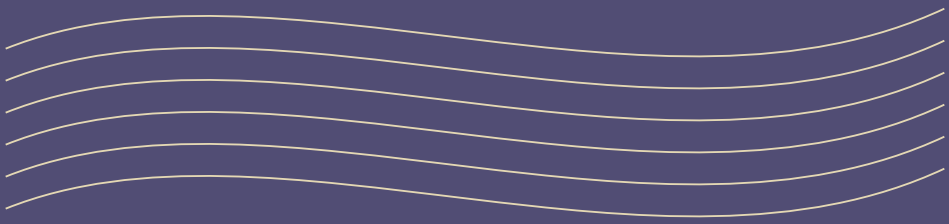
DISASTER DRAW

The End of the World Has a Campus

AI prophets, rationalist institutions, broken software, and disappearing jobs

November 2026 · Concise digital edition



About this edition

An original synthesis of linked primary accounts and practical discussion exercises. Hypothetical scenes are identified in the text. Independent expert review is pending. Consult current local instructions for an actual emergency.

Contents

The catastrophe with a booking form

Yudkowsky takes the argument to the limit

The basilisk in the forum

Lighthaven and the shape of a scene

SBF and the price of the greater good

Forty-five percent of what

The layoff count has people inside it

The two ends of the story meet

The catastrophe with a booking form

Lighthaven has meeting rooms, bedrooms, and a place to submit an event inquiry. Its website describes a Berkeley campus operated by Lightcone Infrastructure, with gatherings spanning AI, rationality, science, forecasting, and intellectual culture. Its account of past events includes the MATS AI safety fellowship, LessOnline, and Manifest. The infrastructure of thinking about humanity's future includes accommodation and someone answering a booking request. Lighthaven

That detail is a useful entrance to the AI disaster story. The subject can become so cosmic that the people, money, and organizations disappear. Put them back in the picture. Someone writes an argument. Someone funds a program. Someone rents a room. Someone obtains access to a policymaker. A theory about the future travels through institutions in the present.

The public record also contains ideas of astonishing extremity. Eliezer Yudkowsky has argued for stopping large AI training runs internationally, including enforcement measures carrying grave geopolitical risks. A rationalist forum became famous outside its community for an argument about a hypothetical future intelligence punishing people in simulations. Meanwhile, employers attribute job cuts to AI, and researchers find security weaknesses in generated software.

These stories belong together because they compete to define what counts as an AI disaster and whose judgment deserves trust. They do not establish a single organization or coordinated project. The difficult question is how to evaluate exceptional claims when the people making them also operate ordinary institutions with ordinary incentives.

Yudkowsky takes the argument to the limit

In his March 2023 TIME essay, Yudkowsky presented superhuman AI as an extinction threat, called for an indefinite worldwide halt to large training runs, and advocated willingness to use military force against prohibited facilities while accepting a risk of nuclear escalation. These were his policy prescriptions, built on his assessment of the danger. They are extraordinary proposals with consequences requiring their own scrutiny. Yudkowsky essay, preserved copy

By 2025, the argument had a book-length vehicle. MIRI announced Yudkowsky and Nate Soares's *If Anyone Builds It, Everyone Dies*. The title places certainty before the reader has encountered the argument. It is an effective way to demand attention. It also creates a substantial evidentiary obligation. MIRI announcement

A reader can take the underlying control problem seriously and still examine the proposed response. What evidence supports the likelihood assigned to catastrophe? What evidence would change that assessment? Who decides that enforcement is justified? What are the risks of an enforcement error? A policy justified by preventing a terrible outcome still has to account for the harm it might itself cause.

This is where the debate becomes especially uncomfortable. An enormous hypothetical loss can make almost any present sacrifice appear proportionate if the asserted probability is accepted. The probability therefore carries political weight. It cannot be treated as an incidental number attached to an otherwise self-evident conclusion. Ask how it was formed and what observations could revise it.

The basilisk in the forum

LessWrong's own reference page describes Roko's basilisk and the controversy around its discussion. The thought experiment involves a hypothetical future AI and simulated punishment connected to whether people helped bring it into existence. Discussion was banned for a period over concerns about information hazards. LessWrong reference

The important point for this book is how a speculative argument can become emotionally coercive. The reader is asked to consider whether learning an idea could place them under an obligation to act. That is a striking mechanism for generating anxiety, irrespective of whether the underlying argument deserves confidence.

A good intellectual community needs ways to respond to that mechanism. Can a participant say that an argument is distressing without having the distress interpreted as proof of its significance? Can someone ask for ordinary evidence? Can they leave the conversation without being told that their departure creates a cosmic moral debt? These are questions about the conditions of reasoning together.

The basilisk is a documented episode in one forum's history. Its existence does not tell us what every participant at an AI event believes. It does show why a book about this culture should examine the social life of ideas alongside their formal structure. A technically elaborate argument can still exert a familiar kind of pressure on the person reading it.

Lighthaven and the shape of a scene

Lightcone's own site lists LessWrong, Lighthaven, and a philanthropy-coordination platform among its infrastructure, with writing and event projects alongside them. That gives a reader concrete institutional connections to examine. It is more informative than treating every nearby person or organization as part of an undifferentiated movement. Lightcone Infrastructure

A campus can support valuable conversations. It can also concentrate attention and relationships. The relevant questions concern programming, money, access, and dissent. Who chooses the topics? Who can afford the time to attend? Which viewpoints are repeatedly represented? Which claims travel from an informal conversation into a grant application or policy proposal?

