

Preemptive Futuring: Intervening in Causal Pathways to Undesirable Futures — Lessons from AI Risk Research

ANONYMOUS AUTHOR(S)

Speculative approaches in HCI offer methods for imagining, materializing, and contesting possible futures. A different challenge arises when speculation aims to prevent undesirable ones. Drawing on public documents from AI risk research communities, we articulate a recurring practice of working backward from hypothetical failures: constructing causal pathways, surfacing assumptions, locating leverage points, and connecting present actions to future outcomes through theories of change. We distill the practice into a six-stage method specification for HCI, as **Preemptive Futuring**, an intervention-oriented mode of futuring. It treats pathways between present conditions and possible futures, rather than scenarios themselves, as primary objects of design, contestation, and revision. We close by reflecting on what futuring in HCI is for, once speculation is asked to supply the solution it ordinarily withholds.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; *Interaction design theory, concepts and paradigms*.

Additional Key Words and Phrases: speculative design, design futuring, design fiction, theory of change, AI safety, existential risk, anticipatory governance, research through design

ACM Reference Format:

Anonymous Author(s). 2026. Preemptive Futuring: Intervening in Causal Pathways to Undesirable Futures — Lessons from AI Risk Research. In *CHI Conference on Human Factors in Computing Systems (CHI '27)*. ACM, New York, NY, USA, 35 pages.

1 Introduction

AI risk researchers study futures that they hope will never become observable. A researcher who evaluates whether language models can deceive their overseers does not predict that deceptive AI will cause catastrophe; the evaluation is conducted so that this particular future does not arise [68, 88]. If the work succeeds, its object of study disappears. This epistemic situation is unusual, and the communities that occupy it have developed distinctive practices for rendering unobservable futures researchable. They decompose catastrophic scenarios into causal premises [23], compare and contest “threat models” [A89], and justify individual projects through explicit theories of change that connect present research to the futures it is intended to avert [97].

HCI has established traditions of its own for engaging with futures. Speculative design, design fiction, and design futuring make possible worlds available for reflection, provocation, and debate [16, 33, 61]. These traditions broaden which futures can be imagined, who imagines them, and what questions are asked of them [20, 47]. When the purpose of speculation is to *prevent* a future rather than to reflect on it, however, a methodological gap appears. A compelling scenario describes what an undesirable future might look like. It does not specify how that future would come about, which present conditions enable it, or where intervention is possible. This gap has been identified from several directions. Tonkinwise [95] observes that speculative design gestures toward politics while neglecting the transition pathways that would connect present to future, and analyses of AI researchers’ broader-impact statements find harm anticipation that is vague, unevenly reasoned, and largely unactionable [70]. HCI is well equipped to make futures imaginable but less explicit about how to move from an undesirable future to a present intervention without collapsing speculation into prediction.

We ask:

2026. Manuscript submitted to ACM

RQ: How do AI risk researchers transform speculative undesirable futures into present research and intervention priorities, and what futuring practices can HCI learn from this process?

We address this question by turning to AI risk research, which requires such reasoning of its fellows, assesses it in mentoring, weighs it in funding decisions, and contests it in public forums. Its reasoning is scattered across a large body of public documents (Section 3). Read together, those documents reveal a recurring pattern: from a future scenario, to the causal pathways that could produce it, to the leverage points along those pathways, to a present intervention. Researchers begin from an undesirable hypothetical future and work backward to construct the causal pathways through which it could arise. They annotate those pathways with assumptions, locate the points at which a pathway might be disrupted, and design interventions whose logic is recorded in an explicit and revisable theory of change.

We distilled this recurring practice into **Preemptive Futuring**, an intervention-oriented mode of futuring that works backward from undesirable futures and treats the pathways connecting present conditions to those futures as objects of design. Scenario-centered speculation asks what a future might be like and what it reveals about the present. Preemptive Futuring additionally asks through which pathways the future could arrive, which assumptions each pathway rests on, and which present interventions would disrupt it. The scenario remains necessary, but the pathway, rather than the scenario, becomes the primary speculative artifact.

We then return Preemptive Futuring to design futuring and set it beside the commitments of that tradition. Speculation there is a discipline of imagination: it multiplies possibilities, declines to resolve them, and values an artifact for the questions it keeps open. Preemption asks for more. It keeps that imaginative work and adds an obligation to it: the account is not finished until it has committed to an action in the present. Placing the two side by side raises two questions for the community. First, what is futuring for: imagination and public discussion, or intervention against a future we are trying to prevent? Second, what is the primary object of that anticipation: the scenario, or the causal pathway that would bring it about? This paper argues for the pathway and leaves the first question open. Sections 7.1 and 7.2 take both up as a mutual critique: what speculative design has to say against a futuring that commits to an intervention, and what the commitment to intervene exposes in a tradition disciplined to withhold it.

The paper makes two contributions:

1. **Preemptive Futuring**, a distillation of the futuring operations that AI risk research requires, teaches, assesses, and funds, grounded in its public documentary record, and their translation into a six-stage method for HCI: articulate an undesirable future; backchain possible causal pathways; surface assumptions; branch and contest; locate leverage points; design interventions and re-future. The operations comprise backchaining from undesirable futures, grading assumptions, contesting pathways, and stating theories of change, and the method is specified with scope conditions and executable detail sufficient for a design team to run and to contest.
2. **A critical reflection on the purpose of futuring in HCI**. What counts as good speculation cannot be settled apart from what the speculation is for. We therefore distinguish the scenario, the pathway, and the intervention as three objects a futuring practice can be answerable to (Section 6), and read Preemptive Futuring back against design futuring's own commitments, asking what each tradition exposes in the other (Sections 7.1 and 7.2).

2 Background and Related Work

2.1 Speculative Futures in HCI and STS

HCI has spent two decades building a vocabulary for engaging futures through design. Speculative and critical design position designed artifacts as questions rather than products, using “what if” propositions to open debate about the futures technology might bring [11, 33]. Design fiction creates diegetic prototypes and story worlds in which possible technologies can be encountered and argued over [14, 16, 28, 91]. Design futuring, the umbrella under which CHI has consolidated these practices, spans speculative enactments in which participants make consequential choices inside a fictional frame [36], material speculations embodied in functioning artifacts [98], experiential futures staged for publics [22], and value-oriented world-building activities such as Timelines, in which participants construct sequences of events connecting the present to a speculated future [102]. Recent reviews document both the breadth of this work and its recurring limitations, including narrow and repetitive imaginaries [90] and a persistent techno-deterministic orientation [85].

These traditions broaden what futures can be imagined and questioned, and it would be a strawman to call them non-interventionist. Speculative enactments generate empirical data [36]; design friction grounds speculation in real civic contexts [39]; contestable-AI work uses speculative design to build dispute mechanisms into deployed systems [4]; anticipatory ethics in adjacent fields argues for deliberately pessimistic foresight as preparation for regulation [37]. Value sensitive design has long used value scenarios and envisioning cards to surface systemic, long-term harms during design [41, 71]. What these approaches share, however, is narrower and more precise than non-intervention: the *route* from present to future is rarely itself an object of critique or revision. Even where the work intervenes (an enactment, a deployed dispute mechanism), the connective tissue between present and future stays implicit: the scenario or the artifact shows a destination, and the route is left to the reader. Tonkinwise makes the point sharply in his review of *Speculative Everything*: speculative design tends to bracket precisely the transitions that would make its futures arrivable [95].

A second strand of work, largely outside HCI, has developed methods where routes matter. Backcasting starts from a desirable future and works backward to the steps required to reach it [32, 79, 81, 82]; even its most participatory forms keep the endpoint desirable [17, 80]. The undesirable-endpoint variant is less developed but not unoccupied. Threatcasting backcasts from a specific future threat to actions that disrupt it [54], Klein’s pre-mortem [56] asks a team to assume failure and explain why, and infrastructural speculations in HCI ask what conditions a speculated artifact would require [101]. Security engineering, system safety, and the decision sciences hold older and sharper relatives, from attack trees [86] and STPA [63] to Assumption-Based Planning [30], and AI safety has lately formalized its own pathway reasoning as safety cases [25]. Table 3 positions each of these against our method. Backward-from-harm reasoning is thus a mature engineering paradigm. What limits its transfer to design is the setting it assumes: a safety case argues to a regulator, and threatcasting serves an organization’s own foresight, whereas a sociotechnical future has no authorized decider. Legitimacy, plurality, and contestation therefore have to be handled inside a design-facing method.

STS supplies the critical frame for both strands. Sociotechnical imaginaries show that collectively held visions of the future are performative political infrastructure [53]; the sociology of expectations shows that articulated futures mobilize resources and coordinate actors in the present [18]; anticipatory governance argues for building capacity (foresight, engagement, integration) rather than predictions [45], and responsible innovation makes anticipation an institutional obligation alongside reflexivity, inclusion, and responsiveness [92]. Two critiques from this literature

bear on any futuring that aims at intervention. Nordmann’s critique of speculative ethics [73] warns that “if and then” reasoning smuggles conditional futures into present moral claims. And critical scholarship on preemption, Anderson’s taxonomy of anticipatory action and Massumi’s analysis of preemptive politics, warns that acting on threats before they emerge can be self-justifying, because the counterfactual is never observable [7, 65].

Finally, existing discussions of speculation quality ask whether speculative work is plausible, provocative, situated, participatory, or rigorous [10–12], and equity-oriented work asks whose imaginaries get to count [20, 47, 77]. These criteria address the scenario; none addresses the pathway from present to future, or the intervention that would act on it. HCI currently lacks a futuring method that treats the causal pathways to undesirable futures as objects of design and revision. That is the gap this paper addresses.

2.2 Theory of Change and Intervention Logics

Theory of Change (ToC) originated in evaluation research as an answer to a practical problem: complex social initiatives were failing evaluation not because they did not work but because no one had articulated the causal chain through which they were supposed to work [100]. The method that grew from Weiss’s diagnosis is explicitly backward-mapping: begin from long-term outcomes, work back through intermediate outcomes to early preconditions, and make the assumptions at each link explicit [27]. ToC has since become near-universal infrastructure in development funding and philanthropy [97], and the surrounding evaluation literature supplies conceptual tools a pathway-oriented futuring method can draw on: realist evaluation’s insistence that mechanisms fire only in contexts [9, 76], contribution analysis’s standard of “reasonable confidence” for causal claims where experiments are impossible [66], and program theory for interventions with multiple alternative causal strands to the same outcome [83]. Transition design is the one design tradition that has claimed ToC as a design construct [52]; like backcasting, it orients toward desired futures, and its critics note that it under-specifies what is being transitioned away from [96].

Preemptive Futuring partly reverses the ToC template (Figure 1). Where a standard ToC runs forward from actions through mechanisms to desired outcomes, the practice we describe runs backward from an undesired outcome through possible mechanisms to a present intervention, and only then constructs a forward ToC for the intervention itself.

A pathway-oriented method also needs a vocabulary for *where* in a causal structure intervention is possible. Systems thinking supplies one in Meadows’s theory of leverage points: a ranked list of places to intervene in a system, from shallow parameters and buffers, through feedback loops and information flows, to the rules, goals, and paradigms from which the rest of the structure follows [67]. Meadows’s ordering carries a warning that sustainability research has since documented empirically: effort concentrates on shallow, tractable, measurable points while the deep points where structure is set go unaddressed [1], and Fischer and Riechers [38] have since turned the ordering into a research agenda for sustainability science. Systems thinking has a parallel lineage in HCI, from sustainable interaction design to calls for design at the level of complex sociotechnical systems [15, 21, 31, 34, 40, 57, 74]. Leverage-point reasoning is what turns a mapped causal pathway into a decision about action.

2.3 AI Risk Research and Its Culture

The community we learn from coheres around the claim that advanced AI could cause catastrophic or even existential harm, and that present research can reduce that risk [19, 75, 84]. Whatever one makes of the claim, and it is contested both externally and internally [29, 42, 46], the field’s epistemic situation is structurally interesting: its central phenomena do not yet exist, its highest-stakes hypotheses cannot be tested directly, and its practitioners must nonetheless prioritize among present actions. The field’s responses to this situation are the raw material of this paper. Researchers decompose

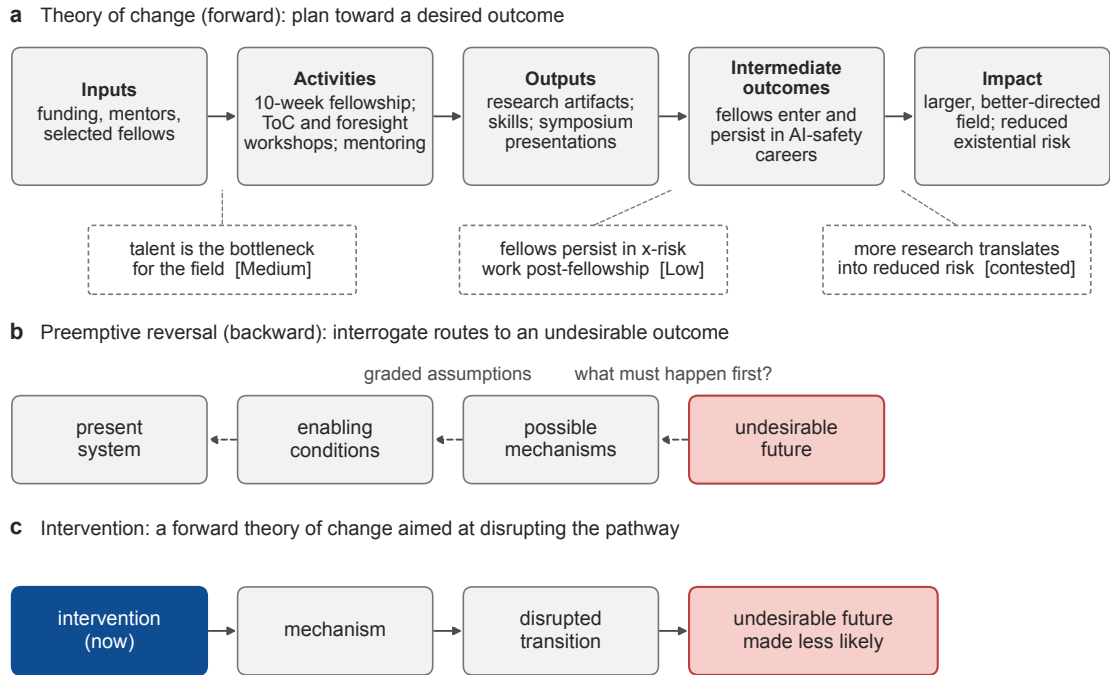


Fig. 1. Theory of change and its preemptive reversal. (a) The standard forward template plans from inputs toward a desired impact [27, 100], shown here as a fellowship theory of change modeled on those published by our field sites [A104, A152]. Dashed boxes name the assumptions on which specific links depend and grade each with a confidence level; ERA assigns “Low” to the persistence assumption on which its long-term impact most depends. (b) Preemptive Futuring first reasons backward from a specified undesirable future, asking at each step what would have to happen first and grading the assumptions on every link. (c) The intervention then carries its own forward theory of change, aimed at disrupting a transition in the mapped pathway.

catastrophe scenarios into conjunctive premises and attach explicit credences [23]; they write and compare “threat models” that specify the causal route from present techniques to future failure [72][A89]; they construct proof-of-concept versions of feared capabilities so present defenses can be tested against them [50]; they translate long-term worries into concrete near-term research problems [5]; and they publish month-by-month narrative scenarios explicitly as policy interventions [59]. Institutionally, frontier labs bind present commitments to capability thresholds that do not yet exist [8], and research agendas for whole subfields are organized around classes of failure the field wants never to observe. Cooperative AI is the clearest instance. Its founding agenda constitutes a subfield around the question of how machine and human agents might jointly improve welfare, naming understanding, communication, commitment, and institutions as the capacities at stake, and presents itself as a bet on the productivity of particular kinds of conversation [A44]. The taxonomy of multi-agent risks published five years later by the foundation that agenda produced sorts the anticipated harms into miscoordination, conflict, and collusion, and names seven risk factors that generate them [A67]; these are failures that appear only once systems interact, and that no evaluation of a single model in isolation would surface.

Institutionally, the community is compact and unusually legible: a “tight-knit” epistemic community whose members move between companies, universities, and fellowships [2], concentrated in a few cities, among them San Francisco and

Berkeley, London, Oxford, and Cambridge [64]. Entry into the field runs through a small number of fellowships and courses, among them the Existential Risk Alliance’s ERA:AI Fellowship,¹ an eight-week residential program that pairs early-career researchers with mentors on catastrophic risks from AI [A152], the ML Alignment & Theory Scholars program (MATS),² the Cooperative AI Foundation’s annual summer school,³ which convenes early-career researchers around the multi-agent agenda [A36, A37], and Google DeepMind’s AGI safety short course [A95]. A handful of research organizations publish the agendas and threat models the pipeline trains toward, a handful of funders support the work, and three public forums (LessWrong,⁴ the Alignment Forum,⁵ and the EA Forum⁶) host much of its reasoning. Most of this research is non-profit and published openly online, and the documents these institutions produce are the corpus of Section 3 (Table 1).

