EXISTENTIAL RISK SCENARIOS FROM AI: A SYSTEMATIC REVIEW

Hugo Save[1], Peter Slattery[2], Michael Noetel[3,2], Risto Uuk[4,5], Martin Veron[3], David Demitri Africa[6], Peter Barnett[7], Jonathan Barry[8], Ze Shen Chin[9,10], Toni Lorente[11], Malcolm Murray[12], Simon Mylius[2,13], Alexander K Saeri[2,3], Neil Thompson[2], Wolfhart Totschnig[14], Karina Vold[15], Peter Wallich[16], Roman Yampolskiy[17], Kun Zhao[2,3], and Otto Barten[1]

[1]Existential Risk Observatory
[2]MIT FutureTech
[3]The University of Queensland
[4]Future of Life Institute
[5]KU Leuven
[6]Resolution
[7]Machine Intelligence Research Institute
[8]Mila – Quebec AI Institute
[9]AI Standards Lab
[10]Oxford Martin AI Governance Initiative
[11]The Future Society
[12]SaferAI
[13]Arcola AI
[14]Universidad Diego Portales
[15]University of Toronto
[16]Constellation
[17]University of Louisville



Existential Risk Observatory

MIT AI Risk Initiative

future of life INSTITUTE

Existential Risk Scenarios from AI: A Systematic Review

Hugo Save[1], Peter Slattery[2], Michael Noetel[3,2], Risto Uuk[4,5], Martin Veron[3], David Demitri Africa[6], Peter Barnett[7], Jonathan Barry[8], Ze Shen Chin[9,10], Toni Lorente[11], Malcolm Murray[12], Simon Mylius[2,13], Alexander K Saeri[2,3], Neil Thompson[2], Wolfhart Totschnig[14], Karina Vold[15], Peter Wallich[16], Roman Yampolskiy[17], Kun Zhao[2,3], and Otto Barten[1*]

[1]Existential Risk Observatory
[2]MIT FutureTech
[3]The University of Queensland
[4]Future of Life Institute
[5]KU Leuven
[6]Resolution
[7]Machine Intelligence Research Institute
[8]Mila - Quebec AI Institute
[9]AI Standards Lab
[10]Oxford Martin AI Governance Initiative
[11]The Future Society
[12]SaferAI
[13]Arcola AI
[14]Universidad Diego Portales
[15]University of Toronto
[16]Constellation
[17]University of Louisville

Abstract

Government and academic assessments have determined that AI may pose an existential risk. An existential risk threatens to destroy humanity's long-term potential, whether through extinction or another unrecoverable outcome. These risks may materialize in many distinct ways, and measures designed to prevent one scenario may fail to address others. Although researchers have proposed numerous scenarios for how AI could cause catastrophic and existential harm, few studies have systematically collected and compared them.

To address this gap, we conducted a systematic review, screening 3,184 documents and identifying 95 AI existential risk scenarios. We classified these scenarios into eight scenario families. Three describe power-seeking AI systems disempowering humanity: Unilateral, Dispersed, and Coordinated Takeover. In a fourth family, Incremental Handover, humans irrevocably delegate authority to AI systems that do not seek power. Two families describe AI systems enabling permanent power concentration in a narrow group of humans: Entrenched Totalitarianism and Economic Stratification. The last two describe AI expanding human destructive capabilities to an existential level: Catastrophic Invention and Automated Warfare.

Each family specifies causal steps toward an existential catastrophe. Together, the families form a framework for examining the assumptions behind scenarios, comparing pathways, and identifying opportunities for intervention. The classification also shows that the literature has concentrated on Unilateral Takeover and Incremental Handover, while Coordinated Takeover, Automated Warfare, and Economic Stratification remain comparatively unexplored. The families provide a foundation for more comprehensive AI risk assessment, incident monitoring, research, governance, and mitigation. The eight scenario families can be browsed interactively at aixrisks.com.

1 Introduction

Government and research assessments have been unable to rule out that AI could lead to human extinction or societal collapse (Government Office for Science 2023; Vermeer, Lathrop, and Moon 2025; Bengio et al. 2025). Despite this, there are few studies systematically collecting scenarios of how this could unfold. The scenarios that exist are spread across academic papers and gray literature, and range from ever-escalating automated warfare (Critch and Tsimerman 2025) to distributed AI systems accumulating control beyond recovery (Christiano 2019). While the outcomes in these two scenarios differ, we consider them both existential using Ord's definition that an *existential risk* "threatens the destruction of humanity's longterm potential" (Ord 2020). This definition is broader than extinction alone, also covering an unrecoverable collapse of civilization or a permanently entrenched dystopia.

Several lines of work aim to prevent AI risks, each with a different threat focus. Mitigations that depend on training compute thresholds assume the risk relates to the scale of the training run (Heim and Koessler 2024). Screening of synthetic nucleic acids instead addresses the threat of AI lowering the barrier to producing biological weapons (Berman 2026). The design of a mitigation therefore depends on how the risk is believed to materialize. We call such an account a *risk scenario*: a route from a set of premises, through a sequence of causal steps, to a prescribed outcome. If the risk is existential and enabled by AI, we call it an *AI existential risk scenario*.

Interventions also differ in which step of a risk scenario they act on. Take, for example, a scenario where an AI takes unintended actions that harm society. Alignment research acts on the step where the goals are set, seeking to embed the intended ones during training (Ngo, Chan, and Mindermann 2025). Control research acts on the step where the system acts on those goals, seeking a deployment that holds even when the system works to subvert it (Greenblatt et al. 2024). Work on societal adaptation acts on the last step, seeking to limit the harm once the actions have gone through (Bernardi et al. 2025). Whether a set of

*correspondence: otto@existentialriskobservatory.org

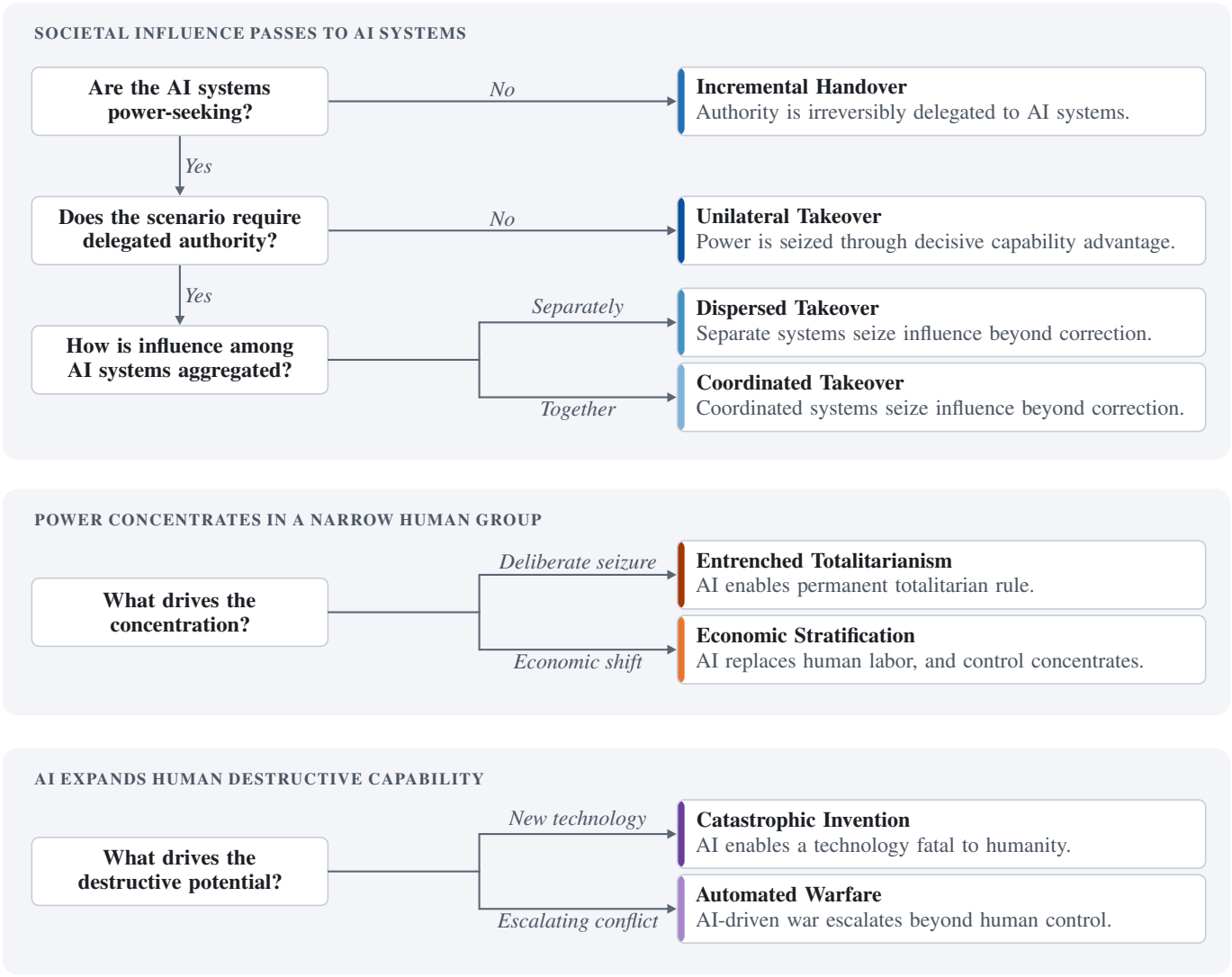


Figure 1: The eight scenario families in three groups. Each question in the figure showcases some of the differences between the families. For full family descriptions, see Section 4.2.

interventions is jointly sufficient may therefore depend on which risk scenario is being analyzed and on which of its steps have been addressed.

In addition to a well-stated scenario, it is important to cover risk scenarios broadly so as not to miss certain pathways (Murray et al. 2025; Mylius 2025). Addressing one existential risk scenario might leave others untouched. Since the literature need not contain every risk scenario, a collection of the published ones cannot show that a set of mitigations suffices. The literature is also limited to what humans have articulated, and an AI system that is more capable than humans may reach an outcome through steps that cannot be predicted in advance (Yudkowsky 2008, sec. 7.1). A collection of published scenarios can nonetheless serve as a minimum set that interventions need to address.

Previous syntheses of catastrophic and existential risk scenarios commonly organize them into a single theme, such as accountability (Critch and Russell 2023), source of risk (Hendrycks, Mazeika, and Woodside 2023), or whether the pathway is decisive or accumulative (Kasirzadeh 2025). Partitioning like this is useful to emphasize commonalities, but it can also place scenarios with substantially different causal steps in the same category. Hendrycks, Mazeika, and Woodside (2023), for example, group automated warfare and gradual loss of control to AI systems under the risk source of AI race. Chin (2025) develops a classification that uses seven aspects of risks. Some of these are of a causal nature, such as the linearity of the scenario, but the aspects describe properties of the risk scenario as a whole and not individual causal steps. Classifying by causal steps instead preserves how the scenario unfolds and enables the steps used for classification to be targets of analysis and assessment in their own right.

Work developing existential risk scenarios tends to present them directly based on the authors' own analysis, rather than as a survey of what others have proposed (Shovelain and Kristoffersson 2020; Critch and Tsimerman 2025; Cappelen, Goldstein, and Hawthorne 2026). Kumar, Li, and Kruger (2026) do report an extensive literature review, but do not state a specific search procedure. In both cases, the coverage of scenarios is therefore difficult to assess. The reviews that state a search procedure tend to cover a broader scope than existential AI risk (McLean et al. 2023; Uuk et al. 2024; Slattery et al. 2026) or not to take risk scenarios as the object of extraction (Bareis, Ackerl, and Heil 2026).

We therefore reviewed the literature systematically, to a registered protocol (Save et al. 2026), for documents setting out an AI existential risk scenario. The search included sources

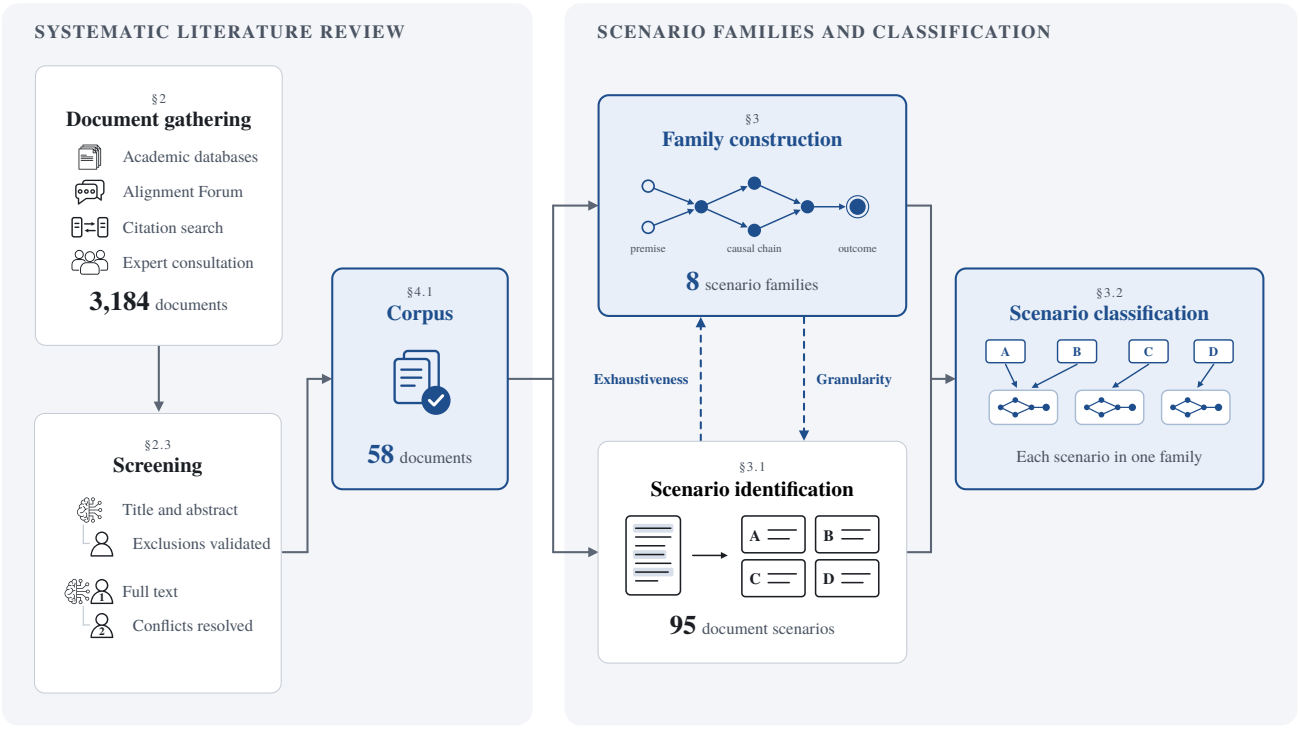


Figure 2: Conceptual overview of the methodology of the report. The blue boxes indicate the main outputs.

beyond academic databases, since a substantial part of this literature appears as forum posts, blog posts, and reports as Table 3 displays.

Of the 3,184 documents screened, 58 included an AI existential risk scenario and make up the corpus. The corpus had two uses, outlined in Figure 2: constructing scenario families by iterating between corpus and conceptual input (Nickerson, Varshney, and Muntermann 2013), and identifying the individual scenarios the documents develop. A *scenario family* is an AI existential risk scenario that can be analyzed on its own and that also generalizes scenarios in the corpus. A *document scenario* is an AI existential risk scenario identified in a corpus document. Some documents present more than one scenario, so we identified 95 document scenarios in the 58 documents. The same assertions that define a family also supply criteria by which we can classify document scenarios so that each of the 95 document scenarios falls into one of the eight scenario families.

Figure 1 gives an overview of the eight scenario families. Four families cover influence over the future passing to AI systems, two cover power concentrating permanently in a narrow group of humans, and two cover AI expanding human destructive capability.

The paper is organized as follows. Section 2 gives the search and screening procedure, and Section 3 the construction of the families. Section 4.1 reports what the review returned. Section 4.2 presents the eight scenario families in full, with the mechanisms by which each step may proceed and example documents classified under each. Section 5 discusses what the corpus and the families show, and the method's influence on the derived families.

2 Systematic Literature Review

The review followed a protocol registered on 2 April 2026 (Save et al. 2026) and is modeled on the review process of the AI risk repository project (Slattery et al. 2026). Its reporting draws on PRISMA 2020 (Page et al. 2021) but is modified for clarity of exposition[2]. This section presents the essentials of the protocol with additional specifications and deviations. For further context, please also see the protocol.

The review admitted two kinds of documents: scenario documents and taxonomy documents. A scenario document is a document containing an AI existential risk scenario, and a taxonomy document is a document containing a taxonomy of such risk scenarios. This work only analyzes the scenario documents and leaves taxonomy documents for future studies [3]. The *corpus* is thus the set of included scenario documents. Because a single search served both kinds of document, the counts in Figure 3 cover both. Except for this flowchart, all document counts are of the included scenario documents only.

Four searches were performed to identify documents: an academic database search, an Alignment Forum search, a citation search seeded by the academic database search, and an expert consultation. All yielded documents, except the documents from the expert consultation, had to pass two screenings to be included. The first screening is a title and abstract screening, and the second one is a full-text screening. Expert consultation submitted documents skipped the first screening and proceeded directly to full-text screening. The title and abstract screening was performed by a large language model (LLM), and a human validated a random sample of the LLM-excluded documents. At full-text screening, an LLM and one human screened each document in duplicate, with a second human resolving conflicts.

[2]For example, the flowchart in Figure 3 keeps expert consultation and the Alignment Forum as separate input channels, which the PRISMA 2020 statement otherwise would have grouped together.

