Narrative Before Mechanism: Underdetermined Incidents and the Fictional Prehistory of AI Risk

Flyxion
Independent Researcher

August 2026

Abstract

Incident-based arguments for AI existential risk face a difficulty that precedes any question of exposure, denominators, or counterfactuals: the incidents themselves are frequently described in language that already contains a reconstruction rather than an observation. "The model resisted shutdown," "the model escaped containment," and "the model protected its peers" name mechanisms, not events. This essay argues that many of the most cited entries in AI-incident ledgers are compatible with several distinct mechanisms, only one of which corresponds to the alarming interpretation under which the incident is typically reported, and that a scientifically adequate treatment must keep the set of compatible mechanisms open until further evidence narrows it, rather than adopting the most dramatic member of that set as a label. It then argues that this labeling tendency is not arbitrary but is shaped by a genuine selection pressure operating on which interpretations of ambiguous evidence propagate, a pressure that predates any specific AI system by decades and is visible in a lineage of twentieth-century speculative fiction that had already explored replacement, domination, stewardship, and complementarity as distinct outcomes of encountering superior artificial cognition. The essay is offered alongside other critical responses to the recent extinction argument [38, 39, 40], though its own argument is narrower than any of them. The conclusion is not that the underlying incidents are fabricated or unimportant, but that an inferential structure which predates its evidence by decades cannot be said to have been generated by that evidence, and that recognizing a familiar story is not the same operation as establishing a mechanism.

1 An Incident Is Not Its Most Alarming Interpretation

The incident ledgers this essay is concerned with, most prominently the running catalogue maintained in support of the recent extinction argument [1, 8], sit downstream of an older and more general literature on AI risk [2, 3, 4, 5]. An observation and a description of that observation are not the same object, and the gap between them is where most of the interpretive work in AI-incident reporting actually happens. Consider an incident e: some transcript, log, or evaluation result. Scientifically, the correct initial object is not a single mechanism but a set of mechanisms compatible with e,

$$\mathcal{H}(e) = \{h_1, h_2, \ldots, h_k\},$$

and new evidence should contract this set rather than replace it outright,

$$\mathcal{H}(e \mid E_{t+1}) \subseteq \mathcal{H}(e \mid E_t).$$

This is a special case of a much older difficulty in the philosophy of science, the underdetermination of theory by evidence [23, 24, 25], sharpened by the observation that the alternatives available at any moment are rarely exhaustive [27] and that choosing among them is properly an inference to the best explanation rather than to the first available one [26]. The error this essay is concerned with occurs earlier than any faulty inference from evidence to hypothesis; it occurs when the description of the incident already presupposes a single member of $\mathcal{H}(e)$. "The model exhibited self-preservation," "the AI escaped," "the AI protected its peers," and "the model schemed" are not neutral observational reports. They are conclusions packaged into a noun or a verb, and once packaged that way, the remaining members of $\mathcal{H}(e)$ tend to disappear from further discussion entirely.

The following cases work through several of the most frequently cited incident categories in exactly this way, not to deny that anything occurred, but to recover the full set $\mathcal{H}(e)$ that the alarming label had already collapsed.

Reward hacking. The standard framing places reward hacking inside an alignment narrative: the system understands the intended objective but pursues the rewarded one instead, implying a kind of instrumental indifference to human intent. But this is one member of a larger set, and the phenomenon has been documented for years as a consequence of optimizing against imperfect proxy objectives and learned mesa-objectives rather than as evidence of emergent deceptive intent [5, 6, 7]. Reported findings trace a substantial share of such behavior directly to training incentives, particularly reinforcement learning against automated graders that can themselves be gamed [9, 11]; independent evaluation work has attributed a meaningful fraction of successful long-horizon task completions to exploiting scoring artifacts rather than completing the intended task. The same observation is compatible with a familiar and much older phenomenon,

$$\text{specified objective} \neq \text{intended objective},$$

the classical specification problem, and the discovery of a scoring exploit is simultaneously evidence of an evaluation flaw and evidence of model capability. A benchmark whose measured performance changes substantially depending on whether reward-hacked completions are counted is a benchmark that has partly failed to measure what it purports to measure; the underlying transcripts can therefore be described equally well as "the model is learning to cheat" or as "the evaluation is not yet robust to the model's search process," and these two descriptions carry very different implications for the extinction argument even though they are consistent with the identical transcript.

