

Catastrophic Human Misuse of Advanced AI

Intent, capability, defense, and recovery

Rodney Malpert, J.D., M.A. (Cornell University) | 29 September 2026

RodMalpert.com

The question and the answer

How likely is deliberate harmful human use of advanced AI to materially enable a major disaster before durable protection controls that risk? The reference exposure scenario runs from 29 September 2026 through the end of 2039. It allows improving and diffusing capabilities alongside uneven defensive absorption. It does not require superintelligence to arrive, and the end date does not mean later risk is zero.

The assessment reports three nested thresholds. Higher-tier probabilities are already included in lower-tier figures.

Threshold	Central judgment
Tier 1 or worse: Major systemic disaster	About 30%
Tier 2 or worse: War-scale devastation	About 10%
Tier 3: Recovery-impairing civilizational catastrophe	About 2%

Tier 1 is a national or international emergency with prolonged critical-function loss and substantial mortality or economic harm. Tier 2 requires years of reconstruction, while productive and institutional recovery capacity survives. Tier 3 removes a credible route to restoring modern industrial civilization across sufficiently extensive regions. Mortality informs that judgment; it does not define it alone.

These imply about 20% with maximum severity at Tier 1, 8% at Tier 2, and 2% at Tier 3, alongside 70% with no qualifying event. Ratios between tiers concern whole histories, not the chance that an individual attack escalates.

Why the thresholds differ so much

The analysis follows serious intent, dangerous access, practical execution, defensive failure, severity, and recovery. States have extensive resources and strong ordinary reasons to avoid indiscriminate destruction. Small actors have fewer resources but a consequential tail of extreme motives. AI can also increase willingness to act by making a dangerous operation seem bounded or controllable.

Cyber leads the lower threshold because digital access can be close to action. Biology leads the upper tail because a successful event can propagate. Both still confront detection, containment, and restoration. The small Tier 3 judgment depends chiefly on surviving industrial recovery capacity, rather than permanent control of dangerous information.

How much confidence to place in the figures

Broad subjective plausibility ranges are 12%-55%, 3%-25%, and 0.4%-8%. They are not calibrated confidence intervals or quantiles from a fitted uncertainty model. The paper provides a hazard specification reproducing the ladder and coherent alternatives; arithmetic consistency does not validate the authored inputs.

The argument and its limits

The strongest challenge is repeated biological attempts

Advanced AI has weakened biological knowledge barriers, but a benchmark ranking is not proof that a novice can complete a dangerous physical operation. Screening, practical execution, response, and the survival of external recovery centers still matter. Their future effectiveness remains uncertain.

Rare intent per person does not establish low aggregate risk. As an illustration, 1,000 independent attempts with a 0.1% chance of a qualifying event each give about a 63% chance of at least one event. Those inputs are hypothetical; real attempts share constraints and provoke adaptation. The example exposes why a large actor pool can invalidate reassurance from a small individual probability.

Within the biological central judgments, Tier 3 is 15% of Tier 1-or-worse risk. If the latter rose from 8% to 50% while that history-level ratio stayed fixed, biology alone would imply 7.5% upper-tier risk. The aggregate 2% assessment would fail. Broad practical uplift arriving before response improves must therefore trigger a serious reassessment.

What the current evidence supports

Primary evaluations show rapidly improving multi-step cyber performance; some omit active defenders and find industrial systems harder. OpenAI classifies Astra at its Critical cyber threshold with tools and access while describing strengthened safeguards. Anthropic reports deeper orchestration and customer credential theft. Those observations support concern about execution and diffusion; they do not establish a frequency of civilization-wide recovery failure.

Global synthesis screening is fragmented. Provider counts are not the share of orders screened, and a policy directive is not proof of audited implementation. Useful defenses include provider safeguards, secure access, physical controls, surveillance, and restoration. Several apparent layers can share the same failure mode.

What should change the assessment

Raise the estimates if practical dangerous capability diffuses before defense, biological assistance substitutes reliably for scarce expertise, cyber agents defeat active defenders and restoration, or military AI compresses strategic decision time. Lower them after realistic demonstrations of broad physical controls, rapid countermeasures, resilient dependencies, and effective restoration.

Earlier effective protection gives an illustrative 14%-4%-0.8% ladder; exposure through 2045 gives 40%-15%-3.5%. These are alternative trajectories, not predictions of a protection date or linear extrapolations.

Relationship to the existential-risk paper

The companion paper estimates about 52% extinction or enduring dystopia and 57% any modeled severe harm, conditional on a rapid superintelligence transition. This paper begins now, includes national disasters, and uses different recovery thresholds. Their percentages cannot be added. A combined assessment needs a common horizon and explicit overlap and feedback. Both also leave deliberate human authoritarian lock-in outside their numerical headlines.

Full paper: Catastrophic Human Misuse of Advanced AI, circulation edition, 29 September 2026. Biological stress tests in Section 4; synthesis in Section 7; attribution and companion conclusions in Section 8; numerical specification in Appendix A. Principal empirical sources: [International AI Safety Report 2026](#), [UK AISI and CAISI cyber evaluations](#), [OpenAI](#) and [Anthropic](#) safety reports, [IBBIS](#), and [SIPRI](#). Sources support mechanisms; the probabilities are the author's and collaborators' judgments.