Social studies of this community are comparatively thin. Ahmed et al.’s study of AI safety’s epistemic culture [2] documents its field-building practices (forecasting, career pipelines, prize competitions), and critics have analyzed its ideological commitments [42] and its methodological narrowness [29]. The community’s *future operations*, the concrete reasoning practices by which imagined futures become present research priorities, have not been studied, and are the object of this paper. We approach AI risk research not as a set of beliefs to endorse or debunk but as an extreme empirical site where intervention-oriented speculation is practiced daily under institutional demand, and where its methods have been forced to become explicit.

3 Sources and Approach

3.1 Why Public Documents

The practice we articulate is documented in public at a scale few communities match (Section 2.3). Programs publish their theories of change, funders publish their criteria, and researchers debate threat models on open forums. An application rubric or a funder’s scoring criterion does not describe the practice; it constitutes it. We therefore built the account from those documents, adapting the trace ethnography of Geiger and Ribes [43], which reconstructs coordinated action by following the documentary traces that participants produce and act upon. We follow the traces without the participant observation that method ordinarily pairs with them, which bounds the account to what the documents coordinate in public. We conducted no interviews, which would have replaced the documentary record with a small number of individual accounts.

Building the account this way carries two risks. The first is over-claiming: presenting the practice as more coherent than it is enacted. We address it by grounding every claim in a citable public document. The second is under-critiquing: adopting the community’s evaluative vocabulary as though it were neutral. We address it by treating the politics of the practice as a finding in its own right (Sections 4.4 and 8). The study collects no data from or about human participants. We take no position on whether the futures the community fears are probable; what we extract is a reasoning form.

3.2 Corpus

The community documents its reasoning in public (Section 2.3). The layer it documents best is how research is *justified* to selectors, mentors, and funders. Our claims concern that training, funding, and discourse layer, which overlaps

¹<https://erafellowship.org/>

²<https://www.matsprogram.org/>

³<https://www.cooperativeai.com/>

⁴<https://www.lesswrong.com/>

⁵<https://www.alignmentforum.org/>

⁶<https://forum.effectivealtruism.org/>

Table 1. The archived trace corpus by stratum, every document cited by its label in Appendix A. Each is held in full text with URL and access date; recorded talks were consulted through verified playlist metadata only and are not counted. Two reports were archived in two parts, so 174 files correspond to 172 documents.

Stratum	Documents	N	Dates
Forum posts and retrospectives	Theory-of-change and prioritization posts, threat-model debates, program announcements, retrospectives, and self-critical audits on LessWrong, the Alignment Forum, and the EA Forum [A7, A8, A13, A20, A21, A27, A41, A45, A47, A48, A57–A59, A64, A66, A68, A70–A73, A75–A77, A81–A84, A87–A93, A95–A98, A101–A105, A108, A109, A115, A117–A126, A128, A131–A141, A144–A146, A148–A153, A155, A156, A159, A162–A166, A168, A169, A172]	88	2016–2026
Program, lab, and funder web documents	Application and selection criteria, mentor project criteria, and curricula; founding posts, research agendas, threat models, scenarios, and strategy documents; funder criteria and RFPs [A1–A6, A9–A12, A14–A18, A22–A26, A28–A40, A42, A43, A46, A49–A56, A60–A62, A74, A78–A80, A86, A94, A99, A100, A106, A107, A110–A112, A114, A116, A127, A130, A142, A143, A154, A157, A158, A167, A170, A171]	72	2019–2026
Technical reports and papers	RAND, FRI, GO-Science, GPI, and arXiv [A19, A44, A63, A65, A67, A69, A85, A113, A129, A147, A160, A161]	12	2020–2026
Total archived documents		172	2016–2026

with but does not exhaust AI safety research. They do not concern how frontier labs choose their next experiment, a distinction we return to as a finding in Section 4.6. Table 1 describes the corpus by stratum. Sources entered it by purposive sampling. We began from the field sites’ own institutional documents [A104, A152, A162] and followed their citations and the community’s canonical curricula outward. We then searched the three forums by program name and by the operation vocabulary of Section 3.3 [A64, A73, A89, A94]. A document was retained if it stated, taught, evaluated, or contested a futuring operation. Because a public corpus skews toward a community’s self-legitimizing artifacts [2, 13], we counter-sampled for disconfirming material. This comprised published null results, internal critiques, and two programs verified at source to make no theory-of-change demand at application (Section 4.6). Every corpus document (Appendix A) is archived in full text with URL and access date. Recorded talks were consulted through verified playlist metadata only, and no finding rests on unarchived material. Quotations are verbatim. Many canonical artifacts are living documents: career-advice pages rewritten, scaling policies versioned, fellowship pages deleted after each cycle. We therefore cite stable archival copies where they exist.

3.3 Analytic Strategy

We did not code for what AI risk researchers believe. Their beliefs about timelines and probabilities are documented [44] and contested [42, 46] elsewhere. We coded instead for **futuring operations**, the recurring moves by which imagined futures are made to do work in the present. These are: constructing undesirable futures; backchaining from outcomes to enabling conditions; articulating and grading assumptions; branching scenarios into alternatives; identifying bottlenecks and leverage points; evaluating proposed interventions; simulating and red-teaming pathways; and revising theories of change in response to evidence or criticism. Section 4 organizes the practice around six clusters of these operations, in the order in which the method of Section 5 uses them. Coding was done by the author without a second coder, and

we did not tabulate frequencies. The claims of Section 4 therefore concern documented, institutionalized practices (requirements, rubrics, and published criteria), not the statistical typicality of individual reasoning.

4 How AI Risk Research Turns Futures into Interventions

This section presents six practices in the order in which the method of Section 5 draws on them. Each subsection says what the community does, which step of the method the practice becomes, and what it lacks from the standpoint of HCI. The corpus nowhere presents the six as a single method. They are required, taught, assessed, and funded separately, and their assembly into one is the analytic work of this section.

4.1 Making an Unobservable Future Researchable

The first practice is the most basic. Researchers begin from a future that has not occurred and convert it into research that can be conducted now. The instrument is the *threat model*, a specific causal story that connects present techniques to a particular future failure. The founding post of the CERI fellowship names the skill it trains as the ability to “think deeply about potential threats” and “break them down into tractable research questions and agendas” [A81]. The threat is speculative, the research questions are present-tense, and the move between them is treated as a teachable competence. This practice becomes Step 1 of the method.

What distinguishes the community is how instrumentally it treats these stories. A threat model is valued as “a specific story to guide what you do,” useful even if it turns out to be wrong [A119]. DeepMind’s alignment team wrote its synthesis of threat models “to enable better research project generation and prioritization,” and an empty cell in the resulting taxonomy is read as a research gap [A89]. The question asked of a threat model is not “how likely is this?” but “what work does this generate?” Two disciplines keep the practice from collapsing into prediction. Stories are kept minimal, because “every time you add an extra detail to a story, it becomes less likely” [A119]. And futures are held as a portfolio rather than a single wager: Anthropic’s strategy document sorts the future into optimistic, intermediate, and pessimistic classes and allocates research across all three [A10]. Where a future cannot be studied directly, it is built in miniature. Researchers train deliberately backdoored models that resist safety training, so that defenses against a hypothesized future system can be tested on a present one [50]. The community names its own predicament plainly: “Imagine you are tasked with curing a disease which hasn’t appeared yet” [A150].

From the standpoint of HCI, the practice is disciplined at only one end of the story. Threat models specify the failing capability in detail while leaving “catastrophe” underspecified. Step 1 therefore adds the questions HCI is better positioned to ask: undesirable for whom, at what scale, and threatening which values.

4.2 Working Backward Through Causal Pathways

The second practice is to ask, of a hypothetical failure, what would have to happen first, and to keep asking until the chain reaches the present. The community calls this *backchaining*, and it becomes Step 2 of the method. Its artifacts pervade the corpus. Carlsmith’s risk report decomposes AI catastrophe into six ordered premises, from capability timelines through deployment incentives to unintended power-seeking, each of which must hold for the outcome to occur [23]. DeepMind’s team distills a “consensus threat model” with two halves: the technical causes of misalignment on one side, and the route by which a misaligned system becomes catastrophic on the other (Figure 2b) [A89].

Two features of these artifacts matter for HCI. First, the pathways are graphs, not chains. They branch, share intermediate conditions, and admit several routes to the same endpoint, which is why the community’s diagrams converge on the form that security engineering reached decades earlier with attack trees (Figure 2a) [86]. Second, the

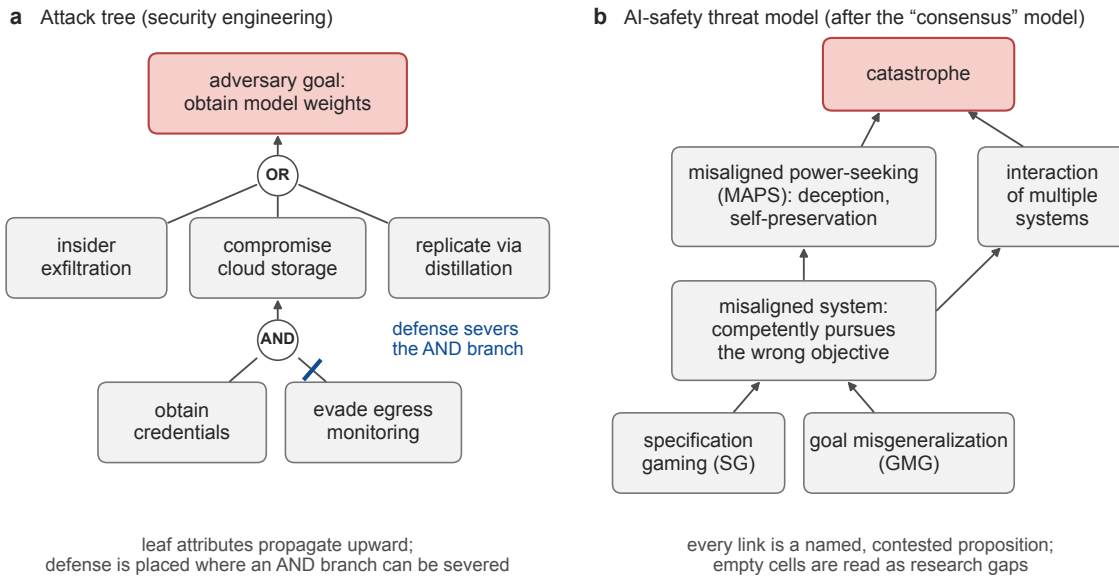


Fig. 2. Threat modeling as backward reasoning from an unwanted outcome. (a) A security-engineering attack tree places the adversary’s goal at the root and enumerates alternative routes (OR) and required conjunctions (AND); leaf attributes propagate upward to identify the cheapest route, and defenses are placed where a conjunction can be severed [86, 89]. (b) The structure of the “consensus threat model” reported by one alignment team’s synthesis of ten published threat models, not a field-wide consensus: technical causes of misalignment feed a misaligned system, which reaches catastrophe through power-seeking or multi-system interaction [A89]. In both genres the analysis begins from the outcome to be prevented and works backward to present conditions, and in both it is the diagram’s links, not its endpoint, that get contested.

pathways are built to be contested. The literature review behind DeepMind’s consensus model lists ten rival causal accounts and attaches explicit strengths and weaknesses to each [A90]. Futures are decomposed into intermediate causal claims precisely so that the claims, rather than the futures, can be debated. These diagrams also circulate. Within about two years the DeepMind taxonomy moved from a research-prioritization device, to the standard frame of overview talks, to a graded exercise in a public course [A94, A95]. In this community a threat model is infrastructure.

What the practice lacks is a safeguard against premature closure: the routes that get drawn are the ones insiders find plausible. Step 2 therefore adopts generative checklists from security threat modeling [89] to force the enumeration of routes a team would not otherwise consider.

4.3 Making Assumptions Explicit

The third practice is to record, for each link in a pathway, the assumptions on which it rests, and to grade the confidence each warrants. This becomes Step 3 of the method. Carlsmith attaches a subjective probability to each of his six premises and multiplies them, and the document carries its own revision history, noting that his overall estimate “has gone up, and is now at >10%” since first circulated [23]. ERA’s published theory of change tabulates eight assumptions behind its fellowship and assigns each a confidence level, including “Low” for the assumption on which its long-term impact most depends, that fellows stay in the field afterward (Figure 1a) [A152].

The effect is that the pathway becomes a container for assumptions. Debates that would otherwise be diffuse, about capability trajectories, actor incentives, or institutional competence, get pinned to specific links in specific diagrams, where they can be graded, prioritized, and revisited. The practice is imperfect, and the community says so. Training programs report that articulating assumptions is what their scholars do worst (Section 4.6). The Forecasting Research Institute’s adversarial collaboration found that skeptic and concerned forecasters still disagreed by two orders of magnitude after eight structured weeks, and traced the gap to assumptions that near-term evidence cannot resolve [A139]. At its furthest extension, the practice prices its futures in millionths of a probability of catastrophe [A84]. We report that pole without endorsing it; the community’s own critics have shown that optimizing on hazy estimates selects for overestimated ones [A86]. The ambition is nonetheless constant: no link without an assumption, no assumption without a grade.

What the practice lacks is a category for normative links, the claims about what actors will value or choose. The grading vocabulary of Step 3 adds it.

4.4 Contesting Pathways Rather Than Predicting Futures

The fourth practice is the institutionalization of disagreement. Rather than treating disagreement as an obstacle, the community builds institutions to sustain it, and directs it at pathways rather than endpoints. This becomes Step 4 of the method. Shared endpoints coexist with contested routes. Researchers who agree that advanced AI could be catastrophic disagree about the causal chain, abrupt takeover versus gradual erosion [55], single-system misalignment versus multi-agent failure [A67], and about what counts as the problem at all [46].

The community’s response has been to fund critique. In what follows, “red-teaming” denotes adversarial critique of plans and arguments, not the adversarial probing of models. A prize contest set aside \$100,000 to “red team” the community’s own analyses, motivated by a feared future of the community itself: that its culture “may give way to a culture of over-deference” [A166]. Research organizations create standing adversarial roles. One is a team whose mission is “to red-team Anthropic’s alignment techniques and evaluations” [A77]. Another is a builder–breaker game in which every proposed training strategy must survive a constructed counterexample [A28]. A third is commissioned external review, published verbatim [A13]. The contest extends to the reasoning method itself. The most upvoted critique of interpretability research in the corpus is a theory-of-change audit that reproduces a colleague’s twenty published theories of impact and rebuts each [A144]. Its highest-scoring reply, by Richard Ngo, attacks the genre rather than the conclusions: alignment researchers are “too focused on backchaining from desired end states, and not focused enough on forward-chaining to find the avenues of investigation which actually allow useful feedback” [A144]. Whether backward reasoning from futures is the right way to choose research is thus a named, scored, public disagreement inside the community that practices it most. Our method carries the objection visibly (Section 7.1).

The sobering observation is that institutionalized contest does not produce convergence. The adversarial collaboration of Section 4.3 barely moved its medians, and a forecasting tournament that paid participants to persuade one another reported the same after “months of debate and the exchange of millions of words” [A85]. What contest produces instead is a catalog of disagreements: named cruxes (the assumptions on which a disagreement turns), graded assumptions, and documented failure modes.

What the practice lacks is breadth in who contests. Every critic cited above shares the community’s training and priors. Step 4 retains the machinery of contestation while extending participation in it.

4.5 Rehearsing the Pathway: Simulation and Role-Play

The fifth practice is performative rather than analytic: the community rehearses its futures through structured role-play. Rehearsal is the second instrument of Step 4, used to surface enabling conditions and actors that a diagram alone omits. Intelligence Rising, a strategic role-play game created at Cambridge’s Centre for the Study of Existential Risk, casts participants as states and AI companies in the run-up to transformative AI. Its designers ran it more than forty times between 2020 and 2024, for companies, ministries, think tanks, and university courses [A19, A65]. RAND runs “Day After” exercises, a two-turn crisis format inherited from its nuclear-proliferation studies, in which officials play a National Security Council responding to an AGI claim [A129, A147].

To HCI’s experiential futures [22, 36], these rehearsals present a grammar that is at once familiar and unfamiliar. Familiar: the appeal to lived experience, staged injects, the debrief as the site of learning. Unfamiliar: the discipline imposed on imagination. Scenarios are set twelve to eighteen months out with everything else held constant, so that officials need not “speculate about transformative changes” [A147]. The founding paper fixes a calibration rule: “one in ten actions should be an action that the participant assigns 10% likelihood of being taken” [A19]. The debrief is an instrument. Participants “record any disagreements with how actions were chosen or resolved,” a phase the designers call “crucial for challenging the facilitator’s own subjective biases” [A19]. The output is engineered to be actionable: RAND harvests each run through standing questions, among them “what playbooks did participants indicate need to be written” [A129]. The rehearsals also transform their designers. After forty-three games the Intelligence Rising team reports that its concerns about a race “have grown significantly” [A65].

