovery. A source that changes one stage does not automatically establish the others. Missing realistic evidence sometimes reflects legal or ethical limits on testing; it should widen uncertainty rather than automatically count as failure.

Observation and review date	Upward implication	Downward implication
Practical biological uplift by end 2030	Reliable substitution for scarce expertise across realistic workflows; raise lower tiers and reassess response failure.	Knowledge gains do not transfer; execution and detection remain substantial constraints.
Uncontrolled access by end 2029	Systems crossing dangerous task thresholds become affordable outside effective provider governance.	Capability gaps, security, or operating costs keep practical dangerous access limited.
Cyber defense and restoration by end 2028	Agents sustain critical disruption against active defenders and repeatedly frustrate restoration.	Realistic exercises demonstrate rapid containment and restoration despite strong agents.

Observation and review date	Upward implication	Downward implication
Screening and surveillance annually	Audited implementation remains fragmented while practical offensive support improves.	Broad, verifiable coverage and faster detection reduce actual execution and response delays.
Strategic decision processes annually	AI compresses warning and commitment decisions or creates false confidence in first-mover options.	Independent verification, deliberation, and crisis communication remain effective.

The dates are review deadlines, not predictions of catastrophic incidents. Failure to observe a major incident by end 2028 is only reassuring to the extent that exposure and detection were sufficient to make one visible. A quiet period with little practical deployment is weaker evidence than repeated demanding defensive tests.

Screening should be assessed by relevant order volume, product type, customer checks, geographical reach, and audit quality. Cyber capability should be assessed with active defense, heterogeneous organizations, realistic permissions, and restoration. Biological uplift should be tested in lawful, safe settings that actually measure substitution for practical skill rather than simply repeat difficult questions.

11 Defensive priorities

The framework favors defenses at the stages where practical risk can still be interrupted. Provider safeguards and access security can suppress casual misuse, identify patterns, and buy time. They should be maintained even when long-run diffusion makes them incomplete. Distribution policy, secure weights, credentials, deployment permissions, and infrastructure security address different channels.

For biology, the priorities are broad and audited synthesis and customer screening, laboratory biosecurity, surveillance, diagnostics, medical response, and continuity of the infrastructure needed to deploy countermeasures. Screening is valuable where it covers a relevant bottleneck. It is not a guarantee against every biological route or a reason to neglect response capacity.

For cyber and infrastructure, the priorities are segmentation, diversity, protected recovery records, tested offline or manual operation, rapid patching, continuity of personnel and supplies, and exercises that measure service restoration. Reducing common dependencies is especially valuable because it can limit the geographic correlation needed for Tier 3. Defender quality and defender diversity are separate requirements.

For military escalation, the priority is usable decision time. Human control must permit independent verification and deliberation rather than become a signature on an automated recommendation. Stable warning channels, credible survivability, and communication can reduce pressure to act on ambiguous information at machine speed.

For finance and trusted records, independent reconciliation and recovery data matter because they preserve coordination after other shocks. Backstops should be tested under loss of

communications or compromised identity rather than assume the ordinary digital environment remains available.

These priorities follow from the causal framework; this paper has not estimated their costs, comparative effectiveness, or optimal allocation. A next policy analysis should compare feasible interventions against specific bottlenecks and test whether their dependencies undermine the intended benefit.

12 Limitations and conclusion

There is no direct historical frequency of civilization-scale AI-enabled misuse. Partial reference classes inform motives, severity, and institutional response, but future capability can alter their relevance. Provider reports are selected observations from systems that can detect and report misuse. They are neither comprehensive global incident counts nor independent replications of every claim.

The central hazards, pathway marginals, recovery judgments, and exposure scenarios are authored assessments. Their arithmetic is reproducible; their empirical calibration is not established. The ranges are plausibility judgments, and the model does not integrate every possible error in its structure. A more elaborate calculation would not remove those limits unless it supplied better evidence about the decisive stages.

Recovery deserves particular caution. Historical restoration after pandemics and wars provides reasons to reject easy civilizational-collapse claims. It does not establish that modern systems withstand a globally correlated novel mechanism, repeated campaigns, or loss of the resources needed for reconstruction. The 2% central judgment is most vulnerable where practical biological uplift, broad access, and weak response coincide.

The conclusion is therefore a conditional severity ladder, not reassurance from a small upper-tier percentage. In the reference scenario through 2039, the assessment is about 30% for at least one major systemic disaster, 10% for war-scale devastation or worse, and 2% for recovery-impairing civilizational catastrophe materially enabled by deliberate human misuse. Cyber leads the lower threshold; biology leads the upper tail.