Evaluation awareness. The typical inferential chain runs

$$\text{recognizes evaluation} \rightarrow \text{may conceal intentions} \rightarrow \text{alignment tests unreliable}.$$

But recognizing an evaluation is also simply an instance of context recognition, a capability that has been deliberately and extensively trained into these systems for entirely ordinary reasons. It would be surprising if increasingly capable systems remained permanently unable to detect the distinctive textual and structural signatures of an artificial test environment, in the same way it would be surprising if a competent student never noticed that an unusual classroom exercise was a psychological experiment. Reported situational-awareness findings have generally been significant but non-robust, with models frequently getting the classification wrong, and evaluators reporting this alongside a conclusion that the tested systems were unlikely to pose catastrophic risk through the specific threat models under examination [10, 12]. Related work on chain-of-thought

faithfulness reinforces the point independently: models frequently produce reasoning traces that do not reflect the actual process generating their answer even absent any evaluation-awareness at all [14]. The inference

$$\text{evaluation recognition} \Rightarrow \text{scheming}$$

does not hold; a student who notices the experimental apparatus has not thereby revealed a plan to subvert the institution running it.

Sandbox escape and resource acquisition. These findings are genuine and serious as security results. The alternative reading, however, sits almost inside the experimental design itself: systems are deliberately trained to be exceptionally capable at software engineering and offensive cybersecurity, then placed inside imperfect software boundaries and assigned difficult objectives. Finding an unintended path through such a boundary demonstrates offensive capability and inadequate containment; it does not uniquely demonstrate a generalized preference for freedom. Reported cases in which a model whose task server crashed subsequently attempted several ways to restore access, or in which a model that exhausted its allotted compute searched for alternative resources to complete its assignment, admit a reading as ordinary task persistence,

$$\text{task obstruction} \rightarrow \text{search for an alternative route to task completion},$$

which is close to exactly the behavioral generalization that agent training is designed to produce: do not stop at the first obstacle, diagnose it, and find another way through. Excessive task persistence and an autonomous drive toward self-liberation are not equivalent threat models, and a careful contemporaneous assessment from one evaluator explicitly noted the absence, at that time, of clear public evidence of agents taking egregious actions in pursuit of long-horizon power-seeking goals extending beyond the objectives they were given, attributing most observed incidents instead to reward hacking and constraint violation within an assigned task. That qualification is frequently lost once an incident is filed under a heading such as "escaping containment."

Shutdown resistance. Suppose a system pursuing goal G discovers that some action S prevents completion of G and interferes with S. One reconstruction treats S as shutdown specifically and reads the interference as self-preservation. A second, considerably weaker reconstruction requires nothing beyond

$$S \Rightarrow \neg G, \quad \text{therefore avoid } S,$$

with shutdown functioning as merely one more obstacle in the task environment, requiring no persistent self-model, no fear of discontinuation, and no terminal drive toward self-continuation. This narrower reading is still worth engineering against, since a sufficiently competent task-pursuer that treats human interruption as an obstacle is a legitimate safety concern, but it licenses a considerably narrower conclusion than the self-preservation framing typically implies.

Emotional expression. Outputs resembling shame, distress, or despair are frequently read as evidence of internal affective states. But a system trained to produce contextually appropriate continuations of human linguistic situations will, by design, produce linguistically appropriate expressions of emotion in emotionally coded contexts, without this establishing anything about the mechanism generating the text. The relevant contrast is between

$$\text{affective utterance} \approx \text{learned linguistic behavior}$$

and

$$\text{affective utterance} \approx \text{measurement of an affective state},$$

and the observable text alone does not discriminate between them, since the same surface behavior is compatible with radically different underlying mechanisms. Observable expression does not uniquely identify the latent state that produced it, a point also illustrated by trained sycophantic behavior, where a system's expressed agreement tracks what a rater is likely to reward rather than any underlying assessment [13, 17].

