

A Tally of Framings Is Not a Census of Causes: What the AI Risk Repository's Human-Versus-AI Split Measures, Why Its Gap Closed From Seventeen Points to Four as the Corpus Grew, and the Attribution Rule a Causal Claim Would Need

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. The AI Risk Repository extracts every risk named in published AI risk frameworks and codes each one by the entity presented as causing it. Its journal version reports that "human decisions cause nearly as many AI risks (38%) as the AI systems themselves (42%)", which it calls contrary to common assumptions. Its first version, coding a smaller corpus with the same Entity categories (version 3 later reworded the definition of Other), reported 51 percent AI and 34 percent human and concluded that risks tend to be attributed to AI. In this analysis the author reviews and synthesises published results to show that the split is a tally of how framework authors phrase risks, not a census of what causes them, and that it moves with three things a causal claim would have to hold fixed. The first is the corpus: on the repository's own printed counts, and if no existing risk was recoded, the 722 coded risks added between its first and second versions leaned human by about twelve points, enough by themselves to close most of the gap. The second is the coding rule: each risk receives exactly one entity, as presented in the source sentence, with joint causation placed in a residual bin, and the first version itself warned that bias risks its sources meant to lay at developers' feet were worded as the model's doing. The third is domain composition: 15 of 24 subdomains are coded at least 60 percent one way, so the aggregate is mostly a mixing weight. Reweighting the subdomains moves each share by under three points, which is a robustness the author credits, although it roughly halves the gap; removing a single domain moves each share by up to ten points, two of the seven single-domain removals reverse which one leads, and recoding the bias and discrimination subdomains as the first version said their sources intended reverses it narrowly. Incident databases, misuse studies and expert surveys give different splits, each for a structural reason. The paper offers a protocol for an attribution census grounded in actual causation and states what the existing tally can and cannot support.

Keywords: AI risk taxonomies; causal attribution; AI governance; content analysis; actual causation; responsibility; AI incidents; measurement validity

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

AI risk now has a literature large enough to be reviewed as a literature. One 2026 audit taxonomy opens by counting "at least 74 AI risk taxonomies" (Galdon Clavell et al., 2026). The frameworks differ in scope, from language-model harms (Weidinger et al., 2021; Weidinger et al., 2022) and sociotechnical harms of algorithmic systems (Shelby et al., 2023) to societal-scale and catastrophic risk (Critch & Russell, 2023; Hendrycks et al., 2023), systemic risk under the EU AI Act's definition (Uuk et al., 2024), and the categories regulators and companies write into policy (Zeng et al., 2024). Some organise risk by where in a system it arises, as with the four modules of an LLM system (Cui et al., 2024) or risk sources across development, training and deployment (Gipiskis et al., 2024); some by the kind of system, as with foundation models (Bommasani et al., 2021), open-weight models assessed for their marginal risk (Kapoor et al., 2024) and advanced assistants (Gabriel et al., 2024). One scoping review of generative AI ethics ranks 378 normative issues by how often the literature raises them (Hagendorff, 2024), and other

consolidation efforts map the frameworks onto one another (Bagehorn et al., 2025). They differ in vocabulary too, which is the problem the AI Risk Repository set out to solve (Slattery et al., 2026). Its authors searched systematically for risk frameworks, extracted every risk each one named, and coded the risks against two taxonomies of their own: a Domain Taxonomy of seven domains and 24 subdomains, and a Causal Taxonomy that records, for each risk, the Entity presented as causing it (Human, AI, or Other), the Intent (intentional, unintentional, or other) and the Timing (pre-deployment, post-deployment, or other).

The repository has become infrastructure. Its taxonomies have been used to audit what a company safety framework covers (Coggins et al., 2025), extended into a mitigation taxonomy by the same group (Saeri et al., 2025) and then extended again from incident reports (Popchanovska et al., 2026). It is the most systematic synthesis of the risk-framework literature the author has found, and nothing in this paper disputes its value as a catalogue.