The estimate should change when evidence changes the joint distribution of motives, access, execution, defense, and recovery. A low probability per individual cannot justify complacency if practical capability reaches a very large actor pool. A frightening model score cannot establish catastrophe if execution and restoration remain decisive. Making those assumptions visible is the paper's principal contribution.

About the Human Author

Rodney Malpert holds a J.D. and M.A. from Cornell University, completed all Ph.D. requirements in Government except the dissertation, and undertook substantial Cornell graduate coursework in analytic philosophy. He practiced law for many years in Silicon Valley and taught law as an adjunct professor at Southern Methodist University and Arizona State University. He achieved the rank of Senior Master in chess with the U.S. Chess Federation. His current research examines AI risk, strategic behavior, and the capacity of institutions to preserve human agency and recover from severe technological shocks.

Website and current papers: RodMalpert.com

Contributions and disclosure

Rodney Malpert established the project's scope and required the distinction between deliberate human misuse and autonomous loss of control. He emphasized the intermediate human-control window, the expansion of the small-actor tail, weak global control of biological access, and recovery capacity as the demanding upper-tier criterion. He subsequently required lower severity thresholds so that a low civilizational-collapse estimate would not conceal more probable national or war-scale disasters. He also insisted that the main estimate address AI-enabled events rather than only the increment from future AI progress.

ChatGPT supplied literature retrieval, structural analysis, probability synthesis, and drafting. Earlier work used GPT-5.6 and GPT-6.0, as documented in the supplied project material. Claude contributed structural redrafting, criticism of attribution and diffusion assumptions, and source review. Their judgments are not independent votes that can be averaged as evidence; the systems can share sources and failure modes.

For this circulation edition, ChatGPT rechecked the principal contemporary capability claims against primary sources, clarified the fixed reporting horizon and functional stopping rule, distinguished overlapping pathway marginals from causal shares, corrected the interpretation of the tier ratios, removed an unsupported screening-coverage percentage and unverified legislative-status claims, specified the aggregate hazard calculation, and added repeated-attempt and recovery stress tests. It also reconciled both risk papers and their executive summaries. The central 30%-10%-2% assessment remains an evidence-informed proposal for criticism, not a numerical conclusion attributed to a cited institution.

The human author is responsible for circulation. The presence of AI assistance and extensive sources does not constitute peer review or independent validation. The record of contribution is intended to make the division of labor and the origin of probability judgments inspectable.

Appendix A Numerical specification and audit

A1 Central nested hazards

Define H1, H2, and H3 as cumulative hazards of qualifying events at or above the three thresholds. The central scenario uses $H1 = -\ln(0.70)$, $H2 = -\ln(0.90)$, and $H3 = -\ln(0.98)$. Under a marked Poisson construction conditional on this reference setting, use mutually exclusive event severities with hazards H1 - H2, H2 - H3, and H3. All are nonnegative, so event counts and maximum-severity outcomes are coherent.

Quantity	Cumulative hazard	Implied cumulative probability
Tier 1 or worse	0.35667494	30%
Tier 2 or worse	0.10536052	10%
Tier 3	0.02020271	2%
Tier 1-only event process	0.25131443	Not a maximum-severity outcome probability
Tier 2-only event process	0.08515781	Not a maximum-severity outcome probability

The maximum observed severity is Tier 1 only with probability $\exp(-H2) - \exp(-H1) = 20\%$; Tier 2 only with probability $\exp(-H3) - \exp(-H2) = 8\%$; and Tier 3 with probability $1 - \exp(-H3) = 2\%$. No qualifying event has probability $\exp(-H1) = 70\%$. These sum to 100%. Event-process probabilities differ from maximum-severity probabilities because more than one event can occur.

A2 A transparent time allocation

Reference exposure interval	H1 increment	H2 increment	H3 increment
29 September 2026 through 2029	0.07000000	0.02000000	0.00350000
2030 through 2034	0.14000000	0.04000000	0.00700000
2035 through 2039	0.14667494	0.04536052	0.00970271

Each interval preserves H1 greater than or equal to H2 greater than or equal to H3. The increments sum to the central totals. This allocates exposure to later diffusion while allowing continuing defense; it does not infer a smooth annual trend or a precise catastrophe date. A different time allocation can give the same headline while implying different near-term priorities. The allocation is an elicitation device, not new empirical evidence.

A3 Checks and a reproduction rule

Reproduction needs only the three total hazards and $P = 1 - \exp(-H)$. Summing the interval hazards must return those totals; the maximum-severity probabilities must be nonnegative and sum to one; and every scenario must satisfy $P1$ greater than or equal to $P2$ greater than or equal to $P3$. The independence diagnostic for overlapping pathway marginals is 1 minus the product of $(1 - p_d)$. It must not replace the separately assessed aggregate.

