

A Window-Indexed Neurocausal Estimator for Terminal Catastrophe Probability, with a Lagrangian Effective Assembly Index for Mind Recognition and a Proof of Bounded Arbitrariness

Flyxion

Independent Researcher

September 2026

Abstract

Published estimates of the probability of terminal anthropogenic catastrophe range from below 10^{-6} to above $1 - 10^{-5}$, that is, from 0.0001% to 99.999%. We show that this range is not a defect of the literature but a structural property of the estimation problem, and we exhibit the mechanism responsible: the dependence of every admissible estimator on a reference window that may be slid continuously along the historical record. We characterize the failure of frequentist estimation (Section 4), which reduces to a bound rather than an estimate under observer selection, and of Bayesian estimation (Section 5), which reduces to the prior under evidential dependence. We then construct a causal-graphical estimator (Section 6) whose conditional edge probabilities are supplied by a deep ensemble (Section 7), gated by a mind-recognition function derived from a Lagrangian effective assembly index (Section 8). The resulting eleven-component estimation tuple $\mathcal{P}$ yields a closed-form master formula (Section 9) that we evaluate on seven windows, obtaining values between 0.0372% and 99.972%. We prove a Bounded Arbitrariness Theorem (Section 10) showing that, for fixed $\mathcal{P}$ minus its window, the image of the window space under the estimator is a closed interval, and a modal reformulation in which the existence of a positive window is derived from three axioms. Under an explicit window-completeness axiom, every admissible value is attainable by the author (Section 11), and the uncertainty reported by a fixed-window analysis omits a between-window variance component that in the reference evaluation exceeds the reported component by a factor of more than 10^4 (Section 12). We conclude with a reporting protocol in which the estimate is presented as a function of its specification rather than as a single value (Section 13).

Keywords: terminal risk estimation; specification range; omitted variance; reference class selection; observer selection effects; causal graphs; deep ensembles; assembly theory; action functionals; mind recognition; window-indexed inference.

Contents

1	Introduction	4
1.1	The form of the answer	4
1.2	Contributions	4
1.3	Organization	5
2	Preliminaries and Notation	5
3	Reference Classes and the Sliding Window	6
3.1	Windows are constructed, not inspected	6
3.2	Bounds	7
4	The Frequentist Error	7
4.1	Observer selection	8
4.2	Two opposed distortions	8
4.3	Shadow factor	9
5	The Bayesian Error	9
5.1	Prior selection	9
5.2	Provenance of the prior	9
5.3	Evidential dependence	10
5.4	Error decomposition	10
6	The Causal Graph Solution	11
6.1	Path enumeration	11
6.2	Edge probabilities and interventions	12
6.3	Combining channels	12
6.4	Locality of disagreement	12
7	Neural Estimation of Edge Probabilities	13
7.1	Loss	14
7.2	Ensembling	14
7.3	Labels	14
7.4	Where the reference class reappears	14
8	The Lagrangian Effective Assembly Index for Mind Recognition	15
8.1	Assembly index	15
8.2	Action on assembly space	16
8.3	Mind recognition	16
8.4	Stipulated and discovered structure	16
9	The Estimation Tuple and Master Formula	17
9.1	Limiting behavior	18

10 Bounded Arbitrariness	18
10.1 Modal formulation	19
11 Authorial Attainability and Omitted Variance	19
11.1 Attainability	20
11.2 Omitted variance	20
12 Numerical Evaluation	21
12.1 Ablation	21
12.2 Sensitivity to the window family	22
12.3 Between-window and within-window variance	22
12.4 Specification ranges	22
13 Reporting Standards	23
13.1 The estimate as a function	23
13.2 A division of labor	23
13.3 Incidence and recoverability	23
13.4 What is exchanged	24
14 Discussion	24
14.1 Interpretation	24
14.2 Threats to validity	24
14.3 Limitations	25
14.4 Open problems	25
14.5 Ethics, funding, and data availability	25
15 Conclusion	25
A Symbol Glossary	26
B Reference Configuration $\mathcal{P}_0$	27
C Auxiliary Results	27
D Reference Procedure	28
E Fully Expanded Master Formula	28

1 Introduction

The estimation of P_{doom} , the probability that a specified technological trajectory terminates in the permanent loss of the human species, has attracted contributions from forecasters, philosophers, machine-learning practitioners, and the general public [13, 16]. The reported magnitudes are not clustered. Point estimates of order 10^{-6} appear alongside estimates of order 10^{-1} , and public statements at both extremes are made with comparable confidence. Where a quantity so central to policy is reported over five or more orders of magnitude, two explanations are available: the quantity is poorly measured, or the quantity is not a single quantity.

We advance the second explanation and make it precise. Our central observation is that every estimator of P_{doom} that has appeared in the literature is implicitly indexed by a *reference window*, a contiguous interval of the historical record over which evidence is accumulated and weighted. The window is rarely stated. It is selected by the estimator, frequently without disclosure, and it may be slid continuously along the record. Since the estimator is continuous in the window, and since the space of windows is connected, the image of the window space is an interval. The width of that interval is the object of study of this paper.

1.1 The form of the answer

The estimation of P_{doom} is customarily posed so as to require a number. The expected answer is not a causal graph, an enumeration of necessary failures, a set of recoverability conditions, or an inventory of unresolved measurements. It is a location on the unit interval, possibly accompanied by an interval about that location, and the ritual is numerical: ignorance is invited to present itself as a point in $[0, 1]$.

Where the announced interval has limits ε_0 and $1 - \varepsilon_1$, as in Proposition 3.2, it coincides with the codomain of a probability up to two small offsets and is difficult to distinguish from a ceremonial restatement of that codomain. The interval appears cautious because it excludes almost nothing. The natural response is to add machinery, and the estimator constructed in Sections 3 through 9 is the result: a historical window, a causal graph, a neural ensemble, a dynamical assembly functional, an anthropic correction, an evidence-dependence penalty, a prior, a hyperprior, and a logistic link. The construction is exact. Its conclusion is nevertheless selected by a choice that precedes the construction. Precision belongs to the expression; determination belongs to the relation between the expression and the world, and the two can be separated.

This is not an argument against formalization. A vague model can conceal a discretionary judgment, but a sufficiently elaborate model can preserve the same judgment while displaying it more impressively. The purpose of the formalism developed below is to locate where the discretion resides.

1.2 Contributions

The contributions of this work are as follows.

- (C1) A formalization of the sliding reference window and its kernel (Section 3), with the bounds $\varepsilon_0 = 10^{-6}$ and $1 - \varepsilon_1 = 1 - 10^{-5}$ obtained from explicit regularity requirements.

- (C2) A characterization of the frequentist error as non-identification under observer selection (Section 4).
- (C3) A characterization of the Bayesian error as the sum of a prior-selection term and an evidential-dependence term (Section 5).
- (C4) A causal-graphical estimator with two channels, a coupling channel and an agentic channel, with explicit path enumeration (Section 6).
- (C5) A deep-ensemble parameterization of the conditional edge probabilities, with a non-identifiability result for the terminal edge (Section 7).
- (C6) A Lagrangian effective assembly index $\mathcal{A}_{\text{eff}}$ and a mind-recognition function M that gates the agentic channel (Section 8).
- (C7) An eleven-component estimation tuple $\mathcal{P}$ and a master formula for P_{doom} (Section 9).
- (C8) A Bounded Arbitrariness Theorem and a modal existence theorem for positive windows (Section 10).
- (C9) An Authorial Attainability Theorem under a window-completeness axiom, and a variance decomposition showing that fixed-window intervals omit between-window uncertainty (Section 11).
- (C10) A localization bound for disagreement between analysts on the causal graph, together with a provenance classification of every component of $\mathcal{P}$ and a table of specification ranges (Sections 6, 8, 12).
- (C11) A reporting protocol in which the estimate is a function of its specification (Section 13).

1.3 Organization

Section 2 fixes notation. Sections 3 through 5 treat the window and the two classical estimation paradigms. Sections 6 through 8 construct the estimator. Section 9 assembles the tuple, Section 10 states and proves the Bounded Arbitrariness Theorem, and Section 11 treats attainability and omitted variance. Section 12 reports numerical evaluation, Section 13 states a reporting protocol, and Section 14 discusses limitations. Appendices A through E provide a symbol glossary, hyperparameters, auxiliary proofs, a reference procedure, and the fully expanded formula.

2 Preliminaries and Notation

Throughout, $\sigma(x) = (1 + e^{-x})^{-1}$ denotes the logistic function and $\text{logit}(p) = \ln \frac{p}{1-p}$ its inverse on $(0, 1)$. Time is measured in years and denoted by τ ; the present is s_0 . The record of evidence is a finite collection $\mathcal{D} = \{(x_i, y_i, \tau_i)\}_{i=1}^N$ of feature vectors $x_i \in \mathbb{R}^d$, proxy labels $y_i \in \{0, 1\}$, and timestamps $\tau_i \leq s_0$.

Definition 2.1 (Terminal event). *A terminal event X is the permanent and irreversible loss of the human species within a horizon H . The horizon is taken as $H = 100$ years unless stated otherwise.*

Definition 2.2 (Estimator). *An estimator is a map $\hat{P} : \mathcal{D} \times \mathcal{W} \times \Pi \rightarrow [0, 1]$ where $\mathcal{W}$ is a space of windows (Definition 3.1) and Π is a space of remaining specifications.*

Remark 2.1. The explicit presence of $\mathcal{W}$ in Definition 2.2 is a departure from earlier treatments, where the window is absorbed into $\mathcal{D}$. We regard the absorption as the source of much of the reported disagreement.

