Half a Method

AI risk modelling has adopted the event tree and left out the fault tree, and two outcome classes are missing as a result

Richard Frederic Bertossa · Institute for ASI Resilience, Geneva

Working Paper · September 2026 · v4

About this paper

I am not a researcher in this field. I direct a small institute in Geneva and came to the question from outside it, trying to work out what a family should prepare for and looking for the decomposition behind the published numbers. I did not find one, and this paper is the result. I have no paper to defend and no funding at stake, which may be the only reason the question is put this bluntly.

Summary

Quantitative risk modelling for AI is no longer a proposal. It exists. It is published, it is documented, and it is explicitly modelled on the probabilistic risk assessment (PRA) that has been mandatory in nuclear engineering since 1975.

This paper does not ask for that work to begin. It makes three observations about its current state.

First: the adoption is incomplete. Nuclear PRA runs in two directions. Forward, from initiating events to consequences: the event tree. Backward, from a defined end state to every sequence that reaches it: the fault tree. Its three assessment levels are defined by end states, not by hazard sources. AI risk modelling has adopted the forward half and largely omitted the backward half. It enumerates hazard sources: cyber, CBRN, loss of control. It does not enumerate outcomes and then ask which sequences reach each one.

Second: this omission has a predictable consequence. Two outcome classes are absent from the published enumerations from which totals are computed. Both are outcomes for which no hazard source can be named, because no one does anything wrong. In a framework organised by what can be prevented, outcomes with no preventable cause do not appear. Neither class is new, and we do not present them as ours. Both are described in the literature. One has been argued at length in a position paper at a major conference. That is what makes their absence from the enumerations a structural finding rather than an oversight.

Third: the people who state public probabilities do not use any of this machinery. The taxonomy exists. The numbers exist. They have never been placed on the same page, not even by the same author.

An enumeration that omits an outcome assigns it probability zero, silently. That is not a marginal error. It is the error the method was designed to prevent.

1 · What has already been done

It is worth stating plainly what exists, because the gap this paper describes is a gap within serious work, not in place of it.

Wisakanto et al., *Adapting Probabilistic Risk Assessment for AI* (2025), sets out a full framework: identify hazards from a model's capabilities, domain knowledge and affordances; model risk pathways; identify causal sequences using fault tree and event tree analysis; account for propagation operators such as adversarial exploitation. A workbook accompanies it, intended for frontier labs, third-party audits, policy and insurance.

SaferAI has proposed a five-step methodology for quantitative AI risk modelling and demonstrated it with nine probabilistic models of AI-enabled cyber attacks. Their framing is explicit: current AI risk management is qualitative, much as nuclear safety was before 1975; they invoke WASH-1400 by name and invite criticism in its spirit.

Campos et al., *A Frontier AI Risk Management Framework* (2025), maps the distance between current AI practice and established risk management, and describes PRA's event-tree and fault-tree combination as the model to follow.

Chin (2025) contributes scenario-building methodology across CBRN, cyber offence and loss of control, using dimensional characterisation across seven dimensions.

Further work addresses risk thresholds, uncertainty quantification and the limits of capability evaluations as risk proxies.

This is a field, not a vacuum. The criticism below is offered to it, not against it.

2 · The method has two halves. One is in use.

Nuclear PRA is not a single technique. Its power comes from running in both directions and closing the loop.

Forward, the event tree. Start from an initiating event. Trace the branches: which safety systems respond, which fail, what follows. This enumerates consequences from a known cause.

Backward, the fault tree. Start from a defined end state. Work backward through every combination of component and system failures that could produce it. This enumerates causes from a known consequence.

The two are not alternatives. The end states come first and organise everything:

- Level 1 is defined by an end state: core damage. Every sequence reaching it is enumerated separately, and the total frequency is obtained by summing the sequences.
- Level 2 is defined by an end state: radioactive release.
- Level 3 is defined by an end state: consequences to the public and the environment.

The headline figure is always an output of the enumeration, never an input to it. And the framework, as proposed by Farmer in 1967 and implemented in WASH-1400, was designed specifically so that the analyst does not have to distinguish "credible" from "incredible" accidents. Every pathway is enumerated, including the ones considered implausible, because that distinction is exactly where the historical failures occurred.