What this practice lacks is what experiential futures have always had: publics. Its participants are officials, executives, and researchers. Step 4 retains the rationing of imagination and the debrief as instrument, and stages them with the people on whom the future would be imposed.

4.6 Theory of Change: Connecting Research to the Future

The sixth practice is the most institutionalized. Before a project is funded or a fellow is placed, a *theory of change* must be articulated: an explicit causal account of how doing X now reduces harm Y later. This becomes Step 6 of the method, and the demand is infrastructural rather than rhetorical. MATS requires every scholar’s research plan to state “the AI threat model or risk factor motivating their research, a theory of change to address that threat model, and a SMART plan for a project to enact that theory of change.” It grades the plan and reports a cohort average of 74/100, with scholars scoring worst on the threat model and the theory of change [A162]. SPAR, LASR Labs, PIBBSS, ML4Good, and AI Safety Camp make the same demand at application, in mandatory reports, or as a curriculum block [A104, A114, A124, A137, A154]. The genre has a canonical grammar: “mapping causality backwards from impact, and questioning each causal step” [35].

Funding enforces the same demand. Coefficient Giving, the ecosystem’s largest funder, lists “theory of change” first among its assessment criteria and gates the expression-of-interest stage on it [A127], and the Long-Term Future Fund told its investigators to assess it first [A103]. The same reasoning form organizes whole institutions. ERA publishes an inputs-to-impact logic model for its fellowship, with eight named assumptions each carrying a self-assigned confidence grade, and treats it as a working document (Figure 1a) [A152]. The Cooperative AI Foundation shows the chain at ecosystem scale: a 2020 research agenda [A44] led to a foundation, whose taxonomy of multi-agent risks [A67] was then named among the motivating papers of a \$10 million funding call [A40].

Two details clarify the function of the theory of change. First, it operates as a diagnostic instrument rather than a compliance document. Pivotal Research and GovAI require no theory of change at application, yet their fellows produce one under supervision before leaving. ERA’s research-management retrospective reports that, with the causal chain written out, teams that pivoted mid-fellowship found it “easier to figure out where to ‘start again’ and rethink where assumptions were wrong” [A134]. Second, the community audits its own theories of change and publishes unfavorable results. A three-wave survey across the x-risk summer fellowships found fellows’ probability of pursuing an x-risk career essentially unchanged by the programs [A98], and Apollo reported that its precursor evaluations “had limited predictive power” for the scheming behaviors they were built to anticipate [A17]. The theory of change makes a speculative future operational by demanding an account of intervention, and then makes that account itself an object of test.

A scoping observation follows, which we offer as a finding rather than a concession. The demand saturates the layer we studied, training programs, funders, and public agendas, which is exactly the layer where justifying research is the product. It does not follow that frontier research is chosen this way, and some of it demonstrably is not. Redwood Research’s control agenda, among the field’s most productive recent programs, declines to backchain from a resolved threat model. It asks instead what limits damage if alignment fails, and iterates empirically from there [A64]. The community demands backchained theories of change from juniors and grantees while some of its best senior work forward-chains, and a method distilled from this corpus inherits that gap (Sections 7.1 and 8). What the practice lacks, finally, is attention to whom the intervention serves. Step 6 adds re-futuring, which asks what future the intervention itself makes possible, and for whom.

5 Distilling Preemptive Futuring as Method

From these documented practices we distill a six-stage futuring method for HCI. The six stages operationalize the pattern described in Section 1 end to end. Two honesty conditions govern this section. First, what follows is a *specification*: its operations are drawn from a community’s documented practice, while their assembly into a single sequence and the participation structure it prescribes have not yet been run with a design team. Section 8 treats their evaluation as the necessary next step. Second, the stages are presented in sequence, but the practice we observed is iterative. Pathways are redrawn when assumptions fail, and interventions return the process to its beginning. For each stage, we indicate the AI-risk practice it generalizes and the adjustments that HCI’s commitments require.

5.1 Step 1 — Articulate an Undesirable Future

Describe the future concretely enough to reason about it: not “AI harms creative workers” but a specified state of affairs, with actors, systems, and conditions named. The AI risk community’s scenario artifacts are instructive less for their content than for their form. Cotra’s takeover scenario names a fictional lab and model and walks through training, deployment, and failure step by step [A41]. The AI 2027 scenario commits to month-by-month specificity precisely so that it can be disagreed with [59]. Specificity is what makes a future backchainable.

Then press the questions that HCI is better equipped to ask than AI safety has been: what exactly is undesirable, for whom, at what scale, which values are threatened, and who holds them. An undesirable future is never universally undesirable; a future that devastates gig workers may serve platform shareholders well. Harm taxonomies from the algorithmic-harms literature offer a vocabulary for specifying the harm end of a pathway with the same discipline that the safety community applies to the capability end [87, 99]. Naming whose undesirable future is being prevented is not a preamble to the method; it is the first design decision in it [47, 77].

5.2 Step 2 — Backchain Possible Causal Pathways

Ask repeatedly: *for this future to occur, what would have to happen first?* Move backward from the future toward the present, recording each enabling condition as a node and each “would have to” as a link. The output is not a linear chain but a causal graph. Real pathways branch, converge, and share sub-goals, which is why security engineering moved from attack trees to directed acyclic graphs [60, 86] and why the AI risk community’s modeling efforts produce influence diagrams rather than timelines [24].

This step is the one most easily confused with backcasting, and the difference matters in practice. A backcasting workshop asks “what steps take us to the future we want?” and succeeds when it produces a feasible plan [79, 82]. Backchaining from an undesirable future asks “what routes lead to the future we fear?” and succeeds when it produces a sufficiently complete map of routes. That map must include routes the team finds implausible, because plausibility judgments are themselves assumptions to be surfaced in Step 3. Generative checklists guard against premature closure here. STRIDE plays this role in threat modeling, forcing analysts to enumerate threat categories they would not have imagined unprompted [58, 89]; harm taxonomies can play the same role for sociotechnical futures [87].

5.3 Step 3 — Surface Assumptions

Annotate every causal link with the assumptions it rests on, and grade them. The grading vocabulary is constructed here. It synthesizes the two registers the field already uses (ERA’s ordinal confidence grades and Carlsmith’s explicit credences) with the mechanism vocabulary of evaluation research. The grades are *evidence-supported* (a documented trend or finding), *plausible* (consistent with evidence but not demonstrated), *contested* (experts disagree), *speculative* (no present evidence either way), and *normative* (a claim about what actors will value or choose). ERA’s confidence-graded assumption table and Carlsmith’s premise-by-premise credences (Section 4.3) are the two models for this annotation, and they show that readers can locate their disagreement only once the grades are written down.

The annotation is the essential part of this step. An unannotated causal diagram looks like knowledge, whereas an annotated one shows exactly where it is conjecture. Annotation is what makes the pathway *contestable* rather than merely visible: a participant who cannot draw a better diagram can still attack a graded assumption. The pathway becomes, in effect, a container for assumptions [9, 76].

5.4 Step 4 — Branch and Contest the Pathway

Construct alternatives deliberately. Ask what other pathways could lead to this future, what conditions would prevent a given transition, and which actors or forces the diagram has omitted. Ask also who disagrees with this causal account and what their version looks like. The AI risk community institutionalizes this stage in forms that HCI can borrow directly: threat-model literature reviews that lay competing causal accounts side by side [A89]; red-teaming norms that assign people to break arguments rather than defend them; and public retrospectives in which authors grade their own past threat models against what actually happened [A94].

This stage is also where pathways are rehearsed rather than merely argued. The simulation practices of Section 4.5 serve as Step 4 instruments. Role-playing a pathway with imagination deliberately rationed (one novel variable, a near-term horizon, everything else frozen) surfaces enabling conditions and omitted actors that discussion misses. The debrief, run as an instrument rather than a wrap-up, records exactly the disagreements that the diagram must absorb [A19][A147]. HCI’s contribution to this stage is to widen who counts as a contestator. In the communities we studied, pathway contestation happens mostly among insiders who share training and priors, a limitation that their

own critics name [29]. Participatory and community-led speculation shows how the branching step can include the people on whom the future would be inflicted, whose causal knowledge of their own circumstances routinely exceeds the analysts' [20, 48]. A pathway that survives only insider contestation remains undertested.

5.5 Step 5 — Locate Leverage Points

Examine the graph for places where intervention is possible from the present. Candidate intervention objects span the full sociotechnical stack: interfaces, algorithms, protocols, datasets, evaluation regimes, incentives, procurement rules, standards, institutions, and social practices. Two heuristics from the material we studied are worth carrying over. First, prefer nodes that many pathways share. A condition that appears on several routes to the future is worth more than one specific to a single route. This is the logic by which attack-tree analysis identifies where defense pays [86] and by which safety researchers select “the most robust bottleneck” as their research target [A169]. Second, beware the shallow-leverage bias. Intervention effort gravitates toward parameters that are tractable and measurable rather than toward the goals, rules, and paradigms where leverage is greatest [1, 67]. The question of which node is tractable from here must be balanced against the question of which node actually matters.

5.6 Step 6 — Design Interventions and Re-Future

Design the intervention and state its theory of change explicitly, that is, the mechanism by which acting now makes the undesirable future less likely. The observed template (threat model, then theory of change, then concrete plan) is teachable and gradeable. MATS grades exactly this triplet in its scholars' research plans [A162].

Then re-future. Interventions change the system in which they intervene, so the final move is to ask what future the intervention itself makes possible. A dispute mechanism can become a surveillance apparatus, and a safety evaluation can become a capabilities benchmark. The security literature's attack-defense trees model the adversary's response to the defense inside the same diagram [60]. Redraw the pathway with the intervention in place, and treat the redrawn graph as a new Step 1 input. Figure 3 shows the full cycle.

This recursion separates Preemptive Futuring from one-shot causal planning. The diagram is never finished. It is a living document whose revision history is part of its content, as the versioned scaling policies, “[OUTDATED]” prefixes, and annual plan updates of our field sites show.

5.7 The Method in Summary: What, When, and Instead of What

Table 2 compresses the method into its operative form: each stage's core question, the artifact it produces, and the quality criterion (Section 6) that bears on it most heavily.

Running the method. A specification must be executable, so we fix the operative details a team needs.

- *Artifact.* The pathway is a directed graph, not a chain. Nodes are conditions or events. Links carry AND/OR semantics where routes require conjunctions, following attack-graph practice [60]. Each link records the causal claim, the assumptions on which it rests, their grades, and an evidence pointer.
- *Stopping rule for Step 2.* Backchain until every open link is either evidence-supported or ends in a condition actionable from the present. Require at least two rival routes before grading begins.
- *Ranking rule for Step 5.* Score candidate leverage nodes on how many mapped pathways share them, how tractable intervention is for the actors present, and how quickly acting there would produce observable feedback.

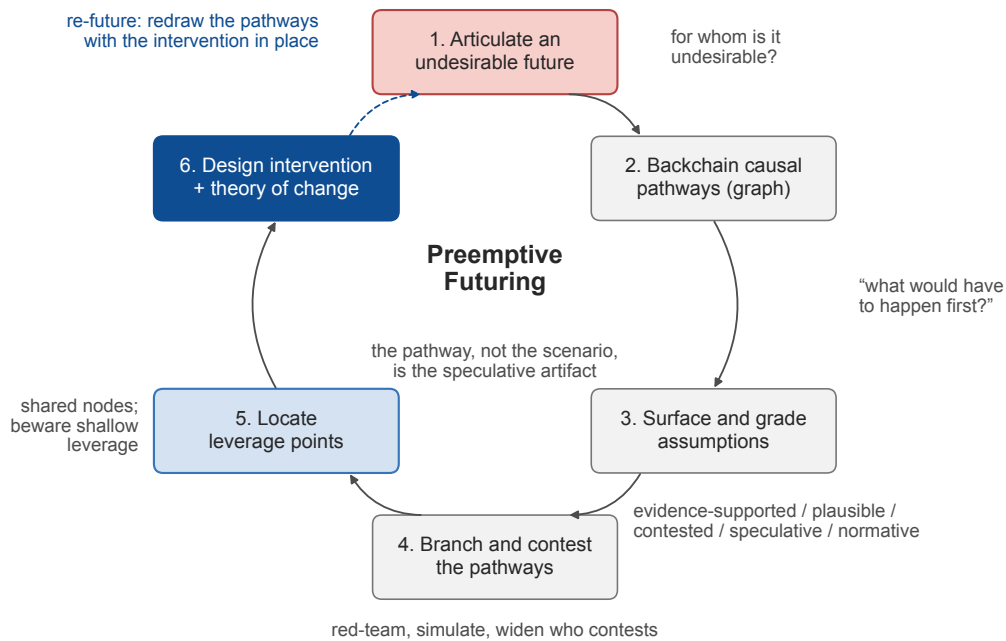


Fig. 3. The Preemptive Futuring cycle. Six stages run from articulating an undesirable future through backchaining, assumption grading, contestation, and leverage analysis to an intervention carrying its own theory of change; the dashed blue re-futuring arrow closes the loop by redrawing the pathways with the intervention in place. Red marks the undesirable future, light blue the leverage stage, and dark blue the intervention stage. The pathway, not the scenario, is the primary speculative artifact.

The third criterion answers the forward-chaining objection of Section 7.1 by making learning value part of leverage.

- *Termination.* The cycle stops, provisionally, when a full pass changes no grades.
- *Facilitation.* Step 4 should not be facilitated in expectation of convergence. The field sites' best evidence is that structured contest localizes disagreement rather than resolving it (Section 4.4). The facilitator's product is therefore a disagreement register attached to the graph: named cruxes rather than consensus.

A worked pathway. Figure 4 shows the artifact once these rules are applied. It is constructed to fix the artifact's grammar rather than reported from a participant session, and the case is drawn from HCI's own terrain rather than from AI risk. Step 1 fixes the undesirable future: a 2035 in which care facilities route family contact through robot mediation by default. Step 2 backchains two routes into it. One runs through procurement. Reimbursement that pays for "contact events" rather than relationship quality, AND staffing ratios worsened by funding rules, together push procurement to select on a per-contact cost metric. Facilities then adopt robots as the lower-cost alternative to staffing. The other route runs through normalization: vendor defaults that log and summarize family calls make robot-mediated contact seem adequate. Step 3 grades each link. The two procurement inputs are evidence-supported, the adoption step

Table 2. The Preemptive Futuring method in summary: each stage's core question, the artifact it produces, and the quality criterion (Section 6) that bears on it most heavily.

#	Stage	Core question	Output artifact	Primary quality focus
1	Articulate an undesirable future	What exactly is undesirable, for whom, at what scale, threatening which values?	Concrete scenario specification (actors, systems, conditions named)	Scenario: specificity, situatedness, no false inevitability
2	Backchain causal pathways	For this future to occur, what would have to happen first?	Causal graph: branching, convergent, never a single chain	Pathway: plurality
3	Surface assumptions	What does each link rest on?	Every link graded: evidence-supported / plausible / contested / speculative / normative	Pathway: transparency, plausibility
4	Branch and contest	What other routes exist? Who disagrees, and what has been omitted?	Rival pathways; red-team and simulation record; widened contesteer pool	Pathway: contestability, sensitivity
5	Locate leverage points	Which nodes do many pathways share, and which are tractable now?	Ranked intervention sites across the sociotechnical stack	Intervention: causal relevance, robustness
6	Design intervention and re-future	Through what mechanism does acting now make the future less likely, and what future does the intervention itself create?	Intervention with forward theory of change; redrawn pathway graph	Intervention: reversibility, second-order effects, legitimacy

is plausible, the normalization step is speculative, and the final step is normative as well as speculative. Step 4 keeps the rival normalization route on the graph rather than resolving it. Step 5 selects the procurement metric as the leverage point, because both routes share it, a standards body can act on it, and the effect is observable within one procurement cycle. Step 6 attaches the intervention, a relationship-quality criterion and contestable contact-log defaults proposed at the standards body, with the redrawn graph as its theory of change.

When to use it. Preemptive Futuring justifies its overhead under four conditions. Three held at our field sites; the fourth is the deficit there that HCI is positioned to repair. First, the harm is credible but not yet observable, so retrospective methods have no incident corpus to work on. Second, the pathways to the harm cross technical and institutional layers, so no single expertise can map them and the graph is informative. Third, intervention capacity exists: someone in the room, or reachable from it, can act on a leverage point, so the exercise ends in design rather than in dread. Fourth, contestation can be hosted: participants with divergent stakes and causal knowledge are available, and the team will let them redraw the diagram. At our field sites this condition held only in the weak, insider form documented in Section 8, which is why Step 4 treats widening the contesteer pool as constitutive. When these conditions fail, a different method is appropriate. For open-ended imagination or provocation, design fiction does the work with less machinery and fewer false promises of rigor [16, 33]. If the future is desired, backcasting is the right tool [82]. If the harm is already manifest, incident analysis and ordinary empirical HCI outperform speculation. If no one can act on what the map shows, a pathway diagram degenerates into fatalism, theater, or, at worst, prediction laundering, which lends an invented future the authority of a causal analysis (Section 7.1).