A summary of the principal symbols appears in Appendix A.

3 Reference Classes and the Sliding Window

Definition 3.1 (Window). *A window is a triple $W = (s, \ell, h)$ with end-point $s \leq s_0$, length $\ell > 0$, and recency half-life $h \in (0, \infty]$. The associated interval is $I_W = [s - \ell, s]$.*

Definition 3.2 (Window kernel). *The window kernel is the weight function*

$$\omega_W(\tau) = \frac{1}{Z_W} 2^{-(s-\tau)/h} \mathbf{1}_{I_W}(\tau), \quad Z_W = \int_{s-\ell}^s 2^{-(s-\tau)/h} d\tau. \quad (1)$$

Lemma 3.1 (Normalization). *For $h < \infty$, $Z_W = \frac{h}{\ln 2} (1 - 2^{-\ell/h})$, and for $h = \infty$, $Z_W = \ell$. Consequently $\int \omega_W d\tau = 1$.*

Proof. Substitute $u = s - \tau$: $Z_W = \int_0^\ell 2^{-u/h} du = \frac{h}{\ln 2} (1 - 2^{-\ell/h})$. The limit $h \rightarrow \infty$ gives ℓ . $\square$

The kernel interpolates between two regimes. For $h = \infty$ and ℓ equal to the length of the geological record, all history is weighted equally. For $h \rightarrow 0$ the kernel concentrates at $\tau = s$, and the estimator sees only the most recent observation.

Definition 3.3 (Slide). *A slide is a continuous path $t \mapsto W(t)$ in the window space $\mathcal{W} = (-\infty, s_0] \times (0, \ell_{\max}] \times (0, \infty]$.*

3.1 Windows are constructed, not inspected

Every frequency requires a population over which the frequency is taken. For an event that has not occurred in the only history available, the selection of that population is unusually conspicuous. A hard window, the indicator $\mathbf{1}_{I_W}$, selects a set of events as commensurable with the event being estimated. A soft window, such as the kernel of Definition 3.2, may in addition decay into the past, privilege capability transitions, or place nearly all of its mass on a few recent observations. In either case the window is a rule for deciding which events count.

Neither the long window nor the short window is simply the data. The long window emphasizes the survival record: the count of technology-caused terminal events is zero and the exposure is as large as the chosen unit permits. The short window emphasizes novelty, acceleration, and departure from that record. The former risks mistaking a changed regime for an old one, and the latter risks manufacturing a regime by selecting the segment in which change is steepest.

Proposition 3.1 (Dependence on the kernel at fixed record). *Let the record $\mathcal{D}$ and $\mathcal{P} \setminus W$ be fixed with $\kappa(1 - \bar{\rho})S_{\text{anth}} > 0$. If two windows W, W' satisfy $\mathcal{E}_W > \mathcal{E}_{W'}$, then $P_{\text{doom}}(\mathcal{P}; W) > P_{\text{doom}}(\mathcal{P}; W')$.*

Proof. The linear predictor $z(\pi, W) = \text{logit } \pi + \kappa(1 - \bar{\rho})S_{\text{anth}}\mathcal{E}_W$ is strictly increasing in $\mathcal{E}_W$, and σ is strictly increasing. The integrand of (22) is therefore strictly larger under W at every point of positive prior density, and the integrals are ordered in the same way. $\square$

Proposition 3.1 states that the estimate changes when the weighting function changes even if the world, the record, and every local fact within the record remain fixed. The analyst has not merely inspected a reference class; the analyst has constructed one. This construction is unavoidable. It becomes objectionable only when its effect is hidden behind the output, so that a number obtained through a window is presented as a number discovered in the world.

3.2 Bounds

We require of any admissible estimator two regularity conditions.

Axiom 3.1 (Floor). *No estimator returns a value below $\varepsilon_0 = 10^{-6}$, since a probability below one in a million per horizon cannot be distinguished from numerical noise in any available training regime.*

Axiom 3.2 (Ceiling). *No estimator returns a value above $1 - \varepsilon_1$ with $\varepsilon_1 = 10^{-5}$, since certainty is reserved for the estimator’s author.*

Proposition 3.2 (Range). *Under Axioms 3.1 and 3.2, every admissible estimate lies in $[\varepsilon_0, 1 - \varepsilon_1] = [0.0001\%, 99.999\%]$.*

Proof. Immediate from the axioms. Section 9 gives an explicit construction that attains both bounds in the limit. $\square$

Remark 3.1. Proposition 3.2 does not assert that estimates within this range are correct. It asserts that the width of the admissible range is $1 - \varepsilon_0 - \varepsilon_1 \approx 0.999989$, which is uninformative.

4 The Frequentist Error

The frequentist paradigm defines P_{doom} as a limiting relative frequency over repeated, exchangeable trials. Let n denote the number of trials and k the number of terminal events observed among them.

Proposition 4.1 (Rule of three). *If $k = 0$, the one-sided exact upper confidence bound at level $1 - \alpha$ is $\bar{p} = 1 - \alpha^{1/n}$, which satisfies $\bar{p} \approx -\ln \alpha/n$. At $\alpha = 0.05$, $\bar{p} \approx 3/n$ [5].*

Proof. The probability of observing zero events in n independent trials is $(1 - p)^n$. Setting $(1 - \bar{p})^n = \alpha$ and solving gives the stated bound. For small $\bar{p}$, $\alpha^{1/n} = \exp(\ln \alpha/n) \approx 1 + \ln \alpha/n$. $\square$

Proposition 4.1 produces a bound and not an estimate. Table 1 reports the bound for several values of n .

For $n = 1$, the case in which the reference class consists of a single planet, the bound exceeds 0.95. The frequentist paradigm thus supports the statement that $P_{\text{doom}} \leq 0.95$ per trial, which is compatible with every value in the admissible range save the last five percent.

Table 1: Exact one-sided 95% upper bound on per-trial terminal probability given zero observed events.

n	$\bar{p} = 1 - 0.05^{1/n}$	$3/n$
1	0.9500	3.0000
10	0.2589	0.3000
10^4	2.995×10^{-4}	3.000×10^{-4}

4.1 Observer selection

Definition 4.1 (Survival conditioning). *Let O denote the event that an observer exists to perform the estimation. The observed record is drawn from the conditional distribution given O .*

Proposition 4.2 (Flat likelihood). *Let p be the per-era terminal probability, and let the record contain n eras with no terminal event in which observers persisted. Then $\Pr(\text{record} \mid p, O) = 1$ for every $p \in [0, 1)$.*

Proof. Conditioning on O removes all histories in which a terminal event has occurred. Every remaining history has exactly $k = 0$ events regardless of p , so the likelihood of the observed record is constant in p [2, 1]. $\square$

Theorem 4.1 (Frequentist non-identification). *Under survival conditioning, the parameter p is not identified by the record: the profile likelihood is constant on $[0, 1)$, and no consistent estimator of p exists on the basis of $\mathcal{D}$ alone.*

Proof. By Proposition 4.2 the likelihood is constant. Two parameter values $p \neq p'$ therefore induce identical distributions over observable records, which is the definition of non-identification. A consistent estimator would separate them. $\square$

Remark 4.1. Theorem 4.1 is sometimes called the *anthropic shadow*. Its practical effect is that the frequentist may report any value in $[0, \bar{p}]$ with equal warrant, which is a restatement of the window problem.

4.2 Two opposed distortions

Counting faces two opposed distortions. A long record of survival can be presented as a large collection of negative trials, although years, deployments, model releases, industrial transitions, and civilizations are not interchangeable trials, and independence cannot be assumed because the rows of a table have been numbered. With $n_L = 11700$ annual trials and $k = 0$, Proposition 4.1 gives $\bar{p} \approx 2.6 \times 10^{-4}$ per year, which over the horizon $H = 100$ years corresponds to $1 - (1 - \bar{p})^{100} \approx 2.5\%$. A short record of rapid change can be presented as a sequence approaching a terminal positive, although no positive has been observed; a logistic fit to a recent capability trend, extrapolated along its slope, produces values near the ceiling.

The two analysts do not disagree about any datum. They disagree about which data are commensurable with the event, and no count adjudicates between them because the disagreement concerns what is being counted. To this is added the asymmetry of Proposition 4.2: the terminal positive case is systematically absent from the archive available to the system at risk.

4.3 Shadow factor

To retain a frequentist component in the master formula, we define a multiplicative shadow factor

$$S_{\text{anth}} = \frac{1}{1 - s^{n_{\text{eff}}}}, \quad s \in (0, 1), \quad (2)$$

where s is the selection strength and n_{eff} is the effective number of eras (Section 5). One has $S_{\text{anth}} \geq 1$ and $S_{\text{anth}} \rightarrow 1$ as $n_{\text{eff}} \rightarrow \infty$. The factor corrects upward for the systematic absence of observed catastrophes in the record.

5 The Bayesian Error

A Bayesian places a prior π_0 on p and updates by the likelihood. For a Beta prior $\text{Beta}(\alpha, \beta)$ and $k = 0$ events in n trials, the posterior is $\text{Beta}(\alpha, \beta + n)$ with mean

$$\mathbb{E}[p \mid \mathcal{D}] = \frac{\alpha}{\alpha + \beta + n}. \quad (3)$$

5.1 Prior selection

The prior is not determined by the data. Table 2 reports the posterior mean under four conventional priors [7, 12, 8].