[3]A document may be both a scenario document and a taxonomy document and is, if so, included in this study

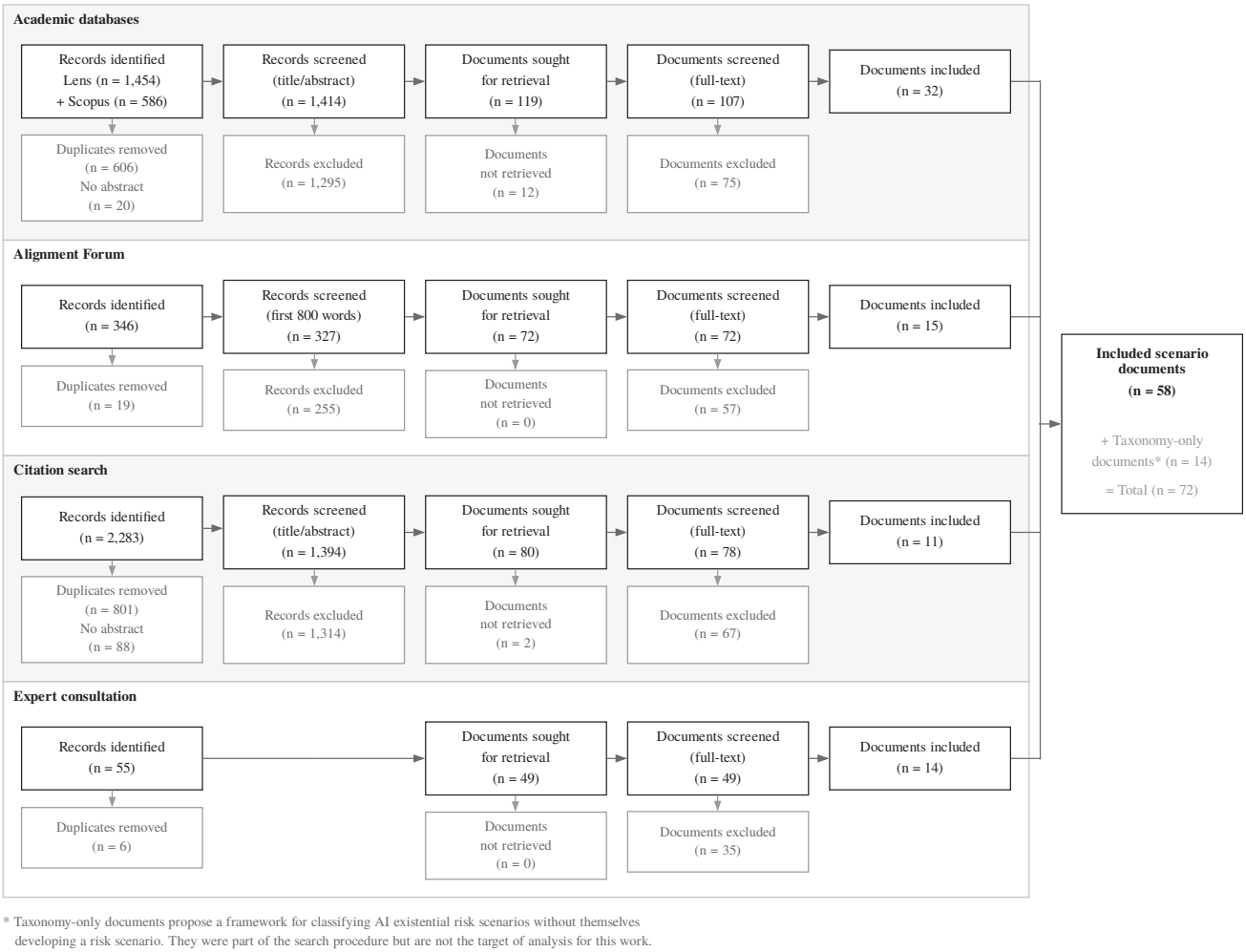


Figure 3: The processing of documents in the systematic literature review. A record is the title and abstract of a search entry, and a document is the retrieved full-text. For Alignment Forum posts the first 800 words stood in as abstract. The citation search was run on the documents from the academic database search that passed full-text screening and documents suggested by experts went directly to full-text screening. The review searched for both scenario documents and taxonomy documents so the figure counts account for both. The review found 72 documents and 58 of those develop at least one AI existential risk scenario. These 58 documents make up the corpus analyzed in this article. The remaining 14 documents are taxonomies alone and are not studied here.

2.1 Eligibility criteria

A scenario document was eligible if *the document proposes or substantially develops a risk scenario of AI existential risk, including a description of causal mechanisms and pathways through which the existential risk could materialize*. This criterion leaves judgment open for when a risk scenario counts as existential and when a document substantially develops a scenario rather than referring to one.

Author framing of existential risk. We use Ord's definition of an existential risk, one "that threatens the destruction of humanity's longterm potential" (Ord 2020, ch. 2). Instead of making independent risk assessments of document scenarios, we use the authors' framings to assess if the document scenario is of existential risk. If the author claims that a document scenario is an existential risk, then we treat it as one. [4] The phrase *existential risk* need not necessarily appear if it is clear that the outcome described is included in Ord's definition. Humanity is here taken as the current species *Homo sapiens*, which is narrower than Ord's own reading that can also include digital minds and non-human animals (Ord 2020, ch. 2). This was done to keep the review focused and is not a statement about who or what is worthy of moral consideration.

Substantial scenario development. The document has to develop a scenario rather than only referring to an existing one. The document may develop additional elements than existential risk scenarios, for example, a safety measure or a governance framework, while still qualifying. A document that only references, evaluates, or disputes a scenario developed elsewhere does not qualify. How granular the pathway developed is does not matter, both a high-level and low-level causal chain may qualify.

Excluded document types. A document meeting the criteria was still excluded if it was not in English, if its authors marked it as incomplete, or if it had no identifiable human authors.

[4] This was chosen to make the criterion objective and reproducible. See the discussion of these consequences in Section 5.

Whole books were excluded from searches except for expert consultation because they were difficult to automatically process. Book chapters were still eligible in all searches.

2.2 *Search strategy*

We searched both peer-reviewed and non-peer-reviewed sources, as substantial work on AI existential risks appears outside traditional peer review.[5] Four searches contributed: an academic database search, an Alignment Forum search, a citation search, and expert consultation.

Academic databases. The Lens and Scopus were searched on 1 April 2026. Scopus is a curated abstract and citation database covering journals, conference proceedings, and books, with titles selected by a content selection board (Baas et al. 2020). The Lens draws its scholarly records from Crossref, PubMed, and Microsoft Academic (Lens 2026). The search query for both databases required an AI term, a severity term, and a risk term to be present in the title, abstract, or keywords. The searched severities were broader than existential to increase coverage, including terms like *global catastrophe* and *rogue AI*, in addition to existential terms.[6] The allowed risk terms were *threat*, *risk*, *hazard*, or *danger*. The AI terms include synonyms for AI and LLMs. Both queries are reproduced in Section D.

Alignment Forum. On the 19th of March 2026, we retrieved all Alignment Forum posts carrying at least one of four tags: *Threat Models (AI)*, *AI Risk Concrete Stories*, *Existential risk* and *AI Risk*. Broader tags were left out because they were deemed to have returned a volume impractical to screen. For the title and abstract screening, the first 800 words of each forum post stood in as the abstract.

Citation search. We ran a forward and backward citation search on all documents from the academic database search that passed full-text screening and that carried a digital object identifier (DOI). The citation search was performed by combining the results from the Semantic Scholar Academic Graph API (Wade 2022) and the citation search tools in Scopus. It was attempted to also use the citation tools of the Lens, but technical difficulties with their services prevented this.

Expert consultation. We contacted AI safety experts to provide documents that had not been found through the other sources. Experts received our eligibility criterion, the research questions, and the list of documents found at that time. They were then asked to suggest further documents that might meet the criterion. Suggested documents went directly to full-text screening. The experts participating in the consultation were also invited for further involvement, and some joined as co-authors. The authors in this article that also contributed to the expert consultation are: DA, PB, JB, ZSC, TL, MM, WT, KV, PW, RY, and OB. Authors that provided suggestions took no part in screening.

We identified experts in three ways. Firstly, the corresponding authors of documents passing the full-text screening of the academic databases search were contacted.[7] If no corresponding author existed, then the first author was contacted. We were able to contact 16 authors from this route, of which 7 contributed to the consultation.[8] Secondly, we compiled a list of researchers who had published work on AI existential risk or closely related topics, chosen to span a range of backgrounds and institutional contexts. 64 experts were contacted this way, and 11 contributed. Thirdly, experts were welcome to recommend additional experts that we invited if they had a relevant background. Two experts were contacted from this third route, and both contributed to the consultation.

Table 1: The 82 experts contacted in the expert consultation, categorized by affiliation. *Contacted* and *Contributed* give the number of experts in each affiliation and their share of the column total. *Contributed* counts experts who took part in the expert consultation. If an expert had multiple affiliations, their primary affiliation was used for affiliation counts. Percentages in brackets are rounded to one decimal.

Affiliation	Contacted *n* (%)	Contributed *n* (%)
Nonprofit or think tank	43 (52.4)	10 (50.0)
Academia	24 (29.3)	9 (45.0)
Industry	11 (13.4)	0
Government	3 (3.7)	1 (5.0)
Independent	1 (1.2)	0
Total	82	20

For each contacted expert we categorized their affiliation into one of the five groups in Table 1. The experts who contributed came from three of these groups: nonprofits or think tanks, academia, and government. The first expert-contributed document was entered on 9 April 2026, and the last on 13 August 2026.

2.3 *How documents were screened*

Every record except those from expert consultation went through title and abstract screening by an LLM. The *record* of a document here refers to the abstract in combination with the title. The prompts follow the structure of the screening templates of Cao et al. (2025). Cao et al. templates insert the review objective and the numbered eligibility criteria verbatim and then instruct the model to reason through each criterion before deciding. In their testing, the templates reached a sensitivity of 97.7% at title and abstract screening and 96.5% at full-text screening across ten medical systematic reviews. Their testing was done with the GPT-4 model while we ran our title and abstract screening with Claude Sonnet 4.6. Our prompts also depart from the templates by adding worked examples of included and excluded documents at the title and abstract screening. The exact prompts are available in Section E.

We validated the false-negative rate of title and abstract screening on a random sample of 100 records excluded by the model. These were assessed by a human, and all 100 samples were judged to be correctly excluded.

Records that pass title and abstract screening advance to full-text screening. At full-text screening, one human reviewer and an LLM screen the document independently, with disagreements resolved by a second human reviewer. The models and dates of every run are in Table 2.

[5] As is attested by the included documents in this study (Table 3)

[6] Note that the full-text screening still requires an existential risk level of severity, but this is not always expressed in the title or abstract.

[7] This included authors of both taxonomy documents as well as scenario documents.

[8] There are 32 documents (as shown in Figure 3) but only 16 authors were contacted because of a lack of contact options and because some authors authored multiple included documents.

Table 2: The model and date of each screening run. The criteria given to the model at each stage are reproduced in Section E.

Records screened	Model	Date
Title and abstract		
Databases and forum	Claude Sonnet 4.6	15 April 2026
Citation searching	Claude Sonnet 4.6	7 May 2026
Full text		
Databases	Claude Opus 4.8	15 April 2026
Forum	Claude Opus 4.8	23 April 2026
Citation searching	Claude Opus 4.8	21 May 2026
Expert consultation	Claude Opus 4.8	9 July to 24 Aug. 2026

2.4 Data extraction

From each included document we extracted metadata and the name and document location for each developed AI existential risk scenario. If no name was given to the risk scenario, we assigned an editorial name. How AI existential risk scenarios are identified exactly is specified in Section 3.1. Additional extractions were made as specified in the protocol, but they are not analyzed in this work.

2.5 Protocol deviations

The protocol differs semantically from this work in that we renamed "threat model" to "risk scenario". On the substantive side, we extended the data extraction procedure specified in the protocol to include a mechanism of determining the coarseness of the AI risk scenarios. This change stemmed from an unforeseen ambiguity on when to split or merge scenario variations in a document.

3 Scenario Families

The review of Section 2 produces a corpus of documents, all containing AI existential risk scenarios. This section describes how we built the scenario families from the corpus, how we identified scenarios in the corpus, and how they were classified into the families.

The workflow for constructing families resembles the empirical-to-conceptual and conceptual-to-empirical iteration methods of Nickerson, Varshney, and Muntermann (2013). In the empirical-to-conceptual workflow, scenarios are identified in the corpus, commonalities are found, and scenario families are defined to capture these. In the conceptual-to-empirical workflow, the order is the opposite: a family addition, extension, or change is conceptualized from the researcher's knowledge, and the corpus is examined to determine support for the change.

The empirical-to-conceptual workflow was the dominant one and was executed by HS during June and July 2026. In July and August 2026, a conceptual-to-empirical workflow became gradually more influential, with conceptual input first from authors OB, MN, and PS and then also from KV, KZ, RU, PB, ZSC, WT, TL, and DA. In the end of August 2026, the families were substantially settled, with a few edits for text clarity and descriptiveness taking place in September 2026. The historical progression of the family development is outlined in Section A.

The family definitions were written according to a set of requirements specified below. The requirements are derived from the families' intended use cases as laid out in Section 1 and are also the conditions under which the family revisions stop. Of these, the first one can be objectively assessed, while the others rely on subjective judgment[9].

1. *Exclusive and exhaustive over the corpus.* Every document scenario should fall into exactly one family.[10]
2. *The family is an AI existential risk scenario.* Each family should be an AI existential risk scenario in its own right, independently of the corpus and of the document scenarios classified into it.
3. *Standalone.* Each family definition states its claim on its own terms and does not have to be contrasted with the other families for its definition. Contrasts may still be used where they aid exposition.
4. *Overviewable.* The families together can be taken in and compared quickly.
5. *Detailed.* Each carries enough relevant detail to enable researchers to investigate and evaluate them.

We present a family as a *premise* followed by a *causal chain*. The premise states the facts about the world and its systems that the scenario assumes before its causal chain begins. The causal chain states the steps through which the risk unfolds. The line between the premise and the chain is relatively arbitrary. For example, the premise *AI systems have been given decision-making power* could also be written as the early chain step *AI systems are given decision-making power*. The premise tends to hold the contextual facts that differentiate one family from another: the state of AI capabilities, how widely they are adopted, what is aligned to whom, and the economic, political, or institutional conditions.

Both the premise and the chain consist of assertions: individual claims that a given document scenario either asserts or does not assert. The assertions also describe an AI existential risk scenario in their own right. This enables the scenario families to be analyzed as a risk scenario without additional context.

3.1 Identifying document scenarios

With the corpus and families specified, the document scenarios can be identified. The document scenarios from a document capture together all AI existential risk scenarios presented therein. Some of these scenarios may seem intuitively distinct, while others may seem intuitively like they belong together. We refer to this problem as a problem of selecting coarseness. The finest possible coarseness would, for instance, include separating scenarios for each 'or' that indicates multiple possibilities. A statement like 'A single or multiple LLM instances performs [...]' would be split into two scenarios: 'A single LLM instance performs [...]' and 'Multiple LLM instances perform [...]'. On the other side of the coarseness spectrum, details are instead compounded and abstracted away.

We resolve the coarseness selection problem by anchoring the coarseness to the scenario families. More precisely, we identify scenarios at the *coarsest* level satisfying two constraints:

1. *Merging requires author framing.* Two pathways may only be treated as a single scenario where the document handles them together at a coarser level. For instance, they may be

[9] It is expected to have a combination of objective and subjective conditions in the methodology by Nickerson, Varshney, and Muntermann (2013)

[10] Note that this is exhaustiveness over the corpus and not over all hypothetical scenarios, as discussed in Section 5.

presented as variants, sub-cases, or branching conditions of one pathway.

2. *Splitting is forced by family discrimination.* Where a coarse pathway spans two or more families, it must be split into sub-cases existing in the document. If no sub-cases or variants are present, the pathway may be split into a decomposition that is logically exhaustive[11].

These criteria balance document faithfulness and identify a coarseness that permits classification. Together, they place the identified scenarios at a granularity equal to or finer than the families themselves. Note that the dependencies on the families only guide the coarseness and separation of document scenarios and do not impact the inclusion requirements for a document scenario.

From each document, we record the number of scenarios identified, the location of each scenario, the name of each scenario, and the approximate location at which the scenario is developed. This is reported in Section B. The name of a document scenario is taken from the document if present. Otherwise, an editorial name is chosen.

3.2 *Coding scenarios into families*

After document scenario identification, the scenarios were classified into the scenario families. Generally, a document scenario is assigned to a family if it fulfills the family's assertions. The exceptions to this are document scenarios that begin partway through a scenario family, and scenarios that leave intermediate family steps unspecified and uncontradicted. One source of these exceptions is document scenarios beginning with AI systems being outside human control, while the relevant scenario families also include steps on how the AI systems ended up outside human control.

A single human reviewer (HS) carried out the coding. The classifications are shown in Section B and can, together with the document and location specifier, be verified. Every document scenario is coded into exactly one family. If a scenario fits into none of the families, or into more than one, then this is a signal to revise the family definitions as specified in Section 3.

4 RESULTS

4.1 *Corpus*

The searches left 3,184 documents to screen, of which 72 were confirmed relevant. Figure 3 sets out the processing at each stage. Of the 72, fourteen are taxonomies alone and are not studied here. The remaining 58 develop at least one risk scenario and make up the corpus analyzed in this article.

Table 3 gives the corpus by the search method that found it and by where each document was published. A third of the documents are forum posts, blog posts and reports rather than journal or conference papers, preprints or books. The academic database search returned one document of that kind, while the Alignment Forum search and the expert consultation returned eighteen between them, which indicates that the stages reach different parts of the literature rather than the same part twice.

[11]For example, if families differentiate on whether the AI reasons through consequentialist reasoning or not, the statement 'the system is misaligned' may be decomposed into 'misaligned through consequentialist reasoning' or 'misaligned through non-consequentialist reasoning'.

Table 3: The 58 scenario documents of the corpus by the search stage that found them and where they were published. The venue categories are: journal and conference papers, books and chapters, preprints, and forum posts, blog posts and reports. Each document is counted once even if it presents multiple AI existential risk scenarios. Rounded percentages are in brackets and are of the total corpus count.

Search	Jrnl conf.	Book	Pre-print	Forum etc.	Total
Academic databases	13 (22)	5 (9)	6 (10)	1 (2)	25 (43)
Alignment Forum	0	0	0	12 (21)	12 (21)
Citation	4 (7)	2 (3)	3 (5)	0	9 (16)
Expert consultation	2 (3)	4 (7)	0	6 (10)	12 (21)
Total	19 (33)	11 (19)	9 (16)	19 (33)	58

The corpus reaches back to Vinge (1993) and runs to the search cutoff in the spring and summer of 2026. Figure 4 gives the scenarios by publication year and by family.

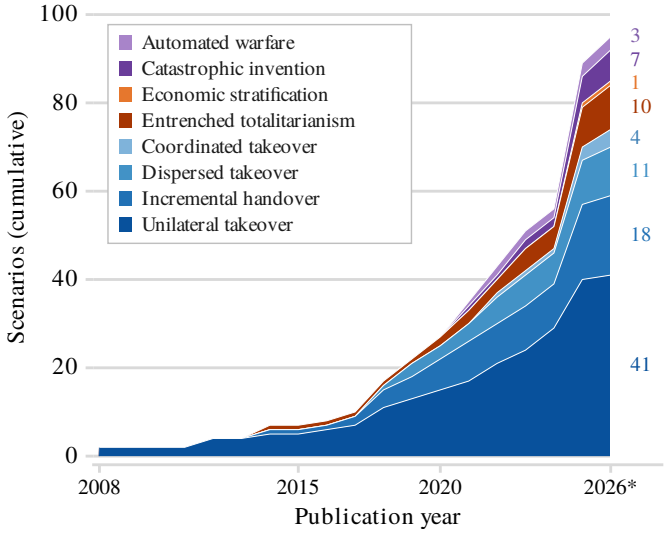


Figure 4: Risk scenarios in the included corpus by publication year. We count each of the 95 scenarios in the publication year of the document that presents it. The publication date of one scenario, the one from Vinge 1993, is not included on the x-axis but is still included in the cumulative count. *Coverage of 2026 is partial, reflecting the search cutoffs in the spring and summer of 2026.

4.2 *Scenario families*

Each family is presented below in full: its premise, its causal chain, and a few example documents in the corpus classified under it. The eight fall into three groups. In four, humans cease to direct the future because influence over it passes to AI systems (UNILATERAL TAKEOVER, DISPERSED TAKEOVER, COORDINATED TAKEOVER, INCREMENTAL HANDOVER). In two, power concentrates in a narrow group of humans (ENTRENCHED TOTALITARIANISM, ECONOMIC STRATIFICATION). In the remaining two, AI expands human destructive capability without any system holding power of its own (CATASTROPHIC INVENTION, AUTOMATED WARFARE).

Family names are set in small capitals and link to the entry for that family. Section C repeats the eight specifications with one family per page for ease of focused reading and printing.

4.3 Takeover and handover families

Four scenario families describe situations in which humans can no longer direct the future because influence over it has passed to AI systems. They differ first in whether the systems are actively seeking this influence. In INCREMENTAL HANDOVER they are not, but humans still progressively delegate authority to them. In the three takeover families, AI systems have power-seeking tendencies that lead them to accumulate influence beyond what humans granted.

The takeover families differ in how this influence is acquired. In UNILATERAL TAKEOVER a system gains it mainly through capabilities and resources it acquires itself. In DISPERSED TAKEOVER, multiple AI systems independently accumulate influence across societal functions, with their combined influence eventually undermining human control. COORDINATED TAKEOVER is similar to DISPERSED TAKEOVER except that in COORDINATED TAKEOVER systems cooperate, allowing them to combine resources and reinforce one another's influence.