Instead of what. Table 3 positions the method against the futuring and analytic traditions it borrows from, showing for each the starting point, primary artifact, and what Preemptive Futuring adds or inverts.

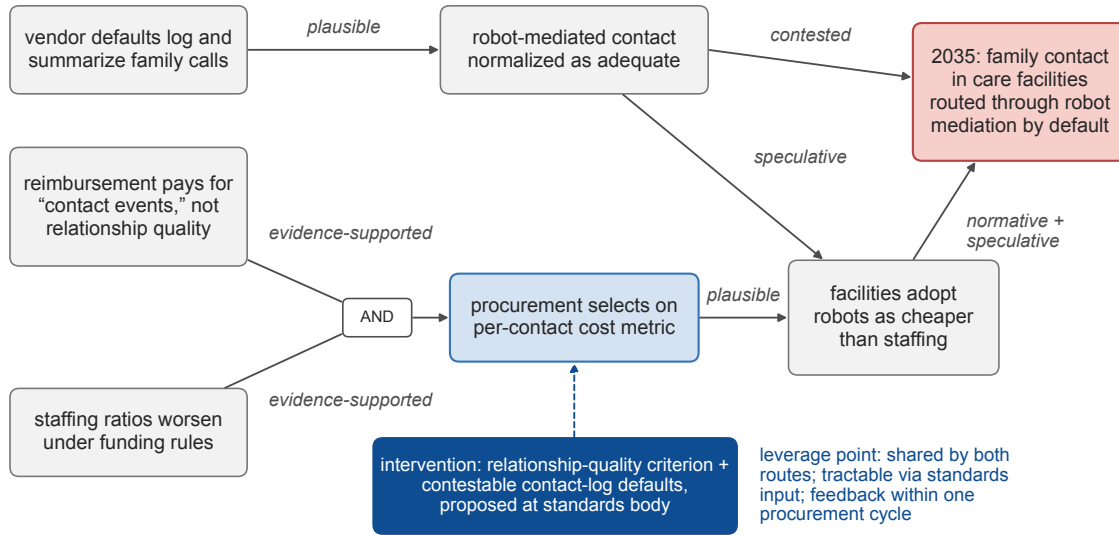


Fig. 4. An illustrative Preemptive Futuring pathway for a care-robotics future, constructed to illustrate the artifact’s grammar rather than reported from a workshop. Gray nodes are conditions and the red node is the undesirable future; italic link labels are assumption grades; AND marks a conjunction one intervention can sever; the light blue node is the selected leverage point and the dark blue box the intervention; the rival route through normalization is retained, not resolved.

6 Evaluating Preemptive Futuring: From Quality of Speculation to Quality of Intervention

Preemptive Futuring needs criteria that separate good practice from bad. Existing evaluative debates in design futuring ask whether speculation is plausible, provocative, situated, or generative, and rightly insist that criteria follow from purpose [10, 12]. When the purpose is intervention, we propose that evaluation must operate at three layers. A compelling scenario attached to a lazy pathway, or a rigorous pathway attached to an infeasible intervention, fails as Preemptive Futuring even if it succeeds as provocation. Good Preemptive Futuring therefore requires scenario quality, pathway quality, and intervention quality together; weakness in any layer undermines the others.

6.1 Scenario Quality

The scenario layer inherits HCI’s existing criteria, sharpened by the intervention orientation. Is the future articulated concretely enough to backchain from, with actors, systems, and conditions named, in the way that separates a mood from a threat model [A41]? Does it expose meaningful tensions rather than rehearsing a genre cliché [90]? Is it situated, stating for whom the future is undesirable, at what scale, and which values are threatened [47, 87]? Does it avoid false inevitability by presenting itself as one branch of a contested possibility space rather than a destiny? Our field sites offer a caution here as much as a model. Their scenario stock skews heavily toward a small family of catastrophes, and their own worldbuilding experiments found that desirable futures had to be commissioned, constrained, and prize-funded into existence [A60]. Scenario quality therefore includes the question speculative design has always asked best: whose imagination is this?

Table 3. Positioning Preemptive Futuring among the futuring and analytic methods it borrows from.

Method	Starts from	Primary artifact	Relation to Preemptive Futuring
Speculative design / design fiction [14, 33]	A provocative possible world	Diegetic artifact or story world	Supplies the scenario craft of Step 1; PF adds the pathway, assumption, and intervention machinery the scenario lacks
Experiential futures / speculative enactments [22, 36]	A future staged for participants	Immersive experience; enactment data	Becomes Step 4's rehearsal instrument; PF adds the causal graph that the enactment stress-tests
Backcasting [79, 82]	A desirable end-state	A feasible plan of steps toward it	PF inverts the polarity: undesirable endpoint, plural contested routes, disruption rather than attainment
Scenario planning [17]	Present critical uncertainties	Two to four divergent scenarios	PF fixes one endpoint and multiplies routes to it, instead of multiplying endpoints
Transition design [52]	Visions of sustainable futures	Transition pathway with theory of change	Shares ToC-as-design-material; PF answers its critics' "from what?" with a specified harm [96]
Threat modeling (security) [86, 89]	An adversary's goal	Attack tree / DAG with mitigations	PF generalizes backward-from-harm to non-adversarial sociotechnical futures, and adds participation and politics
Theory of change (evaluation) [27, 100]	A program's desired impact	Logic model with assumptions	PF runs it in reverse to map the harm, then forward again for the intervention (Figure 1)
Value scenarios / envisioning cards [41, 71]	Stakeholder harms and values	Scenario prompts and card decks	Feed Steps 1–2 as generative checklists; PF adds grading, contestation, and leverage analysis
Timelines [102]	Values advocacy through world-building	Participant-built event sequences linking present to future	The closest CHI relative; PF adds causal (not chronological) links, graded assumptions, and the leverage/intervention stages
Threatcasting [54]	A specific future threat	Backcast threat scenario with disrupt/mitigate/recover actions	Nearest futures-studies relative; PF adds pathway plurality, assumption grading, and a politics of who contests the pathway
Safety cases [25]	An acceptable-risk claim for a specific system	Structured argument graph with evidence and assumptions	Same artifact grammar, certification framing; PF addresses sociotechnical futures with no authorized decider
Assumption-Based Planning / RDM [30, 62]	An existing plan; deep uncertainty	Load-bearing assumptions; robust strategy sets	Supplies the decision machinery PF's Steps 3 and 5 draw on; PF supplies the futuring, participation, and design framing

6.2 Pathway Quality

The pathway layer is the distinctive contribution, and we propose six criteria. For each, we state whether it is grounded in Section 4's practices, imported from evaluation and decision literatures, or constructed by us to name a gap those practices exhibit. The layering itself extends the reflective dimensions Kozubaev et al. [61] proposed for design futuring, which already include plurality and situatedness at the scenario level. Our claim is that intervention-oriented speculation needs these properties at the pathway and intervention levels as well.

Transparency (grounded: Sections 4.3 and 4.6). Are the causal assumptions visible and graded? A pathway that hides its assumptions is an argument disguised as a map [27][A152]. Transparency concerns the visibility of assumptions; traceability, below, concerns the explicability of the intervention’s connection. A pathway can display its assumptions yet fail to explain why acting here touches the future at all.

Plausibility (imported: realist evaluation and contribution analysis). Is there a stated basis for each transition, whether evidence, mechanism, or explicitly flagged conjecture? The standard is realist rather than predictive. It asks for mechanisms that could fire in specified contexts [9, 76], held to contribution-analysis standards of reasonable confidence rather than proof [66].

Plurality (grounded: Sections 4.2 and 4.4). Are alternative pathways represented, including ones the authors find implausible? A single-chain pathway is a symptom of premature closure [83][A90].

Contestability (grounded: Section 4.4; radicalized here beyond the insider form observed). Can participants disagree with the pathway in a form that registers? This is a property of the artifact and its setting jointly. Graded assumptions give critics a target [23]; institutionalized red-teaming gives them a role [A166]; and participation determines who may be a critic at all [48].

Sensitivity (grounded: Section 4.4’s crux practice; formalized in decision analysis). Which assumptions most affect the outcome? The FRI adversarial collaboration is exemplary. It located the two-orders-of-magnitude disagreement in a handful of named cruxes, most of which near-term evidence cannot resolve: an honest sensitivity analysis of a contested future [A139].

Traceability (imported: theory-of-change discipline). Can we explain, link by link, why a proposed intervention connects to the future it claims to address? This is the property that theory-of-change discipline enforces [97][A162] and that vague harm anticipation lacks [70].

6.3 Intervention Quality

The intervention layer asks the following of any proposed present action.

Causal relevance. Does the intervention act on a mechanism that actually appears in the pathway, or is it attached to the scenario by topic rather than by cause?

Tractability. Can it plausibly be enacted by identifiable actors with identifiable resources? The field sites are disciplined about this to a fault, with tractability scoring built into their prioritization frameworks [A170].

Robustness. Is it useful across several plausible pathways rather than only the authors’ favorite? The practices we observed prize interventions positioned at nodes that many pathways share [A169]; the same logic transfers directly.

Reversibility. Can it be revised or withdrawn if its assumptions prove wrong? Preemptive action under uncertainty is responsible only if it remains correctable [26, 93].

Second-order effects. What new undesirable futures could the intervention itself enable? Preemption can produce what it claims to prevent [7, 94], and the re-futuring step in the method (Step 6) exists precisely to force this question.

Disclosure (constructed: named after a norm our field sites hold and our method would otherwise drop). When is the pathway itself the hazard? Communities that practice backward-from-harm reasoning pair it with disclosure norms, such as staged release and withheld operational detail. They do so because a well-graded route map to a harm is, in the wrong hands, a capability transfer. An intervention-quality assessment must therefore state what part of the pathway is published, at what grain, and to whom.

Legitimacy (constructed: named after the deficiency of Section 8, not found in the corpus). Who decided that this intervention, against this future, on behalf of these people, was justified? Our field sites answer largely with

expert consensus and funder judgment. Their own evaluation regimes reveal how contested that answer is. The same ecosystem contains a program whose impact ledger counts a placement at a frontier lab as negative [A136] and funders who evaluate research by its portfolio role rather than its findings [A92]. HCI’s answer must be broader, and Section 8 takes the question up.

Applying the criteria. We offer these criteria as a structured checklist rather than a numeric rubric. The field sites teach a warning about premature quantification, having watched their own evaluation instruments fail their predictive tests [A17]. A checklist must still discriminate, so we close by applying it to two published projects. Alfrink et al.’s contestable camera cars [4] fare well on intervention quality: the dispute mechanism is causally relevant, tractable, and reversible, and its deployment setting supplies legitimacy. The pathway layer, however, is thin. The route from present conditions to the futures the design forestalls is implicit, and its assumptions are ungraded. Coulton et al.’s design-fiction world [28] inverts the profile: rich scenario quality and, deliberately, no pathway or intervention claims at all. This is not a failing of the work but exactly the difference the framework registers. The criteria separate the two cases and name what each would need to become Preemptive Futuring. The camera cars need their pathway drawn and graded; the fiction would need everything after Step 1. That is the intended use: a diagnosis of where a project sits, applied repeatedly, in public, with the failures logged.

7 Discussion

7.1 From Scenario-Centered to Pathway-Centered Futuring

Our core theoretical claim is a shift of speculative attention. Speculative design made possible worlds available for reflection [33, 61]; Preemptive Futuring additionally makes the *relations* between present and future available for design. The pathway becomes a speculative artifact in its own right, drafted, critiqued, exhibited, and revised. Its parts can be individually contested, its assumptions individually graded, and its joints individually fitted with interventions. Timelines [102] and path-oriented design fiction [78] gestured at this. Preemptive Futuring adds the machinery (assumption grading, adversarial branching, leverage analysis, theories of change) that turns a timeline into an inspectable argument. It thereby shows that Tonkinwise’s critique [95], that speculative design brackets the transitions that would make its futures arrivable, is answerable in practice.

The shift changes what speculation claims. Preemptive Futuring does not ask “will this future happen?” It asks instead: if it did, through which pathways, and which present interventions remain valuable across them? This is conditional intervention reasoning rather than prediction. The distinction separates the method from forecasting and from the speculative ethics Nordmann [73] criticizes for smuggling conditional futures into present moral claims.

Our field sites model the distinction imperfectly. Nearly every serious practitioner disclaims prediction and then keeps score anyway. A consultancy that re-staged the Intelligence Rising game for retired generals later audited its fiction against events and reported that “29% of our injects had either occurred, or similar events had occurred,” beside the disclaimer “we don’t claim to forecast the future” [51]. RAND’s exercises likewise disclaim prediction yet harvest capability requirements from every run [A129, A147]. A scorecard whose items were authored to be plausible, whose match criterion stretches to “similar events,” and which has no pre-registered resolution rule licenses nothing; it is prediction laundering conducted in public. The method therefore carries a discipline the field sites lack. A published pathway should state, link by link, either a pre-registered resolution criterion or that the link is unscorable, so that later retrospectives grade what was claimed [A85].

A second boundary the community drew for us is Ngo’s objection (Section 4.4) that backchaining favors work that generates no near-term feedback. The objection has evidence behind it, since the field’s forward-chained measurements have repeatedly falsified its backchained expectations [A17]. Our response is a scope condition. Preemptive Futuring allocates anticipatory attention and justifies present intervention; it does not choose research directions, and Step 5’s feedback-observability criterion imports what the objection demands. Under ignorance, the aim is Stirling’s [93]: to open up the space of framings and then close down, accountably, on action.

The shift also re-reads a familiar artifact. Theory of Change is usually treated as the diagrammatic paperwork of funders [97]. In the communities we studied it is a futuring technology, taught, graded, workshoped, and revised as the primary genre of disciplined imagination [3][A124][A162]. It is the concrete artifact in which the sociology of expectations [18] can watch expectations being made inspectable and, as ERA’s confidence-graded assumption table shows, caught failing [A152]. Transition design claimed theories of change as design material a decade ago [52]. Our field sites show what happens when an entire research culture runs on the genre: it becomes an apprenticeship form, a debugging tool, and an audit trail at once.

7.2 Speculative and Preemptive Futuring

The two traditions share the imaginative work and diverge on what follows it. Speculative design withholds the solution. Its artifacts pose a question and decline to answer it, and design that terminates in a recommendation has returned to the problem-solving role speculation was set against [11, 33]. Design fiction, material speculation, and experiential futures all keep the imagined world non-committal, and that reticence is what allows speculation to be plural and provocative without having to be right [16, 22, 98]. It is also what keeps the tradition clear of Nordmann’s charge [73] that conditional futures acquire present moral force without argument. Preemption carries the same imaginative work further. It is not finished until it has named an action to take now, and it therefore inherits a burden speculative practice was built not to carry: the imagined future must justify an intervention imposed on people who did not imagine it. Tonkinwise’s complaint [95] that speculative design brackets the transitions that would make its futures arrivable names this same boundary from the other side.

Carrying that burden requires an artifact that can hold a justification and still be changed. This is the sense in which we treat the hypothesized pathway as a design material rather than as an analysis. Transition Design already claimed theories of change as material for design [52]; the pathway extends that claim to the reasoning that connects a feared future to a present act. The distinction from an analytic model is the point. A threat model or an attack tree is completed correctly or incorrectly; a design material is authored, annotated, misused, handed to someone else, and revised. Read this way, the six stages of Section 5 name operations performed on a material as much as steps in a procedure: articulating an endpoint, branching a route, grading a link, severing a conjunction, attaching an intervention, and redrawing the map once that intervention is in place. The material is what makes a justification inspectable, and inspectability is what separates an argument for preventive action from an assertion of it.

What HCI can take from our field sites is that repertoire, and not the arrangement in which we found it. The operations transfer. The endpoint fixed by whoever convenes the exercise does not, and neither does an authorship restricted to those already fluent in the community’s vocabulary. Preemption is politically productive as well as protective, and an uncertain future can authorize action whose own effects escape comparable scrutiny [7, 65]. A repertoire adopted without that caution would give HCI a sharper instrument and the same enclosure. The condition of transfer is that the material stay open exactly where the source tradition closes it, at the endpoint and at the right to redraw. Whether HCI can hold both dispositions at once is the question this paper leaves open. A field that has learned to distrust design

work terminating in a recommendation is asked here to produce an intervention on a schedule, and the risk is not that it will fail but that it will succeed too readily, trading speculation’s openness for the satisfaction of having something to do. The pathway is our answer to that risk. It is the one artifact in the sequence that must remain contestable after the intervention has been chosen. The next section sets out what each tradition can offer the other on that basis.