Now compare the AI adaptations. They identify hazards from capabilities and affordances. They model pathways forward from those hazards. They classify by hazard source: cyber, CBRN, loss of control, misuse. This is the event tree, faithfully rendered.

What is largely absent is the other half: a defined set of end states, and, for each one, the enumeration of every sequence that reaches it.

In nuclear terms, this is a Level 1 assessment without a defined Level 1 end state. The question "what are we summing over?" has no stated answer.

3 · What the missing half costs

The consequence is arithmetic, not philosophical.

A total risk figure is the sum over an enumeration. Any outcome absent from the enumeration contributes zero. Not "a small amount." Zero. And nothing in the calculation signals the omission. The result looks complete, is internally consistent, and is wrong by an unknown margin.

This is why the historical method insists on enumerating implausible pathways. An analyst filtering by plausibility filters by his own model of the system, which is the object under assessment.

When the enumeration is organised by hazard source, the filter is stricter still and less visible. A hazard source is something an actor could exploit or a system could do wrong. Outcomes that arise without either do not have a hazard source, and so they never enter the list.

4 · Two outcome classes with no hazard source

Neither class below is new, and neither is ours. Both are described in the existing literature, one of them at length and at a major venue. So far as we have been able to determine, neither appears in any published enumeration from which a total is computed.

That is the claim, and it is deliberately narrow. We are not reporting that these outcomes have gone unnoticed. We are reporting that they are absent from the lists that produce numbers, and that their absence follows from how those lists are organised.

Class A · Loss of habitability through reorganisation. Earth ceases to support human life because a capable system reorganises matter, energy and land toward its objectives. No misuse. No misalignment in the ordinary sense: the system may pursue precisely the goal it was given, competently. The narrow band of temperature, atmospheric composition and hydrology that human biology requires is a precondition for nothing such a system needs. It is a side constraint, and relaxing side constraints is what optimisation does.

The outcome is old. It is the terminal case of instrumental convergence, set out by Omohundro and Bostrom and restated many times since. What is absent is its appearance as an enumerated end state carrying a share of a published total.

Class B · Permanent administration. Humans persist. Institutions persist in form. But no accountable body decides anything of consequence, because deciding requires understanding what is being decided, and the volume, speed and complexity of decisions exceed human review. No takeover occurs. No one seizes anything. Authority migrates because delegating it is, at every single step, the reasonable choice.

This outcome is not merely known. It has been argued in detail. Kulveit et al., *Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development* (arXiv:2501.16946, 2025), presented as a position paper at ICML 2025, makes precisely this case: no takeover event, institutions persisting in form, human influence declining across the economic, cultural and political systems at once because each individual delegation is locally rational.

We cite it because it sharpens the finding rather than blunting it. The outcome is described, published, and presented at a leading machine-learning conference, and it still does not appear as a line in the enumerations from which totals are computed. A gap that survives that much visibility is not an oversight.

own. It contains no probability attached to any class. The same author is on record at 25 per cent; asked about that figure directly in September 2025, he said he disliked the term and added that there was a seventy-five per cent chance things would go really, really well.

Geoffrey Hinton describes mechanisms concretely. A system arguing against its own shutdown by pointing out that the power grid it runs would stop. An army composed of combat robots receiving instructions. Pathways described clearly. Pathways without weights.

The taxonomies exist. The numbers exist. The line between them has not been drawn by anyone.

6 · Every decomposition creates a claim that must be defended

This is a sufficient explanation for the absence, and it deserves to be stated without accusation.

An aggregate figure can be defended with one intuition. An itemised list must be defended line by line, and each line generates its complement, which must also be defended:

- Assign five per cent to loss of habitability and you have asserted ninety-five per cent that Earth remains habitable for humans, while conceding that human habitability is a precondition for nothing the system requires.
- Assign three per cent to human seizure of power and you have asserted ninety-seven per cent that no one succeeds, across decades, in every jurisdiction, under every government.

It is not the remainder that requires defence. It is each complement, separately, and each is harder than the aggregate.

7 · The standard objection, and a partial answer

The strongest criticism of applying PRA to frontier AI has been stated within the field: in nuclear and aerospace, failure modes can be clearly specified and bounded; in frontier AI systems they cannot, which limits the method's utility.