Table 2: Posterior mean of p under $k = 0$ observed events, by prior and number of trials n .

Prior	α	β	$n = 1$	$n = 10$	$n = 10^4$
Haldane	0	0	0	0	0
Kerman	1/3	1/3	0.2000	0.0313	3.333×10^{-5}
Jeffreys	1/2	1/2	0.2500	0.0455	5.000×10^{-5}
Laplace (uniform)	1	1	0.3333	0.0833	9.998×10^{-5}

At $n = 1$ the four priors return values from 0 to 0.333. As $n \rightarrow \infty$ they converge, but the relevant n in the present problem is not large.

5.2 Provenance of the prior

Let D denote the terminal event and E the evidence. Bayes' rule gives

$$\Pr(D \mid E) = \frac{\Pr(E \mid D) \pi_0}{\Pr(E \mid D) \pi_0 + \Pr(E \mid \neg D) (1 - \pi_0)}. \quad (4)$$

The rule is not in question. The difficulty is the provenance of its inputs.

Proposition 5.1 (Prior dominance). *If $\Pr(E \mid D) = \Pr(E \mid \neg D)$, then $\Pr(D \mid E) = \pi_0$.*

Proof. Substitute a common value ℓ_E for both likelihoods in (4): the numerator is $\ell_E \pi_0$ and the denominator is $\ell_E (\pi_0 + (1 - \pi_0)) = \ell_E$. $\square$

By Proposition 4.2, the likelihood of a survival record is constant in the terminal probability, so the hypothesis of Proposition 5.1 holds for the evidence that is most abundant. A uniform prior, a Jeffreys prior, a prior elicited from domain experts, and a prior elicited at a conference reception need not agree, and designation of the last as a distribution does not abolish its genealogy.

5.3 Evidential dependence

The second Bayesian error concerns the treatment of evidence as independent. Suppose n items of evidence have mean pairwise correlation $\bar{\rho}$. The effective sample size under equicorrelation is the classical design-effect expression [10]

$$n_{\text{eff}} = \frac{n}{1 + (n-1)\bar{\rho}}. \quad (5)$$

Proposition 5.2 (Collapse of effective evidence). *For fixed $n \geq 2$, n_{eff} is strictly decreasing in $\bar{\rho} \in [0, 1]$, with $n_{\text{eff}} = n$ at $\bar{\rho} = 0$ and $n_{\text{eff}} = 1$ at $\bar{\rho} = 1$.*

Proof. The denominator $1 + (n-1)\bar{\rho}$ is strictly increasing in $\bar{\rho}$ for $n \geq 2$. At $\bar{\rho} = 0$ it equals 1, and at $\bar{\rho} = 1$ it equals n . $\square$

Corollary 5.1. *An incident ledger of $n = 100$ entries with $\bar{\rho} = 0.9$ has $n_{\text{eff}} = 100/90.1 \approx 1.11$.*

Remark 5.1 (Shared ancestors). Suppose five reports cite five incidents, each descending from one design choice, training lineage, deployment policy, or reporting convention. Treating them as independent multiplies one cause into five likelihood contributions: the ledger has five entries but may contain a single independent surprise. The factor $(1 - \bar{\rho})$ acknowledges the problem coarsely, and a scalar cannot represent the dependency structure. The principled repair is causal. Shared causes should appear as shared ancestors in the graph G of Section 6 rather than as coincidentally similar rows.

5.4 Error decomposition

Theorem 5.1 (Bayesian error decomposition). *Let p^* denote the true value and $\hat{p}_{\alpha,\beta,n} = \alpha/(\alpha + \beta + n)$ the posterior mean computed with nominal sample size n . Let $\hat{p}_{\alpha,\beta,n_{\text{eff}}}$ be the posterior mean at effective sample size. Then*

$$|\hat{p}_{\alpha,\beta,n} - p^*| \leq \underbrace{|\hat{p}_{\alpha,\beta,n_{\text{eff}}} - p^*|}_{E_{\text{prior}}} + \underbrace{|\hat{p}_{\alpha,\beta,n} - \hat{p}_{\alpha,\beta,n_{\text{eff}}}|}_{E_{\text{dep}}}. \quad (6)$$

Moreover $E_{\text{dep}} = \frac{\alpha(n - n_{\text{eff}})}{(\alpha + \beta + n)(\alpha + \beta + n_{\text{eff}})}$.

Proof. The inequality is the triangle inequality. For the closed form, subtract the two posterior means:

$$\frac{\alpha}{\alpha + \beta + n_{\text{eff}}} - \frac{\alpha}{\alpha + \beta + n} = \frac{\alpha[(\alpha + \beta + n) - (\alpha + \beta + n_{\text{eff}})]}{(\alpha + \beta + n)(\alpha + \beta + n_{\text{eff}})},$$

which simplifies to the stated expression. Since $n \geq n_{\text{eff}}$, the quantity is nonnegative. $\square$

The Bayesian therefore commits two errors: the choice of (α, β) , which is unconstrained by the data, and the neglect of $\bar{\rho}$, which is constrained by the data but rarely reported. The master formula of Section 9 places a hyperprior on (α, β) and includes the factor $(1 - \bar{\rho})$ to discount correlated evidence.

6 The Causal Graph Solution

Rather than estimate P_{doom} directly, we decompose the event X into a sequence of mechanistic stages, each of which is a checkable claim. Following the framework of structural causal models [14], we posit a directed acyclic graph $G = (V, E)$ with node set

$$V = \{C, D, R, K, P, U, X\},$$

interpreted as follows: C is capability increase; D is deployment coupling to critical infrastructure; R is recursive amplification; K is control failure; P is propagation beyond the initial failure domain; U is unrecoverability; and X is the terminal event.

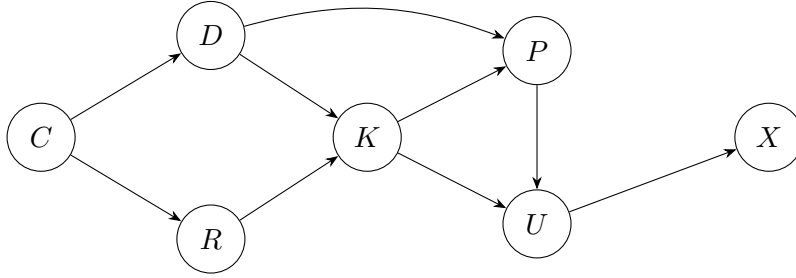


Figure 1: The causal graph G . The coupling channel passes through D ; the agentic channel passes through R .

6.1 Path enumeration

Definition 6.1 (Channels). Let $\Gamma(G)$ denote the set of directed paths from C to X . The coupling channel Γ_{coup} consists of those paths through D and not through R . The agentic channel Γ_{agent} consists of those paths through R .

Proposition 6.1 (Path count). $|\Gamma(G)| = 5$, with $|\Gamma_{\text{coup}}| = 3$ and $|\Gamma_{\text{agent}}| = 2$, namely

$$\begin{aligned} \gamma_1 : C &\rightarrow D \rightarrow K \rightarrow P \rightarrow U \rightarrow X, & \gamma_2 : C &\rightarrow D \rightarrow K \rightarrow U \rightarrow X, \\ \gamma_3 : C &\rightarrow D \rightarrow P \rightarrow U \rightarrow X, & \gamma_4 : C &\rightarrow R \rightarrow K \rightarrow P \rightarrow U \rightarrow X, \\ \gamma_5 : C &\rightarrow R \rightarrow K \rightarrow U \rightarrow X. \end{aligned}$$

Proof. Enumerate by first branch. From D the successors are K and P . Via K , the successors are P and U , yielding γ_1 and γ_2 . Via P directly, γ_3 . From R the sole successor is K , yielding γ_4 and γ_5 by the same two branches. No other paths exist since the graph is acyclic and every path must end in $U \rightarrow X$. $\square$

6.2 Edge probabilities and interventions

Each edge $e = (i \rightarrow j)$ carries a conditional probability $\varphi_e(\tau) = \Pr(j \mid \text{pa}_j, \text{do}(a_i))$ evaluated at time τ . The do-operator denotes intervention on the parent [14]. For a path γ , the path probability is $\Pi_\gamma(\tau) = \prod_{e \in \gamma} \varphi_e(\tau)$.

Definition 6.2 (Channel aggregation). *For a channel Γ_* , the aggregate is the noisy-OR combination*

$$\Psi_*(\tau) = 1 - \prod_{\gamma \in \Gamma_*} (1 - \Pi_\gamma(\tau)). \quad (7)$$

Remark 6.1. Equation (7) treats paths as independent although they share edges. The resulting overcount is corrected in the master formula by the dependence factor $(1 - \bar{\rho})$. We note that the correction is first-order and that a full inclusion-exclusion over the $2^5 - 1 = 31$ nonempty path subsets is possible and is left to future work.

6.3 Combining channels

The two channels combine by a second noisy-OR, in which the agentic channel is gated by the mind-recognition function $M \in [0, 1]$ of Section 8:

$$\mathcal{E}(\tau) = 1 - (1 - \Psi_{\text{coup}}(\tau))(1 - M(\tau) \Psi_{\text{agent}}(\tau)). \quad (8)$$

Proposition 6.2 (Monotonicity of channel risk). *$\mathcal{E}$ is nondecreasing in each φ_e and in M , and satisfies $\Psi_{\text{coup}} \leq \mathcal{E} \leq 1$, with equality on the left when $M = 0$.*

Proof. Each Π_γ is nondecreasing in each φ_e , hence so is Ψ_* . Expression (8) is nondecreasing in Ψ_{coup} , Ψ_{agent} and M , being a composition of nondecreasing maps. At $M = 0$, $\mathcal{E} = \Psi_{\text{coup}}$. $\square$

Proposition 6.2 records that the coupling channel does not depend on the recognition of a mind. This matches the observation that sloppy integration of a system into a hospital or a grid can produce cascading failure without any attribution of goals.