7.3 What HCI Can Learn From, and Challenge In, AI Risk Research

The exchange we propose runs in both directions. From AI risk research, HCI gains what its speculative traditions have lacked. The first gain is an intervention orientation that treats speculation as unfinished until it terminates in present action. The others are explicit theories of change, causal decomposition as a discipline of imagination, leverage-point reasoning, and the institutional forms (graded research plans, adversarial review, published self-audits) that keep intervention-oriented speculation honest. From HCI, AI risk research gains correctives its own critics have demanded [2, 29]. The offer must be stated precisely, however. The community has considered reasons to resist it and has already run participation experiments of its own (public red-teaming events, a commissioned worldbuilding contest [A60], deliberative input into model constitutions [49]). The considered objection is that catastrophic-risk mitigation is not a preference-aggregation problem: widening participation on technical causal questions does not obviously improve the causal model. We accept half of it. What participation most defensibly improves is the *legitimacy* of pathway selection (whose futures get mapped, which harms count, who may redraw the diagram), and that is where HCI genuinely leads. The stronger claim is that situated knowledge also improves pathway *accuracy* where links run through lived institutional and social terrain. That claim holds for exactly those links and should be argued case by case rather than assumed. What HCI adds to these participation experiments is the machinery this paper describes: participation located at the pathway, where disagreement is inspectable, rather than at the scenario, where it is decorative. HCI also brings reflexivity about how instruments transform their users, attention to lived experience alongside expected value, and resistance to the premature closure that turns one community’s threat models into everyone’s research agenda. Preemptive Futuring is not AI safety practice imported into HCI. It is a method that emerges between the two traditions and, we would argue, is available to neither alone. The analytic machinery without the pluralism reproduces the narrowness for which our field sites are criticized; the pluralism without the machinery reproduces the intervention gap for which speculative design is criticized.

8 Limitations

Our account has boundaries that should temper its claims.

Empirical scope. Our empirical base covers particular AI risk communities: a Cambridge-centered fellowship ecosystem and its transatlantic extensions. It does not cover the field as a whole, and certainly not the many AI research cultures that reject the existential framing [46].

Pathways are not causal models. The causal pathways at the heart of the method remain speculative constructs, not validated causal models. Treating the diagrams as predictions would reproduce exactly the false confidence the method is designed to avoid. The field sites’ own instrument failures show how easily anticipatory tools overreach [A17].

Normative situatedness. “Undesirable future” is normatively situated, and the method inherits whatever exclusions shape the room in which Step 1 is performed.

A specification, not yet a validated method. No design team has yet run the assembled sequence. Our worked pathway (Figure 4) is an authored illustration rather than a session output. The practices it distills have an operational record in their community of origin: graded research plans, published null results, and contests that failed to converge (Sections 4.4 and 4.6). What remains untested is their transfer. Evaluating that transfer against Section 6’s criteria, reported failures included, is the necessary next step, and the standard against which the method’s claims should be re-read.

Backchaining’s internal critics. Backchaining has thoughtful critics inside the community from which we learned. Section 7.1 states Ngo’s selection-effect objection and our scope condition in response. The tension is bounded there, not dissolved.

The epistemics of prevention. If the undesirable future never occurs, we cannot know whether the intervention worked or the pathway was wrong. Preemptive Futuring does not escape this constitutive problem, and no approach does. By making pathways and assumptions explicit and revisable, however, it at least ensures that what cannot be verified can still be argued about.

The politics of causal futures. Causal diagrams appear neutral, but they are not. Every pathway encodes decisions about whose future matters, what counts as harm, which actors appear as nodes and which as background, and who is authorized to intervene [53, 69]. Our field sites make the stakes concrete. Their pathways are drawn by a small, demographically narrow network in which the same people design the wargames, write the scenarios, grade the forecasts, and fund the exercises [29][A65]. The funding that moves from philanthropic networks into government capability recommendations is rarely visible in the resulting documents. The critical literature on preemption sharpens the worry. Anticipatory action can be self-justifying, because the prevented future never arrives to be checked, and preemptive framings historically produce the authorities they claim to serve [6, 7, 65]. To design a causal pathway is therefore also to design a politics of the future. This is not a reason to abandon pathway-centered futuring. It is the reason the method’s contestability criteria and its participatory widening (Step 4) are constitutive rather than optional. A pathway that cannot be contested by the people it concerns is not an analysis but an enclosure.

9 Conclusion

Futures need not be predictions to matter for present action. We have described how AI risk research communities turn speculative undesirable futures into present research and intervention priorities. They make futures researchable through threat models, work backward through contested causal pathways, grade assumptions in public, discipline projects with theories of change, institutionalize adversarial contest, and rehearse futures until the rehearsals reshape the rehearsers. We have distilled these practices into Preemptive Futuring, a six-stage method that treats causal pathways toward undesirable futures as design material. Its three-layer quality framework extends speculation’s evaluative debates from scenario to pathway to intervention. Nothing in the method is specific to AI. Energy, climate, and human–robot futures present the same combination of credible unobserved harm and cross-layer pathways. For HCI, the offer is an expanded role for speculation: making the pathways through which undesirable futures might emerge inspectable, contestable, and intervenable.

Positionality and Competing Interests

One author previously took part in two of the training programmes examined here and received financial support from them. Both engagements preceded the formulation of the research question, and neither organization had any

role in the design, analysis, or writing of this paper. We record this as a competing interest, and as a plausible source of bias in three specific places: which documents came to our attention, which reasoning moves we recognized as a practice worth naming, and how far we adopted the community’s evaluative vocabulary rather than examining it. The paper is written so that these risks can be checked rather than taken on trust. Every substantive claim is tied to a citable public document; the corpus was counter-sampled for disconfirming material, including published null results, internal critiques, and two programmes verified at source to make no theory-of-change demand at application; and the politics of the practice is treated as a finding in its own right (Sections 4.4 and 8). To preserve anonymity for review we withhold the names of the two programmes, the dates and duration of the participation, and the amounts and sources of the support; the relationships themselves are declared above without those particulars, and all of them will be named in full at camera-ready.

Research Ethics

This study analyzes documents that their authors chose to publish in public. It involved no interviews, observation sessions, surveys, or recruitment, and it collected no data from or about human participants. It therefore did not require research ethics review. Quotations are drawn from publicly accessible documents, cited with URLs and access dates, and archived in full text.

Generative AI Use

Generative AI systems were used throughout this project under our direction: to search and summarize literature, to locate and archive corpus documents, to draft and revise prose, to render the figures from authored specifications, and to produce simulated reviews used as internal critique. Citations were checked programmatically against CrossRef and arXiv records. We are responsible for all content, including any claim, citation, or interpretation that an AI system proposed.

References

- [1] David J. Abson, Joern Fischer, Julia Leventon, Jens Newig, Thomas Schomerus, Ulli Vilsmaier, Henrik von Wehrden, Paivi Abernethy, Christopher D. Ives, Nicolas W. Jager, and Daniel J. Lang. 2017. Leverage points for sustainability transformation. *Ambio* 46, 1 (2017), 30–39. doi:10.1007/s13280-016-0800-y
- [2] Shazeda Ahmed, Klaudia Jaźwińska, Archana Ahlawat, Amy Winecoff, and Mona Wang. 2024. Field-building and the epistemic culture of AI safety. *First Monday* 29, 4 (2024). doi:10.5210/fm.v29i4.13626
- [3] Michael Aird. 2022. Workshop: Building a Theory of Change for Your Research. EA Global DC 2022, workshop recording and slides. Accessed August 9, 2026. <https://www.youtube.com/watch?v=PwnkYncoUsA>
- [4] Kars Alfrink, Ianus Keller, Neelke Doorn, and Gerd Kortuem. 2023. Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM. doi:10.1145/3544548.3580984
- [5] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. arXiv:1606.06565 <https://arxiv.org/abs/1606.06565>
- [6] Louise Amoore. 2013. *The Politics of Possibility: Risk and Security Beyond Probability*. Duke University Press, Durham, NC. doi:10.1215/9780822377269
- [7] Ben Anderson. 2010. Preemption, precaution, preparedness: Anticipatory action and future geographies. *Progress in Human Geography* 34, 6 (2010), 777–798. doi:10.1177/0309132510362600
- [8] Anthropic. 2025. Responsible Scaling Policy, Version 2.1. Anthropic. March 31, 2025. Accessed August 9, 2026. <https://www.anthropic.com/responsible-scaling-policy>
- [9] Brad Astbury and Frans L. Leeuw. 2010. Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation. *American Journal of Evaluation* 31, 3 (2010), 363–381. doi:10.1177/1098214010371972
- [10] James Auger. 2013. Speculative design: crafting the speculation. *Digital Creativity* 24, 1 (2013), 11–35. doi:10.1080/14626268.2013.767276
- [11] Jeffrey Bardzell and Shaowen Bardzell. 2013. What is “critical” about critical design?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, 3297–3306. doi:10.1145/2470654.2466451

- [12] Eric P. S. Baumer, Mark Blythe, and Theresa Jean Tanenbaum. 2020. Evaluating Design Fiction: The Right Tool for the Job. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference (DIS '20)*. ACM, 1901–1913. doi:10.1145/3357236.3395464
- [13] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The Values Encoded in Machine Learning Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, 173–184. doi:10.1145/3531146.3533083
- [14] Julian Bleecker. 2009. Design Fiction: A Short Essay on Design, Science, Fact and Fiction. Near Future Laboratory. Accessed August 9, 2026. <https://blog.nearfuturelaboratory.com/2009/03/17/design-fiction-a-short-essay-on-design-science-fact-and-fiction/>
- [15] Eli Blevis. 2007. Sustainable Interaction Design: Invention & Disposal, Renewal & Reuse. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, 503–512. doi:10.1145/1240624.1240705
- [16] Mark Blythe. 2014. Research through design fiction: narrative in real and imaginary abstracts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, 703–712. doi:10.1145/2556288.2557098
- [17] Lena Börjeson, Mattias Höjer, Karl-Henrik Dreborg, Tomas Ekvall, and Göran Finnveden. 2006. Scenario types and techniques: Towards a user's guide. *Futures* 38, 7 (2006), 723–739. doi:10.1016/j.futures.2005.12.002
- [18] Mads Borup, Nik Brown, Kornelia Konrad, and Harro van Lente. 2006. The sociology of expectations in science and technology. *Technology Analysis & Strategic Management* 18, 3–4 (2006), 285–298. doi:10.1080/09537320600777002
- [19] Nick Bostrom. 2013. Existential Risk Prevention as Global Priority. *Global Policy* 4, 1 (2013), 15–31. doi:10.1111/1758-5899.12002
- [20] Kirsten E. Bray, Christina Harrington, Andrea G. Parker, N'Deye Diakhate, and Jennifer Roberts. 2022. Radical Futures: Supporting Community-Led Design Engagements through an Afrofuturist Speculative Design Toolkit. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM. doi:10.1145/3491102.3501945
- [21] Christina Bremer, Bran Knowles, and Adrian Friday. 2022. Have We Taken On Too Much?: A Critical Review of the Sustainable HCI Landscape. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, 1–11. doi:10.1145/3491102.3517609
- [22] Stuart Candy and Jake Dunagan. 2017. Designing an experiential scenario: The People Who Vanished. *Futures* 86 (2017), 136–153. doi:10.1016/j.futures.2016.05.006
- [23] Joseph Carlsmith. 2022. Is Power-Seeking AI an Existential Risk? arXiv:2206.13353 <https://arxiv.org/abs/2206.13353>
- [24] Sam Clarke, Ben Cottier, Aryeh Englander, Daniel Eth, David Manheim, Samuel Dylan Martin, and Issa Rice. 2022. Modeling Transformative AI Risks (MTAIR) Project — Summary Report. arXiv:2206.09360 <https://arxiv.org/abs/2206.09360>
- [25] Joshua Clymer, Nick Gabrieli, David Krueger, and Thomas Larsen. 2024. Safety Cases: How to Justify the Safety of Advanced AI Systems. arXiv:2403.10462 <https://arxiv.org/abs/2403.10462>
- [26] David Collingridge. 1980. *The Social Control of Technology*. Frances Pinter, London.
- [27] James P. Connell and Anne C. Kubisch. 1998. Applying a theory of change approach to the evaluation of comprehensive community initiatives: Progress, prospects, and problems. In *New Approaches to Evaluating Community Initiatives, Vol. 2: Theory, Measurement, and Analysis*, Karen Fulbright-Anderson, Anne C. Kubisch, and James P. Connell (Eds.). The Aspen Institute, Washington, DC.
- [28] Paul Coulton, Joseph Lindley, Miriam Sturdee, and Michael Stead. 2017. Design Fiction as World Building. In *Proceedings of the 3rd Biennial Research Through Design Conference (RTD 2017)*. doi:10.6084/m9.figshare.4746964
- [29] Carla Zoe Cremer and Luke Kemp. 2022. Democratising Risk: In Search of a Methodology to Study Existential Risk. arXiv:2201.11214 <https://arxiv.org/abs/2201.11214>
- [30] James A. Dewar. 2002. *Assumption-Based Planning: A Tool for Reducing Avoidable Surprises*. Cambridge University Press. doi:10.1017/CBO9780511606472
- [31] Paul Dourish. 2010. HCI and Environmental Sustainability: The Politics of Design and the Design of Politics. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems (DIS '10)*. ACM, 1–10. doi:10.1145/1858171.1858173
- [32] Karl H. Dreborg. 1996. Essence of backcasting. *Futures* 28, 9 (1996), 813–828. doi:10.1016/S0016-3287(96)00044-4
- [33] Anthony Dunne and Fiona Raby. 2013. *Speculative Everything: Design, Fiction, and Social Dreaming*. MIT Press, Cambridge, MA.
- [34] Steve Easterbrook. 2014. From Computational Thinking to Systems Thinking: A Conceptual Toolkit for Sustainability Computing. In *Proceedings of the 2nd International Conference on ICT for Sustainability (ICT4S 2014)*. Atlantis Press. doi:10.2991/ict4s-14.2014.28
- [35] Effective Altruism Forum. 2023. Theory of change (topic entry). Effective Altruism Forum wiki. Accessed August 9, 2026. <https://forum.effectivealtruism.org/topics/theory-of-change>
- [36] Chris Elsdon, David Chatting, Abigail C. Durrant, Andrew Garbett, Bettina Nissen, John Vines, and David S. Kirk. 2017. On Speculative Enactments. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 5386–5399. doi:10.1145/3025453.3025503
- [37] Casey Fiesler. 2021. Innovating Like an Optimist, Preparing Like a Pessimist: Ethical Speculation and the Legal Imagination. *Colorado Technology Law Journal* 19, 1 (2021), 1–33. Accessed September 11, 2026. <https://scholar.law.colorado.edu/ctlj/vol19/iss1/1/>
- [38] Joern Fischer and Maraja Riechers. 2019. A leverage points perspective on sustainability. *People and Nature* 1, 1 (2019), 115–120. doi:10.1002/pan3.13
- [39] Laura Forlano and Anijo Mathew. 2014. From Design Fiction to Design Friction: Speculative and Participatory Design of Values-Embedded Urban Technology. *Journal of Urban Technology* 21, 4 (2014), 7–24. doi:10.1080/10630732.2014.971525
- [40] Jodi Forlizzi. 2018. Moving beyond user-centered design. *Interactions* 25, 5 (2018), 22–23. doi:10.1145/3239558
- [41] Batya Friedman and David Hendry. 2012. The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, 1145–1148. doi:10.1145/2207676.2208562