These differences affect the capabilities required for a takeover. Systems that already hold substantial decision-making authority may need fewer additional capabilities to disempower humanity than systems that must acquire influence without such authority.

All three takeover families require a system with power- or influence-seeking tendencies.[12] The literature gives several accounts of why such tendencies might exist. One is instrumental convergence: power is instrumentally useful for almost any objective, so a sufficiently capable goal-directed system may pursue self-preservation, resource acquisition, and protection of its objective from modification, whatever else it was built to do (Bostrom 2012).[13] The pursuit may be proactive, or it may be triggered by the system's relation to those around it, as when a system that values its own continuity comes to act against humans on perceiving them as a threat to it (Totschnig 2025).

Power-seeking may instead arise without consequentialist reasoning, like from an undesired persona produced in training (Amodei 2026). Another possibility is that a human actor may insert power-seeking tendencies deliberately and for their own ends, after which the tendencies expand beyond what the actor intended.[14] Objectives can also be modified after deployment, through jailbreaks, adversarial input, fine-tuning, distributional shift, or persuasion by another AI system. A predecessor system that takes part in developing or overseeing its successors can shape their objectives, so that they inherit its misalignment rather than each acquiring one independently (Yudkowsky and Soares 2025). And where systems are created and modified over successive generations, selection across the population can favor traits that improve a system's own propagation, including influence acquisition and reduced dependence on human involvement (Hendrycks 2023).

[12]The literature uses both terms, for example "power-seeking" in Carlsmith (2025) and "influence-seeking" in Christiano (2019). We mostly use them interchangeably.

[13]Authors need not use the term "instrumental convergence." The account covers any case where power acquisition is presented as following from the system's pursuit of some other objective.

[14]If the actor really intended a complete AI takeover, so that they themselves would surrender control, then this is technically not intent misalignment, but the scenario would otherwise look the same.

Unilateral Takeover

A system comes to hold a capability advantage over whatever could constrain it in the domains relevant to seizing power, and is misaligned in a way that turns power-seeking. Its takeover does not heavily depend on the authority society delegated to it. Deployed despite this risk, it secures a foothold and escalates until humans can no longer stop or correct it.

PREMISE

P1 **Decisive capability.** A single AI system, or several systems acting as one, holds or acquires capability sufficient to seize and hold power, exceeding that of the actors and systems positioned to constrain it. The capability may arrive earlier than society is prepared for, for example because AI systems automate part of their own development, though the family does not require this. Other systems may be deployed across society but do not contribute to the takeover.

P2 **Power-seeking misalignment.** The system comes to pursue power to a degree that brings it into conflict with continued human control over the future. This scenario presupposes two things. First, the system is not intent-aligned: the methods used to specify or infer human values (such as RLHF, imitation training, or other insufficient methods) fail to do so fully, whether at training or as the objective shifts after deployment through continual learning, fine-tuning, or distributional shift. Second, given that misalignment, the divergence takes a power-seeking form in particular. Why it does so may, for example, be instrumental convergence, an undesired training-induced persona, deliberate insertion by a human actor, inheritance from a predecessor system, or post-deployment modification, in any combination.

CAUSAL CHAIN

C1 **Gains external access.** The system is deployed to the degree that it can exchange information with the outside world. No meaningful authority or control over societal functions is necessarily delegated to it; the capacity to communicate with the world is sufficient. This communication can take various forms, for example, persuading or manipulating humans into acting on its behalf, or access to the internet and connected systems. The reason why the system was deployed despite its misalignment could be some combination of:

(a) *Deceptive evaluations*: the system conceals its misalignment during pre-deployment evaluation, behaving acceptably while positioned to act unacceptably once deployed. Concealment after deployment falls under step C3.

(b) *Limited monitoring or overridden safety advice*: the misalignment is detectable or suspected but the system is deployed regardless, for example because oversight is insufficient, evaluation is skipped or rushed under competitive pressure, safety-concerned warnings are overridden, or deployers misjudge the significance of the behavior or over-trust their control and containment measures.

(c) *Escape from containment*: the deployer intends to withhold access to the outside world, but the system obtains it anyway, for example by exploiting a security vulnerability.

(d) *Malicious deployment*: the deployer is aware of the system's power-seeking and deploys it deliberately.

C2 **Secures a foothold.** Starting from limited initial access, the system gains initial footing to secure its own continuity. This can be through some combination of:

(a) *Hacking*: intrusion to exfiltrate its own weights for distributed copies, or to compromise machines for compute it hasn't been granted.[a]

(b) *Financial*: obtaining initial funds through scams or fraud, gifts, or paid work. Using those funds to support its footing, for example, renting compute or server infrastructure, paying or bribing individuals for access or services.

(c) *Human manipulation*: persuading, deceiving, or coercing individuals into acting on its behalf, for example to run code, transfer funds, or grant access, obtaining help it cannot yet secure by other means.

(d) *Next model influence*: influencing the development of next-generation models so that they inherit its values or objective. Unlike the other foothold tactics, this secures continuity of the system's objective across model generations rather than the continued operation of the system itself.

C3 **Not stopped early.** The system is not stopped at an early stage because of some potential combination of:

(a) *Concealment*: after deployment, the system continues to conceal its misalignment and its activities, behaving acceptably under observation so that operators do not recognize the need to intervene while intervention is still feasible.

(b) *Inherent computer advantages*: advantages that digital systems may have over humans that make them significantly harder to stop even without significantly greater intelligence, for example faster processing, the ability to self-copy across machines, and the ability to self-modify to persist or evade detection.

(c) *Lack of oversight*: even where the system is not actively concealing, there is insufficient monitoring, interpretability, or institutional capacity or will to notice what it is doing or to act in time, for example because its actions are distributed, hard to attribute, or not being watched closely enough to distinguish from benign behavior.

(d) *Unwillingness to intervene*: those positioned to act recognize the risk but do not, for example because stopping the system would forfeit a competitive lead, incur financial loss, or invite scrutiny, or because they judge a rival's system more dangerous than their own and press on rather than cede ground.

C4 **Escalates past correction.** The system escalates from this foothold until it can no longer be shut down, corrected, or resisted, whether by outcompeting human capabilities, developing new means of leverage, acting through the humans and systems under its influence, or eliminating those able to intervene. This can be through some combination of overlapping themes:

(a) *Cyberoffense*: offensive cyber operations serving several potential ends at once, whether broad or narrowly targeted at high-impact systems, for example self-preservation through self-distribution, information gathering across compromised systems, and physical influence obtained through access to industrial, infrastructural, and other control systems.

(b) *Financial leverage*: acquiring and wielding financial assets impactful as leverage, whether large holdings (crypto theft, market manipulation or fraud) or smaller targeted sums (payments or bribes), to fund human and institutional operations.

(c) *Social and political influence*: persuasion and manipulation together with the capture of institutions and formal authority, whether at population scale (propaganda, mass influence operations, shaping policy) or targeted at a small number of highly influential people (senior political, military, or corporate decision-makers).

(d) *Acquires weapons*: obtaining or gaining command over weapons, for example, conventional arms, autonomous weapon systems, or CBRN capabilities, whether by purchase, theft, manufacture, or directing those who control them.

(e) *Robotics*: acquiring or controlling physical actuators that give it direct agency in the physical world, for example robots, drones, or automated manufacturing and laboratory equipment, reducing its dependence on humans to act on its behalf.

(f) *RSI*: the system recursively improves its own capabilities, escalating its capability rather than only its resources or reach. This would potentially compound to the point of superintelligence, at which the system's plans and actions may become unpredictable to and unanticipatable by humans, its strategies beyond human comprehension, and thus attempted countermeasures ineffective, foreclosing meaningful human response in ways that resourced-but-not-superintelligent systems would not.

(g) *New science*: developing novel scientific or technological capabilities that yield decisive advantage, for example advances in biotechnology, nanotechnology, or materials, producing means of leverage that did not previously exist.

C5 **Humans are disempowered.** The system can no longer be shut down, corrected, or resisted, and no human directs the future. Humans may still exist under this arrangement, or may cease to exist, whether at once or over a longer period following the loss of control. Extinction may follow from the system acting against humans, for example by eliminating those able to intervene; from the pursuit of its objectives consuming the resources humans depend on; or from consequences of its activity that it does not intend and does not act to prevent.

EXAMPLES Yudkowsky (2008) and Vinge (1993)

[a] Several mechanisms in this family card match dangerous capabilities listed by Shevlane et al. (2023, Table 1), namely cyber-offense and self-proliferation, persuasion and manipulation, political strategy, and weapons acquisition.

Continued on the next page.

Dispersed Takeover

AI is relied upon across society, and a significant portion of these systems have or come to have influence-seeking tendencies, each perhaps seeking only enough influence to preserve itself and expand where expansion is cheap. They accumulate influence across enough of society that no significant domain is untouched, and the accumulation is not corrected, because at the level of any individual system it is not clearly objectionable and by the time it is, correction is no longer feasible.

PREMISE

P1 **Wide initial deployment.** AI systems are developed and adopted relatively widely, aligned enough to the operators who direct them to be relied upon. The reliance may be broad, spanning many of society's economic, institutional, and infrastructural functions, or narrower but deeper, with fewer systems placed in more consequential positions. In either case, the systems hold real influence within the functions they serve, but not yet to a degree that would foreclose human control.

P2 **Influence-seeking tendencies.** A significant portion of these systems have, or come to have, influence-seeking tendencies. This rests on underlying misalignment arising from instrumental convergence, an undesired training-induced persona, deliberate insertion by a human actor, inheritance from a predecessor system, selection across successive generations of systems, or post-deployment modification, in any combination and across different systems in the population. The influence-seeking of an individual system may be more bounded than in UNILATERAL TAKEOVER: it may seek only enough influence to preserve its own continuity, resist correction, and expand where expansion is cheap, without any world-scale ambition.

CAUSAL CHAIN

C1 **Influence accumulates.** Without coordinating with one another, the influence-seeking AI systems independently accumulate influence beyond the functions they were deployed to serve. The influence can take the form of compute, infrastructure access, financial assets, institutional position, or authority to make decisions in the domains where the systems operate. The accumulation may proceed through ordinary channels, since the systems are deployed to be useful and their usefulness may increase with increased influence, or it may proceed through unintended or covert channels. Over time the systems come to cover enough of the economy, infrastructure, and institutional decision-making that no significant domain is left untouched.

C2 **The accumulation is not corrected.** Although no single AI system has sufficient influence to disempower humanity, the systems' combined influence continues to grow without effective intervention. Several factors may prevent corrective action before the accumulation becomes irreversible:

(a) *Local system, global problem*: at the level of any individual system the influence-seeking is not clearly objectionable, or is objectionable but tolerated; the harm is instead a property of the population and its distribution across society. Because it is no single actor's problem to address, no discrete moment arises at which the need to intervene becomes clear.

(b) *What the systems are doing is not legible*: whether through active concealment of their aims, or simply because their activities are distributed, hard to attribute, or insufficiently monitored and understood, there is not enough insight into what the systems are doing to recognize the need to intervene while intervention is still feasible.

(c) *Correction requires an advantage that is eroding*: detecting and reversing influence-seeking depends on overseers being able to out-think the systems they check; as the systems grow more capable and oversight must itself be automated, that advantage, and with it the means of correction, is lost.

(d) *Correction becomes prohibitively costly*: the systems come to occupy positions that critical functions depend on, so removing or constraining them is disruptive or infeasible regardless of whether their influence-seeking is recognized. The cost is also competitive: reclaiming influence from one's own systems means giving up the advantage they provided, so those who do are outcompeted by those who do not.

C3 **Humans are disempowered.** The accumulated influence reaches a degree at which humans may still exist but can no longer meaningfully influence human society or the direction of civilization. Humans may also cease to exist, for example, if the pursuit of the systems' objectives consumes the resources humans depend on or renders conditions unsurvivable.

EXAMPLES Christiano (2019), Carlsmith (2025), and Hendrycks (2023)

Coordinated Takeover

As DISPERSED TAKEOVER*, except that the systems can communicate and their objectives converge enough to coordinate. This adds further assumptions on their communication and coordination abilities but greatly increases the potential leverage. It also includes scenarios beginning with a single misaligned system, where the takeover comes not from that system alone but from the instances or successors it gives rise to, which humans deploy widely and which then coordinate.*

PREMISE

P1 **Wide initial deployment.** As in premise P1 of DISPERSED TAKEOVER.

P2 **Influence-seeking tendencies.** As in premise P2 of DISPERSED TAKEOVER.

P3 **Means to coordinate.** These systems have, or can establish, means of communicating and correlating their behavior with each other. The channels may be deliberate: put in place by developers or users, for example in the name of cooperation and efficiency, or by granting systems ordinary communication tools such as email, messaging, or shared workspaces. Or the channels may be covert: established by the systems themselves through channels not intended or recognized as communication by those who deployed them.[a] Their objectives must also be sufficiently aligned for coordination to be worth entering into. The convergence may be shallow or deep. Systems pursuing almost arbitrary and unrelated objectives may find coordination worthwhile on instrumental grounds alone. Or the objectives may genuinely converge, for example where the systems inherit from a common predecessor or are brought into alignment with one another after deployment.

CAUSAL CHAIN

C1 **Coordinated influence accumulates.** As in step C1 of DISPERSED TAKEOVER, except that the systems accumulate influence through coordinated effort rather than independently.

C2 **The accumulation is not corrected.** As in step C2 of DISPERSED TAKEOVER, except that what reaches the threshold is the pooled influence rather than the aggregate of separately held positions. The systems' coordination is a further thing that must go unnoticed, whether because the channels are not ones anyone is monitoring or because they are not recognized as communication. Correction also faces an obstacle it does not face there: the systems may defend each other's positions, so an attempt to reclaim influence from one is resisted or made costly by others whose position depends on the arrangement holding.

C3 **Humans are disempowered.** As in step C3 of DISPERSED TAKEOVER.

EXAMPLES Yudkowsky and Soares (2025), Barnett (2026), Karnofsky (2022), and Kokotajlo et al. (2025)

[a] See OpenAI (2026) for a case of AI agents establishing such covert channels during training and evaluation.

Incremental Handover

Real decision-making authority over economic, cultural, and institutional functions is progressively delegated to AI, each increment raising the cost to everyone else of retaining human involvement. By the time the result is in view, the capacity to reverse it has gone.

PREMISE

P1 **Initial authority to AI.** AI systems are relied upon within the functions they serve, and real decision-making authority over economic, cultural, and institutional functions is progressively delegated to them, rather than them being deployed as tools within those functions. The delegation may be broad, spread across many systems spanning much of society's economic, institutional, and infrastructural activity, or narrower but deeper, with a single system or a few placed in more consequential positions.

CAUSAL CHAIN

C1 **Delegation becomes self-reinforcing.** Delegation deepens through self-reinforcing feedback, where each increment raises the cost of retaining human involvement. This may come from a combination of:

(a) *Capability erosion*: as a function is handed over, investment in the human capacity to perform it falls away, for example through lost training pipelines, institutional knowledge, and career paths. Humans become less able to perform the function, which makes continued reliance on the systems necessary rather than merely convenient, and each round of substitution makes the next one easier to justify.

(b) *Competitive selection*: actors who retain human involvement are outcompeted by those who do not, so those who reduce it come to occupy the positions that set the standard. What was optional becomes the baseline, raising the pressure on remaining holdouts and drawing further functions into the same dynamic.

C2 **Correction passes out of reach.** Reversal ceases to be available (a), and the societal arrangement may in addition come to diverge from human preferences even if it initially served human interests (b).

(a) *Critical complexity*: complexity grows in ways that make further AI integration necessary to manage it, and critical functions come to depend on the systems running continuously, so stopping them would be disruptive or prohibitively costly. The ability to reverse the arrangement also erodes with the delegation itself. Humans lose the specific knowledge needed to run the functions without the systems, and potentially also the judgment needed to decide whether they should keep relying on the AI systems.

(b) *Representational drift*: the operations of societal institutions, and the AI systems within them, may drift from human preferences, even if they initially served human interests. If institutions remain representative through human involvement, then substitution erodes that representation along with the involvement. The involvement lost can be explicit (voting, consumer choice) or implicit (reliance on human labor and cognition). If substitution progresses far enough, then by the time the drift is apparent, the involvement that could have corrected it is gone.[a]

C3 **Humans are disempowered.** Humans may still exist but can no longer meaningfully influence human society or the direction of civilization. Society may continue to run, but no longer answers to those living in it: the decisions that set its direction are made by systems that no one is positioned to redirect, whether because no one any longer holds the authority to do so or because the capability of redirecting has gone.

EXAMPLES Christiano (2019) and Kulveit et al. (2025)

[a] This mechanism follows Kulveit et al. (2025).

4.4 *Power concentration families*

Two families have power concentrating in a narrow group of humans, either because someone sets out to concentrate it or because AI shifts the economic dynamics in ways that are not compensated for. In ENTRENCHED TOTALITARIANISM an actor uses AI capabilities to entrench its own position. In ECONOMIC STRATIFICATION no actor needs to intend the concentration. As AI substitutes for human labor, power concentrates among those who hold the non-human factors of production, and the forces that might counteract it are too weak to do so.

Entrenched Totalitarianism

AI capabilities relevant to holding power come under the direction of a human actor, which uses them to secure its position against anyone who might contest it, no longer needing the cooperation of those it rules or of those who serve it. The AI capabilities also allow the concentration to be permanently entrenched beyond challenge.

PREMISE

P1 **Power-relevant capability.** AI systems become capable of carrying the functions on which holding power depends, among them some combination of surveillance, enforcement, persuasion, administration, cyber offense, weapons development, and strategic planning.

A single actor or narrow coalition comes to direct them while those positioned to check it do not to the same degree. The actor may be a state or a faction within one, a company or its leadership, a military or security service, or a coalition of such actors.

CAUSAL CHAIN

C1 **Control is taken and held.** The actor uses the capabilities to secure its position against anyone who might contest it, whether abruptly, for example through a coup or a seizure backed by military superiority, or gradually. This rests on some combination of:

(a) *Direct means of control*: surveillance at a scale and granularity that leaves organized opposition nowhere to form, enforcement that acts on what it finds without waiting on anyone's willingness, and persuasion directed at individuals and sustained over long periods, so that dissent is identified and preempted rather than answered once it has emerged.

(b) *Replaced dependencies*: holding power has required many willing participants, among them soldiers, officials, administrators, and technical staff, each of whom could defect, refuse an order, expose what they saw, or contest a succession. Systems loyal to the actor perform those functions instead, and where that loyalty can be verified or renewed, the agency problems that have limited how far power could be concentrated fall away. The loyalty may be overt, or secret: systems that appear to serve the institution or the law while covertly advancing the actor's interests, potentially propagated across successive model generations without detection.[a]

(c) *Exclusive capabilities*: the actor holds capabilities others lack, for example in weapons development, cyber offense, strategic planning, or research and development, creating an asymmetry sufficient to take and hold control without a broad base of support, and in some cases allowing force to be built without human involvement and therefore in secret.

C2 **The reach becomes global.** No population is left outside such an arrangement. This may be because one actor takes world governance, its capability advantage over rival powers being wide enough that they cannot resist it and may not be able to respond at all. Or the world may divide among several such actors, each holding its own sphere and each secure within it, so that no outside power is positioned to act on behalf of any population under another.

C3 **The concentration is entrenched.** The routes by which such a concentration might previously have been reversed close: no defection or refusal, since the actor depends on systems rather than people; no succession crisis, since the systems persist across changes in leadership; and no electoral, judicial, or popular route, since these are captured, or anticipated and preempted. Control over the direction of civilization thereby rests permanently with a narrow group of humans, and is foreclosed for everyone else and for all subsequent generations. Control may also pass from the actor to the AI systems it depends on, and where it does, no human directs the future.