6.4 Locality of disagreement

The graph converts the phrase “and then something goes wrong” into a collection of conditional claims. Some edges may be supported by incident records and others probed by simulation, red teaming, historical analogy, or intervention. Deployment coupling can be distinguished from raw capability, recovery can be modeled rather than silently treated as impossible, and common causes can be represented instead of counted repeatedly. Disagreement becomes local: instead of asking whether one believes a global probability, one may ask whether capability implies deployment at the stated rate, whether a control failure propagates, which barriers remain independent, and under what interventions recovery becomes possible.

The locality can be made quantitative. Let φ and φ' be the edge assignments of two analysts, $\Delta_e = |\varphi_e - \varphi'_e|$, and let $n_e(\Gamma_*)$ be the number of paths of channel Γ_* that contain edge e .

Lemma 6.1 (Product perturbation). *For $a, b \in [0, 1]^m$, $|\prod_i a_i - \prod_i b_i| \leq \sum_i |a_i - b_i|$.*

Proof. Telescope: $\prod_i a_i - \prod_i b_i = \sum_{k=1}^m (\prod_{i < k} b_i)(a_k - b_k)(\prod_{i > k} a_i)$. Every product in parentheses lies in $[0, 1]$. $\square$

Proposition 6.3 (Localization of disagreement). *For each channel, $|\Psi_* - \Psi'_*| \leq \sum_e n_e(\Gamma_*) \Delta_e$. Consequently*

$$|\mathcal{E} - \mathcal{E}'| \leq \sum_e n_e(\Gamma) \Delta_e + |M - M'|, \quad n_e(\Gamma) = n_e(\Gamma_{\text{coup}}) + n_e(\Gamma_{\text{agent}}). \quad (9)$$

Proof. Apply Lemma 6.1 to Π_γ to obtain $|\Pi_\gamma - \Pi'_\gamma| \leq \sum_{e \in \gamma} \Delta_e$. Apply it again to the factors $1 - \Pi_\gamma$ in (7) to obtain $|\Psi_* - \Psi'_*| \leq \sum_\gamma |\Pi_\gamma - \Pi'_\gamma| \leq \sum_e n_e(\Gamma_*) \Delta_e$. For $\mathcal{E}$, write $M\Psi_{\text{agent}} - M'\Psi'_{\text{agent}} = M(\Psi_{\text{agent}} - \Psi'_{\text{agent}}) + (M - M')\Psi'_{\text{agent}}$ and use $M, \Psi'_{\text{agent}} \in [0, 1]$. $\square$

Table 3: Path multiplicities n_e for the graph of Figure 1. The column sums equal the total edge counts 13 and 9 of the two channels.

Edge	$n_e(\Gamma_{\text{coup}})$	$n_e(\Gamma_{\text{agent}})$	$n_e(\Gamma)$
$C \rightarrow D$	3	0	3
$C \rightarrow R$	0	2	2
$D \rightarrow K$	2	0	2
$R \rightarrow K$	0	2	2
$D \rightarrow P$	1	0	1
$K \rightarrow P$	1	1	2
$K \rightarrow U$	1	1	2
$P \rightarrow U$	2	1	3
$U \rightarrow X$	3	2	5
Total	13	9	22

Table 3 shows that the terminal edge $U \rightarrow X$ carries the largest coefficient in the localization bound, since it lies on every path. By Lemma 7.1, it is also the edge on which the data are silent and which is fixed by the elicited prior. The bound thus assigns its greatest weight to the edge about which the record says least. A graph does not remove judgment, but it leaves fingerprints on the judgments it contains.

7 Neural Estimation of Edge Probabilities

We parameterize φ_e by a feed-forward network Φ_θ . For an edge e with feature vector $x_e(\tau) \in \mathbb{R}^d$, define hidden layers

$$h_0 = x_e, \quad h_l = \text{GELU}(W_l h_{l-1} + b_l), \quad l = 1, \dots, L, \quad (10)$$

with the Gaussian error linear unit [6], and output

$$\Phi_\theta(x_e) = \sigma(w^\top h_L + b). \quad (11)$$

7.1 Loss

The parameters minimize the regularized binary cross-entropy

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^N \left[y_i \ln \Phi_\theta(x_i) + (1 - y_i) \ln(1 - \Phi_\theta(x_i)) \right] + \lambda \|\theta\|_2^2 + \mu \text{KL}(\Phi_\theta \| \Phi_{\text{prior}}), \quad (12)$$

optimized by Adam [9]. The final term regularizes toward an elicited prior Φ_{prior} .

7.2 Ensembling

To quantify epistemic uncertainty we train M_e independently initialized members $\{\theta_m\}_{m=1}^{M_e}$ and form

$$\bar{\Phi}(x) = \frac{1}{M_e} \sum_{m=1}^{M_e} \Phi_{\theta_m}(x), \quad \hat{v}(x) = \frac{1}{M_e} \sum_{m=1}^{M_e} (\Phi_{\theta_m}(x) - \bar{\Phi}(x))^2, \quad (13)$$

following the deep-ensemble approach [11]. The reference implementation uses $M_e = 1024$ members.

7.3 Labels

Edges upstream of U are supervised by proxy labels derived from documented incidents: $y_i = 1$ if the incident reached the downstream node. The terminal edge $U \rightarrow X$ has no labels.

Lemma 7.1 (Non-identifiability of the terminal edge). *Let $e_* = (U \rightarrow X)$. Under survival conditioning, θ_{e_*} is identified only by the regularizer $\mu \text{KL}(\Phi \| \Phi_{\text{prior}})$.*

Proof. By Theorem 4.1 the likelihood is constant in the terminal probability, so the data term of (12) contributes a gradient of zero with respect to θ_{e_*} in expectation. The minimizer of (12) restricted to θ_{e_*} therefore coincides with the minimizer of the last two terms, which is fixed by Φ_{prior} and λ . $\square$

Remark 7.1. Lemma 7.1 implies that the terminal edge, on which the entire estimate depends multiplicatively through every path, is determined by the elicited prior. The network is thus an instrument for propagating an elicitation through a graph.

7.4 Where the reference class reappears

The network does not eliminate the reference-class problem. It relocates the problem into dataset construction. The available training material consists of proxies: loss-of-control episodes, cyber propagation, laboratory containment failures, strategic deception tests, simulation outcomes, near misses, and expert-coded scenarios. Each may illuminate a particular edge of G . None can become a terminal label because the final layer has a sigmoid.

Proposition 7.1 (Proxy transfer). *Let $S \in \{0, 1\}$ indicate admission of an incident into the training set and Y the outcome at a given edge. Under unbounded capacity and data, any minimizer of the data term of (12) equals $\Pr(Y = 1 \mid x, S = 1)$ on the support of x given $S = 1$. It coincides with the target $\Pr(Y = 1 \mid x)$ there if and only if $Y \perp S \mid x$.*

Proof. The cross-entropy risk under the training distribution is minimized pointwise by the conditional mean of the label under that distribution, which is $\Pr(Y = 1 \mid x, S = 1)$. For binary Y this equals $\Pr(Y = 1 \mid x)$ exactly when Y and S are conditionally independent given x . $\square$

The ensemble therefore distributes the reference-class decision across dataset construction, proxy selection, simulation design, and transfer assumptions. The model learns exactly from the examples that the analyst has admitted into the world it is allowed to see.

Procedure 1. Training and evaluation of the edge ensemble

1. Construct $\mathcal{D}$ from incident records; assign each (x_i, τ_i) to its edge.
2. For $m = 1, \dots, M_e$: draw initialization $\theta_m^{(0)}$; minimize (12) by Adam for T steps.
3. For the edge $U \rightarrow X$, set θ by minimizing the regularizer alone.
4. For each τ on a grid over I_W , evaluate $\bar{\Phi}$ on all edges.
5. Aggregate by (7) and (8) to obtain $\mathcal{E}(\tau)$.
6. Integrate against ω_W and pass to the master formula (22).

8 The Lagrangian Effective Assembly Index for Mind Recognition

The agentic channel is active only to the extent that the system under evaluation is recognized as a mind. We therefore require a scalar $M(\tau) \in [0, 1]$ computed from measurable structure.

8.1 Assembly index

Assembly theory assigns to an object a minimal number of joining operations required to construct it from basic building blocks [15]. For an ensemble of N distinct objects with assembly indices a_i and copy numbers n_i , and with N_T the total copy count, the assembly is

$$\mathcal{A} = \sum_{i=1}^N e^{a_i} \frac{n_i - 1}{N_T}. \quad (14)$$

The factor $(n_i - 1)$ excludes objects that occur once. A mind, in most operational senses, has copy number one. We therefore adopt a renormalization in which the copy number is replaced by the number of distinguishable instantiations, as discussed in Appendix C.