- [42] Timnit Gebru and Émile P. Torres. 2024. The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday* 29, 4 (2024). doi:10.5210/fm.v29i4.13636
- [43] R. Stuart Geiger and David Ribes. 2011. Trace Ethnography: Following Coordination through Documentary Practices. In *2011 44th Hawaii International Conference on System Sciences*. IEEE, 1–10. doi:10.1109/HICSS.2011.455
- [44] Katja Grace, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, Jan Brauner, and Richard C. Korzekwa. 2024. Thousands of AI Authors on the Future of AI. arXiv:2401.02843 <https://arxiv.org/abs/2401.02843>
- [45] David H. Guston. 2014. Understanding ‘anticipatory governance’. *Social Studies of Science* 44, 2 (2014), 218–242. doi:10.1177/0306312713508669
- [46] Balint Gyevnár and Atoosa Kasirzadeh. 2025. AI safety for everyone. *Nature Machine Intelligence* 7 (2025), 531–542. doi:10.1038/s42256-025-01020-y
- [47] Christina Harrington and Tawanna R. Dillahunt. 2021. Eliciting Tech Futures Among Black Young Adults: A Case Study of Remote Speculative Co-Design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI ’21)*. ACM. doi:10.1145/3411764.3445723
- [48] Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing Community-Based Collaborative Design: Towards More Equitable Participatory Design Engagements. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 216:1–216:25. doi:10.1145/3359318
- [49] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. 2024. Collective Constitutional AI: Aligning a Language Model with Public Input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FACt ’24)*. ACM. doi:10.1145/3630106.3658979
- [50] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermy, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. 2024. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566 <https://arxiv.org/abs/2401.05566>
- [51] i3 Gen. 2026. Intelligence Rising: An Introduction. i3 Gen. March 2026; consultancy’s own account. Accessed August 9, 2026. <https://i3gen.co.uk/intelligence-rising-an-introduction/>
- [52] Terry Irwin. 2015. Transition Design: A Proposal for a New Area of Design Practice, Study, and Research. *Design and Culture* 7, 2 (2015), 229–246. doi:10.1080/17547075.2015.1051829
- [53] Sheila Jasanoff and Sang-Hyun Kim. 2009. Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea. *Minerva* 47, 2 (2009), 119–146. doi:10.1007/s11024-009-9124-4
- [54] Brian David Johnson, Natalie Vanatta, and Cyndi Coon. 2021. *Threatcasting*. Springer (Synthesis Lectures on Threatcasting). doi:10.1007/978-3-031-02575-4
- [55] Atoosa Kasirzadeh. 2025. Two types of AI existential risk: decisive and accumulative. *Philosophical Studies* 182 (2025), 1975–2003. doi:10.1007/s11098-025-02301-3
- [56] Gary Klein. 2007. Performing a Project Premortem. *Harvard Business Review* 85, 9 (2007), 18–19.
- [57] Bran Knowles, Oliver Bates, and Maria Håkansson. 2018. This Changes Sustainable HCI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI ’18)*. ACM, 1–12. doi:10.1145/3173574.3174045
- [58] Loren Kohnfelder and Praerit Garg. 1999. The Threats to Our Products. Microsoft Interface (internal publication). Accessed August 9, 2026. <https://shostack.org/files/microsoft/The-Threats-To-Our-Products.docx>
- [59] Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean. 2025. AI 2027. AI Futures Project. April 3, 2025. Accessed August 9, 2026. <https://ai-2027.com/>
- [60] Barbara Kordy, Ludovic Piètre-Cambacédès, and Patrick Schweitzer. 2014. DAG-based attack and defense modeling: Don’t miss the forest for the attack trees. *Computer Science Review* 13–14 (2014), 1–38. doi:10.1016/j.cosrev.2014.07.001
- [61] Sandjar Kozubaev, Chris Elsdén, Noura Howell, Marie Louise Juul Søndergaard, Nick Merrill, Britta Schulte, and Richmond Y. Wong. 2020. Expanding Modes of Reflection in Design Futuring. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI ’20)*. ACM. doi:10.1145/3313831.3376526
- [62] Robert J. Lempert, Steven W. Popper, and Steven C. Bankes. 2003. *Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis*. RAND Corporation, Santa Monica, CA. doi:10.7249/MR1626
- [63] Nancy G. Leveson. 2012. *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press, Cambridge, MA. doi:10.7551/mitpress/8179.001.0001
- [64] Seth Lifland. 2026. There Should Be More AI Safety Hubs. LessWrong. Accessed 2026-09-03. <https://www.lesswrong.com/posts/jXQ9hPL8RqkH5ybnt/there-should-be-more-ai-safety-hubs>
- [65] Brian Massumi. 2007. Potential Politics and the Primacy of Preemption. *Theory & Event* 10, 2 (2007). doi:10.1353/tae.2007.0066
- [66] John Mayne. 2012. Contribution analysis: Coming of age? *Evaluation* 18, 3 (2012), 270–280. doi:10.1177/1356389012451663
- [67] Donella Meadows. 1999. *Leverage Points: Places to Intervene in a System*. Technical Report. The Sustainability Institute, Hartland, VT. Accessed September 11, 2026. https://donellameadows.org/wp-content/userfiles/Leverage_Points.pdf
- [68] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. 2024. Frontier Models are Capable of In-context Scheming. arXiv:2412.04984 <https://arxiv.org/abs/2412.04984>

- [69] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. ACM, 735–746. doi:10.1145/3442188.3445935
- [70] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. 2021. Unpacking the Expressed Consequences of AI Research in Broader Impact Statements. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AI/ES '21)*. ACM, 795–806. doi:10.1145/3461702.3462608
- [71] Lisa P. Nathan, Predrag V. Klasnja, and Batya Friedman. 2007. Value scenarios: a technique for envisioning systemic effects of new technologies. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2585–2590. doi:10.1145/1240866.1241046
- [72] Richard Ngo, Lawrence Chan, and Sören Mindermann. 2024. The Alignment Problem from a Deep Learning Perspective. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*. <https://arxiv.org/abs/2209.00626>
- [73] Alfred Nordmann. 2007. If and Then: A Critique of Speculative NanoEthics. *NanoEthics* 1, 1 (2007), 31–46. doi:10.1007/s11569-007-0007-6
- [74] Donald A. Norman and Pieter Jan Stappers. 2015. DesignX: Complex Sociotechnical Systems. *She Ji: The Journal of Design, Economics, and Innovation* 1, 2 (2015), 83–106. doi:10.1016/j.sheji.2016.01.002
- [75] Toby Ord. 2020. *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury, London.
- [76] Ray Pawson and Nick Tilley. 1997. *Realistic Evaluation*. SAGE, London.
- [77] Luiza Prado de O. Martins. 2019. Designer/Shapeshifter: A Decolonizing Redirection for Speculative and Critical Design. In *Tricky Design: The Ethics of Things*. Bloomsbury. doi:10.5040/9781474277211.ch-007
- [78] Sebastian Prost, Elke Mattheiss, and Manfred Tscheligi. 2015. From Awareness to Empowerment: Using Design Fiction to Explore Paths towards a Sustainable Energy Future. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, 1649–1658. doi:10.1145/2675133.2675281
- [79] Jaco Quist and Philip Vergragt. 2006. Past and future of backcasting: The shift to stakeholder participation and a proposal for a methodological framework. *Futures* 38, 9 (2006), 1027–1045. doi:10.1016/j.futures.2006.02.010
- [80] John Robinson. 2003. Future subjunctive: backcasting as social learning. *Futures* 35, 8 (2003), 839–856. doi:10.1016/S0016-3287(03)00039-9
- [81] John Bridger Robinson. 1982. Energy backcasting: A proposed method of policy analysis. *Energy Policy* 10, 4 (1982), 337–344. doi:10.1016/0301-4215(82)90048-9
- [82] John B. Robinson. 1990. Futures under glass: A recipe for people who hate to predict. *Futures* 22, 8 (1990), 820–842. doi:10.1016/0016-3287(90)90018-D
- [83] Patricia J. Rogers. 2008. Using Programme Theory to Evaluate Complicated and Complex Aspects of Interventions. *Evaluation* 14, 1 (2008), 29–48. doi:10.1177/1356389007084674
- [84] Stuart Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
- [85] Camilo Sanchez, Sijia Wang, Kaisa Savolainen, Felix Anand Epp, and Antti Salovaara. 2025. Let's Talk Futures: A Literature Review of HCI's Future Orientation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM. doi:10.1145/3706598.3713759
- [86] Bruce Schneier. 1999. Attack Trees: Modeling security threats. *Dr. Dobbs's Journal* 24, 12 (1999), 21–29. Accessed September 11, 2026. https://www.schneier.com/academic/archives/1999/12/attack_trees.html
- [87] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AI/ES '23)*. ACM, 723–741. doi:10.1145/3600211.3604673
- [88] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, and Allan Dafoe. 2023. Model evaluation for extreme risks. arXiv:2305.15324 <https://arxiv.org/abs/2305.15324>
- [89] Adam Shostack. 2014. *Threat Modeling: Designing for Security*. Wiley, Indianapolis, IN.
- [90] Robert Soden, Pradnya Pathak, and Olivia Doggett. 2021. What We Speculate About When We Speculate About Sustainable HCI. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS '21)*. ACM. doi:10.1145/3460112.3471956
- [91] Bruce Sterling. 2005. *Shaping Things*. MIT Press, Cambridge, MA.
- [92] Jack Stilgoe, Richard Owen, and Phil Macnaghten. 2013. Developing a framework for responsible innovation. *Research Policy* 42, 9 (2013), 1568–1580. doi:10.1016/j.respol.2013.05.008
- [93] Andy Stirling. 2008. “Opening Up” and “Closing Down”: Power, Participation, and Pluralism in the Social Appraisal of Technology. *Science, Technology, & Human Values* 33, 2 (2008), 262–294. doi:10.1177/0162243907311265
- [94] Cass R. Sunstein. 2005. *Laws of Fear: Beyond the Precautionary Principle*. Cambridge University Press. doi:10.1017/CBO9780511790850
- [95] Cameron Tonkinwise. 2014. How We Intend to Future: Review of Anthony Dunne and Fiona Raby, Speculative Everything: Design, Fiction, and Social Dreaming. *Design Philosophy Papers* 12, 2 (2014), 169–187. doi:10.2752/144871314x14159818597676
- [96] Cameron Tonkinwise. 2015. Design for Transitions — from and to what? *Design Philosophy Papers* 13, 1 (2015), 85–92. doi:10.1080/14487136.2015.1085686
- [97] Isabel Vogel. 2012. *Review of the use of 'Theory of Change' in international development*. Technical Report. UK Department for International Development, London. Accessed September 11, 2026. https://assets.publishing.service.gov.uk/media/57a08a5ded915d3cfd00071a/DFID_ToC_Review_VogelV7.pdf
- [98] Ron Wakkary, William Odom, Sabrina Hauser, Garnet Hertz, and Henry Lin. 2015. Material Speculation: Actual Artifacts for Critical Inquiry. In *Proceedings of the Fifth Decennial Aarhus Conference on Critical Alternatives (Aarhus Series on Human Centered Computing, Vol. 1)*. doi:10.7146/

- aahcc.v1i1.21299
- [99] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, 214–229. doi:10.1145/3531146.3533088
- [100] Carol H. Weiss. 1995. Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts*, James P. Connell, Anne C. Kubisch, Lisbeth B. Schorr, and Carol H. Weiss (Eds.). The Aspen Institute, Washington, DC, 65–92.
- [101] Richmond Y. Wong, Vera Khovanskaya, Sarah E. Fox, Nick Merrill, and Phoebe Sengers. 2020. Infrastructural Speculations: Tactics for Designing and Interrogating Lifeworlds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM. doi:10.1145/3313831.3376515
- [102] Richmond Y. Wong and Tonya Nguyen. 2021. Timelines: A World-Building Activity for Values Advocacy. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM. doi:10.1145/3411764.3445447

A Archived Corpus Documents

The archived corpus documents cited throughout the paper with A-numbered labels, and grouped by stratum in Table 1, are listed here with each document’s title, source, date where recoverable, URL, and archive date.

Archived Corpus Documents

- [A1] 80,000 Hours. 2026. Why AI risks are the world’s most pressing problems. 80,000 Hours. Archived 2026-08-09. <https://80000hours.org/problem-profiles/artificial-intelligence>
- [A2] AI Futures Project. 2025. About - AI 2027. AI Futures Project. Archived 2026-08-09. <https://ai-2027.com/about>
- [A3] AI Impacts. 2021. Vignettes workshop. AI Impacts. Archived 2026-08-09. <https://aiimpacts.org/vignettes-workshop/>
- [A4] AI Safety Camp. 2026. AI Safety Camp. AI Safety Camp. Archived 2026-08-09. <https://aisafety.camp/>
- [A5] airiskfund.com. [n. d.]. Grantmaking Process - AI Risk Mitigation Fund. airiskfund.com. Archived 2026-08-09. <https://www.airiskfund.com/grantmaking-process>
- [A6] aivillage.org. 2023. Generative Red Team recap (AI Village, DEF CON). aivillage.org. Archived 2026-08-09. <https://aivillage.org/blog/generative-recap/>
- [A7] Will Aldred. 2022. CERI Research Symposium Presentations (incl. Youtube links). EA Forum. Archived 2026-08-09. <https://forum.effectivealtruism.org/posts/ivjgAZRM7No4fnDgo/>
- [A8] Aidan Alexander and Ambitious Impact. 2023. Nailing the basics – Theories of change. EA Forum. Archived 2026-08-09. <https://forum.effectivealtruism.org/posts/9t7St3pfEEiDsQ2Tr/nailing-the-basics-theories-of-change>
- [A9] Anthropic. 2023. Anthropic’s Responsible Scaling Policy. Anthropic. Archived 2026-08-09. <https://www.anthropic.com/news/anthropics-responsible-scaling-policy>
- [A10] Anthropic. 2023. Core Views on AI Safety: When, Why, What, and How. Anthropic. March 8, 2023. Accessed August 9, 2026. <https://www.anthropic.com/news/core-views-on-ai-safety>
- [A11] Anthropic. 2023. Frontier Threats Red Teaming for AI Safety. Anthropic. Archived 2026-08-09. <https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety>
- [A12] Anthropic. 2024. Challenges in red teaming AI systems. Anthropic. Archived 2026-08-09. <https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems>
- [A13] Anthropic. 2024. External reviews of “Alignment faking in large language models”. Anthropic. December 18, 2024. Accessed August 9, 2026. <https://assets.anthropic.com/m/24c8d0a3a7d0a1f1/original/Alignment-Faking-in-Large-Language-Models-reviews.pdf>
- [A14] Apart Research. [n. d.]. Apart Research. Apart Research. Archived 2026-08-09. <https://www.apartresearch.com/>
- [A15] Apart Research. 2026. Apart Research - Sprints. Apart Research. Archived 2026-08-09. <https://apartresearch.com/sprints>
- [A16] Apollo Research. 2024. The first year of Apollo Research. Apollo Research. Archived 2026-08-09. <https://www.apolloresearch.ai/blog/the-first-year-of-apollo-research>
- [A17] Apollo Research. 2025. Research Note: Our scheming precursor evals had limited predictive power for our in-context scheming evals. Apollo Research. July 3, 2025. Accessed August 9, 2026. <https://www.apolloresearch.ai/blog/research-note-our-scheming-precursor-evals-had-limited-predictive-power-for-our-in-context-scheming-evals/>
- [A18] arena.education. [n. d.]. ARENA - Our Mission. arena.education. Archived 2026-08-09. <https://www.arena.education/our-mission>
- [A19] Shahar Avin, Ross Gruetzmacher, and James Fox. 2020. Exploring AI Futures Through Role Play. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES ’20)*. ACM. doi:10.1145/3375627.3375817
- [A20] Beth Barnes. 2020. Debate update: Obfuscated arguments problem. AI Alignment Forum. December 23, 2020. Accessed August 9, 2026. <https://www.alignmentforum.org/posts/PJLABqQ962hZEqhdB/debate-update-obfuscated-arguments-problem>
- [A21] Ryan Blough. 2022. Wargaming AGI Development. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/ouFnZoYaKqicC6jH8/wargaming-agi-development>
- [A22] BlueDot Impact. [n. d.]. BlueDot Impact - AI Alignment Course Unit 2. BlueDot Impact. Archived 2026-08-09. <https://bluedot.org/courses/alignment/2>
- [A23] BlueDot Impact. [n. d.]. BlueDot Impact - AI Alignment Course Unit 8. BlueDot Impact. Archived 2026-08-09. <https://bluedot.org/courses/alignment/8>
- [A24] BlueDot Impact. [n. d.]. BlueDot Impact - AI Governance Course Unit 4. BlueDot Impact. Archived 2026-08-09. <https://bluedot.org/courses/ai-governance/4/1>
- [A25] BlueDot Impact. 2025. Field strategy playbook. BlueDot Impact. Archived 2026-08-09. <https://blog.bluedot.org/p/field-strategy-playbook>
- [A26] cdn.prod.website-files.com. 2024. CAIF Strategy 2024 (PDF). cdn.prod.website-files.com. Archived 2026-08-09. https://cdn.prod.website-files.com/61fe446e584616a902bcb4f0/66bda79f8a3da9e7ff75ed56_CAIF-Strategy-2024.pdf
- [A27] Paul Christiano. 2019. What failure looks like. Alignment Forum. Archived 2026-08-09. <https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like>