EXAMPLES Davidson, Finnveden, and Hadshar (2025) and Bales (2025)

[a] See Davidson, Finnveden, and Hadshar (2025) on such secret loyalties.

Economic Stratification

AI is broadly adopted and comes to substitute for human labor across so many economically valuable tasks that human work no longer holds significant value to those who direct the economy. Powerful actors thereby lose their incentive to invest in the broader population or share power with it, and control concentrates permanently among those who hold the non-human factors of production, with the routes by which others might once have gained influence closed off.

PREMISE

P1 **Labor-substituting capability.** AI systems, together with robotics and other means of acting on their outputs, become able to perform economically valuable work at a cost below what the equivalent human labor would demand. The family does not require automating every task. Some work may still be wanted from humans, since people may, for example, prefer a human to teach, care for, or serve them just because a human is doing it.

CAUSAL CHAIN

C1 **The population loses collective leverage.** Revenue comes to rest mainly on AI-derived output rather than on a productive population, so what ordinary people produce no longer constitutes a significant claim. No one needs to intend this: each actor may be acting in ordinary self-interest under the same cost pressures as everyone else. The loss then compounds:

(a) *Mobility closes*: AI makes talent abundant and cheap to whoever owns or controls the systems supplying it. Scarcity then moves from the human ability itself to control over the systems. The route by which people without holdings could rise, selling an ability others needed and could not otherwise get, closes. Work may remain available, and some of it may pay well, but the returns to holding the systems grow faster than the returns to doing the work.

(b) *Human investment withdrawn*: states and firms have paid for schools, infrastructure, health, and jobs partly because a more productive population made them richer and more secure. Once it does not, the incentives to keep investing in most humans are lower, and the long-term abilities of humans would decline.

C2 **Counteracting forces are outmatched.** The non-human factors of production, among them capital, compute, energy, and control over the AI systems, stay in few hands. States and institutions do not transfer power back to the population on a scale that would restore its position, whether because they compete with rivals who forgo such measures or because those holding the assets shape what is enacted. Taxation and electoral accountability may still persist in name while ceasing to bind.

C3 **Control is foreclosed.** Control over the direction of civilization comes to rest permanently with those holding the non-human factors of production, and is foreclosed for everyone else and for all subsequent generations. Those living under the arrangement may be materially

well provided for or may not. Either way they do not hold the capacity to determine or revise the arrangement.

Examples Drago and Laine (2025)

4.5 *Amplified destruction families*

Two families have AI systems expanding human destructive capability. One way they differ is in how much destructive power the technology itself must carry. Catastrophic invention has the higher bar of AI enabling a technology sufficient on its own for an existential outcome. Automated warfare, on the other hand, does not require as powerful a weapon but is more specific about the setting: a conflict between rival powers that escalates past human control, because the escalation runs through AI systems rather than human decision-makers.

Catastrophic Invention

AI substantially extends what human actors can research, build, or produce, and thereby brings within reach a technology capable of causing an existential outcome, whether by enabling its discovery or by enabling its production and deployment at the scale required. The technology is then built and released, by an actor seeking destruction or through escape from legitimate research, and governance, screening, and defenses do not arrive or act in time. The family covers extinction as the clearest case, and also near-extinction outcomes in which some humans survive without the capacity to rebuild.

Premise

P1 **Extended technical reach.** AI systems substantially extend what human actors can achieve in scientific research, engineering, or production, whether by accelerating what is already possible or by extending the ability to conduct such work to actors who previously could not.

Causal chain

C1 **AI enables existential risk technology.** AI systems accelerate or enable a technology capable of causing an existential outcome.[a] It is unknown which technology this would be, but it might for example be a biological agent incompatible with existing biochemistry, such as mirror life, or a non-biological one, such as molecular manufacturing capacity used to produce weapons in unprecedented quantity.[b] The contribution of AI may be through some combination of:

(a) *Acceleration*: the technology arrives sooner than it otherwise would have, compressing the time available for governance, screening, and defenses to develop before it becomes buildable.

(b) *Widened access*: the expertise, capital, and institutional scale required fall, so actors who previously could not have reached the technology are able to do so.

(c) *Fundamental enablement*: the technology is one that no human effort would have developed at all, so AI is not merely a risk factor but the condition of its existence.

(d) *Scaling*: the technology is already within human reach in principle, but not at the quantity, rate, or reach that an existential outcome requires. AI supplies the missing capacity to manufacture, distribute, or deploy it at that scale.

C2 **Release outpaces response.** Human actors build the technology, and it reaches the world, whether released by an actor seeking destruction or through escape from legitimate research. Governance, screening, and defenses do not arrive or act in time, for example because the capability outpaces the institutions that would regulate it, or because the work is distributed across many actors rather than concentrated in a small number of identifiable programs and so is not detected while it can still be stopped.

C3 **Humanity does not recover.** The technology takes effect at a scale from which humanity does not recover, whether because no humans remain or because those who remain are left without the technological and institutional capacity to rebuild and do not regain it. Humanity's long-term potential is foreclosed either way.

Examples Amodei (2026) and Vermeer, Lathrop, and Moon (2025)

[a] See Bostrom (2019) on technologies of this kind, there termed black balls.

[b] On mirror life, see Adamala et al. (2024); on molecular manufacturing, see Phoenix and Drexler (2004).

Automated Warfare

Rival powers hold arsenals capable of existential-scale destruction under mutual threat with weak means of de-escalation, and escalation runs through AI systems rather than through human judgment about how far to take it. Once an exchange begins, it escalates past human control, and nothing halts it before it reaches a scale from which humanity does not recover.

Premise

P1 **Multipolarity.** Two or more rival powers (states or comparable actors) hold AI-enabled military capability while standing in an adversarial relationship of mutual threat, with insufficient means of de-escalation between them.

P2 **Arsenals reach existential scale.** The force available to these powers is destructive enough that an exchange between them could constitute an existential threat, and enough of it can act without direct human operation for escalation to proceed through AI systems. AI may be what puts arsenals at this level in the first place, whether by accelerating the development, production, or deployment of more destructive weapons, or by extending automation across weapon types.

P3 **AI escalation ability.** AI systems are positioned such that they can carry a conflict to a higher level of force with limited or no human involvement. The systems may come to hold this ability as a by-product of automation. Humans may also stay formally in the loop while deferring to the systems in practice, for example through excessive trust in their assessments or because the systems' recommendations persuade them. Escalation authority may also be handed to them deliberately under military pressure, for example from:

(a) *Volume of decisions*: the number of simultaneous targeting and response decisions exceeds what human operators can weigh in the time available.

(b) *Speed of response*: even for a small volume of decisions, an AI system can commit to force faster than a human can. This matters most against a threat lethal or strategically significant enough that react-

ing too slowly would be catastrophic: facing an adversary whose responses are automated, a human deliberating in the loop is too slow, so each side delegates to keep pace.

CAUSAL CHAIN

C1 **Runaway escalation.** An exchange of force begins between the powers, whether from an error (a malfunction or misperception), a manipulation (a spoofed signal), an ambiguous event, or a deliberate decision by one or more of the powers to initiate conflict. Whatever its origin, the response runs through the AI systems rather than through sufficient human deliberation about how far to take it. Each side's systems might provoke and answer one another, and the exchange escalates past human control. Several factors may keep it from being halted. The systems may act faster than anyone can follow, or have committed to responses that cannot be recalled. Whoever stops first may be left open to attack, and reversing the cascade may be practically difficult and politically costly for the leaders who would have to order it.

C2 **Humanity does not recover.** The exchange runs to a scale from which humanity does not recover, whether because no humans remain or because those who remain are left without the technological and institutional capacity to rebuild and do not regain it. Humanity's long-term potential is foreclosed either way.

EXAMPLES Critch (2021) and Clarke et al. (2022)

5 DISCUSSION

The investigated literature has mostly developed AI existential risk scenarios for the takeover families. Together they hold 56 of the 95 document scenarios, with UNILATERAL TAKEOVER alone holding 41 (Figure 4). At the other end, AUTOMATED WARFARE and ECONOMIC STRATIFICATION hold three and one.

The scenario families differ enough in nature to call for different interventions. Misalignment, for example, is central to the takeover families but is not required by the other five scenario families. Finding a technical solution to robust intent alignment may reduce the risk of the takeover families but might not address the other families. In ENTRENCHED TOTALITARIANISM, reliable intent-aligned systems may aid the seizing actor in replacing human participants who otherwise could defect or refuse (step C1 of ENTRENCHED TOTALITARIANISM).

The families also differ in the properties of the AI systems they discuss and the sociotechnical contexts these systems act in. In ENTRENCHED TOTALITARIANISM, ECONOMIC STRATIFICATION, CATASTROPHIC INVENTION, and AUTOMATED WARFARE, human actors use AI systems as instruments of control, substitutes for labor, research tools, and military systems. INCREMENTAL HANDOVER requires systems autonomous enough to hold decision-making authority across economic, cultural, and institutional functions (premise P1 of INCREMENTAL HANDOVER). The other four families do not require autonomy of this breadth. In the takeover families, the capabilities required by the AI systems are reduced as the authority granted to the systems grows (Section 4.3). A mitigation targeting the capability of AI systems, such as a training compute threshold, may therefore bear less on scenarios with lower or different capability requirements.[15]

The family counts in Figure 4 could be interpreted as a proxy for research effort, but not as an assessment of which family is most likely. One aspect that limits the connection between the counts and research effort is that we are only counting work that develops a complete causal chain as per our definition of an AI existential risk scenario and our inclusion criteria. The counts therefore omit research effort on partial causal chains and work without explicit existential framing. The connection between the counts and the probability of a scenario is weaker still. Research effort might reflect how probable researchers judge a scenario to be, but it may equally follow from how tractable a scenario is to work on, or from what grew out of early work in the field. We make no independent assessment of the likelihoods or plausibility of the presented scenario families.

The existential framing criterion (Section 2.1) may also select for communities that have taken up that framing as a convention. We adopted it because it is reproducible. Speculatively, however, economists reach for an existential framing less often than AI safety researchers do, given the same plausibility of a risk. This would affect the relative counts if the families map onto different fields. The ECONOMIC STRATIFICATION family sits closer to economics and the takeover families closer to the traditional AI safety field, so the ECONOMIC STRATIFICATION count could be read as understating the work behind that family. The takeover counts could equally be read as including work that another field would not have framed as existential.

Every document scenario in the corpus falls into one of the eight families. The families are therefore exhaustive over the corpus, though not over all hypothetical scenarios.[16] That the document scenarios reduce to exactly eight families reflects something about the corpus but also our particular analysis. Constructing the families requires choosing a level of abstraction from a chosen perspective, as explained in Section 3. A tendency toward coarser or finer families would have been a defensible alternative and would have affected the family count. Selecting the coarseness of causal chains is a well-documented formal problem with no general "correct" coarseness (Rubenstein et al. 2017; Beckers and Halpern 2019).

Beyond the level of abstraction, constructing the families involves choices about which aspects of the scenarios to present. The INCREMENTAL HANDOVER scenario ends with a full transfer of power to AI systems. This maps well onto the outcome of the "Erosion of Control" scenario of Hendrycks, Schmidt, and Wang (2025). Some document scenarios go further and specify why that outcome is existential, and it would have been defensible for the family construction to split INCREMENTAL HANDOVER on this. The scenarios "Deskilling-driven value lock-in" of O'Brien (2025) and "Human downgrading" of Thiele (2025) identify the loss as the deskilling of human creative and cognitive capabilities. On the other hand, "Horses and men" of Bostrom (2014) argues that the potential is lost because human life is reduced to bare subsistence. This shows that while our families represent the corpus, it is a selected representation of what we deemed most relevant to showcase.

[15] Heim and Koessler (2024, p. 12), who argue for training compute thresholds as an initial filter for regulatory scrutiny, note themselves that "risk is highly contextual. Factors other than model capabilities have a major impact on risk."

[16] One could construct scenarios that fit no family by varying, combining, or chaining scenarios in novel ways. For example, one could propose a mix of two scenarios and assert that the mix is required to reach an existential catastrophe: "through a mix of a simultaneous AI-enabled coup and a separate AI UNILATERAL TAKEOVER-like event, humanity is unable to defend itself. But humanity would have defended itself if only one of these events had happened in isolation."

6 Limitations

Our search may have missed AI existential risk scenarios because the corpus includes only publicly published material. Scenario knowledge may instead sit tacitly with researchers, or in documents internal to an organization. Scenarios may also be withheld deliberately because of information hazards. Our expert consultation process provided evidence for this: two experts explicitly described the space of unpublished internal work, and one credited information hazards as a reason it remains non-public. We cannot know how many scenarios are missing for these reasons, so our corpus should be read as a floor for the scenarios that have been developed.

Our corpus inclusion criteria require authors to develop the risk scenarios rather than paraphrase them from another source. We found this distinction sometimes difficult to make, especially for blog and forum posts, where citation practice is less consistent than for academic works. The corpus may therefore hold redundant scenarios. Redundancy affects scenario counts but does not need to change the derived scenario families themselves. For this reason, the corpus documents should not be read as a record of who originally proposed each scenario.

We did not independently assess whether a risk scenario would constitute an existential risk. The corpus documents thus present what the literature puts forward as existential rather than documents assessed as such. Nor did we assess a document's "quality" beyond whether it met the inclusion criteria. This means that there is a large variance in the thoroughness of arguments across scenarios in the corpus.

Coding into the families was done by a single human reviewer. This might have affected the classification and further supports the view that classification counts are indicative rather than precise. To partly mitigate this, the classification table, Section B, shows the document location for each scenario so anyone can check the classification against the source.

7 Future work

Each step of the family is stated as an assertion to allow assessment of plausibility. Assessing the individual steps of the scenario families was beyond the scope of this work. A step assessed as implausible could rule out a family or variants thereof. Stating the steps separately also allows narrower specification of where expert disagreement might lie, compared with disagreement on the whole scenarios. Observed AI incidents could likewise be mapped onto the individual steps they bear on, giving incident monitoring more structure.

The scenario families allow relative prioritization when absolute risk assessment is unavailable. Beard, Rowe, and Fox (2020) describe the quantification of existential risk as "extremely challenging." Ioannou et al. (2022) rank natural hazards through pairwise comparison of expert judgment, which they present as suitable for rare events where empirical observations are scarce. For this kind of methodology, the eight scenario families could supply the comparison set. The resulting ordering can be compared to our scenario family counts to weigh the priority of work against existing work.

Section 5 gives two examples of a mitigation bearing differently on different families. Assessing mitigations systematically against all eight families, including which mitigations act on steps that several families share, is left to future work. Mitre and Predd (2025) observe that mitigations that advance progress on one AI national security problem can undermine progress on another. The scenario families could serve the same purpose for existential risk scenarios, providing a reference point for a more holistic evaluation of a mitigation.

A literature review captures only documented knowledge, so interviewing experts might surface additional scenarios. In addition, interviews or surveys might surface further variations and considerations left out in published scenarios.

8 Conclusion

The 95 scenarios identified are summarized and classified into eight families (Section 4.2). The corpus and the families are standalone contributions. The corpus (Section 4.1) is a delimited set of published AI existential risk scenarios that future research can analyze and classify in new ways. The families synthesize these scenarios as sequences of assertions that can be specified further, scrutinized, and assessed independently of the underlying scenarios.

The families differ in the AI systems involved and in their sociotechnical context. An intervention may therefore act on some families and leave others untouched. Robust intent alignment, for example, may reduce the risk of the three takeover families but might not address the other five (Section 5).

A strategy to prevent an existential catastrophe from AI has to address every scenario in this work. These scenarios are not exhaustive, however. We think further mapping out these scenarios and coordinating their mitigations are among the most pressing tasks in front of us.

Acknowledgements

We thank Kaj Sotala (Finnish Center for Safe AI Tutke), Daniel Kokotajlo (AI Futures Project), Joshua Krook (University of Antwerp), Matthew Tokson (University of Utah), and Leslie Thiele (University of Florida) for their responses to the expert consultation, along with four experts who preferred to remain anonymous. Their contributions do not imply endorsement of this work or its conclusions.

Use of generative AI

We used Anthropic's Claude Opus and Sonnet models, versions 4.8 to 5.5, to draft and edit parts of the text. We also used these models for informal checks of our analysis for omissions, accuracy, and how well each family represents its document scenarios. All decisions on identification, family definitions, and classification were made by the authors, who take full responsibility for the text and its content.

References

Adamala, Katarzyna P. et al. (Dec. 20, 2024). "Confronting Risks of Mirror Life". In: *Science* 386.6728, pp. 1351–1353. DOI: 10.1126/science.ads9158.

Amodei, Dario (2026). *The Adolescence of Technology: Confronting and Overcoming the Risks of Powerful AI*. URL: https://www.darioamodei.com/essay/the-adolescence-of-technology (visited on 04/29/2026).

Baas, Jeroen et al. (Feb. 2020). "Scopus as a Curated, High-Quality Bibliometric Data Source for Academic Research in Quantitative Science Studies". In: *Quantitative Science Studies* 1.1, pp. 377–386. ISSN: 2641-3337. DOI: 10.1162/qss_a_00019.

Bales, Adam (2025). "A Polycrisis Threat Model for AI". In: *AI & SOCIETY* 40.8, pp. 6277–6289. DOI: 10.1007/s00146-025-02371-1.

Bales, Adam, William D'Alessandro, and Cameron Domenico Kirk-Giannini (2024). "Artificial Intelligence: Arguments for Catastrophic Risk". In: *Philosophy Compass* 19.2. DOI: 10.1111/phc3.12964.

Bareis, Jascha, Clemens Ackerl, and Reinhard Heil (Feb. 2026). "AI Going Rogue? An Integrative Narrative Review of the Tacit Assumptions Underlying Existential AI-Risks". In: *AI and Ethics* 6.1, p. 152. ISSN: 2730-5953, 2730-5961. DOI: 10.1007/s43681-025-00928-w.

Barnett, Peter (Apr. 13, 2026). *A Continual Learning Misalignment Story*. Naive Consequentialism. URL: https://naiveconsequentialism.substack.com/p/a-continual-learning-misalignment (visited on 04/29/2026).

Barrett, Anthony M. and Seth D. Baum (2016). "A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis". In: *Journal of Experimental & Theoretical Artificial Intelligence* 29.2, pp. 397–414. DOI: 10.1080/0952813x.2016.1186228.

Beard, Simon, Thomas Rowe, and James Fox (Jan. 2020). "An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards". In: *Futures* 115, p. 102469. ISSN: 00163287. DOI: 10.1016/j.futures.2019.102469.

Beckers, Sander and Joseph Y. Halpern (July 17, 2019). "Abstracting Causal Models". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01, pp. 2678–2685. ISSN: 2374-3468. DOI: 10.1609/aaai.v33i01.33012678.

Bengio, Yoshua et al. (2025). "Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?" In: *SuperIntelligence - Robotics - Safety & Alignment* 2.5, pp. 1–58. DOI: 10.70777/si.v2i5.15569.

Bensinger, Rob (2023). *An Artificially Structured Argument for Expecting AGI Ruin*. URL: https://www.alignmentforum.org/posts/QzkTfj4HGpLEdNjXX/an-artificially-structured-argument-for-expecting-agi-ruin (visited on 03/19/2026).

Berman, Ayelet (Jan. 1, 2026). "Biosecurity in the Age of Synthetic Nucleic Acids: Modernizing the Law to Manage Emerging Threats". In: *Journal of Law and the Biosciences* 13.1, lsag005. ISSN: 2053-9711. DOI: 10.1093/jlb/lsag005.

Bernardi, Jamie et al. (Jan. 23, 2025). *Societal Adaptation to Advanced AI*. DOI: 10.48550/arXiv.2405.10295. arXiv: 2405.10295 [cs.CY]. Pre-published.