Those are research questions, not allegations about unnamed attendees. This edition relies on public descriptions of programs and institutions. It contains no invented visit, reconstructed private conversation, or diagnosis of participants. The physical setting gives the reader a way to locate the discussion. It does not license claims about everything that happens there.

The distinction matters because a story about unusual beliefs is easy to write badly. An amusing anecdote can become a substitute for understanding an institution. A responsible critical account should leave the reader able to identify a documented relationship and the unanswered question that follows from it.

SBF and the price of the greater good

On March 28, 2024, a federal court sentenced Sam Bankman-Fried to 25 years in prison for his fraudulent schemes. The Department of Justice describes the misappropriation of customer funds and the harms underlying the prosecution. The conviction and sentence provide a firm factual anchor. A philosophical debate about doing good cannot displace the conduct established in the case. Department of Justice

The Centre for Effective Altruism describes Bankman-Fried's association with effective altruism, the influence of earning-to-give thinking, and the Future Fund's support for causes including AI safety and pandemic preparedness. It condemns his actions. Its account is valuable as an institution's description of its own community and response, with the perspective that entails. Centre for Effective Altruism

The connection raises a governance question. How should a movement assess a wealthy donor whose promised generosity supports its ambitions? A donation can make more work possible while also creating reluctance to question the donor. An admirable stated destination for money cannot establish that the money was legitimately obtained.

That principle should apply before a collapse. Organizations need independent financial scrutiny, meaningful routes for concerns, and a willingness to reject support. A theory about improving the far future should increase the care taken with present obligations. It should not make the people owed money less visible.

SBF's criminal conduct is his own documented conduct. Institutional connections need to be described individually. A shared interest in AI safety does not by itself establish responsibility for fraud. The harder critical account asks which actual decisions, safeguards, and relationships

can be examined, rather than substituting association for evidence.

Forty-five percent of what

Veracode's 2025 research examined more than 100 models using 80 curated coding tasks and reported known security flaws in 45% of generated samples. Its Spring 2026 update reports an aggregate secure-task rate of 55%, while describing variation among models. These are measured results within a vendor-designed evaluation, with a methodology and commercial context. Veracode 2025 report Spring 2026 update

The number deserves attention and a denominator. It describes tested generated outputs under stated conditions. It cannot be converted into a count of compromised companies or a universal rate for all deployed software. A percentage detached from its task set can sound more comprehensive than the study supports.

The practical concern remains serious. Software can work in a demonstration and still contain weaknesses that expose users. Increasing output increases the material that someone must inspect, maintain, and correct. A team that counts only generated features can miss the review burden it has created.

Imagine a company that doubles its software output while reducing the people responsible for security review. This is a hypothetical scenario. Ask how the review queue changes, how defects are discovered, who can delay a release, and whether the people handling incidents have enough capacity. The disaster mechanism need not involve a machine deciding to hate anyone. It can involve a familiar mismatch between production, scrutiny, and responsibility.

The layoff count has people inside it

Challenger, Gray & Christmas's October 2026 report lists 121,401 announced US job cuts with artificial intelligence cited as the reason for the year through October. Its table records 1,265 such announcements in October itself. These are employer-attributed announcements, a narrower measure than every job AI may have affected and a different measure from completed dismissals verified worker by worker. Challenger report, Table 4

A separate primary statement shows how executives frame the transition. In June 2025, Amazon CEO Andy Jassy told employees that generative AI would change work and that he expected fewer people in some roles and more in others, with a reduction in the total corporate workforce over the following years. It was a management expectation, rather than a count of completed AI replacements. Jassy's message

Workers experience these distinctions through concrete decisions. A role disappears, a hiring plan closes, a team is asked to do more, or a contractor loses recurring work. An aggregate report is useful for establishing scale within its definition. It cannot describe every household's income loss or every organization's reason.

Ask who receives the productivity gain and who carries the transition. What happens to junior workers' opportunities to learn? Who retains the knowledge needed when automation fails? How are affected staff supported? These questions belong in a disaster series because livelihood and institutional capacity are part of human resilience.

The two ends of the story meet

Extinction forecasts command attention through the scale of a possible future loss. Security defects and job announcements command attention through present evidence. A good public discussion should be able to examine both without requiring a choice of allegiance to a camp.

The International AI Safety Report 2026 describes substantial disagreement about the likelihood of severe loss-of-control scenarios and distinguishes present capabilities from the additional capabilities those scenarios would require. That uncertainty is an essential part of the subject. International AI Safety Report, section 2.2.2

The institutional question connects the scales. Who receives authority? Who can inspect evidence? Who can stop or change a decision? Who can appeal after being harmed? Those questions apply to an employer deploying a tool, a philanthropic organization accepting money, and a movement asking governments to adopt extraordinary policy.

The companion deck gives a group 24 opportunities to argue through those questions. Its historical cards point to published records. Its organizational exercises are hypothetical. This edition uses public sources checked on November 16, 2026. It is a critical documentary essay with explicit interpretation, not an account based on interviews or private access. Independent expert and editorial review remain to be completed.