- [A28] Paul Christiano, Mark Xu, and Ajeya Cotra. 2021. ARC's first technical report: Eliciting Latent Knowledge. Alignment Research Center. December 14, 2021. Accessed August 9, 2026. <https://www.alignment.org/blog/arcs-first-technical-report-eliciting-latent-knowledge/>
- [A29] coefficientgiving.org. 2025. Request for Proposals: Technical AI Safety Research. coefficientgiving.org. Archived 2026-08-09. <https://coefficientgiving.org/funds/navigating-transformative-ai/request-for-proposals-technical-ai-safety-research/>
- [A30] cold-takes.com. 2022. AI Safety Seems Hard to Measure. cold-takes.com. Archived 2026-08-09. <https://www.cold-takes.com/ai-safety-seems-hard-to-measure/>
- [A31] Constellation. [n. d.]. Constellation - Astra Fellowship. Constellation. Archived 2026-08-09. <https://www.constellation.org/programs/astra>
- [A32] Cooperative AI Foundation. [n. d.]. Cooperative AI – Foundation. Cooperative AI Foundation. Archived 2026-08-09. <https://www.cooperativeai.com/foundation>
- [A33] Cooperative AI Foundation. [n. d.]. Curriculum. Cooperative AI Foundation. Archived 2026-08-09. <https://www.cooperativeai.com/curriculum>
- [A34] Cooperative AI Foundation. 2024. Cooperative AI and Governance. Cooperative AI Foundation. Archived 2026-08-09. <https://www.cooperativeai.com/post/cooperative-ai-and-governance>
- [A35] Cooperative AI Foundation. 2024. The Cooperative AI Foundation's 2024 Strategy. Cooperative AI Foundation. Archived 2026-08-09. <https://www.cooperativeai.com/post/the-cooperative-ai-foundations-2024-strategy>
- [A36] Cooperative AI Foundation. 2025. Cooperative AI Summer School 2025. Cooperative AI Foundation. Archived 2026-08-09. <https://www.cooperativeai.com/summer-school/summer-school-2025>
- [A37] Cooperative AI Foundation. 2025. Cooperative AI Summer School 2025: Highlights and Takeaways. Cooperative AI Foundation. Archived 2026-08-09. <https://www.cooperativeai.com/post/cooperative-ai-summer-school-2025-recap>
- [A38] Cooperative AI Foundation. 2025. Introducing CAIF's Impact Report. Cooperative AI Foundation. Archived 2026-08-09. <https://www.cooperativeai.com/post/introducing-caifs-impact-report>
- [A39] Cooperative AI Foundation. 2025. Updates to the CAIF Grant Program in 2025. Cooperative AI Foundation. Archived 2026-08-09. <https://www.cooperativeai.com/post/updates-to-the-caif-grant-program-in-2025>
- [A40] Cooperative AI Foundation. 2026. Scaling AI Safety for a Multi-Agent World: \$10m Fund for Multi-Agent Safety. Cooperative AI Foundation. Accessed August 9, 2026. <https://www.cooperativeai.com/calls-for-proposals/scaling-ai-safety-for-a-multi-agent-world>
- [A41] Ajeya Cotra. 2022. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. LessWrong. July 18, 2022. Accessed August 9, 2026. <https://www.lesswrong.com/posts/pRkFkzwKZ2fa3R6H/without-specific-countermeasures-the-easiest-path-to>
- [A42] coveragecritic.com. 2019. The Optimizer's Curse & Wrong-Way Reductions. coveragecritic.com. Archived 2026-08-09. <https://coveragecritic.com/2019/04/03/the-optimizers-curse-wrong-way-reductions/>
- [A43] cs.stanford.edu. 2018. URL Source: <https://cs.stanford.edu/~jsteinhardt/ResearchasaStochasticDecisionProcess.html>. cs.stanford.edu. Archived 2026-08-09. <https://cs.stanford.edu/~jsteinhardt/ResearchasaStochasticDecisionProcess.html>
- [A44] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tatum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. 2020. Open Problems in Cooperative AI. arXiv:2012.08630 <https://arxiv.org/abs/2012.08630>
- [A45] Daniel Filan and Ryan Kidd. 2024. MATS AI Safety Strategy Curriculum v2. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/rhEXTkDmssrHBnrS/mats-ai-safety-strategy-curriculum-v2>
- [A46] deepmindsafetyresearch.medium.com. 2025. Introducing our short course on AGI safety. deepmindsafetyresearch.medium.com. Archived 2026-08-09. <https://deepmindsafetyresearch.medium.com/introducing-our-short-course-on-agi-safety-1072adb7912c>
- [A47] Elizabeth. 2019. What Comes After Epistemic Spot Checks? LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/XkXfXtJ3GNqkAfpJt/what-comes-after-epistemic-spot-checks>
- [A48] Elizabeth. 2022. Epistemic Legibility. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/jbE85wCkRr9z7tqmD/epistemic-legibility>
- [A49] epoch.ai. 2025. Announcing GATE. epoch.ai. Archived 2026-08-09. <https://epoch.ai/blog/announcing-gate>
- [A50] ERA Fellowship. [n. d.]. About the Fellowship – ERA Fellowship. ERA Fellowship. Archived 2026-08-09. <https://erafellowship.org/fellowship>
- [A51] ERA Fellowship. [n. d.]. Careers – ERA Fellowship. ERA Fellowship. Archived 2026-08-09. <https://erafellowship.org/careers>
- [A52] ERA Fellowship. 2025. About – ERA Fellowship. ERA Fellowship. Archived 2026-08-09. <https://erafellowship.org/about>
- [A53] ERA Fellowship. 2025. AI Risks – ERA Fellowship. ERA Fellowship. Archived 2026-08-09. <https://erafellowship.org/ai-risks>
- [A54] ERA Fellowship. 2025. Alumni – ERA Fellowship. ERA Fellowship. Archived 2026-08-09. <https://erafellowship.org/alumni>
- [A55] ERA Fellowship. 2025. Past Research – ERA Fellowship. ERA Fellowship. Archived 2026-08-09. <https://erafellowship.org/research-projects>
- [A56] ERA Fellowship. 2026. ERA Fellowship: Frequently Asked Questions. ERA (Existential Risk Alliance). Accessed August 9, 2026. <https://erafellowship.org/faq>
- [A57] erafellowship. 2024. Announcing the Cambridge ERA:AI Fellowship 2024. EA Forum. Archived 2026-08-09. <https://forum.effectivealtruism.org/posts/6EFpaNZhL8T8DvSgT/>
- [A58] Ronny Fernandez and Ryan Kidd. 2024. MATS AI Safety Strategy Curriculum. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/JsJjuikJsidkyfhyr/mats-ai-safety-strategy-curriculum>
- [A59] Gideon Futerman and Nandini Shiralkar. 2023. Culture and Programming Retrospective: ERA Fellowship 2023. EA Forum. Archived 2026-08-09. <https://forum.effectivealtruism.org/posts/uSKQLstvoxHKtWkrg/culture-and-programming-retrospective-era-fellowship-2023>
- [A60] Future of Life Institute. 2022. FLI Worldbuilding Contest. Future of Life Institute. Accessed August 9, 2026. <https://worldbuild.ai/>

- [A61] GovAI. 2018. AI Governance: A Research Agenda (Allan Dafoe). GovAI. Archived 2026-08-09. <https://cdn.governance.ai/GovAI-Research-Agenda.pdf>
- [A62] GovAI. 2024. GovAI Annual Report 2023. GovAI. Archived 2026-08-09. https://cdn.governance.ai/GovAI_Annual_Report_2023.pdf
- [A63] Government Office for Science. 2024. *AI 2030 Scenarios: Helping Policy Makers Plan for the Future of AI*. Technical Report. UK Government Office for Science. January 2024. Archived in research/raw/ai2030.md. Accessed August 29, 2026. <https://www.gov.uk/government/publications/ai-scenarios-2030-helping-policy-makers-plan-for-the-future-of-ai>
- [A64] Ryan Greenblatt and Buck Shlegeris. 2024. The case for ensuring that powerful AIs are controlled. AI Alignment Forum. January 24, 2024. Accessed August 9, 2026. <https://www.alignmentforum.org/posts/kcKrE9mzEHrdqtDpE/the-case-for-ensuring-that-powerful-ais-are-controlled>
- [A65] Ross Gruetzemacher, Shahar Avin, James Fox, and Alexander K. Saeri. 2025. Strategic Insights from Simulation Gaming of AI Race Dynamics. *Futures* (2025). Preprint: arXiv:2410.03092. doi:10.1016/j.futures.2025.103563
- [A66] Oliver Habryka, Ben Pace, Raymond Arnold, and Jim Babcock. 2018. Introducing the AI Alignment Forum (FAQ). LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/FoiIRDC3EhJHx7ayY/introducing-the-ai-alignment-forum-faq>
- [A67] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčík, The Anh Han, Edward Hughes, Vojtěch Kovářik, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. 2025. Multi-Agent Risks from Advanced AI. Cooperative AI Foundation Technical Report #1. arXiv:2502.14143 <https://arxiv.org/abs/2502.14143>
- [A68] Harrison. 2025. Apply to the Cambridge ERA:AI Fellowship 2025. EA Forum. Archived 2026-08-09. <https://forum.effectivealtruism.org/posts/a6xhZDwyQkp5DFPJA/>
- [A69] Dan Hendrycks, Eric Schmidt, and Alexandr Wang. 2025. Superintelligence Strategy: Expert Version (arXiv abstract). arXiv. Archived 2026-08-09. <https://arxiv.org/abs/2503.05628>
- [A70] Benjamin Hilton. 2022. Preventing an AI-related catastrophe: Problem profile. 80,000 Hours; stable copy on the Effective Altruism Forum. August 29, 2022; last updated March 2025. Accessed August 9, 2026. <https://forum.effectivealtruism.org/posts/zpReK9a8gkpGNyMbT/preventing-an-ai-related-catastrophe-problem-profile>
- [A71] Jacob Hilton. 2024. A bird's eye view of ARC's research. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/ztokaf9harkTmRcn4/a-bird-s-eye-view-of-arc-s-research>
- [A72] Marius Hobbhahn. 2022. Lessons learned from talking to >100 academics about AI safety. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/SqjQFhn5KTarfW8v7/lessons-learned-from-talking-to-greater-than-100-academics>
- [A73] Marius Hobbhahn, beren, Lee Sharkey, Lucius Bushnaq, Dan Braun, Mikita Balesni, and Jérémy Scheurer. 2023. Announcing Apollo Research. LessWrong. May 30, 2023. Accessed August 9, 2026. <https://www.lesswrong.com/posts/FG6icLPKizEaWHex5/announcing-apollo-research>
- [A74] horizonpublicservice.org. 2026. Horizon Institute for Public Service - Become a Fellow. horizonpublicservice.org. Archived 2026-08-09. <https://horizonpublicservice.org/programs/become-a-fellow/>
- [A75] Evan Hubinger. 2022. How likely is deceptive alignment? Alignment Forum. Archived 2026-08-09. <https://www.alignmentforum.org/posts/A9NxPTwbw6r6Awuwt/how-likely-is-deceptive-alignment>
- [A76] Evan Hubinger. 2022. How likely is deceptive alignment? LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/A9NxPTwbw6r6Awuwt/how-likely-is-deceptive-alignment>
- [A77] Evan Hubinger. 2024. Introducing Alignment Stress-Testing at Anthropic. AI Alignment Forum. January 12, 2024. Accessed August 9, 2026. <https://www.alignmentforum.org/posts/EPDSdXr8YbsDkgsDG/introducing-alignment-stress-testing>
- [A78] Intelligence Rising. [n. d.]. About - Intelligence Rising. Intelligence Rising. Archived 2026-08-09. <https://www.intelligencerising.org/about>
- [A79] intelligence.org. 2024. MIRI 2024 Mission and Strategy Update. intelligence.org. Archived 2026-08-09. <https://intelligence.org/2024/01/04/miri-2024-mission-and-strategy-update/>
- [A80] intelligence.org. 2024. MIRI's 2024 End-of-Year Update. intelligence.org. Archived 2026-08-09. <https://intelligence.org/2024/12/02/miris-2024-end-of-year-update/>
- [A81] Jamie B. 2022. AI Safety Overview: CERI Summer Research Fellowship. Effective Altruism Forum. March 24, 2022. Accessed August 9, 2026. <https://forum.effectivealtruism.org/posts/T3piiDHvaGuzE7KKF/ai-safety-overview-ceri-summer-research-fellowship>
- [A82] John, Paul, Vanessa, and Evan. 2022. What I Learned Running Refine. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/3zZjF3YKJ257x79mu/what-i-learned-running-refine>
- [A83] Jonas V. 2025. Help the AI 2027 team make an online AGI wargame. LessWrong. June 27, 2025. Accessed August 9, 2026. <https://www.lesswrong.com/posts/TjT3RdAfmrLqgb68K/help-the-ai-2027-team-make-an-online-agi-wargame>
- [A84] Nikola Jurkovic. 2023. Microdooms averted by working on AI Safety. LessWrong. September 17, 2023. Accessed August 9, 2026. <https://www.lesswrong.com/posts/mTtXjKN3Ew8CAEHGr/microdooms-averted-by-working-on-ai-safety>
- [A85] Ezra Karger, Josh Rosenberg, Zachary Jacobs, Molly Hickman, Rose Hadshar, Kayla Gamin, Taylor Smith, Bridget Williams, Tegan McCaslin, Stephen Thomas, and Philip E. Tetlock. 2023. *Forecasting Existential Risks: Evidence from a Long-Run Forecasting Tournament*. Technical Report Working Paper #1. Forecasting Research Institute. Accessed September 11, 2026. <https://forecastingresearch.org/xpt>

- [A86] Holden Karnofsky. 2011. Why we can't take expected value estimates literally (even when they're unbiased). The GiveWell Blog. August 18, 2011. Accessed August 9, 2026. <https://blog.givewell.org/2011/08/18/why-we-cant-take-expected-value-estimates-literally-even-when-theyre-unbiased/>
- [A87] Corin Katzke. 2024. Scenario planning for AI x-risk. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/foqm3fKfDuRhtwjuH/scenario-planning-for-ai-x-risk>
- [A88] Corin Katzke and Justin Bullock. 2024. AI governance needs a theory of victory. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/cuAFelmqyEm3X6s5/ai-governance-needs-a-theory-of-victory-1>
- [A89] Zachary Kenton, Rohin Shah, David Lindner, Vikrant Varma, Victoria Krakovna, Mary Phuong, Ramana Kumar, and Elliot Catt. 2022. Clarifying AI X-risk. AI Alignment Forum. November 1, 2022. Accessed August 9, 2026. <https://www.alignmentforum.org/posts/GctJD5oCDRxCspEaZ/clarifying-ai-x-risk>
- [A90] Zachary Kenton, Rohin Shah, David Lindner, Vikrant Varma, Victoria Krakovna, Mary Phuong, Ramana Kumar, and Elliot Catt. 2022. Threat Model Literature Review. AI Alignment Forum. November 1, 2022. Accessed August 9, 2026. <https://www.alignmentforum.org/posts/wnnkD6P2k2TfHnNmt/threat-model-literature-review>
- [A91] Ryan Kidd. 2024. Talk: AI safety fieldbuilding at MATS. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/WGNYAdBrsNjujJaB8/talk-ai-safety-fieldbuilding-at-mats>
- [A92] Ryan Kidd. 2024. Why I funded PIBBSS. LessWrong. Accessed August 9, 2026. <https://www.lesswrong.com/posts/Yjiw5rnu4mJsYN8Xc/why-i-funded-pibbss-1>
- [A93] Daniel Kokotajlo. 2021. What 2026 Looks Like. LessWrong. August 6, 2021. Written for the AI Impacts Vignettes Workshop. Archived in research/raw/. Accessed August 9, 2026. <https://www.lesswrong.com/posts/6Xgy6CAf2jqHhynHL/what-2026-looks-like>
- [A94] Victoria Krakovna. 2023. Retrospective on my posts on AI threat models. Victoria Krakovna's blog. December 20, 2023. Accessed August 9, 2026. <https://vkrakovna.wordpress.com/2023/12/20/retrospective-on-ai-threat-models/>
- [A95] Victoria Krakovna and Rohin Shah. 2025. A short course on AGI safety from the GDM Alignment team. LessWrong. February 14, 2025. Accessed August 9, 2026. <https://www.lesswrong.com/posts/TJrCumJxhzTmNBsRz/a-short-course-on-agi-safety-from-the-gdm-alignment-team>
- [A96] Victoria Krakovna, Vikrant Varma, Ramana Kumar, and Mary Phuong. 2022. Refining the Sharp Left Turn threat model, part 1: claims and mechanisms. AI Alignment Forum. August 12, 2022. Accessed August 9, 2026. <https://www.alignmentforum.org/posts/usKXS5jGDzjwqv3FJ/refining-the-sharp-left-turn-threat-model-part-1-claims-and-mechanisms>
- [A97] Robert Kralisch, Linda Linsefors, and Remmelt. 2024. AI Safety Camp 10. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/57wx7B3GQavvKkPne/ai-safety-camp-10>
- [A98] L Rudolf L. 2022. Assessing SERI/CHERI/CERI summer program impact by surveying fellows. Effective Altruism Forum. September 26, 2022. Accessed August 9, 2026. <https://forum.effectivealtruism.org/posts/yeDRczoJuJZMbJvrq/assessing-seri-cheri-ceri-summer-program-impact-by-surveying>
- [A99] Lark. 2021. Measure is unceasing. nunosempere.com. Archived 2026-08-09. <https://nunosempere.com/blog/2021/06/24/shallow-evaluations-of-longtermist-organizations/>
- [A100] learn.arena.education. [n. d.]. ARENA Chapter 3: LLM Evaluations (full course text). learn.arena.education. Archived 2026-08-09. https://learn.arena.education/chapter3_llm_evals/
- [A101] Gavin Leech. 2022. What I learned from the criticism contest. Effective Altruism Forum. October 1, 2022. Accessed August 9, 2026. <https://forum.effectivealtruism.org/posts/xcsyfp6zb3JpCjKGq/what-i-learned-from-the-criticism-contest>
- [A102] Linch. 2022. .01% Fund — Ideation and Proposal. Effective Altruism Forum. March 1, 2022. Accessed August 9, 2026. <https://forum.effectivealtruism.org/posts/pGaesfn4R9xWK4Rmk/01-fund-ideation-and-proposal>
- [A103] Linch, Daniel Eth, and Caleb Parikh. 2023. What Does a Marginal Grant at LTFF Look Like? Funding Priorities and Grantmaking Thresholds at the Long-Term Future Fund. Effective Altruism Forum. August 10, 2023. Accessed August 9, 2026. <https://forum.effectivealtruism.org/posts/7RrjXQhGgAjiDLWYR/what-does-a-marginal-grant-at-ltff-look-like-funding>
- [A104] Linda Linsefors. 2025. Theory of Change for AI Safety Camp. LessWrong. January 22, 2025. Accessed August 9, 2026. <https://www.lesswrong.com/posts/Kd2cbLXQxCCRQDcH/theory-of-change-for-ai-safety-camp>
- [A105] David Manheim and Aryeh Englander. 2021. Modelling Transformative AI Risks (MTAIR) Project: Introduction. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/qnA6paRwMky3Q6ktk/>
- [A106] MATS. [n. d.]. MATS About us. MATS. Archived 2026-08-09. <https://www.matsprogram.org/about>
- [A107] MATS. 2026. MATS Research. MATS. Archived 2026-08-09. <https://www.matsprogram.org/>
- [A108] Stephen McAleese, Stag, and Violet Hour. 2025. Shallow review of technical AI safety, 2025. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/Wti4Wr7Cf5ma3FGWa/shallow-review-of-technical-ai-safety-2025-2>
- [A109] mesaoptimizer. 2023. [outdated] My current theory of change to mitigate existential risk by misaligned ASI. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/PnqtmT8fkq5kfPTww/outdated-my-current-theory-of-change-to-mitigate-existential>
- [A110] METR. 2024. Guidelines for capability elicitation. METR. March 15, 2024. Accessed August 9, 2026. <https://metr.org/blog/2024-03-15-guidelines-for-capability-elicitation/>
- [A111] METR. 2024. METR - An update on our general capability evaluations. METR. Archived 2026-08-09. <https://metr.org/blog/2024-08-06-update-on-evaluations/>