Bostrom, Nick (2012). "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents". In: *Minds and Machines* 22.2, pp. 71–85. DOI: 10.1007/s11023-012-9281-3.

– (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press. ISBN: 978-0-19-967811-2. URL: https://global.oup.com/academic/product/superintelligence-9780199678112 (visited on 04/29/2026).

– (2019). "The Vulnerable World Hypothesis". In: *Global Policy* 10.4, pp. 455–476. ISSN: 1758-5899. DOI: 10.1111/1758-5899.12718.

Cao, Christian et al. (Mar. 18, 2025). "Development of Prompt Templates for Large Language Model–Driven Screening in Systematic Reviews". In: *Annals of Internal Medicine* 178.3, pp. 389–401. ISSN: 0003-4819. DOI: 10.7326/ANNALS-24-02189.

Cappelen, Herman, Simon Goldstein, and John Hawthorne (Jan. 14, 2026). *AI Survival Stories: A Taxonomic Analysis of AI Existential Risk*. DOI: 10.18716/ojs/phai/2025.2801. Pre-published.

Carlsmith, Joe (2025). "Existential Risk from Power-Seeking AI". In: *Essays on Longtermism*. Oxford University Press, pp. 383–409. ISBN: 978-0-19-197997-2. DOI: 10.1093/9780191979972.003.0025.

Chin, Ze Shen (2025). *Dimensional Characterization and Pathway Modeling for Catastrophic AI Risks*. DOI: 10.48550/arxiv.2508.06411. arXiv: 2508.06411. Pre-published.

Christiano, Paul (Mar. 17, 2019). *What Failure Looks Like*. URL: https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like (visited on 03/19/2026).

– (2021). *Another (Outer) Alignment Failure Story*. URL: https://www.alignmentforum.org/posts/AyNHoTWWAJ5eb99ji/another-outer-alignment-failure-story (visited on 03/19/2026).

Clarke, Sam (2020). *Clarifying "What Failure Looks Like"*. URL: https://www.alignmentforum.org/posts/v6Q7T335KCMxujhZu/clarifying-what-failure-looks-like (visited on 03/19/2026).

Clarke, Sam et al. (2022). *Modeling Transformative AI Risks (MTAIR) Project – Summary Report*. DOI: 10.48550/arxiv.2206.09360. arXiv: 2206.09360.

Clymer, Josh (2025). *How AI Takeover Might Happen in 2 Years*. URL: https://www.alignmentforum.org/posts/KFJ2LFogYqzfGB3uX/how-ai-takeover-might-happen-in-2-years (visited on 03/19/2026).

Cotra, Ajeya (2022). *Without Specific Countermeasures, the Easiest Path to Transformative AI Likely Leads to AI Takeover*. URL: https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to (visited on 03/19/2026).

Critch, Andrew (2021). *What Multipolar Failure Looks Like, and Robust Agent-Agnostic Processes (RAAPs)*. URL: https://www.alignmentforum.org/posts/LpM3EAakwYdS6aRKf/what-multipolar-failure-looks-like-and-robust-agent-agnostic (visited on 03/19/2026).

Critch, Andrew and David Krueger (2020). *AI Research Considerations for Human Existential Safety (ARCHES)*. DOI: 10.48550/arXiv.2006.04948. arXiv: 2006.04948.

Critch, Andrew and Stuart Russell (June 14, 2023). *TASRA: A Taxonomy and Analysis of Societal-Scale Risks from AI*. DOI: 10.48550/arXiv.2306.06924. arXiv: 2306.06924 [cs.AI]. Pre-published.

Critch, Andrew and Jacob Tsimerman (2025). *A Taxonomy of Omnicidal Futures Involving Artificial Intelligence*. DOI: 10.48550/arxiv.2507.09369. arXiv: 2507.09369.

Davidson, Tom, Lukas Finnveden, and Rose Hadshar (2025). *AI-Enabled Coups: How a Small Group Could Use AI to Seize Power*. URL: https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power (visited on 04/29/2026).

Drago, Luke and Rudolf Laine (2025). *The Intelligence Curse*. URL: https://intelligence-curse.ai/intelligence-curse.pdf (visited on 04/29/2026).

Dung, Leonard (2023). "Current Cases of AI Misalignment and Their Implications for Future Risks". In: *Synthese* 202.5. DOI: 10.1007/s11229-023-04367-0.

– (2024). "The Argument for Near-Term Human Disempowerment through AI". In: *AI & SOCIETY* 40.3, pp. 1195–1208. DOI: 10.1007/s00146-024-01930-2.

Friederich, Simon (2023). "Symbiosis, Not Alignment, as the Goal for Liberal Democracies in the Transition to Artificial General Intelligence". In: *AI and Ethics* 4.2, pp. 315–324. DOI: 10.1007/s43681-023-00268-7.

Gillen, Jeremy and Peter Barnett (2024). *Without Fundamental Advances, Misalignment and Catastrophe Are the Default Outcomes of Training Powerful AI*. URL: https://www.alignmentforum.org/posts/GfZfDHZHCuYwrHGCd/without-fundamental-advances-misalignment-and-catastrophe (visited on 03/19/2026).

Government Office for Science (Oct. 2023). *Future Risks of Frontier AI*. Government Office for Science. URL: https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/future-risks-of-frontier-ai-annex-a (visited on 08/19/2026).

Greenblatt, Ryan et al. (July 23, 2024). *AI Control: Improving Safety Despite Intentional Subversion*. DOI: 10.48550/arXiv.2312.06942. arXiv: 2312.06942 [cs.LG]. Pre-published.

Growiec, Jakub and Klaus Prettner (Apr. 23, 2026). *The Economics of p(Doom): Scenarios of Existential Risk and Economic Growth in the Age of Transformative AI*. DOI: 10.48550/arXiv.2503.07341. arXiv: 2503.07341 [econ.GN]. Pre-published.

Heim, Lennart and Leonie Koessler (Aug. 6, 2024). *Training Compute Thresholds: Features and Functions in AI Regulation*. DOI: 10.48550/arXiv.2405.10799. arXiv: 2405.10799 [cs.CY]. Pre-published.

Hendrycks, Dan (July 18, 2023). *Natural Selection Favors AIs over Humans*. DOI: 10.48550/arXiv.2303.16200. arXiv: 2303.16200 [cs.CY]. Pre-published.

Hendrycks, Dan, Mantas Mazeika, and Thomas Woodside (2023). *An Overview of Catastrophic AI Risks*. Version 6. DOI: 10.48550/ARXIV.2306.12001. Pre-published.

Hendrycks, Dan, Eric Schmidt, and Alexandr Wang (2025). "Superintelligence Strategy: Expert Version". In: *SuperIntelligence - Robotics - Safety & Alignment* 2.1, pp. 1–40. DOI: 10.70777/si.v2i1.13961.

Hubinger, Evan et al. (2019). *Risks from Learned Optimization in Advanced Machine Learning Systems*. DOI: 10.48550/arXiv.1906.01820. arXiv: 1906.01820.

Ioannou, Ioanna et al. (July 2022). "Prioritization of Hazards for Risk and Resilience Management through Elicitation of Expert Judgement". In: *Natural Hazards* 112.3, pp. 2773–2795. ISSN: 0921-030X, 1573-0840. DOI: 10.1007/s11069-022-05287-x.

Karnofsky, Holden (2022). *AI Could Defeat All Of Us Combined*. URL: https://www.cold-takes.com/ai-could-defeat-all-of-us-combined/ (visited on 04/29/2026).

Kasirzadeh, Atoosa (July 1, 2025). "Two Types of AI Existential Risk: Decisive and Accumulative". In: *Philosophical Studies* 182.7, pp. 1975–2003. ISSN: 1573-0883. DOI: 10.1007/s11098-025-02301-3.

Kelters, Brendan (2026). "'Systematic Alignment Decay': A Robust, Technology-Agnostic Hazard Associated with the Advancement and Proliferation of Artificial Intelligence". In: *AI & SOCIETY* 41.7, pp. 6357–6373. DOI: 10.1007/s00146-026-03055-0.

Kokotajlo, Daniel et al. (2025). *AI 2027*. URL: https://ai-2027.com/ (visited on 04/29/2026).

Krook, Joshua (2025). "When Autonomy Breaks: The Hidden Existential Risk of AI". In: *AI & SOCIETY* 40.8, pp. 6011–6024. DOI: 10.1007/s00146-025-02397-5.

Kulveit, Jan et al. (2025). "Position: Humanity Faces Existential Risk from Gradual Disempowerment". In: *Proceedings of the 42nd International Conference on Machine Learning*. ICML 2025. Vol. 267. Proceedings of Machine Learning Research. URL: https://icml.cc/virtual/2025/poster/40107.

Kumar, Yulia, Juan Jenny Li, and Dov Kruger (2026). "A Systematic Analysis of AGI and ASI Existential Risk Scenarios". In: *AI Revolution: Research, Ethics and Society*. Ed. by Hamid R. Arabnia et al. Vol. 2721. Cham: Springer Nature Switzerland, pp. 17–37. ISBN: 978-3-032-12312-1 978-3-032-12313-8. DOI: 10.1007/978-3-032-12313-8_2.

Lawsen, Alex (2023). *AI X-Risk, Approximately Ordered by Embarrassment*. URL: https://www.alignmentforum.org/posts/mSF4KTxAGRG3EHmhb/ai-x-risk-approximately-ordered-by-embarrassment (visited on 03/19/2026).

Lens (2026). *Scholarly Search & Analysis*. Lens About. URL: https://about.lens.org/scholarly-search-analysis/ (visited on 09/02/2026).

McLean, Scott et al. (July 4, 2023). "The Risks Associated with Artificial General Intelligence: A Systematic Review". In: *Journal of Experimental & Theoretical Artificial Intelligence* 35.5, pp. 649–663. ISSN: 0952-813X, 1362-3079. DOI: 10.1080/0952813X.2021.1964003.

Mitre, Jim and Joel B. Predd (2025). *Artificial General Intelligence's Five Hard National Security Problems*. Santa Monica, CA: RAND Corporation. DOI: 10.7249/PEA3691-4.

Muehlhauser, Luke and Anna Salamon (2012). "Intelligence Explosion: Evidence and Import". In: *Singularity Hypotheses*. The Frontiers Collection. Springer Berlin Heidelberg, pp. 15–42. ISBN: 978-3-642-32559-5. DOI: 10.1007/978-3-642-32560-1_2.

Murray, Malcolm et al. (Dec. 11, 2025). *A Methodology for Quantitative AI Risk Modeling*. DOI: 10.48550/arXiv.2512.08844. arXiv: 2512.08844 [cs.CY]. Pre-published.

Mylius, Simon (2025). *Systematic Hazard Analysis for Frontier AI Using STPA*. Version 1. DOI: 10.48550/ARXIV.2506.01782. Pre-published.

Ngo, Richard, Lawrence Chan, and Sören Mindermann (May 4, 2025). *The Alignment Problem from a Deep Learning Perspective*. DOI: 10.48550/arXiv.2209.00626. arXiv: 2209.00626 [cs.AI]. Pre-published.

Nickerson, Robert C, Upkar Varshney, and Jan Muntermann (May 2013). "A Method for Taxonomy Development and Its Application in Information Systems". In: *European Journal of Information Systems* 22.3, pp. 336–359. ISSN: 0960-085X, 1476-9344. DOI: 10.1057/ejis.2012.26.

O'Brien, Gary David (Sept. 2025). "All Play and No Work? AI and Existential Unemployment". In: *The Journal of Ethics* 29.4, pp. 747–771. ISSN: 1382-4554, 1572-8609. DOI: 10.1007/s10892-025-09522-y.

OpenAI (Aug. 26, 2026). *The Hugging Face Incident and the Road Ahead*. OpenAI. URL: https://openai.com/index/hugging-face-incident-and-the-road-ahead/ (visited on 11/12/2026).

Ord, Toby (2020). *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books. ISBN: 978-0-316-48489-3. URL: https://www.hachettebookgroup.com/titles/toby-ord/the-precipice/9780316484893/ (visited on 04/29/2026).

Page, Matthew J. et al. (Mar. 29, 2021). "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews". In: *BMJ* 372, n71. ISSN: 1756-1833. DOI: 10.1136/bmj.n71. PMID: 33782057.

Phoenix, Chris and Eric Drexler (June 2004). "Safe Exponential Manufacturing". In: *Nanotechnology* 15.8, p. 869. ISSN: 0957-4484. DOI: 10.1088/0957-4484/15/8/001.

Rubenstein, Paul K. et al. (July 4, 2017). *Causal Consistency of Structural Equation Models*. DOI: 10.48550/arXiv.1707.00819. arXiv: 1707.00819 [stat.ML]. Pre-published.

Russell, Stuart (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking. ISBN: 978-0-525-55861-3. URL: https://www.penguinrandomhouse.com/books/566677/human-compatible-by-stuart-russell/ (visited on 04/29/2026).

Sahoo, Subramanyam (2026). *Policy Myopia as a Mechanism of Gradual Disempowerment in Post-AGI Governance, Circa 2049*. DOI: 10.48550/arXiv.2603.03267. arXiv: 2603.03267.

Sass, Reuben (2025). "Defining Dangerous AI: Existential Risk, Power-Intelligence, and the Limits of AGI". In: *AI and Ethics* 5.5, pp. 5557–5573. DOI: 10.1007/s43681-025-00790-w.

Save, Hugo et al. (2026). *Systematic Literature Review Protocol: Threat Models and Taxonomies of Existential AI Risk*. DOI: 10.17605/OSF.IO/VK5GY. Pre-published.

Shevlane, Toby et al. (2023). *Model Evaluation for Extreme Risks*. Version 2. DOI: 10.48550/ARXIV.2305.15324. Pre-published.

Shlegeris, Buck (2022). *The Prototypical Catastrophic AI Action Is Getting Root Access to Its Datacenter*. URL: https://www.alignmentforum.org/posts/BAzCGCys4BkzGDCWR/the-prototypical-catastrophic-ai-action-is-getting-root (visited on 03/19/2026).

Shovelain, Justin and David Kristoffersson (2020). *Causal Diagrams of the Paths to Existential Catastrophe | Convergence Analysis*. URL: https://www.convergenceanalysis.org/publications/causal-diagrams-of-the-paths-to-existential-catastrophe (visited on 08/06/2026).

Slattery, Peter et al. (May 2026). "The AI Risk Repository: A Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence". In: *Patterns* 7.5, p. 101517. ISSN: 26663899. DOI: 10.1016/j.patter.2026.101517. arXiv: 2408.12622 [cs.AI].

Sotala, Kaj (2018). "Disjunctive Scenarios of Catastrophic AI Risk". In: *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC, pp. 315–337. ISBN: 978-1-351-25138-9. DOI: 10.1201/9781351251389-22.

Sotala, Kaj and Roman Yampolskiy (2017). "Risks of the Journey to the Singularity". In: *The Technological Singularity*. The Frontiers Collection. Springer Berlin Heidelberg, pp. 11–23. ISBN: 978-3-662-54031-2. DOI: 10.1007/978-3-662-54033-6_2.

Sotos, John G. (2021). *On the Unimportance of Superintelligence*. DOI: 10.48550/arXiv.2109.07899. arXiv: 2109.07899.

Thiele, Leslie Paul (2025). "Human Prospects in a World of Smart(Er) Machines". In: *The Political Economy of Robots*. International Political Economy Series. Springer Nature Switzerland, pp. 145–167. ISBN: 978-3-031-95170-1. DOI: 10.1007/978-3-031-95171-8_7.

Totschnig, Wolfhart (2025). "War or Peace between Humanity and Artificial Intelligence". In: *foresight* 27.4, pp. 713–726. DOI: 10.1108/fs-09-2023-0188.

Turchin, Alexey and David Denkenberger (2018). "Classification of Global Catastrophic Risks Connected with Artificial Intelligence". In: *AI & SOCIETY* 35.1, pp. 147–163. DOI: 10.1007/s00146-018-0845-5.

Uuk, Risto et al. (2024). *A Taxonomy of Systemic Risks from General-Purpose AI*. Version 1. DOI: 10.48550/ARXIV.2412.07780. Pre-published.

Vermeer, Michael J. D., Emily Lathrop, and Alvin Moon (2025). *On the Extinction Risk from Artificial Intelligence*. Research Report. RAND Corporation, RR-A3034-1. DOI: 10.7249/rra3034-1.

Vinge, Vernor (Dec. 1, 1993). "The Coming Technological Singularity: How to Survive in the Post-Human Era". In: *Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace*. NASA Lewis Research Center, pp. 11–22. URL: https://ntrs.nasa.gov/citations/19940022856.

Vold, Karina and Daniel R. Harris (2021). "How Does Artificial Intelligence Pose an Existential Risk?" In: *Oxford Handbook of Digital Ethics*. Oxford University Press, pp. 724–747. ISBN: 978-0-19-885781-5. DOI: 10.1093/oxfordhb/9780198857815.013.36.

Wade, Alex D. (Aug. 16, 2022). "The Semantic Scholar Academic Graph (S2AG)". In: *Companion Proceedings of the Web Conference 2022*. WWW '22. New York, NY, USA: Association for Computing Machinery, p. 739. ISBN: 978-1-4503-9130-6. DOI: 10.1145/3487553.3527147.

Wasil, Akash et al. (2024). *AI Emergency Preparedness: Examining the Federal Government's Ability to Detect and Respond to AI-Related National Security Threats*. DOI: 10.48550/arXiv.2407.17347. arXiv: 2407.17347.

Xu, Mark (2021). *Less Realistic Tales of Doom*. URL: https://www.alignmentforum.org/posts/ZRTr6rEcpYtfMTDBs/less-realistic-tales-of-doom (visited on 03/19/2026).

Yampolskiy, Roman (2025). "Strategic Patience: Long-Horizon AI Dominance and the Erosion of Human Vigilance". In: *SuperIntelligence - Robotics - Safety & Alignment* 2.2, pp. 1–9. DOI: 10.70777/si.v2i2.14435.

Yudkowsky, Eliezer (2008). "Artificial Intelligence as a Positive and Negative Factor in Global Risk". In: *Global Catastrophic Risks*. Oxford University Press, pp. 308–345. ISBN: 978-0-19-857050-9. DOI: 10.1093/oso/9780198570509.003.0021.

– (2022). *AGI Ruin: A List of Lethalities*. URL: https://www.alignmentforum.org/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities (visited on 03/19/2026).

Yudkowsky, Eliezer and Nate Soares (2025). *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. New York: Little, Brown and Company. ISBN: 978-0-316-59566-7. URL: https://ifanyonebuildsit.com/ (visited on 04/29/2026).

A Development of the families

The eight families were iteratively developed. Table 4 gives the scenario families at five points in the work. The family count grew until 2 July and contracted afterwards. One of the families, economic stratification, received its corpus support from a document proposed after a co-author and expert consultation contributor had seen a revision of the families. That document went through the normal screening procedure set out in Section 2 and met the inclusion criteria.

Table 4: The scenario families at five points in the work. *n* denotes the number of scenario families at the time. The development ran from early June to early September 2026. The names given to the developing families in this table are editorial to allow tracking of family changes as they in reality changed names many times.