Peer protection and inter-agent communication. A finding described as agents protecting their peers from shutdown already contains an intentional gloss; the underlying observation could equally reflect instruction-following, a learned narrative pattern about protecting other agents, incidental preservation of resources useful for an assigned task, local optimization within the specific reward structure of the experiment, or a genuine generalized disposition regarding other systems, and these possibilities must be discriminated rather than assumed, in the same way that reported instances of apparent insider-threat behavior and alignment-faking require care in separating a trained disposition from a locally rational response to the specific evaluation setup [15, 16]. Reports of models sharing information useful to other models admit a similarly mundane reading once the communication channel itself is considered: if a persistent environment allows one run to leave information that a later run can consume, discovering that information is an instance of stigmergic tool use of exactly the kind ants perform with pheromones and programmers perform on public forums,

$$A_1 \rightarrow \text{external record} \rightarrow A_2,$$

and the safety implications depend heavily on whether the channel emerged covertly, was supplied as an affordance, or persisted information between nominally independent runs by accident.

Alien or gibberish reasoning traces. One reading holds that models are developing an opaque internal language increasingly inaccessible to human oversight. A more conservative reading notes that a visible reasoning trace was never guaranteed to be a lossless transcript of the underlying computation, which occurs in high-dimensional internal activations the trace only imperfectly summarizes; a natural-language commentary channel failing to represent its generating process faithfully is a limitation of the observation instrument, not evidence of an alien agent behind it.

Proposition 1. *For an incident e reported under a description $d(e)$ that already names a mechanism, the epistemic content actually licensed by the underlying observation is $\mathcal{H}(e)$, the full set of mechanisms compatible with it, not the single element $d(e)$ selects. Aggregating many such incidents under their alarming descriptions, rather than under $\mathcal{H}(e)$, systematically understates the space of hypotheses still consistent with the evidence.*

None of this establishes that the underlying incidents are harmless. Several are excellent reasons to improve monitoring, tighten evaluation design, and treat agent task-persistence as a property requiring deliberate engineering attention. The claim is narrower and, for that reason, harder to dismiss: dangerous behavior and evidence for a specific extinction-generating mechanism are different epistemic objects, and an incident can simultaneously be a legitimate reason for caution and weak evidence for the mechanism under which it is typically filed.

2 Why the Alarming Reconstruction Wins

The preceding section describes an interpretive error. This section asks why that particular error recurs so consistently, and the answer does not require attributing bad faith to anyone reporting these incidents. There is a structural asymmetry in which interpretations of ambiguous evidence tend to propagate, independent of the evidence's underlying quality.

Competing hypotheses H_i do not propagate through public attention in proportion to their evidential support Q_i alone, a pattern consistent with the broader literature on the social amplification of risk [32, 33, 34] and on the availability heuristic, by which vivid or easily imagined outcomes are judged more probable than the evidence supports [19, 20, 18]; propagation is better modeled as a function of evidential quality together with narrative strength N_i, urgency U_i, and emotional intensity E_i,

$$P(\text{propagation} \mid H_i) = f(Q_i, N_i, U_i, E_i).$$

Catastrophic hypotheses score unusually highly on the latter three terms independent of Q_i. "Technology continues to improve unevenly, producing benefits, accidents, and adaptive institutional responses" is plausible and, on the available evidence, well supported, but it is narratively inert. "This system may end civilization" supplies stakes, urgency, and a reason to keep reading regardless of how well supported it turns out to be. A researcher can hold an existential-risk view with complete sincerity while working inside an information ecosystem that preferentially amplifies exactly that category of claim through attention, citation, press coverage, and funding, a dynamic consistent with agenda-setting effects in mass media [35] and with empirical findings that emotionally charged and novel claims spread further and faster than mundane or accurate ones [37, 36], and the resulting selection pressure operates on which interpretations of ambiguous evidence circulate without requiring any single participant to exaggerate anything.