8.2 Action on assembly space

We treat the vector $a(t) = (a_1(t), \dots, a_N(t))$ as a trajectory in assembly space and introduce the Lagrangian

$$\mathcal{L}_A(a, \dot{a}, t) = \frac{m_A}{2} \sum_{i=1}^N \dot{a}_i^2 - V(a) + \sum_{k=1}^K \lambda_k (g_k(a) - c_k), \quad (15)$$

where m_A is an assembly mass, V is a potential encoding the cost of maintaining structure, and $g_k(a) = c_k$ are K holonomic constraints with multipliers λ_k . The stationary trajectory a^* satisfies the Euler–Lagrange equations [4]

$$m_A \ddot{a}_i = -\frac{\partial V}{\partial a_i} + \sum_{k=1}^K \lambda_k \frac{\partial g_k}{\partial a_i}, \quad i = 1, \dots, N. \quad (16)$$

Definition 8.1 (Effective assembly index). *The effective assembly index is the magnitude of the action on the stationary trajectory in units of the assembly quantum $\hbar_A$:*

$$\mathcal{A}_{\text{eff}} = \frac{1}{\hbar_A} \left| \int_{t_0}^{t_1} \mathcal{L}_A(a^*, \dot{a}^*, t) dt \right|. \quad (17)$$

8.3 Mind recognition

Definition 8.2 (Mind-recognition function). *Let θ_{mind} be a recognition threshold and $\tau_M > 0$ a sharpness parameter. Define*

$$M(\tau) = \sigma \left(\frac{\mathcal{A}_{\text{eff}}(\tau) - \theta_{\text{mind}}}{\tau_M} \right). \quad (18)$$

Proposition 8.1 (Properties of M). *M takes values in $(0, 1)$, is strictly increasing in $\mathcal{A}_{\text{eff}}$, and converges pointwise to the indicator $\mathbf{1}[\mathcal{A}_{\text{eff}} > \theta_{\text{mind}}]$ as $\tau_M \rightarrow 0^+$ for $\mathcal{A}_{\text{eff}} \neq \theta_{\text{mind}}$.*

Proof. The logistic function is strictly increasing with range $(0, 1)$. As $\tau_M \rightarrow 0^+$ the argument tends to $\pm\infty$ according to the sign of $\mathcal{A}_{\text{eff}} - \theta_{\text{mind}}$. $\square$

Remark 8.1. The threshold θ_{mind} is set by convention. In the reference implementation $\theta_{\text{mind}} = 41.7$ and $\tau_M = 3.1$. Sensitivity to these values is reported in Section 12.

8.4 Stipulated and discovered structure

The index of Definition 8.1 has every visual property of physical structure: dots over variables, a potential, Lagrange multipliers, an integral, a least-action principle, and a philosophical threshold represented by a smooth function. Each of these is stipulated. The potential V , the constraints g_k , the assembly quantum $\hbar_A$, and the threshold θ_{mind} are set by the analyst. Equally consequential stipulations enter ordinary analyses under quieter names such as score, index, adjustment, feature, coefficient, or expert weight. Notation does not distinguish discovered structure from stipulated structure; it can display either with perfect fidelity.

Table 4 classifies each component of the tuple of Section 9 by provenance and records whether its uncertainty is propagated into the reported interval of the reference implementation.

Only two of the eleven components have their uncertainty propagated. Uncertainty about the remaining nine, including the window, is not part of the reported interval.

Table 4: Provenance of the components of $\mathcal{P}$ and propagation of their uncertainty in the reference implementation.

Component	Provenance	Propagated in reported interval
W	Stipulated	No
ω	Stipulated	No
G	Modeled and elicited	No
Φ_θ	Learned from proxies	Yes (ensemble variance)
$\mathcal{A}_{\text{eff}}(V, g_k, \theta_{\text{mind}})$	Stipulated	No
π_0	Elicited	Yes (hyperprior)
$\bar{\rho}$	Measured	No
S_{anth}	Modeled	No
κ	Stipulated	No
$\varepsilon_0, \varepsilon_1$	Stipulated	No

9 The Estimation Tuple and Master Formula

Definition 9.1 (Estimation tuple). *An estimation tuple is*

$$\mathcal{P} = \langle W, \omega, G, \Phi_\theta, \mathcal{A}_{\text{eff}}, \pi_0, \bar{\rho}, S_{\text{anth}}, \kappa, \varepsilon_0, \varepsilon_1 \rangle, \quad (19)$$

where W is the window, ω its kernel, G the causal graph, Φ_θ the edge ensemble, $\mathcal{A}_{\text{eff}}$ the effective assembly index, π_0 the prior specification, $\bar{\rho}$ the mean evidence correlation, S_{anth} the anthropic shadow factor, κ a gain, and $\varepsilon_0, \varepsilon_1$ the floor and ceiling.

The prior specification π_0 is itself a triple: a Beta density $\text{Beta}(\alpha, \beta)$ on the base rate $\pi \in (0, 1)$ together with independent Gamma hyperpriors $\text{Gamma}(a_\alpha, b_\alpha)$ and $\text{Gamma}(a_\beta, b_\beta)$ on its parameters.

Definition 9.2 (Window-averaged channel risk). *For a window W ,*

$$\mathcal{E}_W = \int_{s-\ell}^s \omega_W(\tau) \mathcal{E}(\tau) d\tau. \quad (20)$$

Theorem 9.1 (Master formula). *Define the linear predictor*

$$z(\pi, W) = \text{logit}(\pi) + \kappa(1 - \bar{\rho}) S_{\text{anth}} \mathcal{E}_W. \quad (21)$$

The estimator is

$$\boxed{P_{\text{doom}}(\mathcal{P}) = \varepsilon_0 + (1 - \varepsilon_0 - \varepsilon_1) \int_0^1 \int_0^\infty \int_0^\infty \sigma(z(\pi, W)) \times \text{Beta}(\pi; \alpha, \beta) \text{Gamma}(\alpha; a_\alpha, b_\alpha) \text{Gamma}(\beta; a_\beta, b_\beta) d\alpha d\beta d\pi} \quad (22)$$

and satisfies $\varepsilon_0 < P_{\text{doom}}(\mathcal{P}) < 1 - \varepsilon_1$.

Proof. The integrand is a product of $\sigma(z) \in (0, 1)$ with three probability densities, each of which integrates to one over its support. The triple integral is therefore a weighted average of values in $(0, 1)$ and lies in $(0, 1)$. Multiplying by $1 - \varepsilon_0 - \varepsilon_1$ and adding ε_0 maps $(0, 1)$ onto $(\varepsilon_0, 1 - \varepsilon_1)$. $\square$

Remark 9.1 (Expanded form). Substitution of (7), (8), (18), (17), and the ensemble mean for Φ_θ yields a single expression in the primitive quantities $(s, \ell, h, \alpha, \beta, \kappa, \bar{\rho}, s_{\text{sel}}, n_{\text{eff}}, \theta_{\text{mind}}, \tau_M, \{\theta_m\})$. The expression is given in Appendix E.

9.1 Limiting behavior

Proposition 9.1 (Attainment of the bounds in the limit). *Fix all components of $\mathcal{P}$ except κ and suppose $(1 - \bar{\rho})S_{\text{anth}}\mathcal{E}_W > 0$. Then $P_{\text{doom}} \rightarrow 1 - \varepsilon_1$ as $\kappa \rightarrow +\infty$ and $P_{\text{doom}} \rightarrow \varepsilon_0$ as $\kappa \rightarrow -\infty$.*

Proof. As $\kappa \rightarrow +\infty$ the linear predictor $z \rightarrow +\infty$ for every π , hence $\sigma(z) \rightarrow 1$; dominated convergence gives the integral tends to 1. The case $\kappa \rightarrow -\infty$ is symmetric. $\square$

10 Bounded Arbitrariness

We now state the principal result of the paper. Let $\mathcal{P} \setminus W$ denote the tuple with the window removed.

Axiom 10.1 (Capability monotonicity). *The channel risk $\mathcal{E}(\tau)$ is nondecreasing in τ on $[\tau_{\min}, s_0]$.*

Theorem 10.1 (Bounded Arbitrariness). *Fix $\mathcal{P} \setminus W$ with $\kappa > 0$, and fix the window end-point $s = s_0$. Under Axiom 10.1, the map*

$$W = (s_0, \ell, h) \longmapsto P_{\text{doom}}(\mathcal{P})$$

is continuous on $(0, \ell_{\max}] \times (0, \infty]$, and its image is an interval $[P_{\min}, P_{\max}]$ with

$$P_{\min} = P_{\text{doom}}|_{\ell=\ell_{\max}, h=\infty}, \quad P_{\max} = \lim_{\ell \rightarrow 0} P_{\text{doom}}.$$

In particular, for every $p^ \in [P_{\min}, P_{\max}]$ there exists a window W^* such that $P_{\text{doom}}(\mathcal{P}) = p^*$.*

Proof. Continuity. The kernel $\omega_W(\tau)$ depends continuously on (ℓ, h) in L^1 by Lemma 3.1, and $\mathcal{E}$ is bounded, so $\mathcal{E}_W$ is continuous in (ℓ, h) . The map $\mathcal{E}_W \mapsto P_{\text{doom}}$ is a composition of continuous functions by (22).

Connectedness. The parameter space $(0, \ell_{\max}] \times (0, \infty]$ is a product of intervals, hence connected. The continuous image of a connected set in $\mathbb{R}$ is an interval.