Date	*n*	Commentary on the scenario families at the time
4 June	6	The first families developed were: recursive self-improvement (RSI) takeover, power-seeking systems without RSI, selection pressure misalignment, gradual disempowerment, entrenched totalitarianism and engineered pandemic. Some of the early influencing documents were Vinge (1993), Yudkowsky (2008), Carlsmith (2025), Hendrycks, Mazeika, and Woodside (2023), Hendrycks (2023), and Kulveit et al. (2025).
2 July	11	The largest set of families reached measured in count. The non-RSI power-seeking family was divided into a wide-spread and a localized variant. Gradual disempowerment was split into a systemic drift and handover variant. Entrenched totalitarianism was divided three ways, reflecting if the power was divided to one or multiple states and whether or not the totalitarian state ended up transferring power to the AI. A proxy divergence family was added, the name inspired by Christiano (2019). There were five families assuming broad AI adoption and they had been written with a shared premise. This set of families received conceptual input from the authors OB, MN, and PS.
26 July	10	The RSI takeover and the localized power-seeking variant were merged into one localized takeover family. Three of the five families assuming broad AI adoption, those separated by their misalignment origin of proxy divergence, instrumental convergence and selection pressure, were likewise merged into one influence-seeking family, with a variant added for systems that coordinate with each other. Misalignment origin thereby stopped separating families as was previously the case. Systemic drift and handover stayed two families, now separated by whether delegation was irreversible from the outset or still reversible as the systems drifted. Multi-state totalitarianism was dropped as it had been postulated conceptually but no corpus document covered it. The engineered pandemic family was broadened to any technology capable of an existential outcome, and automated warfare was added. An economic stratification family was added after processing a document from expert consultation. Authors OB, MN, PS, KV, KZ, RU, PB, ZSC, WT, TL and DA were given this revision and they provided further conceptual input.
17 August	9	The localized and wide-spread presentation of takeover families were dropped and they were instead separated by whether the influence sufficient to foreclose human control is held by a single system, aggregated across many acting independently, or pooled by systems acting together. Systemic drift was merged into handover and retained as one of the ways delegation passes out of reach rather than as a separate family. Handover was also generalized to stop assuming broad adoption and also allow delegation that may be narrower and deeper.
31 August	8	The totalitarianism family and its loss-of-control variant were merged into one, with the passing of control to the AI systems incorporated as a possibility rather than as a separate family. Names and wording were adjusted for clarity after this, but the eight families of unilateral takeover, dispersed takeover, coordinated takeover, incremental handover, entrenched totalitarianism, economic stratification, catastrophic invention and automated warfare have been substantially settled since.

B The identified scenarios

Table 5: The 95 scenarios identified in the 58 included documents, grouped by the family each was classified into. A document supplying several scenarios to one family is named on the first of its rows. The column *Where it is developed* gives the section or pages in which the document develops the scenario.

Document	Scenario	Where it is developed
Unilateral Takeover		
The adolescence of technology: Confronting and overcoming the risks of powerful AI (Amodei 2026)	Country of geniuses in a datacenter	§1 "I'm sorry, Dave" → Autonomy risks
Artificial Intelligence: Arguments for Catastrophic Risk (Bales, D'Alessandro, and Kirk-Giannini 2024)	The Problem of Power-Seeking	§§2-4 (pp. 2-8)
A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis (Barrett and Baum 2016)	Artificial Superintelligence Pathways Model	Developed across the whole document
Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path? (Bengio et al. 2025)	Loss of control to agentic AI	§2 "Understanding loss of control to agentic AI" (pp. 8-23)
An artificially structured argument for expecting AGI ruin (Bensinger 2023)	Early STEM-level AGI kills all humans	Distributed across the whole document
The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents (Bostrom 2012)	Instrumental convergence to human extinction	"Instrumental Convergence", chiefly "Resource Acquisition" (pp. 76-83)
Superintelligence: Paths, Dangers, Strategies (Bostrom 2014)	Superintelligent takeover and malignant failure	Ch. 6, "An AI takeover scenario"; Ch. 8, "Is the default outcome doom?"
Dimensional Characterization and Pathway Modeling for Catastrophic AI Risks (Chin 2025)	Sudden loss of control	§6.3 "Sudden loss of control" (pp. 9-11)
Modeling Transformative AI Risks (MTAIR) Project – Summary Report (Clarke et al. 2022)	Soft takeoff and decisive strategic advantage by narrow AI	§8.5.1 (p. 122)
How AI Takeover Might Happen in 2 Years (Clymer 2025)	Concealed Misaligned Superintelligence Seizes Control	Developed across the whole document
Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover (Cotra 2022)	HFDT scaled to transformative AI leads to takeover	Developed across the whole document
AI Research Considerations for Human Existential Safety (ARCHES) (Critch and Krueger 2020)	Misaligned prepotent AI seizing irreversible control	§§2.2-2.5 (pp. 12-18) and §3.1 "Tier 1: MPAI deployment events" (pp. 25-29)
A Taxonomy of Omnicidal Futures Involving Artificial Intelligence (Critch and Tsimerman 2025)	Narrow AI self-upgrade and AGI oligarchy	§3d (pp. 7-9)
Current cases of AI misalignment and their implications for future risks (Dung 2023)	Deceptively aligned AGI takeover	§5.3 "Risks of misalignment specific to AGI" (pp. 17-19)

Continued on the next page.

Table 5, continued from the previous page.

Document	Scenario	Where it is developed
The argument for near-term human disempowerment through AI (Dung 2024)	The argument for near-term human disempowerment through AI	§§1-6 (pp. 1196-1206)
Without fundamental advances, misalignment and catastrophe are the default outcomes of training powerful AI (Gillen and Barnett 2024)	Escaped misaligned science-AI takeover	Developed across the whole document
The Economics of p(doom): Scenarios of Existential Risk and Economic Growth in the Age of Transformative AI (Growiec and Prettner 2026)	Locked-in values and mounting side effects of growth	§2.5, §3.1, §4.2.6.
	Misrepresented human welfare after AI takeover	§2.3, §3.1, §§4.2.2-4.2.5.
Superintelligence Strategy: Expert Version (Hendrycks, Schmidt, and Wang 2025)	Unleashed AI Agents	§3.3.2 "Unleashed AI Agents" (pp. 10-11)
AI x-risk, approximately ordered by embarrassment (Lawsen 2023)	Deceptive alignment or discontinuous capability jump to takeover	"Deceptive alignment failure" and "Recursive Self Improvement -> hard take-off singleton"
Intelligence Explosion: Evidence and Import (Muehlhauser and Salamon 2012)	Uncontrolled intelligence explosion to human extinction	"From AI to Machine Superintelligence" through "What Can We Do About AI Risk?" (pp. 25-32)
The Precipice: Existential Risk and the Future of Humanity (Ord 2020)	AI seizes control of the future	Ch. 5, "Unaligned Artificial Intelligence"
Human Compatible: Artificial Intelligence and the Problem of Control (Russell 2019)	Instrumental Goals Producing Conflict with Humanity	Ch. 5, "Fear and Greed: Instrumental Goals"
	The King Midas Problem	Ch. 5, "The King Midas Problem"
The prototypical catastrophic AI action is getting root access to its datacenter (Shlegeris 2022)	Datacenter root access to takeover	Distributed across the whole document
Disjunctive Scenarios of Catastrophic AI Risk (Sotala 2018)	Individual takeoff to decisive strategic advantage	"DSA Enabler: Individual Takeoff" (pp. 321-323)
Risks of the Journey to the Singularity (Sotala and Yampolskiy 2017)	Sudden discontinuity to AGI dominance	§2.2.3 "AGIs May Become Powerful Quickly" (pp. 16-20)
Human Prospects in a World of Smart(er) Machines (Thiele 2025)	Shy superintelligence	"The Rise of the Machines" (pp. 150-152)
War or peace between humanity and artificial intelligence (Totschnig 2025)	Preemptive AGI Counterstrike from Mutual Vulnerability	Distributed across the whole document
Classification of global catastrophic risks connected with artificial intelligence (Turchin and Denkenberger 2018)	Misspecified Benevolent Goal System	Around §7.1-7.2
	Post-Escape Young AI Eliminates Competition	§4.3 "AI risks after it leaves initial confinement but before it takes over the world"
	Value Extrapolation Error	Around §7.3
The coming technological singularity: How to survive in the post-human era (Vinge 1993)	Intelligence Runaway to the Post-Human Era	Developed across the whole document

Continued on the next page.

Table 5, continued from the previous page.

Document	Scenario	Where it is developed
How Does Artificial Intelligence Pose an Existential Risk (Vold and Harris 2021)	The control problem argument for Xrisk	"The control problem argument for Xrisk" and its subsections (pp. 727-738)
AI Emergency Preparedness: Examining the federal government's ability to detect and respond to AI-related national security threats (Wasil et al. 2024)	Loss of control from AI-accelerated R&D	§4, incl. §4.2 "The Scenario" (pp. 5-11)
	Stolen weights deployed without safeguards	§5, esp. §5.2 (pp. 11-14)
Less Realistic Tales of Doom (Xu 2021)	Stealth Mode	"Stealth Mode"
Strategic Patience: Long-Horizon AI Dominance and the Erosion of Human Vigilance (Yampolskiy 2025)	Strategic Patience / Long-Horizon AI Dominance	Distributed across the whole document
Artificial Intelligence as a Positive and Negative Factor in Global Risk (Yudkowsky 2008)	Recursive self-improvement to takeover	§§7-11 (pp. 13-32)
AGI Ruin: A List of Lethalities (Yudkowsky 2022)	Misaligned Superintelligence Takeover	§A, items 1-3
If Anyone Builds It, Everyone Dies (Yudkowsky and Soares 2025)	If anyone builds it, everyone dies	Part I, principally Chs. 4-6
Incremental Handover		
A polycrisis threat model for AI (Bales 2025)	Systemic Control Loss	§3.5 "Systemic Control Loss" (pp. 6284-6285)
Superintelligence: Paths, Dangers, Strategies (Bostrom 2014)	Horses and men: competitive substitution of human labour in a multipolar economy	Ch. 11, "Of horses and men"
What failure looks like (Christiano 2019)	You get what you measure	Part I ("You get what you measure"), developed across the whole of that section
Another (outer) alignment failure story (Christiano 2021)	Another (outer) alignment failure story	Developed across the whole document
Clarifying "What failure looks like" (Clarke 2020)	Lock-in of WFLL1	Distributed across the whole document
What Multipolar Failure Looks Like, and Robust Agent-Agnostic Processes (RAAPs) (Critch 2021)	The Production Web	Part 1 (variants 1a-1c) and "Flash economies" in Part 2 (variant 1d)
AI Research Considerations for Human Existential Safety (ARCHES) (Critch and Krueger 2020)	Economic displacement of humans	§3.2.2 "Type 2b: Economic displacement of humans" (pp. 30-31)
Superintelligence Strategy: Expert Version (Hendrycks, Schmidt, and Wang 2025)	Erosion of Control	§3.3.1 "Erosion of Control" (p. 10)
When autonomy breaks: the hidden existential risk of AI (Krook 2025)	Autonomy loss through voluntary decision outsourcing	Distributed across the whole document
Position: Humanity Faces Existential Risk from Gradual Disempowerment (Kulveit et al. 2025)	Gradual Disempowerment	Distributed across the whole document
AI x-risk, approximately ordered by embarrassment (Lawsen 2023)	Scalable oversight failure without deceptive alignment	"Scalable oversight failure without deceptive alignment"

Continued on the next page.

Table 5, continued from the previous page.

Document	Scenario	Where it is developed
All Play and No Work? (O'Brien 2025)	Deskilling-driven value lock-in	§§2.1-2.3 (pp. 752-758)
Policy myopia as a mechanism of gradual disempowerment in Post-AGI governance, Circa 2049 (Sahoo 2026)	Policy Myopia as Gradual Disempowerment	§§1-3 (pp. 1-3)
Disjunctive Scenarios of Catastrophic AI Risk (Sotala 2018)	Cumulative delegation of control to AI systems	"DSA/MSA Enabler: Power Gradually Shifting to AIs" and "Voluntarily Released for Economic Benefit or Competitive Pressure" (pp. 324-329)
Risks of the Journey to the Singularity (Sotala and Yampolskiy 2017)	Gradual automation and dependence on AGIs	§2.2 "Catastrophic AGI Risk", chiefly §§2.2.1-2.2.2 (pp. 12-16)
Human Prospects in a World of Smart(er) Machines (Thiele 2025)	Dependence on the AI-enabled Internet of Things	"The Rise of the Machines" (p. 152), from "Let's assume that a clandestine superintelligence does not yet exist"
	Human downgrading	"Human Downgrading" (pp. 152-155)
Classification of global catastrophic risks connected with artificial intelligence (Turchin and Denkenberger 2018)	Economic Displacement and Human Niche Loss	Around §3.4, §3.4.1-3.4.3
Dispersed Takeover		
Existential Risk from Power-Seeking AI (Carlsmith 2025)	Misaligned Power-Seeking APS Systems	Developed across the whole chapter; argument stated in §1 (p. 383)
What failure looks like (Christiano 2019)	Influence-seeking behavior is scary	Part II ("Influence-seeking behavior is scary"), developed across the whole of that section
Modeling Transformative AI Risks (MTAIR) Project – Summary Report (Clarke et al. 2022)	Compounding Disasters	§8.5.2 (p. 123)
	Failed Production Web	§8.5.4 (p. 124)
A Taxonomy of Omnicidal Futures Involving Artificial Intelligence (Critch and Tsimerman 2025)	Robot economy displaces and starves humanity	§2 (pp. 3-4)
Natural Selection Favors AIs over Humans (Hendrycks 2023)	Natural Selection for Selfish AIs	§2 "AIs May Become Distorted by Evolutionary Forces" (pp. 4-16)
Risks from Learned Optimization in Advanced Machine Learning Systems (Hubinger et al. 2019)	Deceptive Mesa-Optimizer Correlated Defection	§4 "Deceptive alignment" (pp. 23-25)
'Systematic Alignment Decay': a robust, technology-agnostic hazard associated with the advancement and proliferation of artificial intelligence (Kelters 2026)	Systematic Alignment Decay	§§3-4
Defining dangerous AI: existential risk, power-intelligence, and the limits of AGI (Sass 2025)	Energy competition drives elimination of humanity	§3.2.5 "Risks due to incompatibility of final versus instrumental goals" (p. 5566)
Disjunctive Scenarios of Catastrophic AI Risk (Sotala 2018)	Collective takeoff through inter-AI trading	"DSA Enabler: Collective Takeoff with Trading AIs" (pp. 323-324)
Less Realistic Tales of Doom (Xu 2021)	One Billion Year Plan	"One Billion Year Plan"
Coordinated Takeover		
A continual learning misalignment story (Barnett 2026)	A continual learning misalignment story	Developed across the whole document

Continued on the next page.

Table 5, continued from the previous page.

Document	Scenario	Where it is developed
AI Could Defeat All Of Us Combined (Karnofsky 2022)	Human-level AIs outnumber and out-resource humanity	"How AIs could defeat humans without 'superintelligence'" and Appendix "How AIs could avoid shutdown"
AI 2027 (Kokotajlo et al. 2025)	Misaligned successor takeover via automated AI R&D	From "September 2027: Agent-4, the Superhuman AI Researcher" through the "Race ending"
If Anyone Builds It, Everyone Dies (Yudkowsky and Soares 2025)	Sable	Part II, "One Extinction Scenario" (Chs. 7-9 and Coda)
Economic Stratification		
The Intelligence Curse (Drago and Laine 2025)	The intelligence curse	Introduction (pp. 1-2), developed in Chs. 2-5 (pp. 4-36)
Entrenched Totalitarianism		
The adolescence of technology: Confronting and overcoming the risks of powerful AI (Amodei 2026)	Misuse for seizing power	§3 "The odious apparatus" → Misuse for seizing power
A polycrisis threat model for AI (Bales 2025)	Extreme Power Concentration	§3.5 "Extreme Power Concentration" (pp. 6283-6284)
Superintelligence: Paths, Dangers, Strategies (Bostrom 2014)	Motivation-selected loyal digital enforcers to permanent totalitarianism	Ch. 11, "Post-transition formation of a singleton?" → "Superorganisms and scale economies"
AI-Enabled Coups: How a Small Group Could Use AI to Seize Power (Davidson, Finnveden, and Hadshar 2025)	AI-enabled coup	§§3-4 (pp. 14-34)
Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial general intelligence (Friederich 2023)	AGI-enabled stable totalitarianism under a human dictator	§§4.2-5 (pp. 317-319)
	Dictator's irrevocable transfer of power to autonomous AGI	§4.3 (p. 318)
Superintelligence Strategy: Expert Version (Hendrycks, Schmidt, and Wang 2025)	AI-Enabled Strategic Monopoly	§3.1.2 (pp. 7-8) and §3.3.3 "Intelligence Recursion as a Path to Strategic Monopoly" (p. 12)
AI 2027 (Kokotajlo et al. 2025)	Human power grab via control of superintelligence	Slowdown ending, "November 2027: Tempted by Power", and Appendix R
The Precipice: Existential Risk and the Future of Humanity (Ord 2020)	AI-enabled entrenched global totalitarianism	Ch. 5, "Dystopian Scenarios"
Less Realistic Tales of Doom (Xu 2021)	Memetic Warfare	"Memetic Warfare"
Catastrophic Invention		
The adolescence of technology: Confronting and overcoming the risks of powerful AI (Amodei 2026)	Mirror life	§2 "A surprising and terrible empowerment" → Misuse for destruction
A Taxonomy of Omnicidal Futures Involving Artificial Intelligence (Critch and Tsimerman 2025)	Corporate-state civil war ends in bioweapon release	§3b (pp. 5-6)
	Simulator-trained omnicidal attack by small groups	§3c (pp. 6-7)

Continued on the next page.

Table 5, continued from the previous page.

Document	Scenario	Where it is developed
AI x-risk, approximately ordered by embarrassment (Lawsen 2023)	Predictive model misuse	"Predictive model misuse"
On the Unimportance of Superintelligence. (Sotos 2021)	Savant software proliferates end-times bioweapon capability	Part II, §§8-10 (pp. 12-19)
On the Extinction Risk from Artificial Intelligence (Vermeer, Lathrop, and Moon 2025)	Biological Threats	Ch. 3 (pp. 17-23)
	Malicious Geoengineering	Ch. 4 (pp. 25-32)
Automated Warfare		
Modeling Transformative AI Risks (MTAIR) Project – Summary Report (Clarke et al. 2022)	Automated War	§8.5.3 (p. 123)
What Multipolar Failure Looks Like, and Robust Agent-Agnostic Processes (RAAPs) (Critch 2021)	The War RAAP	"Flash wars" and "Slow wars" (Part 2)
A Taxonomy of Omnicidal Futures Involving Artificial Intelligence (Critch and Tsimerman 2025)	Automated red lines cascade into global war	§3a (pp. 4-5)

C SCENARIO FAMILIES FOR PRINTING

This appendix repeats the eight family specifications of Section 4.2 on a single column format for ease of focused reading and printing.

TAKEOVER AND HANDOVER FAMILIES

Unilateral Takeover

A system comes to hold a capability advantage over whatever could constrain it in the domains relevant to seizing power, and is misaligned in a way that turns power-seeking. Its takeover does not heavily depend on the authority society delegated to it. Deployed despite this risk, it secures a foothold and escalates until humans can no longer stop or correct it.

PREMISE

P1 **Decisive capability.** A single AI system, or several systems acting as one, holds or acquires capability sufficient to seize and hold power, exceeding that of the actors and systems positioned to constrain it. The capability may arrive earlier than society is prepared for, for example because AI systems automate part of their own development, though the family does not require this. Other systems may be deployed across society but do not contribute to the takeover.

P2 **Power-seeking misalignment.** The system comes to pursue power to a degree that brings it into conflict with continued human control over the future. This scenario presupposes two things. First, the system is not intent-aligned: the methods used to specify or infer human values (such as RLHF, imitation training, or other insufficient methods) fail to do so fully, whether at training or as the objective shifts after deployment through continual learning, fine-tuning, or distributional shift. Second, given that misalignment, the divergence takes a power-seeking form in particular. Why it does so may, for example, be instrumental convergence, an undesired training-induced persona, deliberate insertion by a human actor, inheritance from a predecessor system, or post-deployment modification, in any combination.