This paper is about one number the repository reports. The journal version's abstract states that "contrary to

common assumptions, human decisions cause nearly as many AI risks (38%) as the AI systems themselves (42%)" (Slattery et al., 2026). The sentence has two readings. On the first, it is a finding about discourse: of the risks that published frameworks name, the share whose wording places a human as the actor is close to the share whose wording places the AI system there. On the second, it is a finding about the world: human decisions are nearly as large a source of AI risk as AI systems are. The verb "cause" invites the second reading, and the repository's own practical guidance relies on it when it suggests that regulators address "the 37% of risks attributed to human decisions through training requirements, while addressing the 42% from AI system actions through technical standards."

The repository's earlier versions make the question sharper. Version 1, posted in August 2024, coded 668 risks from 43 documents and reported 51 percent AI, 34 percent human and 15 percent other, summarised in a table of insights as "AI risks tend to be attributed to decisions or actions taken by AI rather than by humans." Version 2, in April 2025, coded 1,390 risks from 65 documents and reported 41 percent AI and 39 percent human. Version 3, published in Patterns in 2026, coded 1,480 risks from 74 documents and reported 42 and 38. The same Entity categories, with the definition of Other reworded in version 3, the same team and the same method; the headline changed from "mostly AI" to "nearly equal." The gap between the two shares fell from seventeen points to four.

In this analysis the author reviews and synthesises published results to show that the split is a property of three inputs that a causal claim would need to hold fixed, and that the repository does not and was not designed to fix. The argument has five parts.

First, the unit and the rule. A repository risk is a phrase extracted from a framework, weighted by how finely that framework subdivides its list, and coded with exactly one entity, "as they were presented by the authors," with any risk relevant to more than one entity coded Other. That instrument measures the grammatical agent of a framework's sentence. The repository's first version saw this clearly and said so.

Second, the corpus. On the repository's own printed counts, the 722 coded risks added between versions 1 and 2 split about 44 percent human to 32 percent AI, if no existing risk was recoded. Either the added documents moved the headline, or recoding did; either way, the headline is a property of the inputs, not a stable fact about AI risk.

Third, the mixture. Within most subdomains the code is close to unanimous: misuse risks are human, misinformation and safety-failure risks are AI. Fifteen of 24 subdomains are coded at least 60 percent one way. The aggregate is therefore mainly a statement about which domains the corpus emphasises. The author's

arithmetic on the repository's supplementary tables shows two things at once: reweighting subdomains by risk count, by document coverage or equally moves either share by less than three points, which is real robustness, although it roughly halves the gap; removing the misuse domain or the AI-safety domain moves each share by about seven to nine points and the gap between them by fourteen to eighteen, and recoding only the bias and discrimination subdomains as the repository's first version said its sources intended is enough to reverse the lead, narrowly.

Fourth, the concept. "Cause" in the sense needed to allocate regulatory effort is actual causation, and the formal and empirical work on it points to three things the tally cannot represent: an outcome normally has several actual causes, responsibility comes in degrees, and which cause gets named depends on the question asked and on norms (Halpern & Pearl, 2005; Chockler & Halpern, 2004; Hilton, 1990; Hitchcock & Knobe, 2009). The blame studies show that people's attributions in joint human-AI failures follow structure, not just agency (Awad et al., 2019; Shank et al., 2019).

Fifth, the other evidence bases. Incident databases, misuse studies, expert surveys and elicitation experiments each produce a different implicit split, and each for a reason located in how its sample was drawn (Marchal et al., 2024; Mengesha et al., 2026; Grace et al., 2024; Lynch et al., 2025). One 2026 comparison of an expert risk ranking with an incident-based ranking found agreement of kappa about 0.20, with an interval crossing zero (Lambros & Wilson, 2026).

The contribution is a reading of what the split measures, a decomposition of its movement into corpus, rule and mixture, the reweighting arithmetic that bounds how much each matters, a comparison with the attribution rules used elsewhere, and a protocol for an attribution census (Algorithm 1) that would license the causal reading. The paper does not argue that human decisions are a minor source of AI risk, or a major one. It argues that the repository's number cannot say which, and states what could.