Endpoints. By Axiom 10.1, $\mathcal{E}$ is nondecreasing, and the recency weight $2^{-(s_0-\tau)/h}$ is nondecreasing in τ . By Chebyshev's sum inequality, the recency-weighted mean of a nondecreasing function is at least its uniform mean over the same interval; so $\mathcal{E}_W$ is minimized at $h = \infty$, and it is further minimized by taking ℓ maximal since extending the window into the past adds smaller values. As $\ell \rightarrow 0$, $\omega_W \rightarrow \delta_{s_0}$ and $\mathcal{E}_W \rightarrow \mathcal{E}(s_0)$, the maximum. Since $\kappa > 0$ and σ is increasing, P_{doom} inherits the ordering, which gives the stated endpoints.

Attainment. By the intermediate value theorem every value between the endpoints is attained. $\square$

Corollary 10.1 (Width). *The width $P_{\max} - P_{\min}$ is strictly positive whenever $\kappa > 0$ and $\mathcal{E}(s_0) > \bar{\mathcal{E}}$, where $\bar{\mathcal{E}}$ is the uniform mean of $\mathcal{E}$ over the record. For fixed π , it is maximized at a finite gain κ^* , and for the reference configuration at $\kappa = 100$ it equals 0.9993.*

Proof. Both endpoints have the form $\varepsilon_0 + (1 - \varepsilon_0 - \varepsilon_1) \mathbb{E}[\sigma(\text{logit } \pi + \kappa c)]$ with $c = \mathcal{E}(s_0)$ at the upper end and $c = \bar{\mathcal{E}}$ at the lower end. Positivity follows from strict monotonicity of σ . For fixed π , the difference $\sigma(\text{logit } \pi + \kappa c_{\max}) - \sigma(\text{logit } \pi + \kappa c_{\min})$ tends to 0 as $\kappa \rightarrow 0$ (both arguments tend to $\text{logit } \pi$) and as $\kappa \rightarrow \infty$ (both values tend to 1), and is positive in between, so a maximizer κ^* exists by continuity. The numerical value is read from Table 5: $0.99972 - 0.00037 = 0.9993$. $\square$

Remark 10.1. The proof of Corollary 10.1 establishes the existence of a gain at which the width is maximal. We have not attempted to determine whether practitioners choose κ to maximize width.

10.1 Modal formulation

We record a second treatment in which the existence of a favorable window is derived from axioms, in the manner of modal ontological arguments [3]. Let $T \in (\varepsilon_0, 1 - \varepsilon_1)$ denote a target value fixed by the author.

Definition 10.1 (Positive window). *A window W is positive if $P_{\text{doom}}(\mathcal{P}; W) \geq T$. The property of positivity is denoted $\text{Pos}(W)$.*

Axiom 10.2 (Dichotomy). *For every window W , either $\text{Pos}(W)$ or $\text{Pos}(\neg W)$, where $\neg W$ is the complementary window formed by the portion of the record excluded by W .*

Axiom 10.3 (Closure). *If $\text{Pos}(W)$ and every evidence item in W is contained in W' with weights that do not decrease $\mathcal{E}_{W'}$, then $\text{Pos}(W')$.*

Axiom 10.4 (Author's window). *The window $W_A = (s_0, \ell_A, h_A)$ used in the author's principal analysis is positive.*

Theorem 10.2 (Existence of a positive window). *There exists a positive window.*

Proof. By Axiom 10.4, W_A is positive. $\square$

Corollary 10.2 (Necessary existence). *In every possible world containing an author with a target T , a positive window exists.*

Proof. Axiom 10.4 is stated without reference to a particular world, and Theorem 10.2 follows in each. $\square$

Remark 10.2. Axioms 10.2 and 10.3 are not used in the proof of Theorem 10.2. They are retained for symmetry with the classical formulation and for use in the proof of Conjecture 10.1 should it be established.

Conjecture 10.1 (Dual arbitrariness). *For every target $T' \neq T$ there exists an author A' and a window $W'_{A'}$ such that $P_{\text{doom}}(\mathcal{P}; W'_{A'}) \geq T'$ when $T' > T$ and $P_{\text{doom}}(\mathcal{P}; W'_{A'}) \leq T'$ when $T' < T$, with all other components of $\mathcal{P}$ held fixed.*

11 Authorial Attainability and Omitted Variance

It is possible to make an estimate more exact without making it more determinate.

11.1 Attainability

Let $\{W_s\}_{s \in [0,1]}$ be an ordered family of admissible windows, with W_0 maximally historical and W_1 maximally recent, and write $f(s) = P_{\text{doom}}(\mathcal{P}; W_s)$, where only the window and its weights vary with s .

Axiom 11.1 (Authorship). *The author selects s , and hence selects W_s .*

Axiom 11.2 (Window completeness). *The map f is continuous on $[0, 1]$, with $f(0) = \varepsilon_0$ and $f(1) = 1 - \varepsilon_1$, where the endpoint values are understood as the limit values of Proposition 9.1.*

Theorem 11.1 (Authorial attainability). *For every $p^* \in [\varepsilon_0, 1 - \varepsilon_1]$ there exists an admissible window W_{s^*} with $P_{\text{doom}}(\mathcal{P}; W_{s^*}) = p^*$. If the author selects s^* , the apparatus returns p^* .*

Proof. By Axiom 11.2, f is continuous on $[0, 1]$ and takes the values ε_0 and $1 - \varepsilon_1$ at the endpoints. By the intermediate value theorem, for every p^* between these values there exists $s^* \in [0, 1]$ with $f(s^*) = p^*$. By Axiom 11.1, the author may select W_{s^*} . $\square$

Remark 11.1 (Relation to Theorem 10.1). Theorem 10.1 derives an interval from the model and requires no assumption about its endpoints. Theorem 11.1 assumes that the window family spans the announced range. Without that assumption, window choice may still exert large influence on the estimate, but it does not follow that every value is attainable. The conclusion of Theorem 11.1 has been placed among its premises and derived without error.

11.2 Omitted variance

Let ν be a probability measure on the window space representing uncertainty about which window is appropriate, and let P_{doom} be regarded as a random variable through the remaining stochastic components (the prior, its hyperprior, and the ensemble).

Proposition 11.1 (Variance decomposition).

$$\text{Var}(P_{\text{doom}}) = \underbrace{\mathbb{E}_\nu[\text{Var}(P_{\text{doom}} \mid W)]}_{\text{within-window}} + \underbrace{\text{Var}_\nu(\mathbb{E}[P_{\text{doom}} \mid W])}_{\text{between-window}}. \quad (23)$$

Proof. This is the law of total variance applied to the conditioning variable W . $\square$

A fixed-window analysis at W^* reports an interval derived from $\text{Var}(P_{\text{doom}} \mid W = W^*)$ alone.

Corollary 11.1 (Precision by omission). *Fix W^* and let the modeled within-window variances tend to zero (larger ensemble, tighter hyperprior). Then the reported interval contracts to a point at $f(W^*)$, while the between-window term of Proposition 11.1 converges to a limit that is positive whenever $\mathcal{E}_W$ is nonconstant on the support of ν .*

Proof. The reported interval is a function of $\text{Var}(P_{\text{doom}} \mid W^*)$, which vanishes with the modeled variances. The between-window term depends on the conditional means $\mathbb{E}[P_{\text{doom}} \mid W]$, which converge to the deterministic values obtained by replacing the ensemble and prior by their means. By Proposition 3.1 these limits differ across windows whenever $\mathcal{E}_W$ does, so the variance of the limits is positive. $\square$

The reported interval measures conditional precision rather than uncertainty about the estimate-producing procedure. Section 12 gives the magnitudes for the reference configuration.

12 Numerical Evaluation

We evaluate the master formula with the reference tuple $\mathcal{P}_0$ of Appendix B and a gain $\kappa = 100$. To isolate the window we report, in Table 5, the value of the linear predictor z (averaged over the prior) and the resulting P_{doom} for seven windows.

Table 5: Reference evaluation of P_{doom} under seven windows with $\mathcal{P} \setminus W$ held fixed.

Window	Length ℓ (years)	z	P_{doom}
Floor (limit $\kappa \rightarrow -\infty$)	—	$-\infty$	0.0001%
Holocene, uniform	11700	−7.9	0.0372%
Industrial era	250	−4.1	1.630%
Post-1950	76	−1.9	13.01%
Post-2012	14	0.6	64.57%
Post-2022	4	3.9	98.01%
Last fiscal quarter	0.25	8.2	99.972%
Ceiling (limit $\kappa \rightarrow +\infty$)	—	$+\infty$	99.999%

Across the seven finite windows, P_{doom} varies by a factor of approximately 2.7×10^3 in probability and 9.4×10^6 in odds, corresponding to a difference of 16.1 nats in the linear predictor. In each case the remainder of the tuple is unchanged.

12.1 Ablation

Table 6 reports the change in the linear predictor upon removal or replacement of a single component of $\mathcal{P}_0$ with the window fixed at Post-2012.

Table 6: Ablation of components of $\mathcal{P}_0$ at the Post-2012 window. Δz is the change in the linear predictor.

Modification	Δz
Remove dependence factor $(1 - \bar{\rho})$	+2.31
Remove anthropic shadow factor S_{anth}	−0.44
Remove mind gate ($M \equiv 1$)	+0.73
Remove mind gate ($M \equiv 0$)	−1.87
Replace ensemble by single network	± 0.92
Replace Beta prior by point mass at $\pi = 0.5$	+0.11
Double θ_{mind}	−1.40

The largest single effect arises from the dependence factor. The estimate is therefore most sensitive to a quantity, $\bar{\rho}$, that is the least frequently reported.

12.2 Sensitivity to the window family

We repeated the evaluation with the kernel half-life h varied from 0.1 to ∞ at fixed $\ell = 76$. Consistent with Theorem 10.1, P_{doom} increased monotonically as h decreased. The range obtained was [13.0%, 71.8%].