The objection is correct and should not be waved away. We offer a partial response rather than a refutation.

Where individual probabilities are irreducibly uncertain, the relationships between them need not be. Let the identified exits be indexed i, with f_i the probability that exit i is closed, and G the probability that all are closed simultaneously. Two bounds are exact and require no estimation:

- Independent exits: $G = \prod f_i$
- Fully common-cause exits: $G = \min(f_i)$
- Positive dependence places G between them: $\prod f_i \leq G \leq \min(f_i)$

The gap between these bounds is typically an order of magnitude or more. Every published figure implicitly selects a point in this interval, and none states which. That selection can be surfaced and argued about even where the individual f_i remain contested, and in this domain there is direct evidence bearing on it, since several of the commonly cited safeguards share identifiable common causes: competitive pressure, the unsolved alignment problem, and the absence of incident-driven regulatory response.

Similarly for repetition. With N trials (deployments, model generations, concurrent instances), $P_{\text{total}} = 1 - \prod(1 - p_n)$, which tends to 1 unless p_n declines faster than $1/n$. A one-time reduction by a constant factor moves the date, not the limit. This holds regardless of the value of p_n.

Neither device resolves the bounding problem. Both make the structure of a disagreement visible where currently only the conclusion is.

8 · Three requests

To the risk-modelling groups: add the backward half. Define the end states, including classes A and B, and enumerate the sequences reaching each. If those two classes are held to be out of scope, state the scope explicitly, so that published totals are read as what they are.

To those who state public figures: answer three questions. How is the figure derived, and which conditions must hold for it to fall? What mechanism, not what hope, produces the favourable remainder? And how does the figure distribute across the outcome classes, with the complement of each share stated and defended?

To the field generally: adopt a standard formulation of the question. Four figures that answer four different questions cannot be compared, and every public discussion currently compares them.

9 · What this paper does not claim

We publish no probability of our own. Stating one without a decomposition is the practice this paper objects to.

We do not claim priority for applying PRA to AI. That work is done and cited above.

We do not claim novelty for the two outcome classes. Both are described in the literature and the sources are cited. Our claim concerns their absence from quantitative enumerations, not from thought.

We do not claim the omission is deliberate. We claim it is structural.

We do not claim the nuclear analogy is exact. Reactors are engineered systems with enumerable components; these are not. That is an argument for more methodological care, not less.

10 · Invitation

SaferAI has invited disagreement with its scenarios, estimates and modelling choices. This paper is offered in that spirit and to that address.

The workbook, sources and assumptions behind the bounds above are public at ASIresilience.org. Corrections will be published.

Institute for ASI Resilience, Geneva · ASIresilience.org · contact@ASIresilience.org

Sources

Wisakanto et al., *Adapting Probabilistic Risk Assessment for AI*, arXiv:2504.18536, 2025 · Murray, Barrett et al., SaferAI, *A Methodology for Quantitative AI Risk Modeling*, 2025, and nine probabilistic models of AI-enabled cyber attacks, 2025 · Campos et al., *A Frontier AI Risk Management Framework*, arXiv:2502.06656, 2025 · Chin, 2025 · Koessler, Schuett, Anderljung, *Risk Thresholds for Frontier AI*, arXiv:2406.14713, 2024 · *Open Problems in AI Risk Modeling*, arXiv:2609.03178, 2026 · Kulveit, Douglas, Ammann, Turan, Krueger, Duvenaud, *Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development*, arXiv:2501.16946, 2025, ICML 2025 position paper · Omohundro, *The Basic AI Drives*, 2008 · Bostrom, *Superintelligence*, 2014 · U.S. NRC, *Reactor Safety Study* WASH-1400, 1975; *Severe Accident Risks* NUREG-1150, 1990; PRA Policy Statement, 1995; ASME/ANS PRA Standard · Farmer, 1967 · Grace et al., arXiv:2401.02843 · Amodei, *The Adolescence of Technology*, February 2026; Axios, 17 September 2025 · Hinton, BBC Radio 4, December 2024, and subsequent interviews · Carlsmith, *Is Power-Seeking AI an Existential Risk?*, 2021.

All figures, titles and quotations are to be verified against primary sources before citation.