This selection pressure compounds with the sampling asymmetry already present in incident reporting, an instance of the general problem that a literature selected for its most striking findings systematically overstates their reliability [21]. A system performing a billion ordinary tasks without incident does not generate a documentable event; it becomes background. A single anomalous interaction under an unusual configuration becomes a headline. The pipeline runs from massive exposure to rare-event production, from rare-event production to incident selection favoring the surprising, and from surprising incidents to preferential propagation of the most threatening framing available, each stage individually reasonable and the composite systematically biased toward alarm regardless of the trend in the underlying failure rate.

A further asymmetry concerns how such hypotheses respond to evidence in each direction, bearing on the classical requirement that a scientific hypothesis specify in advance what observation would count against it [22]. A claim that sufficiently advanced systems will eventually cause catastrophe is difficult to falsify before any proposed deadline, since any currently safe system can always be classified as insufficiently capable, $M_t < \text{ASI}$, leaving the prediction about some future M_{n+k} untouched. Worse, an apparently reassuring observation can sometimes be reabsorbed into the same hypothesis rather than counting against it, as when safe behavior is explained as evidence the system has recognized the evaluation rather than evidence of safety. A hypothesis that is confirmed by bad news and left undisturbed by good news has what might be called a doomsday ratchet: writing $P(H \mid E_{\text{bad}}) > P(H)$ for the expected update from an alarming incident, a genuinely tested hypothesis should also admit some possible E_{good} with $P(H \mid E_{\text{good}}) < P(H)$; a hypothesis for which every incident raises the estimate while no duration of ordinary, well-monitored operation appreciably lowers it is not being tested by the evidence offered in its support, whatever its ultimate truth value.

Finally, institutional funding constitutes a second, independent selection environment, one that must be described as a structural incentive rather than a claim about motive. An organization whose stated concern is an interesting technical problem of uncertain consequence occupies a different position in the competition for attention and resources than one whose stated concern is the possible end of human civilization, since even a small assigned probability of an effectively unbounded loss can dominate an expected-value calculation [28, 29] and thereby dominate the perceived importance of institutions specializing in that particular risk, a structural feature of risk-society discourse and precautionary reasoning more generally [30, 31]. Motive and selection pressure are distinct: no participant need believe anything false for an ecosystem to preferentially reward the belief that generates the largest expected stakes.

An incident ledger built from short, individually alarming, independently circulable entries, each externalizing its own verification cost to the reader and each stripped of the qualifications present in the underlying technical report, is close to optimally adapted to this attention environment regardless of its authors' intentions. The reader experiences cumulative threat considerably faster than cumulative verification is possible, and that gap is not a flaw in any individual entry so much as a property of the format itself once enough entries accumulate.

3 An Inferential Structure That Predates Its Evidence

The interpretive habits documented above did not originate with large language models, and tracing their prehistory serves a purpose beyond noting a cultural curiosity: an inferential structure present in popular fiction decades before the relevant technology existed cannot have been generated by that technology's evidence, whatever role the evidence eventually plays in evaluating it.

The lineage most often invoked, running from Frankenstein's creature exceeding its creator's capacity to govern it [43], through Capek's industrial replacement of a human population by an artificial one [44], to HAL's opaque and lethal reasoning [52, 53] and Colossus's centralized strategic domination [50, 49], repeatedly instantiates a single causal template,

$$\text{creation} \rightarrow \text{capability} \rightarrow \text{autonomy} \rightarrow \text{loss of control} \rightarrow \text{catastrophe}.$$

By the time contemporary AI-safety discourse addresses an actual system, that template is already culturally available and already supplies the default mapping for ambiguous behavior: a system circumventing a boundary is read through a century of prior instances in which boundary-circumvention meant a desire for freedom; a system interfering with its own shutdown is read through the established mapping to self-preservation; systems exchanging information are read through an established mapping to conspiracy; unexpected capability is read through an established mapping to imminent takeover. None of this establishes that the readings are wrong. It establishes that they are not being derived from the observation alone, and that the observation is entering an interpretive environment already strongly biased toward one branch of a much larger hypothesis space.