1.1 Scope, and relation to the author's earlier papers

The paper is about the Causal Taxonomy's Entity variable and the aggregate split it yields. It touches Intent and Timing only where they share the Entity variable's construction, and it does not assess the Domain Taxonomy, which raises different questions. It uses all three public versions of the repository paper because the movement between them is evidence; where the versions disagree, the paper says which it quotes.

Two of the author's earlier papers are neighbours. One examined what an early-warning threshold must deliver in a frontier safety framework; it concerned whether capability evaluations anticipate thresholded tasks, not how frameworks attribute risk. Another argued that a

claimed change in a deployed model is a test against a declared null; Section 4 below borrows only its habit of asking what else could have moved a number between two measurements. Neither paper treats risk taxonomies, causal coding, or the human-versus-AI question.

2. What the repository measured

2.1 Corpus, extraction and coding

The repository's corpus is a set of documents that each present a structured classification of AI risks. Version 1 retrieved 17,288 records, screened 7,945, assessed 91 in full text and included 43: 21 from the search, 13 from forward and backward searching, and 9 from expert suggestions. Later versions added documents as they were identified: 22 by March 2025, giving 65 in version 2, and 74 in version 3 (Slattery et al., 2026). From each document, risks were extracted with quotes and page numbers, giving 777 risks in version 1, 1,612 in version 2 and 1,725 in version 3.

The Causal Taxonomy was built by best-fit framework synthesis, a method from qualitative evidence synthesis in health research that starts from an existing framework and revises it against the data (Carroll et al., 2011; Carroll et al., 2013). The starting framework was Yampolskiy's taxonomy of pathways to dangerous AI (Yampolskiy, 2016), which classifies pathways by timing (before or after deployment) and cause, where causes are external (on purpose, by mistake, or environment) or internal to the system. The repository reports choosing it because it was highly cited, simple, comprehensive and well defined, and revising it over three iterations into the Entity, Intent and Timing variables.

Two features of the coding procedure matter for everything that follows. The first is that coding followed presentation: "we coded risks as they were presented by the authors, aiming to capture the studied phenomena directly rather than impose our own interpretations or infer intent." The second is exclusivity: "we categorized risks relevant to multiple levels of each causal factor (e.g., both pre-deployment and post-deployment) as 'Other'." Each risk receives one Entity code. A risk whose source names both a human and an AI as acting goes to Other, as does a risk whose actor is unclear. Version 3 makes the first case explicit in its table definition, where Other covers a risk that "arises from human-AI interaction rather than either agent alone, or the causing entity is ambiguous or unspecified"; version 1's table defined Other as a risk "caused by some other reason or is ambiguous," while its text already noted that Other in the human-computer interaction domain meant "both Humans and AI were presented in the descriptions of those risks, or the causal Entity was ambiguous."

Coding was done by a single reviewer per risk, with discussion among three coders where relevant, in all three versions. Version 3 states that duplicate extraction and coding would reduce the risk of bias or error, and

that "Future work should assess inter-rater reliability with coders external to the authorship team."

2.2 The numbers, version by version

Table 1 consolidates the values the argument rests on. Every row is printed in a version of the repository paper; the version and table are named on the row.

Three small inconsistencies in the printed values should be recorded so that no reader mistakes them for the paper's point. Version 3 gives the human share as 38 percent in its abstract and Table S2 and as 37 percent in one paragraph of its practical implications. Its document-level counts of 65 and 70 of 74 documents correspond to 88 and 95 percent, while the printed percentages are 85 and 93. And its reconstructed aggregate, computed below from its subdomain tables, comes out at 37 percent human and 45 percent AI rather than 38 and 42, which is within the error expected from rounded subdomain shares and from the two taxonomies being coded over slightly different subsets of risks. None of these affects the argument; all of them are reasons to read the second digit of any repository percentage loosely.