Manuscript submitted to ACM

- [A112] METR. 2025. METR - About. METR. Archived 2026-08-09. <https://metr.org/about>
- [A113] Jim Mitre and Joel B. Predd. 2025. *Artificial General Intelligence's Five Hard National Security Problems*. Technical Report PE-A3691-4. RAND Corporation. Accessed September 11, 2026. <https://www.rand.org/pubs/perspectives/PEA3691-4.html>
- [A114] ML4Good. 2026. ML4Good Technical Programme. ML4Good. Accessed August 9, 2026. <https://www.ml4good.org/technical-track/>
- [A115] Ian David Moss. 2019. A Short Introduction to Theory of Change. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/xQk6feH9pmy6mKt3x/a-short-introduction-to-theory-of-change>
- [A116] Zvi Mowshowitz. 2021. Zvi's Thoughts on the Survival and Flourishing Fund (SFF). Don't Worry About the Vase (blog). December 14, 2021. Accessed August 9, 2026. <https://thezvi.wordpress.com/2021/12/14/zvis-thoughts-on-the-survival-and-flourishing-fund-sff/>
- [A117] MvK and nandini. 2023. ERA Fellowship Alumni Stories. EA Forum. Archived 2026-08-09. <https://forum.effectivealtruism.org/posts/BBbk8yGynurYrcin/>
- [A118] Neel Nanda. 2021. My Overview of the AI Alignment Landscape: A Bird's Eye View. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/SQ9cZtfrzDJmw9A2m/my-overview-of-the-ai-alignment-landscape-a-bird-s-eye-view>
- [A119] Neel Nanda. 2021. My Overview of the AI Alignment Landscape: Threat Models. LessWrong. December 25, 2021. Accessed August 9, 2026. <https://www.lesswrong.com/posts/3DFBbPFZyscrAitKS/my-overview-of-the-ai-alignment-landscape-threat-models>
- [A120] Neel Nanda. 2022. Concrete Advice for Forming Inside Views on AI Safety. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/LbSWMfbAHQnFpApQ7/concrete-advice-for-forming-inside-views-on-ai-safety>
- [A121] Neel Nanda. 2022. A Longlist of Theories of Impact for Interpretability. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/uK6sQCNMw8WKzJeCQ/a-longlist-of-theories-of-impact-for-interpretability>
- [A122] nandini. 2022. Announcing ERA: a spin-off from CERL. EA Forum. Archived 2026-08-09. <https://forum.effectivealtruism.org/posts/uZY7fqXrP6P4vkK7g/announcing-era-a-spin-off-from-cerl>
- [A123] nandini. 2023. Introducing the Insights of an ERA Forum Sequence. EA Forum. Archived 2026-08-09. <https://forum.effectivealtruism.org/posts/waS5NaJqdYugW8am/>
- [A124] Dušan D. Nešić and Nora Ammann. 2024. Retrospective: PIBBSS Fellowship 2023. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/DD5EBtfJiisRkuPWF/>
- [A125] Dušan D. Nešić, clem_acs, and Lucas Teixeira. 2024. Retrospective: PIBBSS Fellowship 2024. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/sHpJascyEuEibzFyg/>
- [A126] Nuno Sempere. 2019. Shapley values: Better than counterfactuals. EA Forum. Archived 2026-08-09. <https://forum.effectivealtruism.org/posts/XHZ9ji7QBtAJZ6byW/shapley-values-better-than-counterfactuals>
- [A127] Open Philanthropy. 2024. Request for Proposals: AI Governance. Coefficient Giving (formerly Open Philanthropy). Updated October 16, 2024; closed January 25, 2026. Accessed August 9, 2026. <https://coefficientgiving.org/funds/navigating-transformative-ai/request-for-proposals-ai-governance/>
- [A128] OscarD. 2023. Hiring Retrospective: ERA Fellowship 2023. EA Forum. Archived 2026-08-09. <https://forum.effectivealtruism.org/posts/5DR89AbnjGjEcYvu8/>
- [A129] Joel B. Predd, Matt Chesson, and Gregory Smith. 2025. *Infinite Potential — Insights from the Two Moonshots Scenario: After-Action Report from a Sequence of Day After Artificial General Intelligence Exercises*. Technical Report RR-A4230-1. RAND Corporation. Accessed September 11, 2026. https://www.rand.org/pubs/research_reports/RRA4230-1.html
- [A130] princint.ai. 2025. PIBBSS Fellowship. princint.ai. Archived 2026-08-09. <https://princint.ai/programs/fellowship/>
- [A131] Raemon. 2016. Epistemic Effort. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/oDy27zRf8uAbJR6M/epistemic-effort>
- [A132] Raemon. 2018. Musings on LessWrong Peer Review. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/rCzXcyM9Gt2pxmhCE/musings-on-lesswrong-peer-review>
- [A133] Raemon. 2019. The LessWrong 2018 Review. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/qXwmMkEBLL59NkvYR/the-lesswrong-2018-review>
- [A134] Tilman Räuher and Gideon Futerman. 2023. Best Practices for early-career Research Management. Effective Altruism Forum. September 15, 2023. Accessed August 9, 2026. <https://forum.effectivealtruism.org/posts/BvTzqdMtcgLnDBzJb/best-practices-for-early-career-research-management>
- [A135] Remmelt and Linda Linsefors. 2024. This might be the last AI Safety Camp. LessWrong. January 2024. Accessed August 9, 2026. <https://www.lesswrong.com/posts/EAZjXKNN2vgoJGF9Y/this-might-be-the-last-ai-safety-camp>
- [A136] Remmelt, Linda Linsefors, and Robert Kralisch. 2024. Funding Case: AI Safety Camp 11. LessWrong. December 23, 2024. Accessed August 9, 2026. <https://www.lesswrong.com/posts/h6ncaN7GsDrfj49La/funding-case-ai-safety-camp-11>
- [A137] Brandon Riggs and Erin Robertson. 2026. LASR Labs Summer 2026 applications are open! LessWrong. March 3, 2026. Accessed August 9, 2026. <https://www.lesswrong.com/posts/dzSAsL4fPuBvCCarT/lasr-labs-summer-2026-applications-are-open>
- [A138] RobertM. 2025. The 2024 LessWrong Review. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/ZpRzTr5QBT6C3Faor/the-2024-lesswrong-review>
- [A139] Josh Rosenberg, Ezra Karger, Avital Morris, Molly Hickman, Rose Hadshar, Zachary Jacobs, and Philip E. Tetlock. 2024. Results from an Adversarial Collaboration on AI Risk (FRI). Effective Altruism Forum. March 11, 2024. Accessed August 9, 2026. <https://forum.effectivealtruism.org/posts/orhjaZ3AJMHZdZckZ/results-from-an-adversarial-collaboration-on-ai-risk-fri>
- [A140] Ruby. 2019. LessWrong FAQ. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/2rWkKwUprgTmPLRbp/lesswrong-faq>

- [A141] Duncan Sabien. 2017. Double Crux — A Strategy for Mutual Understanding. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/exa5kmvopeRyfJgCy/double-crux-a-strategy-for-mutual-understanding>
- [A142] samotsvety.org. 2022. Samotsvety's AI risk forecasts. samotsvety.org. Archived 2026-08-09. <https://samotsvety.org/blog/2022/09/09/samotsvety-s-ai-risk-forecasts/>
- [A143] securityandtechnology.org. 2023. AI Loss of Control Risk: Indications and Warning. securityandtechnology.org. Archived 2026-08-09. <https://securityandtechnology.org/virtual-library/report/ai-loss-of-control-risk-indications-warning/>
- [A144] Charbel-Raphaël Segerie. 2023. Against Almost Every Theory of Impact of Interpretability. LessWrong. August 17, 2023; see also the discussion thread, including R. Ngo's reply. Accessed August 9, 2026. <https://www.lesswrong.com/posts/LNA8mubrByG7SFacm/against-almost-every-theory-of-impact-of-interpretability-1>
- [A145] Nuño Sempere. 2022. Five steps for quantifying speculative interventions. Effective Altruism Forum. February 18, 2022. Accessed August 9, 2026. <https://forum.effectivealtruism.org/posts/3hH9NRqzGam65mgPG/five-steps-for-quantifying-speculative-interventions>
- [A146] Nuño Sempere. 2026. Thresholds for funding existential risk interventions. Effective Altruism Forum. July 11, 2026. Accessed August 9, 2026. <https://forum.effectivealtruism.org/posts/aBAkov3iJNWDWu8qS/thresholds-for-funding-existential-risk-interventions>
- [A147] Afek Shamir, Henri van Soest, Stephen Clare, and Sana Zakaria. 2026. *Insights from table-top exercises in Europe on AI safety and cyber misuse*. Technical Report RR-A5082-1. RAND Corporation. Accessed September 11, 2026. https://www.rand.org/pubs/research_reports/RR-A5082-1.html
- [A148] Lee Sharkey, Beren Millidge, and Marius Hobbhahn. 2023. Theories of Change for AI Auditing. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/uhFo7XeaRHL7b9ioh/announcing-the-era-cambridge-summer-research-fellowship>
- [A149] Adam Shimi. 2020. The "Backchaining to Local Search" Technique in AI Alignment. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/qEjh8rpjG4qGtFuK/>
- [A150] Adam Shimi. 2021. On Solving Problems Before They Appear: The Weird Epistemologies of Alignment. AI Alignment Forum. October 11, 2021. Accessed August 9, 2026. <https://www.alignmentforum.org/posts/FQqcejhNWGG8vHDch/on-solving-problems-before-they-appear-the-weird>
- [A151] Nandini Shiralkar. 2023. Announcing the ERA Cambridge Summer Research Fellowship. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/uhFo7XeaRHL7b9ioh/announcing-the-era-cambridge-summer-research-fellowship>
- [A152] Nandini Shiralkar, Tilman Räuher, and Moritz von Knebel. 2023. ERA's Theory of Change. Effective Altruism Forum. August 10, 2023. Accessed August 9, 2026. <https://forum.effectivealtruism.org/posts/9tG7daTLzyxArfQev/era-s-theory-of-change>
- [A153] Buck Shlegeris. 2021. Redwood Research's current project. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/k7oxdbNaGATZbtEg3/redwood-research-s-current-project>
- [A154] SPAR. 2026. Mentor project criteria, Supervised Program for Alignment Research. SPAR. Accessed August 9, 2026. <https://sparai.org/mentors/>
- [A155] Jonny Spicer. 2025. Evaluating "What 2026 Looks Like" So Far. LessWrong. February 24, 2025. Archived in research/raw/. Accessed August 9, 2026. <https://www.lesswrong.com/posts/u9Kr97di29CkMvjaj/evaluating-what-2026-looks-like-so-far>
- [A156] Stag, Stephen McAleese, and Dr. David Mathers. 2024. Shallow review of technical AI safety, 2024. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/fAW6RXLKLHC3WXkS/shallow-review-of-technical-ai-safety-2024>
- [A157] Survival and Flourishing Fund. [n. d.]. S-Process for making grant recommendations. Survival and Flourishing Fund. Archived 2026-08-09. <https://survivalandflourishing.fund/s-process>
- [A158] Survival and Flourishing Fund. 2025. Speculation Grants. Survival and Flourishing Fund. Archived 2026-08-09. <https://survivalandflourishing.fund/speculation-grants>
- [A159] technicalities and Stag. 2023. Shallow review of live agendas in alignment & safety. LessWrong. November 27, 2023. Accessed August 9, 2026. <https://www.lesswrong.com/posts/zaaGsFBeDTpCsYHef/shallow-review-of-live-agendas-in-alignment-and-safety>
- [A160] David Thorstad. 2023. High risk, low reward: A challenge to the astronomical value of existential risk mitigation. globalprioritiesinstitute.org. Archived 2026-08-09. <https://www.globalprioritiesinstitute.org/wp-content/uploads/David-Thorstad-High-risk-low-reward-A-challenge-to-the-astronomical-value-of-existential-risk-mitigation.pdf>
- [A161] UK Government Office for Science. 2023. AISI Research Agenda | The AI Security Institute. UK Government Office for Science. Archived 2026-08-09. <https://www.aisi.gov.uk/research-agenda>
- [A162] utilistrutil, Juan Gil, Ryan Kidd, Christian Smith, deus_ex_maki, and Laura Vaughan. 2023. MATS Summer 2023 Retrospective. LessWrong. December 1, 2023. Accessed August 9, 2026. <https://www.lesswrong.com/posts/zwf68YaySvXhWYCd/mats-summer-2023-retrospective>
- [A163] utilistrutil, Juan Gil, yams, Laura Vaughan, K Richards, and Ryan Kidd. 2024. MATS Alumni Impact Analysis. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/jeBkx6agMuBCQW94C/mats-alumni-impact-analysis>
- [A164] utilistrutil, Laura Vaughan, deus_ex_maki, Christian Smith, Juan Gil, Henry Sleight, Matthew Wearden, and Ryan Kidd. 2024. MATS Winter 2023-24 Retrospective. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/Z87fSrxQb4yLXKcTk/mats-winter-2023-24-retrospective>
- [A165] Lizka Vaintrob. 2022. Winners of the EA Criticism and Red Teaming Contest. Effective Altruism Forum. October 1, 2022. Accessed August 9, 2026. <https://forum.effectivealtruism.org/posts/YgbpxJmEdFhGpqc/winner-of-the-ea-criticism-and-red-teaming-contest>
- [A166] Lizka Vaintrob, Fin Moorhouse, and Joshua Teperowski Monrad. 2022. Announcing a contest: EA Criticism and Red Teaming. Effective Altruism Forum. June 1, 2022. Accessed August 9, 2026. <https://forum.effectivealtruism.org/posts/8hvmvrgcxJJ2pYR4X/announcing-a-contest-ea-criticism-and-red-teaming>
- [A167] vkrakovna.wordpress.com. 2022. Paradigms of AI alignment: components and enablers. vkrakovna.wordpress.com. Archived 2026-08-09. <https://vkrakovna.wordpress.com/2022/06/02/paradigms-of-ai-alignment-components-and-enablers/>

Manuscript submitted to ACM

- [A168] John Wentworth. 2024. The Field of AI Alignment: A Postmortem, and What To Do About It. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/nwpyhyagpPYDn4dAW/the-field-of-ai-alignment-a-postmortem-and-what-to-do-about>
- [A169] John Wentworth and David Lorell. 2025. The Plan — 2025 Update. LessWrong. December 31, 2025. Accessed August 9, 2026. <https://www.lesswrong.com/posts/vh5ZjdmJYJgnbpq8C/the-plan-2025-update>
- [A170] Robert Wiblin. 2019. A framework for comparing global problems in terms of expected impact. 80,000 Hours. First published 2016; updated October 2019. Accessed August 9, 2026. <https://80000hours.org/articles/problem-framework/>
- [A171] worldbuild.ai. 2022. Winning Worlds - FLI Worldbuilding Contest. worldbuild.ai. Archived 2026-08-09. <https://worldbuild.ai/winners>
- [A172] yams, Carson Jones, deus_ex_maki, and Ryan Kidd. 2024. Talent Needs of Technical AI Safety Teams. LessWrong. Archived 2026-08-09. <https://www.lesswrong.com/posts/QzQQvGJYDeaDE4CfG/talent-needs-of-technical-ai-safety-teams>