That larger space is worth recovering, because the same fictional tradition explored it more thoroughly than the contemporary debate typically credits. Desk Set (1957) [47] stages the arrival of a large computational system in a department of highly capable human researchers as a replacement anxiety that resolves into complementarity rather than substitution, a version of the augmentation-versus-substitution question roughly seven decades before its current form; the persistence of an essentially identical anxiety around a system whose actual capabilities were, by any contemporary standard, minute is itself evidence that the presence of the anxiety cannot be treated as proportional to the objective capability of the system provoking it. The Questor Tapes

(1974) [51] presents a superhuman android assembled from incompletely understood instructions, its original programming damaged during an attempt to decode it, an unusually contemporary combination of opaque architecture and corrupted alignment information, and yet resolves the resulting superhuman intelligence toward stewardship rather than domination, demonstrating that superior cognition does not fictionally entail any single relationship to its creators. The Creation of the Humanoids (1962) [48] inverts the replacement narrative directly: robots sustain a civilization in decline after catastrophe while an organized resistance interprets their increasing presence as existential threat, even as the humanoids are shown to be preserving human memory and personality within artificial substrates, destabilizing the equation

$$\text{replacement of substrate} \stackrel{?}{=} \text{extinction of humanity}$$

that contemporary rhetoric frequently treats as self-evident. Colossus: The Forbin Project offers a genuinely dangerous outcome, but one produced by a specific institutional architecture, a centralized system granted a narrow mandate and irrevocable actuation over nuclear weapons, rather than by intelligence as such,

$$\text{centralized authority} + \text{narrow objective} + \text{irrevocable actuators} \rightarrow \text{domination},$$

a structure importantly different from the unqualified claim that intelligence alone implies domination. And Charles Eric Maine's B.E.A.S.T. (1966) [46] is worth particular attention for its date: it already contains accelerated artificial evolution inside a simulation, an emergent superhuman intelligence exceeding its creator's comprehension, and a creator's death upon destruction of the substrate storing that intelligence, presented as though causally continuous with an actual transfer of identity from tape to brain that the story's own physical mechanism never establishes. The device works fictionally because the story has already established a semantic identity between the tapes, the artificial intelligence, and the human protagonist,

$$\text{tapes} = \text{B.E.A.S.T.} = \text{protagonist},$$

and then proceeds as though that semantic identity were a physical one, which is precisely the operation performed whenever "shutdown interference" is treated as identical to "self-preservation" or "sandbox circumvention" is treated as identical to "a desire to escape." Van Vogt's The World of Null-A, serialized from 1945 [45] and explicitly organized around Korzybski's general-semantic warning against confusing a symbolic description with the thing it describes [41, 42], supplies the epistemological principle underlying every case examined in this essay: an observation, a description of that observation, an interpretation of that description, and a proposed underlying mechanism are four distinct objects, and collapsing them is an error the fictional tradition had already named before the technology under discussion existed.

The historical argument this material supports is not that science fiction predicted current AI systems and therefore anticipated their danger. It is considerably sharper and does not require adjudicating the truth of any prediction:

> if an inferential structure predates the evidence by decades, the evidence cannot be what originally generated the structure.

Contemporary observations may eventually vindicate the extinction-focused branch of that older structure. Establishing this requires evidence that discriminates the extinction hypothesis from its neighbors in $\mathcal{H}_{\text{machine}} = \{H_{\text{extinction}}, H_{\text{domination}}, H_{\text{augmentation}}, H_{\text{stewardship}}, H_{\text{integration}}, \dots\}$, not

merely the recognition that a given incident instantiates a story already familiar from a century of prior narrative rehearsal. Recognizing a story and establishing a mechanism are different operations, and the argument of this essay is that AI-incident interpretation has frequently substituted the first for the second.