For timing or defense scenarios, apply the same conversion to the stated target probabilities. Effective protection by 2032 corresponds to total hazards $-\ln(0.86)$, $-\ln(0.96)$, and $-\ln(0.992)$. Exposure through 2045 corresponds to $-\ln(0.60)$, $-\ln(0.85)$, and $-\ln(0.965)$. These are alternative trajectory judgments. They are not obtained merely by truncating or extending the reference increments.

The model does not assume real-world attacks are an unconditional homogeneous Poisson process. Regime uncertainty, learning, clustering, response, and common failures can invalidate that approximation. The construction makes the central forecast's accounting explicit while leaving those structural limitations visible. A next-generation model should represent the shared capability-and-defense regime and update it after serious incidents.

References and source status

The numbered sources support mechanisms, observations, and definitions. None supplies the headline probabilities or the hazard inputs. Contemporary claims were checked for this edition against the primary pages listed below; legal implementation and global screening coverage are not inferred from an announced policy or a provider count.

- [1] International AI Safety Report. International AI Safety Report 2026, February 2026. Biological assistance, evaluation limitations, open-weight safeguards, and the distinction between capability and real-world uplift. [Primary report](#).
- [2] UK AI Security Institute. How do frontier AI agents perform in multi-step cyber-attack scenarios, March 2026. Capability evaluations without active defenders; inference-budget and industrial-control limitations. [Evaluation](#).
- [3] UK AI Security Institute and U.S. Center for AI Standards and Innovation. Preliminary Assessment of Kimi K3's Cyber Capabilities, July 2026. Selected preliminary tests and deployment-safeguard qualifications. [Joint assessment](#).
- [4] OpenAI. Safety overview GPT-6 Astra, 3 September 2026. Developer assessment of Critical cyber capability with tools and access; strengthened safeguards and adversarial-evaluation limitations. [Safety overview](#).
- [5] Anthropic. Countering misuse of AI September 2026, 10 September 2026. Provider investigations, deeper orchestration, customer credential theft, and stated limits on observed harm and autonomy. [Threat report](#).
- [6] IBBIS. Major Updates to the Global DNA Synthesis Map, 11 June 2026. Dated assessment of screening commitments, intermediaries, and policy gaps. Provider counts are not order-volume coverage. [Map report announcement](#). See also [Common Mechanism](#).
- [7] The White House. Improving the Safety and Security of Biological Research, 5 May 2025, Executive Order 14292. Directive concerning the synthesis-screening framework; not proof of subsequent implementation. [Executive order](#).
- [8] World Health Organization. 14.9 million excess deaths associated with the COVID-19 pandemic in 2020 and 2021, 5 May 2022. Severity calibration and reported-versus-excess mortality distinction. [WHO estimate](#).
- [9] Cybersecurity and Infrastructure Security Agency. Infrastructure Dependency Primer. Conceptual account of dependencies and resilience, not an AI catastrophe estimate. [Primer](#).
- [10] Saalman, Lora. AI in Chinese, Indian and US Nuclear Postures, Norms and Systems. SIPRI, February 2026. DOI 10.55163/QZBL7397. [Publication](#).
- [11] Malpert, Rodney, with ChatGPT and critical AI review. AI Risk Extinction, Severe Dystopia, and Extreme Catastrophe. Circulation edition, 29 September 2026. Separate conditional transition model: approximately 52% extinction or enduring dystopia and 57% any modeled severe harm. Not a directly combinable probability.
- [12] Malpert, Rodney, with ChatGPT and review by Claude. AI Takeoff The First Thirty-Six Months of Recursive Self-Improvement RSI. September 2026 circulation scenario. A conditional March 2028-to-March 2031 chronology, not a forecast of durable protection.
- [13] Committee on Payments and Market Infrastructures and International Organization of Securities Commissions. Guidance on cyber resilience for financial market infrastructures. Bank for International Settlements, 29 June 2016. [Guidance](#).
- [14] Fearon, James D. Rationalist Explanations for War. International Organization 49(3), 379-414, 1995. DOI 10.1017/S0020818300033324. Bargaining and commitment mechanisms; not an AI event frequency.
- [15] Kydd, Andrew H., and Barbara F. Walter. The Strategies of Terrorism. International Security 31(1), 49-80, 2006. DOI 10.1162/isec.2006.31.1.49. Strategic motive framework; not evidence that every actor is instrumentally rational.
- [16] Shapiro, Jacob N. The Terrorist's Dilemma Managing Violent Covert Organizations. Princeton University Press, 2013. Organizational control and principal-agent constraints.
- [17] Brundage, Miles, and colleagues. The Malicious Use of Artificial Intelligence Forecasting, Prevention, and Mitigation. 2018. [Foundational analysis](#). Its threat framework informs the scaffold; present capability claims rely on newer sources.