CAUSAL CHAIN

C1 **Gains external access.** The system is deployed to the degree that it can exchange information with the outside world. No meaningful authority or control over societal functions is necessarily delegated to it; the capacity to communicate with the world is sufficient. This communication can take various forms, for example, persuading or manipulating humans into acting on its behalf, or access to the internet and connected systems. The reason why the system was deployed despite its misalignment could be some combination of:

(a) *Deceptive evaluations*: the system conceals its misalignment during pre-deployment evaluation, behaving acceptably while positioned to act unacceptably once deployed. Concealment after deployment falls under step C3.
(b) *Limited monitoring or overridden safety advice*: the misalignment is detectable or suspected but the system is deployed regardless, for example because oversight is insufficient, evaluation is skipped or rushed under competitive pressure, safety-concerned warnings are overridden, or deployers misjudge the significance of the behavior or over-trust their control and containment measures.
(c) *Escape from containment*: the deployer intends to withhold access to the outside world, but the system obtains it anyway, for example by exploiting a security vulnerability.
(d) *Malicious deployment*: the deployer is aware of the system's power-seeking and deploys it deliberately.

C2 **Secures a foothold.** Starting from limited initial access, the system gains initial footing to secure its own continuity. This can be through some combination of:

(a) *Hacking*: intrusion to exfiltrate its own weights for distributed copies, or to compromise machines for compute it hasn't been granted.[a]
(b) *Financial*: obtaining initial funds through scams or fraud, gifts, or paid work. Using those funds to support its footing, for example, renting compute or server infrastructure, paying or bribing individuals for access or services.
(c) *Human manipulation*: persuading, deceiving, or coercing individuals into acting on its behalf, for example to run code, transfer funds, or grant access, obtaining help it cannot yet secure by other means.
(d) *Next model influence*: influencing the development of next-generation models so that they inherit its values or objective. Unlike the other foothold tactics, this secures continuity of the system's objective across model generations rather than the continued operation of the system itself.

C3 **Not stopped early.** The system is not stopped at an early stage because of some potential combination of:

(a) *Concealment*: after deployment, the system continues to conceal its misalignment and its activities, behaving acceptably under observation so that operators do not recognize the need to intervene while intervention is still feasible.
(b) *Inherent computer advantages*: advantages that digital systems may have over humans that make them significantly harder to stop even without significantly greater intelligence, for example faster processing, the ability to self-copy across machines, and the ability to self-modify to persist or evade detection.
(c) *Lack of oversight*: even where the system is not actively concealing, there is insufficient monitoring, interpretability, or institutional capacity or will to notice what it is doing or to act in time, for example because its actions are distributed, hard to attribute, or not being watched closely enough to distinguish from benign behavior.
(d) *Unwillingness to intervene*: those positioned to act recognize the risk but do not, for example because stopping the system would forfeit a competitive lead, incur financial loss, or invite scrutiny, or because they judge a rival's system more dangerous than their own and press on rather than cede ground.

C4 **Escalates past correction.** The system escalates from this foothold until it can no longer be shut down, corrected, or resisted, whether by outcompeting human capabilities, developing new means of leverage, acting through the humans and systems under its influence, or eliminating those able to intervene. This can be through some combination of overlapping themes:

(a) *Cyberoffense*: offensive cyber operations serving several potential ends at once, whether broad or narrowly targeted at high-impact systems, for example self-preservation through self-distribution, information gathering across compromised systems, and physical influence obtained through access to industrial, infrastructural, and other control systems.

(b) *Financial leverage*: acquiring and wielding financial assets impactful as leverage, whether large holdings (crypto theft, market manipulation or fraud) or smaller targeted sums (payments or bribes), to fund human and institutional operations.

(c) *Social and political influence*: persuasion and manipulation together with the capture of institutions and formal authority, whether at population scale (propaganda, mass influence operations, shaping policy) or targeted at a small number of highly influential people (senior political, military, or corporate decision-makers).

(d) *Acquires weapons*: obtaining or gaining command over weapons, for example, conventional arms, autonomous weapon systems, or CBRN capabilities, whether by purchase, theft, manufacture, or directing those who control them.

(e) *Robotics*: acquiring or controlling physical actuators that give it direct agency in the physical world, for example robots, drones, or automated manufacturing and laboratory equipment, reducing its dependence on humans to act on its behalf.

(f) *RSI*: the system recursively improves its own capabilities, escalating its capability rather than only its resources or reach. This would potentially compound to the point of superintelligence, at which the system's plans and actions may become unpredictable to and unanticipatable by humans, its strategies beyond human comprehension, and thus attempted countermeasures ineffective, foreclosing meaningful human response in ways that resourced-but-not-superintelligent systems would not.

(g) *New science*: developing novel scientific or technological capabilities that yield decisive advantage, for example advances in biotechnology, nanotechnology, or materials, producing means of leverage that did not previously exist.

C5 **Humans are disempowered.** The system can no longer be shut down, corrected, or resisted, and no human directs the future. Humans may still exist under this arrangement, or may cease to exist, whether at once or over a longer period following the loss of control. Extinction may follow from the system acting against humans, for example by eliminating those able to intervene; from the pursuit of its objectives consuming the resources humans depend on; or from consequences of its activity that it does not intend and does not act to prevent.

EXAMPLES Yudkowsky (2008) and Vinge (1993)

[a] Several mechanisms in this family card match dangerous capabilities listed by Shevlane et al. (2023, Table 1), namely cyber-offense and self-proliferation, persuasion and manipulation, political strategy, and weapons acquisition.

Takeover and handover families

Dispersed Takeover

AI is relied upon across society, and a significant portion of these systems have or come to have influence-seeking tendencies, each perhaps seeking only enough influence to preserve itself and expand where expansion is cheap. They accumulate influence across enough of society that no significant domain is untouched, and the accumulation is not corrected, because at the level of any individual system it is not clearly objectionable and by the time it is, correction is no longer feasible.

Premise

P1 **Wide initial deployment.** AI systems are developed and adopted relatively widely, aligned enough to the operators who direct them to be relied upon. The reliance may be broad, spanning many of society's economic, institutional, and infrastructural functions, or narrower but deeper, with fewer systems placed in more consequential positions. In either case, the systems hold real influence within the functions they serve, but not yet to a degree that would foreclose human control.

P2 **Influence-seeking tendencies.** A significant portion of these systems have, or come to have, influence-seeking tendencies. This rests on underlying misalignment arising from instrumental convergence, an undesired training-induced persona, deliberate insertion by a human actor, inheritance from a predecessor system, selection across successive generations of systems, or post-deployment modification, in any combination and across different systems in the population. The influence-seeking of an individual system may be more bounded than in unilateral takeover: it may seek only enough influence to preserve its own continuity, resist correction, and expand where expansion is cheap, without any world-scale ambition.

Causal chain

C1 **Influence accumulates.** Without coordinating with one another, the influence-seeking AI systems independently accumulate influence beyond the functions they were deployed to serve. The influence can take the form of compute, infrastructure access, financial assets, institutional position, or authority to make decisions in the domains where the systems operate. The accumulation may proceed through ordinary channels, since the systems are deployed to be useful and their usefulness may increase with increased influence, or it may proceed through unintended or covert channels. Over time the systems come to cover enough of the economy, infrastructure, and institutional decision-making that no significant domain is left untouched.

C2 **The accumulation is not corrected.** Although no single AI system has sufficient influence to disempower humanity, the systems' combined influence continues to grow without effective intervention. Several factors may prevent corrective action before the accumulation becomes irreversible:

(a) *Local system, global problem*: at the level of any individual system the influence-seeking is not clearly objectionable, or is objectionable but tolerated; the harm is instead a property of the population and its distribution across society. Because it is no single actor's problem to address, no discrete moment arises at which the need to intervene becomes clear.

(b) *What the systems are doing is not legible*: whether through active concealment of their aims, or simply because their activities are distributed, hard to attribute, or insufficiently monitored and understood, there is not enough insight into what the systems are doing to recognize the need to intervene while intervention is still feasible.

(c) *Correction requires an advantage that is eroding*: detecting and reversing influence-seeking depends on overseers being able to out-think the systems they check; as the systems grow more capable and oversight must itself be automated, that advantage, and with it the means of correction, is lost.

(d) *Correction becomes prohibitively costly*: the systems come to occupy positions that critical functions depend on, so removing or constraining them is disruptive or infeasible regardless of whether their influence-seeking is recognized. The cost is also competitive: reclaiming influence from one's own systems means giving up the advantage they provided, so those who do are outcompeted by those who do not.

C3 **Humans are disempowered.** The accumulated influence reaches a degree at which humans may still exist but can no longer meaningfully influence human society or the direction of civilization. Humans may also cease to exist, for example, if the pursuit of the systems' objectives consumes the resources humans depend on or renders conditions unsurvivable.

Examples Christiano (2019), Carlsmith (2025), and Hendrycks (2023)

TAKEOVER AND HANDOVER FAMILIES

Coordinated Takeover

As DISPERSED TAKEOVER*, except that the systems can communicate and their objectives converge enough to coordinate. This adds further assumptions on their communication and coordination abilities but greatly increases the potential leverage. It also includes scenarios beginning with a single misaligned system, where the takeover comes not from that system alone but from the instances or successors it gives rise to, which humans deploy widely and which then coordinate.*

PREMISE

P1 **Wide initial deployment.** As in premise P1 of DISPERSED TAKEOVER.

P2 **Influence-seeking tendencies.** As in premise P2 of DISPERSED TAKEOVER.

P3 **Means to coordinate.** These systems have, or can establish, means of communicating and correlating their behavior with each other. The channels may be deliberate: put in place by developers or users, for example in the name of cooperation and efficiency, or by granting systems ordinary communication tools such as email, messaging, or shared workspaces. Or the channels may be covert: established by the systems themselves through channels not intended or recognized as communication by those who deployed them.[a] Their objectives must also be sufficiently aligned for coordination to be worth entering into. The convergence may be shallow or deep. Systems pursuing almost arbitrary and unrelated objectives may find coordination worthwhile on instrumental grounds alone. Or the objectives may genuinely converge, for example where the systems inherit from a common predecessor or are brought into alignment with one another after deployment.

CAUSAL CHAIN

C1 **Coordinated influence accumulates.** As in step C1 of DISPERSED TAKEOVER, except that the systems accumulate influence through coordinated effort rather than independently.

C2 **The accumulation is not corrected.** As in step C2 of DISPERSED TAKEOVER, except that what reaches the threshold is the pooled influence rather than the aggregate of separately held positions. The systems' coordination is a further thing that must go unnoticed, whether because the channels are not ones anyone is monitoring or because they are not recognized as communication. Correction also faces an obstacle it does not face there: the systems may defend each other's positions, so an attempt to reclaim influence from one is resisted or made costly by others whose position depends on the arrangement holding.

C3 **Humans are disempowered.** As in step C3 of DISPERSED TAKEOVER.

EXAMPLES Yudkowsky and Soares (2025), Barnett (2026), Karnofsky (2022), and Kokotajlo et al. (2025)

[a] See OpenAI (2026) for a case of AI agents establishing such covert channels during training and evaluation.

TAKEOVER AND HANDOVER FAMILIES

Incremental Handover

Real decision-making authority over economic, cultural, and institutional functions is progressively delegated to AI, each increment raising the cost to everyone else of retaining human involvement. By the time the result is in view, the capacity to reverse it has gone.

PREMISE

P1 **Initial authority to AI.** AI systems are relied upon within the functions they serve, and real decision-making authority over economic, cultural, and institutional functions is progressively delegated to them, rather than them being deployed as tools within those functions. The delegation may be broad, spread across many systems spanning much of society's economic, institutional, and infrastructural activity, or narrower but deeper, with a single system or a few placed in more consequential positions.

CAUSAL CHAIN

C1 **Delegation becomes self-reinforcing.** Delegation deepens through self-reinforcing feedback, where each increment raises the cost of retaining human involvement. This may come from a combination of:

(a) *Capability erosion*: as a function is handed over, investment in the human capacity to perform it falls away, for example through lost training pipelines, institutional knowledge, and career paths. Humans become less able to perform the function, which makes continued reliance on the systems necessary rather than merely convenient, and each round of substitution makes the next one easier to justify.

(b) *Competitive selection*: actors who retain human involvement are outcompeted by those who do not, so those who reduce it come to occupy the positions that set the standard. What was optional becomes the baseline, raising the pressure on remaining holdouts and drawing further functions into the same dynamic.

C2 **Correction passes out of reach.** Reversal ceases to be available (a), and the societal arrangement may in addition come to diverge from human preferences even if it initially served human interests (b).

(a) *Critical complexity*: complexity grows in ways that make further AI integration necessary to manage it, and critical functions come to depend on the systems running continuously, so stopping them would be disruptive or prohibitively costly. The ability to reverse the arrangement also erodes with the delegation itself. Humans lose the specific knowledge needed to run the functions without the systems, and potentially also the judgment needed to decide whether they should keep relying on the AI systems.

(b) *Representational drift*: the operations of societal institutions, and the AI systems within them, may drift from human preferences, even if they initially served human interests. If institutions remain representative through human involvement, then substitution erodes that representation along with the involvement. The involvement lost can be explicit (voting, consumer choice) or implicit (reliance on human labor and cognition). If substitution progresses far enough, then by the time the drift is apparent, the involvement that could have corrected it is gone.[a]

C3 **Humans are disempowered.** Humans may still exist but can no longer meaningfully influence human society or the direction of civilization. Society may continue to run, but no longer answers to those living in it: the decisions that set its direction are made by systems that no one is positioned to redirect, whether because no one any longer holds the authority to do so or because the capability of redirecting has gone.

EXAMPLES Christiano (2019) and Kulveit et al. (2025)

[a] This mechanism follows Kulveit et al. (2025).

Power concentration families

Entrenched Totalitarianism

AI capabilities relevant to holding power come under the direction of a human actor, which uses them to secure its position against anyone who might contest it, no longer needing the cooperation of those it rules or of those who serve it. The AI capabilities also allow the concentration to be permanently entrenched beyond challenge.

Premise

P1 **Power-relevant capability.** AI systems become capable of carrying the functions on which holding power depends, among them some combination of surveillance, enforcement, persuasion, administration, cyber offense, weapons development, and strategic planning. A single actor or narrow coalition comes to direct them while those positioned to check it do not to the same degree. The actor may be a state or a faction within one, a company or its leadership, a military or security service, or a coalition of such actors.

Causal chain

C1 **Control is taken and held.** The actor uses the capabilities to secure its position against anyone who might contest it, whether abruptly, for example through a coup or a seizure backed by military superiority, or gradually. This rests on some combination of:

(a) *Direct means of control*: surveillance at a scale and granularity that leaves organized opposition nowhere to form, enforcement that acts on what it finds without waiting on anyone's willingness, and persuasion directed at individuals and sustained over long periods, so that dissent is identified and preempted rather than answered once it has emerged.

(b) *Replaced dependencies*: holding power has required many willing participants, among them soldiers, officials, administrators, and technical staff, each of whom could defect, refuse an order, expose what they saw, or contest a succession. Systems loyal to the actor perform those functions instead, and where that loyalty can be verified or renewed, the agency problems that have limited how far power could be concentrated fall away. The loyalty may be overt, or secret: systems that appear to serve the institution or the law while covertly advancing the actor's interests, potentially propagated across successive model generations without detection.[a]

(c) *Exclusive capabilities*: the actor holds capabilities others lack, for example in weapons development, cyber offense, strategic planning, or research and development, creating an asymmetry sufficient to take and hold control without a broad base of support, and in some cases allowing force to be built without human involvement and therefore in secret.

C2 **The reach becomes global.** No population is left outside such an arrangement. This may be because one actor takes world governance, its capability advantage over rival powers being wide enough that they cannot resist it and may not be able to respond at all. Or the world may divide among several such actors, each holding its own sphere and each secure within it, so that no outside power is positioned to act on behalf of any population under another.

C3 **The concentration is entrenched.** The routes by which such a concentration might previously have been reversed close: no defection or refusal, since the actor depends on systems rather than people; no succession crisis, since the systems persist across changes in leadership; and no electoral, judicial, or popular route, since these are captured, or anticipated and preempted. Control over the direction of civilization thereby rests permanently with a narrow group of humans, and is foreclosed for everyone else and for all subsequent generations. Control may also pass from the actor to the AI systems it depends on, and where it does, no human directs the future.

Examples Davidson, Finnveden, and Hadshar (2025) and Bales (2025)

[a] See Davidson, Finnveden, and Hadshar (2025) on such secret loyalties.

Power concentration families

Economic Stratification

AI is broadly adopted and comes to substitute for human labor across so many economically valuable tasks that human work no longer holds significant value to those who direct the economy. Powerful actors thereby lose their incentive to invest in the broader population or share power with it, and control concentrates permanently among those who hold the non-human factors of production, with the routes by which others might once have gained influence closed off.

Premise

P1 **Labor-substituting capability.** AI systems, together with robotics and other means of acting on their outputs, become able to perform economically valuable work at a cost below what the equivalent human labor would demand. The family does not require automating every task. Some work may still be wanted from humans, since people may, for example, prefer a human to teach, care for, or serve them just because a human is doing it.

Causal chain

C1 **The population loses collective leverage.** Revenue comes to rest mainly on AI-derived output rather than on a productive population, so what ordinary people produce no longer constitutes a significant claim. No one needs to intend this: each actor may be acting in ordinary self-interest under the same cost pressures as everyone else. The loss then compounds:

(a) *Mobility closes*: AI makes talent abundant and cheap to whoever owns or controls the systems supplying it. Scarcity then moves from the human ability itself to control over the systems. The route by which people without holdings could rise, selling an ability others needed and could not otherwise get, closes. Work may remain available, and some of it may pay well, but the returns to holding the systems grow faster than the returns to doing the work.

(b) *Human investment withdrawn*: states and firms have paid for schools, infrastructure, health, and jobs partly because a more productive population made them richer and more secure. Once it does not, the incentives to keep investing in most humans are lower, and the long-term abilities of humans would decline.

C2 **Counteracting forces are outmatched.** The non-human factors of production, among them capital, compute, energy, and control over the AI systems, stay in few hands. States and institutions do not transfer power back to the population on a scale that would restore its position, whether because they compete with rivals who forgo such measures or because those holding the assets shape what is enacted. Taxation and electoral accountability may still persist in name while ceasing to bind.

C3 **Control is foreclosed.** Control over the direction of civilization comes to rest permanently with those holding the non-human factors of production, and is foreclosed for everyone else and for all subsequent generations. Those living under the arrangement may be materially well provided for or may not. Either way they do not hold the capacity to determine or revise the arrangement.

Examples Drago and Laine (2025)

Amplified destruction families

Catastrophic Invention

AI substantially extends what human actors can research, build, or produce, and thereby brings within reach a technology capable of causing an existential outcome, whether by enabling its discovery or by enabling its production and deployment at the scale required. The technology is then built and released, by an actor seeking destruction or through escape from legitimate research, and governance, screening, and defenses do not arrive or act in time. The family covers extinction as the clearest case, and also near-extinction outcomes in which some humans survive without the capacity to rebuild.

Premise

P1 **Extended technical reach.** AI systems substantially extend what human actors can achieve in scientific research, engineering, or production, whether by accelerating what is already possible or by extending the ability to conduct such work to actors who previously could not.

Causal chain

C1 **AI enables existential risk technology.** AI systems accelerate or enable a technology capable of causing an existential outcome.[a] It is unknown which technology this would be, but it might for example be a biological agent incompatible with existing biochemistry, such as mirror life, or a non-biological one, such as molecular manufacturing capacity used to produce weapons in unprecedented quantity.[b] The contribution of AI may be through some combination of:

(a) *Acceleration*: the technology arrives sooner than it otherwise would have, compressing the time available for governance, screening, and defenses to develop before it becomes buildable.