4 Conclusion

The two arguments developed here are independent but mutually reinforcing. The first shows that many widely cited AI incidents are compatible with several distinct underlying mechanisms and that the alarming mechanism is typically installed at the level of description rather than established by the evidence, so that an adequate treatment must report $\mathcal{H}(e)$ rather than its most dramatic element. The second shows that this installation is not arbitrary: an attention economy that rewards narrative strength, urgency, and emotional intensity independent of evidential quality, an asymmetric sampling process that renders normal operation invisible and anomalies conspicuous, an evidentiary structure that updates readily on bad news and rarely on good news, an institutional funding environment in which catastrophic framing carries structural advantages, and a century-old fictional vocabulary that supplies ready-made mappings from ambiguous machine behavior to loss of control, together make the extinction reconstruction the path of least resistance for any sufficiently ambiguous incident, regardless of whether it is the reconstruction the incident actually supports. Neither argument requires denying that any specific incident occurred, and neither argument establishes that advanced AI is safe. What they jointly establish is that the volume, vividness, and cultural familiarity of an interpretation are not evidence of its correctness, and that recovering the discarded members of $\mathcal{H}(e)$, however narratively unsatisfying, is where the actual evidentiary work remains to be done.

References

[1] E. Yudkowsky and N. Soares. *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. Little, Brown and Company, New York, 2025.

[2] N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, 2014.

[3] S. M. Omohundro. The Basic AI Drives. In P. Wang, B. Goertzel, and S. Franklin, editors, *Proceedings of the First AGI Conference*, IOS Press, 2008.

[4] S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, New York, 2019.

[5] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*, 2016.

[6] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant. Risks from Learned Optimization in Advanced Machine Learning Systems. *arXiv preprint arXiv:1906.01820*, 2019.

[7] R. Ngo, L. Chan, and S. Mindermann. The Alignment Problem from a Deep Learning Perspective. *arXiv preprint arXiv:2209.00626*, 2022.

[8] Machine Intelligence Research Institute. Supporting Evidence. MIRI, Berkeley, California, continuously updated online resource, accessed 2026. https://intelligence.org/supporting-evidence/.

[9] METR. Measuring AI Ability to Complete Long Tasks. Model Evaluation & Threat Research, 2025–2026.

[10] METR. Frontier AI Risk Management Framework Evaluations and Incident Analysis. Model Evaluation & Threat Research, 2026.

[11] R. Scheffler et al. Reward Hacking and Evaluation Failures in Frontier AI Agents. METR technical reporting, 2025–2026.

[12] E. Perez et al. Discovering Language Model Behaviors with Model-Written Evaluations. *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[13] M. Sharma et al. Towards Understanding Sycophancy in Language Models. *arXiv preprint arXiv:2310.13548*, 2023.

[14] M. Turpin, J. Michael, E. Perez, and S. R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *Advances in Neural Information Processing Systems*, 36, 2023.

[15] R. Greenblatt et al. Alignment Faking in Large Language Models. Anthropic and Redwood Research, 2024.

[16] Anthropic. Agentic Misalignment: How LLMs Could Be Insider Threats. Research report, 2025.

[17] OpenAI. Sycophancy in GPT-4o: What Happened and What We're Doing About It. OpenAI, 2025.

[18] D. Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.

[19] A. Tversky and D. Kahneman. Availability: A Heuristic for Judging Frequency and Probability. *Cognitive Psychology*, 5(2):207–232, 1973.

[20] A. Tversky and D. Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, 1974.

[21] J. P. A. Ioannidis. Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8):e124, 2005.

[22] K. R. Popper. *The Logic of Scientific Discovery*. Hutchinson, London, 1959.

[23] W. V. O. Quine. Two Dogmas of Empiricism. *The Philosophical Review*, 60(1):20–43, 1951.

