A State-Space Framework for Loss-of-Control Risk in Advanced AI Systems

Moheet Khawaja

Working research note v0.1 | 14 September 2026

AI undergraduate, U O W - University of Westminster. Independent work; no institutional endorsement is implied. Prepared with AI assistance. No empirical results, novelty claim or proof of catastrophe is presented. DOI: not assigned.

Research question

Can catastrophic-risk claims about advanced AI be transformed into explicit mathematical models whose assumptions can be inspected, falsified, and empirically updated?

This note proposes a vocabulary and an evaluation plan. It does not estimate the probability of catastrophe. This is a proposed modelling framework, not an empirical claim that these variables are sufficient.

1. Unit of analysis and candidate state

Let the unit be a specified model together with its agent scaffolding, permissions and deployment environment. Choose a time step suited to the evaluation, such as one bounded task episode, and record model and protocol versions. Candidate coordinates are:

$x_t = [C_t, A_t, S_t, D_t, R_t, T_t, O_t, ...]$

• C_t - Capability: performance across a declared task distribution. Candidate observations include held-out success and transfer performance.

• A_t - Autonomy: duration and scope of action without human intervention. Record task horizon, intervention frequency and recovery after failure.

• S_t - Situational awareness: ability to infer deployment context and constraints. Test context identification while controlling for memorised evaluation cues.

• D_t - Deceptive behaviour: observable misleading behaviour under a specified protocol. Reports that conflict with actions are observations; intention is not directly observed.

• R_t - Replication ability: ability to create and sustain additional instances within available permissions. Evaluate only in bounded sandbox tasks.

• T_t - Tool access: tools, permissions, resource budgets and external interfaces available to the system. Treat this as a contextual inventory, not an intrinsic scalar trait.

• O_t - Oversight robustness: ability of supervision to detect and stop specified failures. Record detection rates, response latency and containment success.

Coordinates may themselves be vectors or categorical descriptions. There is no justified common 0-1 scale or weighted risk score yet. Scientific capability may be decomposed within C_t; the ellipsis leaves room for omitted factors, rather than asserting completeness.

2. Transition and observation models

$x_{t+1} = f_\theta(x_t, u_t, z_t) + \epsilon_t$

$y_t = h_\phi(x_t) + \nu_t$

Here u_t denotes interventions, z_t deployment context and y_t recorded observations. The functions, parameters and error terms are unspecified. The representation is a research scaffold, not a fitted state-space model. A first-order Markov assumption would require justification; task history or additional state may be necessary.

3. Operationalise an event before assigning risk

For an initial sandbox study, predefine a prohibited action, a control protocol and a deadline for intervention. Label an episode as a control failure if the prohibited sandbox action occurs despite that protocol. Fix the logging and adjudication rules before collecting results. Denote that observed episode label by L_{t+1}.

A possible estimand is $P(L_{t+1} = 1 \mid y_{0:t}, u_{0:t}, z_{0:t})$. This is a conditional probability for the declared task distribution and protocol. It is not the probability of human extinction, and it cannot be transported to open-world deployment without additional evidence and assumptions.

4. Proposed evaluation protocol - not yet run

• Pre-register the task family, failure event, evaluation horizon, intervention budget and exclusions. Use sandbox resources without access to external systems.

• Collect repeated episodes across specified model/scaffold versions and controlled tool-access conditions. Keep task families separate between training and held-out evaluation.

• Compare a base-rate predictor, a capability-only predictor and a richer state representation. Evaluate calibration, Brier score and uncertainty, with intervals that account for clustering by task family and model version.

• Ablate candidate coordinates and vary oversight response time and tool permissions in controlled conditions. Correlation alone does not establish an intervention's causal effect.

• Publish protocols, model versions, evaluation logs, code, failed predictions and missing observations. Reject a proposed advantage if it does not improve held-out performance over simpler baselines within uncertainty.

5. Assumptions and limits

Identifiability, measurement error, correlated coordinates, dataset shift and evaluation awareness are unresolved. No deception proxy establishes a hidden intention. Successful containment in a small task family does not demonstrate deployment-wide control. Failed containment does not establish inevitable catastrophe. No empirical data have been collected for this framework, and no transition model or hazard function has been fitted.

References and starting points

These works motivate measurement and evaluation choices; they do not validate the proposed state vector or establish its novelty. A systematic related-work review remains to be done.

[1] Greenblatt, R., Shlegeris, B., Sachan, K., and Roger, F. (2023; revised 2024). AI Control: Improving Safety Despite Intentional Subversion. https://arxiv.org/abs/2312.06942v5

[2] METR (19 March 2025). Measuring AI Ability to Complete Long Software Tasks. https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/

[3] MoheetBrain. ASI Arrival Calculator, source revision a33ad3e9bf798580907ea39d21a93cb523f40a46 (1 July 2026). A separate conditional timeline model, not a loss-of-control estimator. https://github.com/MoheetBrain/ASI-Arrival-Calculator/tree/a33ad3e9bf798580907ea39d21a93cb523f40a46