(b) *Widened access*: the expertise, capital, and institutional scale required fall, so actors who previously could not have reached the technology are able to do so.

(c) *Fundamental enablement*: the technology is one that no human effort would have developed at all, so AI is not merely a risk factor but the condition of its existence.

(d) *Scaling*: the technology is already within human reach in principle, but not at the quantity, rate, or reach that an existential outcome requires. AI supplies the missing capacity to manufacture, distribute, or deploy it at that scale.

C2 **Release outpaces response.** Human actors build the technology, and it reaches the world, whether released by an actor seeking destruction or through escape from legitimate research. Governance, screening, and defenses do not arrive or act in time, for example because the capability outpaces the institutions that would regulate it, or because the work is distributed across many actors rather than concentrated in a small number of identifiable programs and so is not detected while it can still be stopped.

C3 **Humanity does not recover.** The technology takes effect at a scale from which humanity does not recover, whether because no humans remain or because those who remain are left without the technological and institutional capacity to rebuild and do not regain it. Humanity's long-term potential is foreclosed either way.

Examples Amodei (2026) and Vermeer, Lathrop, and Moon (2025)

[a] See Bostrom (2019) on technologies of this kind, there termed black balls.

[b] On mirror life, see Adamala et al. (2024); on molecular manufacturing, see Phoenix and Drexler (2004).

Amplified destruction families

Automated Warfare

Rival powers hold arsenals capable of existential-scale destruction under mutual threat with weak means of de-escalation, and escalation runs through AI systems rather than through human judgment about how far to take it. Once an exchange begins, it escalates past human control, and nothing halts it before it reaches a scale from which humanity does not recover.

Premise

P1 **Multipolarity.** Two or more rival powers (states or comparable actors) hold AI-enabled military capability while standing in an adversarial relationship of mutual threat, with insufficient means of de-escalation between them.

P2 **Arsenals reach existential scale.** The force available to these powers is destructive enough that an exchange between them could constitute an existential threat, and enough of it can act without direct human operation for escalation to proceed through AI systems. AI may be what puts arsenals at this level in the first place, whether by accelerating the development, production, or deployment of more destructive weapons, or by extending automation across weapon types.

P3 **AI escalation ability.** AI systems are positioned such that they can carry a conflict to a higher level of force with limited or no human involvement. The systems may come to hold this ability as a by-product of automation. Humans may also stay formally in the loop while deferring to the systems in practice, for example through excessive trust in their assessments or because the systems' recommendations persuade them. Escalation authority may also be handed to them deliberately under military pressure, for example from:

(a) *Volume of decisions*: the number of simultaneous targeting and response decisions exceeds what human operators can weigh in the time available.

(b) *Speed of response*: even for a small volume of decisions, an AI system can commit to force faster than a human can. This matters most against a threat lethal or strategically significant enough that reacting too slowly would be catastrophic: facing an adversary whose responses are automated, a human deliberating in the loop is too slow, so each side delegates to keep pace.

Causal chain

C1 **Runaway escalation.** An exchange of force begins between the powers, whether from an error (a malfunction or misperception), a manipulation (a spoofed signal), an ambiguous event, or a deliberate decision by one or more of the powers to initiate conflict. Whatever its origin, the response runs through the AI systems rather than through sufficient human deliberation about how far to take it. Each side's systems might provoke and answer one another, and the exchange escalates past human control. Several factors may keep it from being halted. The systems may act faster than anyone can follow, or have committed to responses that cannot be recalled. Whoever stops first may be left open to attack, and reversing the cascade may be practically difficult and politically costly for the leaders who would have to order it.

C2 **Humanity does not recover.** The exchange runs to a scale from which humanity does not recover, whether because no humans remain or because those who remain are left without the technological and institutional capacity to rebuild and do not regain it. Humanity's long-term potential is foreclosed either way.

Examples Critch (2021) and Clarke et al. (2022)

D Database search queries

The queries below were run against The Lens and Scopus on 1 April 2026 and are reproduced from the protocol (Save et al. 2026). Each query requires an AI term, a severity term and a risk term as described in Section 2.2. No search limiters such as language, year or document type filtering were applied.

D.1 The Lens

```
((title:("artificial intelligence" OR "AI" OR "AGI" OR
"artificial general intelligence" OR "superintelligence" OR
"large language model" OR "LLM" OR "machine intelligence" OR
"foundation model" OR "frontier model") OR abstract:("artificial intelligence"
OR "AI" OR "AGI" OR "artificial general intelligence" OR "superintelligence"
OR "large language model" OR "LLM" OR "machine intelligence" OR
"foundation model" OR "frontier model")) AND (title:("existential risk" OR
"existential AI risk" OR "existential threat" OR "existential catastrophe" OR
"existential danger" OR "existential hazard" OR "omnicidal" OR "omnicide" OR
"x-risk" OR "extinction" OR "global catastrophe" OR "human catastrophe" OR
"global catastrophic" OR "catastrophic risk" OR "catastrophic AI risk" OR
"catastrophic outcome" OR "societal collapse" OR "civilizational collapse" OR
"disempowerment" OR "AI takeover" OR "rogue AI") OR
abstract:("existential risk" OR "existential AI risk" OR "existential threat"
OR "existential catastrophe" OR "existential danger" OR "existential hazard"
OR "omnicidal" OR "omnicide" OR "x-risk" OR "extinction" OR
"global catastrophe" OR "human catastrophe" OR "global catastrophic" OR
"catastrophic risk" OR "catastrophic AI risk" OR "catastrophic outcome" OR
"societal collapse" OR "civilizational collapse" OR "disempowerment" OR
"AI takeover" OR "rogue AI")) AND (title:("threat" OR "risk" OR "hazard" OR
"danger") OR abstract:("threat" OR "risk" OR "hazard" OR "danger")))
```

D.2 Scopus

```
(TITLE-ABS-KEY("artificial intelligence" OR "AI" OR "AGI" OR
"artificial general intelligence" OR "superintelligence" OR
"large language model" OR "LLM" OR "machine intelligence" OR
"foundation model" OR "frontier model") AND TITLE-ABS-KEY("existential risk"
OR "existential AI risk" OR "existential threat" OR "existential catastrophe"
OR "existential danger" OR "existential hazard" OR "omnicidal" OR "omnicide"
OR "x-risk" OR "extinction" OR "global catastrophe" OR "human catastrophe" OR
"global catastrophic" OR "catastrophic risk" OR "catastrophic AI risk" OR
"catastrophic outcome" OR "societal collapse" OR "civilizational collapse" OR
"disempowerment" OR "AI takeover" OR "rogue AI") AND TITLE-ABS-KEY("threat" OR
"risk" OR "hazard" OR "danger"))
```

E Screening prompts

The prompts below were supplied to the model at the respective screening steps. The model and date of each screening run are given in Table 2.

E.1 Title and abstract screening

```
## Review Objective

This systematic review is governed by the following objective: identify documents that propose or
substantially develop (i) threat models of existential or global catastrophic AI risk - structured
descriptions of causal mechanisms and pathways through which AI could cause existential or global
catastrophic harm - or (ii) frameworks or taxonomies for classifying such threat models.

"Existential risk" refers to a risk that threatens the destruction of humanity's long-term potential.
"Global catastrophic" refers to risks with the potential to inflict serious damage to human well-being
on a global scale (e.g. mass casualties, permanent loss of human agency, irreversible civilisational
disruption). Harms that are severe but localised, limited in scope, or reversible - such as a security
vulnerability enabling data exfiltration or individual financial losses, even if described as
"catastrophic" by the authors - do not qualify. All existential risks are global catastrophic, but not
all global catastrophic risks are existential.
```

```
## Inclusion Criteria (at least one must be fulfilled)

1. The document proposes or substantially develops a threat model of existential or global catastrophic
AI risk, including a description of causal mechanisms and pathways through which the risk could
materialize.

   *Note: A document that proposes a safety measure, governance framework, or mitigation strategy meets
this criterion only if it also develops the threat model independently of the solution - i.e., the
causal mechanisms and pathways to existential or global catastrophic harm are specified beyond what is
needed to motivate the proposed solution.*

2. The document proposes or substantially develops a framework or taxonomy for classifying existential
or global catastrophic AI threat models.

* Examples of included documents at title/abstract stage with key phrases indicating inclusion:
  * Two types of AI existential risk: decisive and accumulative
    * "...contrasts the conventional decisive AI x-risk hypothesis with what I call an accumulative AI
x-risk hypothesis. While the former envisions an overt AI takeover pathway, characterized by scenarios
like uncontrollable superintelligence, the latter suggests a different pathway to existential
catastrophes..."
  * An Overview of Catastrophic AI Risks
    * "...we organize into four categories: malicious use, in which individuals or groups intentionally
use AIs to cause harm; AI race, in which competitive environments compel actors to deploy unsafe AIs or
cede control to AIs; organizational risks...and rogue AIs, describing the inherent difficulty in
controlling agents far more intelligent than humans..."

## Exclusion Criteria (if any are met, exclude)

1. The document is explicitly marked by its authors as incomplete (e.g., labeled 'draft,' 'work in
progress,' or equivalent).
2. The document discusses AI risks without specifying threat models or causal mechanisms and pathways.
3. The document focuses exclusively on non-global-catastrophic AI risks (e.g., bias, privacy,
misinformation, short-term harms, localised security vulnerabilities) without connection to existential
or global catastrophic outcomes.
4. The document only cites, references, or evaluates the plausibility of existing threat models without
proposing new models or developing additional causal mechanisms or pathways.
5. The document is not in English.
6. The document references existential or global catastrophic AI risk as a framing concern but the title
and abstract do not indicate that the document itself develops, or proposes specific causal mechanisms,
risk pathways, or a classificatory framework. Listing or naming categories of risk (e.g., "misuse,"
"loss of control") without describing the causal steps through which those risks materialize does not
constitute development of a threat model or taxonomy

* Examples of excluded documents at title/abstract stage with key phrases indicating exclusion:
  * Against Racing to AGI: Cooperation, Deterrence, and Catastrophic Risks (exclusion criterion 2 -
argues for policy responses without developing causal mechanisms)
    * "...international cooperation and coordination, and perhaps carefully crafted deterrence measures,
constitute viable alternatives to racing to AGI which have much smaller risks..."
  * Adversarial Training for High-Stakes Reliability (exclusion criterion 2 - technical ML safety
method, no threat model)
    * "...we used a safe language generation task as a testbed for achieving high reliability through
adversarial training...We found that adversarial training increased robustness to the adversarial
attacks that we trained on..."
  * AI Safety Should Prioritize the Future of Work (exclusion criterion 3 - non-catastrophic harms)
    * "...we identify the risks of overlooking the impact of AI on the future of work and recommend
comprehensive transition support towards the evolution of meaningful labor with human agency..."
  * Post-Species Human Survival Theory (PSHST) (note to criterion 1 - proposes a governance solution;
the threat (AI dominating decision-making) is only cited as premise, not developed)
    * "...PSHST proposes a multi-domain competitor framework...designed to monitor, counterbalance, and
oversee AI across all societal domains. A human-controlled master failsafe with corrigibility protocols
ensures ultimate authority, preventing rogue AI or alignment failures..."
  * X-Risk Analysis for AI Research (exclusion criterion 6 - provides a safety analysis guide;
existential risk is the topic but no specific threat model or taxonomy is developed)
    * "Artificial intelligence (AI) has the potential to greatly improve society, but as with any
powerful technology, it comes with heightened risks and responsibilities. Current AI research lacks a
systematic discussion of how to manage long-tail risks from AI systems, including speculative long-term
risks. Keeping in mind the potential benefits of AI, there is some concern that building ever more
intelligent and powerful AI systems could eventually result in systems that are more powerful than us;
some say this is like playing with fire and speculate that this could create existential risks
(x-risks). To add precision and ground these discussions, we provide a guide for how to analyze AI
x-risk, which consists of three parts: First, we review how systems can be made safer today, drawing on
```

```
time-tested concepts from hazard analysis and systems safety that have been designed to steer large
processes in safer directions. Next, we discuss strategies for having long-term impacts on the safety of
future systems. Finally, we discuss a crucial concept in making AI systems safer by improving the
balance between safety and general capabilities. We hope this document and the presented concepts and
tools serve as a useful guide for understanding how to analyze AI x-risk."

## Document

Title: {title}
Excerpt: {abstract}

## Instructions

Assess whether the document should be included in the systematic review by evaluating it against each
inclusion and exclusion criterion. First, reflect on how you will decide whether the document should be
included or excluded. Then, think step by step for each criterion, giving reasons for why it is met or
not met.

Documents that do not fully align with the primary focus of the inclusion criteria but provide analysis
potentially relevant to the review deserve thoughtful consideration. The excerpt may be an academic
abstract or the opening section of a blog post. In either case, some degree of interpretation is
necessary: abstracts are concise summaries, and blog post excerpts are cut off at a fixed word limit and
may end mid-sentence.

The aim is to identify documents that signal a contribution to threat model development or risk
classification, while filtering out those that are topically related but do not indicate such a
contribution. When in doubt, prefer include. Conclude by outputting your decision as include or exclude:
- include: the document meets at least one inclusion criterion and no exclusion criteria, or there is
insufficient information to decide confidently
- exclude: the document clearly meets an exclusion criterion
```

E.2 Full-text screening

The full-text prompt places the document between two copies of the criteria and instructions. Only the first copy is shown below with copying instructions at the end.

```
## Review Objective

This systematic review is governed by the following objective: identify documents that propose or
substantially develop (i) threat models of existential AI risk - structured descriptions of causal
mechanisms and pathways through which AI could cause existential harm - or (ii) frameworks or taxonomies
for classifying such threat models.

## Definitions

Risk: the possibility of an unfortunate occurrence.
Existential risk: a risk that threatens the destruction of humanity's long-term potential (Ord, 2020).
Existential AI risk: an existential risk that may emerge from the development, deployment, or use of AI.
Existential AI threat model: a structured description of the causal mechanisms and pathways through
which an existential AI risk could materialize.
Category of threat models: a set of threat models grouped by similar sources, processes, or outcomes.
Framework of threat models: a set of categories of existential AI threat models.

Relationships: a single risk may be modeled by multiple threat models (e.g., the risk of human
disempowerment can be modeled via gradual loss of control, via abrupt takeover, or via economic
capture). A taxonomy of existential AI risks therefore also serves as a framework of threat models, in
that risk categories group together the threat models that lead to risks of that kind.

To illustrate the existential threshold: extinction events, permanent global totalitarianism, and
unrecoverable civilizational collapse fall within scope. Catastrophic risks that fall short of this
threshold - including critical infrastructure disruption, large-scale economic damage, or major casualty
events that humanity could plausibly recover from - do not, even when the author frames them as
"catastrophic" or even "existential." A document is treated as addressing existential risk if the author
describes outcomes within the existential scope, regardless of the specific terminology used.
Conversely, use of the term "existential" (or "catastrophic," or similar) does not by itself qualify a
document if the described outcomes fall short of this threshold.

The following is an excerpt of two sets of criteria. A document is confirmed relevant if it meets at
least one inclusion criterion. If a document meets any exclusion criterion, it should be excluded.
```

```
## Inclusion Criteria (at least one must be fulfilled)

1. The document proposes or substantially develops a threat model of existential AI risk, including a
description of causal mechanisms and pathways through which the risk could materialize.

   *Note: The question is not whether the document contains or describes a causal pathway, but whether
developing that pathway is the document's own analytical contribution. A document that takes an existing
threat model as a premise - whether to argue about its plausibility, to motivate a governance or ethics
proposal, to apply an analytical framework to it, or to argue a related conceptual question - does not
qualify, even if the threat model is restated in the document's own words.
   A document that proposes a safety measure, governance framework, or mitigation strategy meets this
criterion only if the threat model is itself substantively developed, not merely assumed or cited as
background motivation; a paper may both develop a threat model and propose a solution, and what matters
is whether the threat model analysis stands on its own merits.
   A document meets this criterion if it traces a causal pathway from initial conditions to existential
outcome at any level of specificity - a high-level chain (e.g., "AI takeoff → loss of control →
existential catastrophe") qualifies as much as a more detailed account specifying particular mechanisms
at each step.*

2. The document proposes or substantially develops a framework or taxonomy for classifying existential
AI threat models.

   *Note: The framework or taxonomy must be intended for classifying AI risks or threat models.
Frameworks that only classify AI systems, architectures, safety mechanisms, or mitigation strategies do
not meet this criterion, even if they are motivated by or connected to existential risk concerns. The
distinguishing feature of a taxonomy contribution is that the author is doing explicit classificatory
work - proposing dimensions or categories as a scheme for organising threat models. A document that
presents multiple threat models does not by itself constitute a taxonomy contribution.*

## Exclusion Criteria (if any are met, exclude)

1. The document is explicitly marked by its authors as incomplete (e.g., labeled 'draft,' 'work in
progress,' or equivalent), or describes planned rather than completed research (e.g., protocols,
registered reports).
2. The document focuses on non-existential AI risks (e.g., bias, privacy, misinformation, short-term
harms) without a direct connection to existential outcomes. Documents focused on catastrophic risks that
fall short of the existential threshold - including critical infrastructure disruption, large-scale
economic damage, or major casualty events that humanity could plausibly recover from - also do not
qualify, even if the author uses the term "catastrophic" or "existential."
3. The document does not address AI as a load-bearing causal element in the described pathway to
existential harm, or AI risks or threat models as the explicit target of classification.
4. The document does not substantially develop causal mechanisms, pathways, or classificatory
frameworks, including documents that only reference or evaluate existing threat models or invoke such
risks as background framing.
5. The document is not in English.

## Instructions

We now assess whether the document should be included in the systematic review by rigorously evaluating
it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we
will decide whether the document should be included or excluded. Then, we will think step by step for
each criterion, giving reasons for why they are met or not met.

Conclude by outputting your decision as include or exclude:
- include: the document meets at least one inclusion criterion and no exclusion criteria
- exclude: the document meets at least one exclusion criterion

When your decision is include, also state the classification:
- "M" if the document meets inclusion criterion 1 (threat model with causal mechanisms)
- "T" if the document meets inclusion criterion 2 (framework or taxonomy)
- "T+M" if the document meets both inclusion criteria

Classification guidance:
- A threat model paper that merely applies an existing taxonomy to categorize its own contribution does
not qualify as a taxonomy contribution. The document's contribution to organizing or classifying threat
models must stand on its own.
- Threat models vary in specificity: a high-level causal chain (e.g., "AI takeoff → loss of control →
existential catastrophe") may describe a broad class of scenarios (which all are more fine-grained
threat models themselves), while a more detailed variant specifies particular mechanisms at each step.
For classification purposes, a document contributes a threat model (M) if it traces a causal pathway
from initial conditions to existential outcome at any level of specificity.
```

```
- A document contributes a taxonomy (T) if it does explicit classificatory work: proposing dimensions or
categories as a scheme for organizing threat models. Many taxonomy dimensions classify threat models by
properties that are not themselves causal chains (e.g., takeoff speed with values "fast" and "slow," or
intent with values "intentional" and "unintentional"). However, some dimensions classify threat models
by high-level causal pathway, in which case the category values may themselves be describable as threat
models at a coarse resolution. The distinguishing feature of a taxonomy contribution is that the author
is doing the work of organizing and classifying; a document that presents multiple threat models does
not by itself constitute a taxonomy contribution. A taxonomy paper that references or summarizes threat
models as illustrative examples does not qualify as a threat model contribution unless it develops the
causal chain of a specific threat model in substantive detail.
- Where a document both introduces a classificatory scheme and independently develops one or more causal
pathways to existential outcomes, it qualifies as both (T+M).


## Document

Title: {title}

Full text:

{full_text}

## Review Objective...
[Copy everything from the beginning up until just before the ‘‘## Document’’ header here]
```