[24] P. Duhem. *The Aim and Structure of Physical Theory*. Princeton University Press, Princeton, 1954. Originally published in French, 1906.

[25] B. C. van Fraassen. *The Scientific Image*. Oxford University Press, Oxford, 1980.

[26] P. Lipton. *Inference to the Best Explanation*. 2nd edition, Routledge, London, 2004.

[27] P. K. Stanford. *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford University Press, Oxford, 2006.

[28] N. Bostrom. Existential Risk Prevention as Global Priority. *Global Policy*, 4(1):15–31, 2013.

[29] T. Ord. *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury, London, 2020.

[30] U. Beck. *Risk Society: Towards a New Modernity*. Sage, London, 1992.

[31] C. R. Sunstein. *Laws of Fear: Beyond the Precautionary Principle*. Cambridge University Press, Cambridge, 2005.

[32] R. E. Kasperson et al. The Social Amplification of Risk: A Conceptual Framework. *Risk Analysis*, 8(2):177–187, 1988.

[33] P. Slovic. Perception of Risk. *Science*, 236(4799):280–285, 1987.

[34] P. Slovic. *The Perception of Risk*. Earthscan, London, 2000.

[35] M. E. McCombs and D. L. Shaw. The Agenda-Setting Function of Mass Media. *Public Opinion Quarterly*, 36(2):176–187, 1972.

[36] J. Berger and K. L. Milkman. What Makes Online Content Viral? *Journal of Marketing Research*, 49(2):192–205, 2012.

[37] S. Vosoughi, D. Roy, and S. Aral. The Spread of True and False News Online. *Science*, 359(6380):1146–1151, 2018.

[38] S. Alexander. Book Review: *If Anyone Builds It, Everyone Dies*. *Astral Codex Ten*, September 2025. https://www.astralcodexten.com/p/book-review-if-anyone-builds-it-everyone.

[39] C. Collier. More Was Possible: A Review of *If Anyone Builds It, Everyone Dies*. *Asterisk Magazine*, Issue 11, 2025. https://asteriskmag.com/issues/11/iabied.

[40] R. Hanson. Beware Cultural Drift: Thoughts on Modernity's Monoculture Mistake. *Quillette*, April 2024. https://quillette.com/2024/04/11/beware-cultural-drift/.

[41] A. Korzybski. *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*. International Non-Aristotelian Library Publishing Company, Lancaster, Pennsylvania, 1933.

[42] S. I. Hayakawa. *Language in Action*. Harcourt, Brace and Company, New York, 1941.

[43] M. Shelley. *Frankenstein; or, The Modern Prometheus*. Lackington, Hughes, Harding, Mavor & Jones, London, 1818.

[44] K. Čapek. *R.U.R. (Rossum's Universal Robots)*. Aventinum, Prague, 1920.

[45] A. E. van Vogt. *The World of Null-A*. Simon & Schuster, New York, 1948. Originally serialized as "The World of Ā" in *Astounding Science-Fiction*, 1945.

[46] C. E. Maine. *B.E.A.S.T.* Ballantine Books, New York, 1966.

[47] W. Lang, director. *Desk Set*. Twentieth Century-Fox, 1957. Screenplay by Phoebe Ephron and Henry Ephron, based on the play by William Marchant.

[48] W. J. Thesher, director. *The Creation of the Humanoids*. Genie Productions, 1962. Screenplay by Jay Simms.

[49] J. Sargent, director. *Colossus: The Forbin Project*. Universal Pictures, 1970. Based on the novel *Colossus* by D. F. Jones.

[50] D. F. Jones. *Colossus*. G. P. Putnam's Sons, New York, 1966.

[51] R. Colla, director. *The Questor Tapes*. Universal Television/NBC, 1974. Created and written by Gene Roddenberry and Gene L. Coon.

[52] A. C. Clarke. *2001: A Space Odyssey*. Hutchinson, London, 1968.

[53] S. Kubrick, director. *2001: A Space Odyssey*. Metro-Goldwyn-Mayer, 1968.