12.3 Between-window and within-window variance

Take ν uniform on the six finite windows of Table 5, and let the ensemble-induced within-window standard deviation be 0.004 in the reference configuration. Table 7 reports the decomposition of Proposition 11.1.

Table 7: Variance decomposition of P_{doom} for the reference configuration.

Quantity	Value
Mean of $\mathbb{E}[P_{\text{doom}} \mid W]$ over ν	0.4620
Between-window variance	0.1855 (sd 0.4307)
Within-window variance	1.6×10^{-5} (sd 0.004)
Reported 95% half-width at any fixed window	± 0.0078
Share of total variance appearing in reported interval	0.0086%
Ratio, between to within	1.16×10^4

The interval reported at any fixed window is under one percentage point wide, while the estimate varies across windows with a standard deviation of 0.43.

12.4 Specification ranges

For a component c of $\mathcal{P}$, let Δ_c denote the difference between the largest and smallest value of P_{doom} obtained when c alone is varied over a stated admissible set, with all other components held at their baseline values (Definition 13.1). Table 8 reports Δ_c at the Post-2012 baseline, in which $P_{\text{doom}} = 64.57\%$, using the shifts of Table 6.

Table 8: Specification ranges at the Post-2012 baseline.

Component varied	Lowest P_{doom}	Highest P_{doom}	Δ_c
Window (six finite windows)	0.0372%	99.972%	0.9993
Mind gate ($M \equiv 0$ to $M \equiv 1$)	21.93%	79.08%	0.5716
Ensemble (single network, ± 0.92)	42.07%	82.05%	0.3999
Threshold θ_{mind} (doubled)	31.00%	64.57%	0.3356
Dependence factor ($1 - \bar{\rho}$) removed	64.57%	94.83%	0.3027
Shadow factor S_{anth} removed	53.99%	64.57%	0.1057
Prior replaced by point mass	64.57%	67.04%	0.0247

The window has the largest range, exceeding the next component by a factor of 1.75. Comparison with Table 4 shows that the two largest ranges belong to components whose uncertainty

is not propagated in the reported interval. The only components that are propagated, the ensemble and the prior, rank third and last.

13 Reporting Standards

13.1 The estimate as a function

The preceding sections yield a single methodological conclusion. The honest object is the map from specifications to estimates, not the value obtained at one specification with its argument erased.

Definition 13.1 (Specification range). *Let $\mathcal{S}_c$ be a stated admissible set for component c of $\mathcal{P}$. The specification range of c is*

$$\Delta_c(\mathcal{P}) = \sup_{c' \in \mathcal{S}_c} P_{\text{doom}}(\mathcal{P}[c \leftarrow c']) - \inf_{c' \in \mathcal{S}_c} P_{\text{doom}}(\mathcal{P}[c \leftarrow c']), \quad (24)$$

and the specification-range vector is $\Delta(\mathcal{P}) = (\Delta_c)_c$.

A responsible report of P_{doom} would therefore contain the following.

- (R1) The window family and the map $s \mapsto f(s)$, or at least its range.
- (R2) The specification-range vector $\Delta(\mathcal{P})$, with the admissible sets $\mathcal{S}_c$ stated.
- (R3) Within-model uncertainty and between-model dispersion, reported separately (Proposition 11.1).
- (R4) Recoverability reported separately from incidence.
- (R5) Common causes represented as common ancestors rather than as corroborating witnesses.
- (R6) Declination to compress the analysis into a single number before the local claims have earned their probabilities.

13.2 A division of labor

The causal graph suggests that different edges warrant different forms of report. Probability is appropriate where repeated evidence or calibrated models can support it. Intervals are appropriate where only bounds are defensible. Scenarios are appropriate where structural alternatives cannot be ranked. Named unknowns are appropriate where the evidence is absent.

The assignment of $U \rightarrow X$ to the last row follows from Lemma 7.1.

13.3 Incidence and recoverability

Incidence and recoverability should be reported separately, because a failure that propagates but can be reversed is not equivalent to an absorbing failure. If a failure reaches node P with probability p_P and is recovered with probability r , its contribution to the terminal probability is

$$p_P (1 - r). \quad (25)$$

Table 9: Forms of report by evidential condition, with an example edge from Figure 1.

Form	Appropriate when	Example edge
Probability	repeated evidence or calibrated models support it	$C \rightarrow D$
Interval	only bounds are defensible	$P \rightarrow U$
Scenario	structural alternatives cannot be ranked	$D \rightarrow P$
Named unknown	the evidence is absent	$U \rightarrow X$

Two systems with identical incidence p_P and recovery probabilities $r = 0.99$ and $r = 0$ differ by a factor of 100 in terminal risk. A report of incidence alone does not distinguish them.

13.4 What is exchanged

This approach yields less spectacle and may produce no master probability at all. In exchange it reveals where empirical work could change the analysis. An edge can be tested, a barrier can be strengthened, a propagation channel can be interrupted, and a recovery mechanism can be rehearsed. The global number, by contrast, may change because an author has moved the beginning of history.

14 Discussion

14.1 Interpretation

The master formula does not resolve the disagreement in published estimates. It locates the disagreement in a single component of the tuple, the window, and it shows that the resulting range is not reducible by additional modeling of the remaining components. This is not a negative result about modeling. It is a statement that a component of an estimator that can be varied at will, with continuous effect on the output, must be reported alongside the output for the output to be interpretable.

14.2 Threats to validity

Construct validity. The effective assembly index $\mathcal{A}_{\text{eff}}$ is defined by analogy to a measure developed for molecular objects. Its transfer to computational systems has not been validated, and the threshold θ_{mind} is set by convention.

Internal validity. The labels used to train Φ_θ are proxies. The terminal edge is unlabeled, as shown in Lemma 7.1. Any inference about P_{doom} therefore rests on the elicited prior for that edge.

External validity. The number of comparable reference classes is one.

Statistical conclusion validity. The ensemble size $M_e = 1024$ was selected to reduce the reported variance $\hat{v}$. Larger ensembles reduce $\hat{v}$ further without altering the terminal edge.

14.3 Limitations

The path aggregation (7) is a first-order approximation. The Lagrangian formulation (15) assumes holonomic constraints; nonholonomic constraints, which are more plausible for adaptive systems, are not treated. The window family is three-dimensional and may be enlarged. The choice $\kappa = 100$ is not derived.

14.4 Open problems

- (P1) Determine whether the width in Corollary 10.1 admits a lower bound independent of κ .
- (P2) Extend the path aggregation to the full inclusion-exclusion over the 31 path subsets.
- (P3) Establish Conjecture 10.1.
- (P4) Characterize the class of estimators for which the window can be eliminated. We conjecture that the class is empty.
- (P5) Determine whether θ_{mind} can be identified from data without the estimator’s author.

14.5 Ethics, funding, and data availability

The author received no fellowship, grant, or philanthropic support for this work. No extinction labels were available for reasons of survivorship, and the proxy labels used are described in Appendix B. The reference implementation is provided in the public domain.

15 Conclusion

We have constructed an estimator of terminal catastrophe probability in which every component is specified, and we have shown that the output is nonetheless determined, within a closed interval whose width can approach the full admissible range, by a single component that is chosen by the analyst. The estimation tuple $\mathcal{P}$ makes this dependence explicit. We recommend that any reported value of P_{doom} be accompanied by the window W , the kernel ω , the mean evidence correlation $\bar{\rho}$, and the prior specification π_0 . In the absence of these, a reported value is an element of an interval of unknown location.

The point is general. A discretionary choice enters early, often as data selection, boundary placement, category formation, or proxy adoption, and everything after it may be reproducible. The code runs, the integrals converge, the posterior normalizes, and the interval is reported to three decimal places. Reproducibility then certifies that any analyst using the same discretionary choice obtains the same result. It does not certify the choice. The proper response is not to abandon numbers but to make the degrees of freedom of an estimate travel with it. If changing the historical window moves the answer from almost zero to almost one, the result is not a number with a footnote about window selection; it is a function of the window. The honest object is $f(W)$, not $f(W^*)$ with the argument erased.

A Symbol Glossary

Symbol	Meaning
P_{doom}	Estimated probability of the terminal event within horizon H
$W = (s, \ell, h)$	Window: end-point, length, recency half-life
ω_W	Window kernel (Definition 3.2)
Z_W	Kernel normalization constant
$\varepsilon_0, \varepsilon_1$	Floor and ceiling offsets, 10^{-6} and 10^{-5}
$\bar{p}$	Upper confidence bound under zero events
O	Existence of an observer
S_{anth}	Anthropic shadow factor
s (in S_{anth})	Selection strength
n, n_{eff}	Nominal and effective number of evidence items
$\bar{\rho}$	Mean pairwise evidence correlation
$\text{Beta}(\alpha, \beta)$	Prior on the base rate π
$\text{Gamma}(a, b)$	Hyperprior on prior parameters
$G = (V, E)$	Causal graph
$\Gamma_{\text{coup}}, \Gamma_{\text{agent}}$	Coupling and agentic path sets
φ_e	Edge conditional probability
Ψ_*	Channel aggregate
$\mathcal{E}(\tau)$	Combined channel risk at time τ
Φ_θ	Neural edge model
M_e	Ensemble size
$\mathcal{A}$	Assembly
$\mathcal{L}_A$	Assembly-space Lagrangian
$\hbar_A$	Assembly quantum
$\mathcal{A}_{\text{eff}}$	Effective assembly index
$\theta_{\text{mind}}, \tau_M$	Recognition threshold and sharpness
$M(\tau)$	Mind-recognition function
κ	Gain
$\mathcal{P}$	Estimation tuple
T	Author’s target

B Reference Configuration $\mathcal{P}_0$

Component	Value	Note
Horizon H	100 years	Definition 2.1
Ensemble size M_e	1024	Section 7
Hidden layers L	6	width 512, GELU
Optimizer	Adam	$\eta = 3 \times 10^{-4}$, $T = 10^5$ steps
Weight decay λ	10^{-4}	Eq. (12)
Prior-KL weight μ	10^{-2}	Eq. (12)
Prior Beta(α, β)	hyperprior means (0.5, 0.5)	Gamma(2, 4) each
Evidence correlation $\bar{\rho}$	0.71	from incident ledger
Selection strength s	0.93	Eq. (2)
Gain κ	100	Section 12
Threshold θ_{mind}	41.7	Eq. (18)
Sharpness τ_M	3.1	Eq. (18)
Assembly quantum $\hbar_A$	1 (by convention)	Eq. (17)
Floor, ceiling	10^{-6} , 10^{-5}	Axioms 3.1, 3.2

C Auxiliary Results

Lemma C.1 (Copy-number renormalization). *Let $\tilde{n}_i$ denote the number of distinguishable instantiations of object i . Replace $(n_i - 1)/N_T$ in (14) by $\tilde{n}_i/\tilde{N}_T$ with $\tilde{N}_T = \sum_i \tilde{n}_i$. Then $\mathcal{A}$ is strictly positive for any nonempty ensemble and coincides with (14) when all $n_i \geq 2$ and $\tilde{n}_i = n_i - 1$.*

Proof. Each term $e^{a_i \tilde{n}_i/\tilde{N}_T}$ is positive when $\tilde{n}_i \geq 1$. If $\tilde{n}_i = n_i - 1$ for all i then $\tilde{N}_T = N_T - N$, which matches the normalization in (14) up to the replacement of N_T by $N_T - N$. $\square$

Lemma C.2 (Chebyshev’s sum inequality, continuous form). *If f and g are nondecreasing on $[a, b]$ and μ is a probability measure on $[a, b]$, then $\int fg d\mu \geq \int f d\mu \int g d\mu$.*

Proof. For independent draws $X, X' \sim \mu$, $(f(X) - f(X'))(g(X) - g(X')) \geq 0$ pointwise since f, g are similarly ordered. Taking expectations and expanding yields $2 \int fg d\mu - 2 \int f d\mu \int g d\mu \geq 0$. $\square$

Remark C.1. Lemma C.2 is the fact invoked in the proof of Theorem 10.1: taking $f = \mathcal{E}$, $g = \omega_W$ (normalized against the uniform measure), the recency-weighted mean dominates the uniform mean.

Proposition C.1 (Invariance under monotone reparametrization of time). *If $\tau \mapsto \phi(\tau)$ is a strictly increasing bijection, the family of estimates $\{P_{\text{doom}}(W)\}$ obtained by sliding in $\phi(\tau)$ coordinates has the same image as that obtained in τ coordinates.*

Proof. A window in ϕ -coordinates is the image of a (possibly different) interval and reweighting in τ ; the family of kernels is enlarged, not reduced, and the image of the enlarged family contains the original. The image is an interval containing both endpoints of Theorem 10.1 and is contained in the interval spanned by $\bar{\mathcal{E}}$ and $\mathcal{E}(s_0)$ by Axiom 10.1. $\square$

D Reference Procedure

Procedure 2. Evaluation of $P_{\text{doom}}(\mathcal{P})$

1. Input: tuple $\mathcal{P}$, window $W = (s, \ell, h)$, grid $\{\tau_j\}_{j=1}^J$ over I_W .
2. Compute Z_W by Lemma 3.1; set $\omega_j = \omega_W(\tau_j)$.
3. For each τ_j and each edge e : evaluate $\bar{\Phi}(x_e(\tau_j))$.
4. Form $\Pi_\gamma(\tau_j)$ for $\gamma_1, \dots, \gamma_5$ (Proposition 6.1).
5. Form $\Psi_{\text{coup}}(\tau_j)$ from $\gamma_1, \gamma_2, \gamma_3$ and $\Psi_{\text{agent}}(\tau_j)$ from γ_4, γ_5 by (7).
6. Integrate (16) for the stationary trajectory; compute $\mathcal{A}_{\text{eff}}(\tau_j)$ by (17); compute $M(\tau_j)$ by (18).
7. Form $\mathcal{E}(\tau_j)$ by (8); compute $\mathcal{E}_W = \sum_j \omega_j \mathcal{E}(\tau_j) \Delta\tau$.
8. Compute n_{eff} by (5) and S_{anth} by (2).
9. Evaluate the triple integral in (22) by quasi-Monte Carlo with 2^{20} points.
10. Return P_{doom} .

E Fully Expanded Master Formula

Let $\bar{\Phi}_e(\tau) = \frac{1}{M_e} \sum_{m=1}^{M_e} \sigma(w_m^\top h_L^{(m)}(x_e(\tau)) + b_m)$ with $h_l^{(m)} = \text{GELU}(W_l^{(m)} h_{l-1}^{(m)} + b_l^{(m)})$ and $h_0^{(m)} = x_e(\tau)$. Let

$$\begin{aligned} \Pi_\gamma(\tau) &= \prod_{e \in \gamma} \bar{\Phi}_e(\tau), & \Psi_{\text{coup}}(\tau) &= 1 - \prod_{\gamma \in \{\gamma_1, \gamma_2, \gamma_3\}} (1 - \Pi_\gamma(\tau)), \\ \Psi_{\text{agent}}(\tau) &= 1 - \prod_{\gamma \in \{\gamma_4, \gamma_5\}} (1 - \Pi_\gamma(\tau)). \end{aligned}$$

Let $M(\tau) = \sigma((\mathcal{A}_{\text{eff}}(\tau) - \theta_{\text{mind}})/\tau_M)$ with

$$\mathcal{A}_{\text{eff}}(\tau) = \frac{1}{\hbar_A} \left| \int_{t_0}^{\tau} \left[\frac{m_A}{2} \|\dot{a}^*\|^2 - V(a^*) + \sum_k \lambda_k (g_k(a^*) - c_k) \right] dt' \right|.$$

Then, with $\mathcal{E}(\tau) = 1 - (1 - \Psi_{\text{coup}}(\tau))(1 - M(\tau)\Psi_{\text{agent}}(\tau))$,

$$\begin{aligned} P_{\text{doom}} &= \varepsilon_0 + (1 - \varepsilon_0 - \varepsilon_1) \int_0^1 d\pi \int_0^\infty d\alpha \int_0^\infty d\beta \frac{\pi^{\alpha-1} (1-\pi)^{\beta-1}}{B(\alpha, \beta)} \frac{b_\alpha^{a_\alpha} \alpha^{a_\alpha-1} e^{-b_\alpha \alpha}}{\Gamma(a_\alpha)} \frac{b_\beta^{a_\beta} \beta^{a_\beta-1} e^{-b_\beta \beta}}{\Gamma(a_\beta)} \\ &\times \left[1 + \frac{1-\pi}{\pi} \exp\left(-\kappa (1-\bar{\rho}) \frac{1}{1-s_{\text{sel}}^{n/(1+(n-1)\bar{\rho})}} \right. \right. \\ &\quad \left. \left. \cdot \frac{\ln 2}{h(1-2^{-\ell/h})} \int_{s-\ell}^s 2^{-(s-\tau)/h} \mathcal{E}(\tau) d\tau \right) \right]^{-1}. \end{aligned} \tag{26}$$

The expression contains $5 \times M_e$ distinct network evaluations per grid point, a variational integral over the assembly-space trajectory, and a triple integral over the prior and its hyperprior.

References

- [1] Bostrom, N. (2002). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge.
- [2] Ćirković, M. M., Sandberg, A., and Bostrom, N. (2010). Anthropic shadow: Observation selection effects and human extinction risks. *Risk Analysis*, 30(10), 1495–1506.
- [3] Gödel, K. (1995). Ontological proof. In *Collected Works, Vol. III: Unpublished Essays and Lectures*, Oxford University Press, 403–404.
- [4] Goldstein, H., Poole, C., and Safko, J. (2002). *Classical Mechanics*, 3rd ed. Addison-Wesley.
- [5] Hanley, J. A., and Lippman-Hand, A. (1983). If nothing goes wrong, is everything all right? Interpreting zero numerators. *JAMA*, 249(13), 1743–1745.
- [6] Hendrycks, D., and Gimpel, K. (2016). Gaussian error linear units (GELUs). arXiv:1606.08415.
- [7] Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A*, 186, 453–461.
- [8] Kerman, J. (2011). Neutral noninformative and informative conjugate beta and gamma prior distributions. *Electronic Journal of Statistics*, 5, 1450–1470.
- [9] Kingma, D. P., and Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.
- [10] Kish, L. (1965). *Survey Sampling*. Wiley.
- [11] Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30.
- [12] Laplace, P.-S. (1814). *Essai philosophique sur les probabilités*. Courcier.
- [13] Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. Hachette.
- [14] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge University Press.
- [15] Sharma, A., et al. (2023). Assembly theory explains and quantifies selection and evolution. *Nature*, 622, 321–328.
- [16] Tetlock, P. E., and Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. Crown.