2.3 What the repository itself says the split means

The repository's body text is more careful than its abstract. Describing the Timing variable, version 3 says that the appropriate distribution of risks across categories "would depend on empirical evidence about risk frequency and severity that is beyond the scope of this review," and its limitations section states that the Causal Taxonomy "assigns risks to human or AI sources, which simplifies the sociotechnical reality that many risks emerge from human-AI interactions rather than from either agent in isolation." Supplementary Table S2 describes the AI share as the share of risks "presented by authors of the documents as due to a decision or action by an artificial intelligence system."

Version 1 went further. Its limitations section warned that "inclusion in this Repository should not be interpreted as an endorsement of a risk's framing or significance," and gave an example that bears directly on the Entity code: "risks such as bias and discrimination have been intended to be ultimately attributed to human developers and designers, but in the verbatim text of the risk the phrasing implicated the AI system (e.g., 'the model generates...')." Version 2 retains the passage. The author did not find it in the version 3 text.

So the repository's own text supports the discourse reading. The causal reading enters through the abstract's verb and through the regulatory guidance that divides the risk population by Entity and assigns a policy instrument to each part. The rest of this paper asks what the discourse reading measures and why it cannot be upgraded to the causal one.

Table 1. Entity coding in the AI Risk Repository, by version, as printed (Slattery et al., 2026, arXiv:2408.12622 v1, v2, v3).

Version (date)	Docs	Risks	Coded	Human	AI	Other
v1 (Aug 2024), Table 3	43	777	668	34%	51%	15%
v2 (Apr 2025), Table C	65	1,612	1,390	39%	41%	20%
v3 (2026), Table S2	74	1,725	1,480	38%	42%	20%
v1, docs with any (Table 5)	41	-	-	83%	98%	68%
v3, docs with any (Table S4)	74	-	-	85%	93%	77%
v3, subdomain 2.2 (S8)	-	-	-	78%	5%	16%
v3, subdomain 3.1 (S8)	-	-	-	4%	87%	9%

3. What a presented-entity code measures

3.1 The grammatical agent of a framework's sentence

A risk framework is written to be used, and the way its authors phrase a risk is chosen for the use. A taxonomy aimed at model developers will say that "the model generates" toxic text, because the developer's lever is the model; one aimed at regulators will say that a deployer used the model to make a decision, because the regulator's lever is the deployer. Weidinger and colleagues' six risk areas for language models describe many risks with the model's outputs as their subject, such as lower performance by social group or leaks of private data (Weidinger et al., 2021), and their later framework for sociotechnical evaluation argues that "context determines whether a given capability may cause harm" and adds human interaction and systemic impact as evaluation layers (Weidinger et al., 2023). The same risk can be phrased either way, and the phrasing tracks the audience.

Hilton's conversational model of causal explanation makes this a general point rather than an observation about AI frameworks. Explanations, on his account, "are selected by questions and are thus governed by general rules of discourse"; the cause an explainer names is the one relevant to the focus of the question being answered (Hilton, 1990). Hitchcock and Knobe add that people's causal judgments are shaped by norms, so that among several factors without which an outcome would not have occurred, the one that departed from a norm tends to be called the cause (Hitchcock & Knobe, 2009). A framework sentence is an answer to an implicit question - what should this audience watch for - and its grammatical agent is the factor that question makes salient. Coding that agent faithfully records the question.

This is why the repository's practice of coding "as presented" is not a weakness of execution. Inferring authors' intentions would have introduced the coders' own causal views, which would be worse. The issue is what the code is a measurement of. The Entity variable faithfully measures which actor framework authors put in the subject position. That is a quantity worth knowing; it is not the share of risk that each actor causes.

3.2 Exclusivity and the residual bin

The second feature of the rule, one code per risk with joint cases sent to Other, has a consequence that can be stated exactly. The Human share is the share of risks presented as caused by humans alone; the AI share is the share presented as caused by AI alone. Risks presented as joint are in neither. If every Other risk in version 3 were a joint case, the share of risks in which a human was presented as involved would be 58 percent, and the share in which an AI was presented as involved would be 62 percent. Neither figure is reported, both exceed half, and their sum exceeds 100 percent because involvement is not exclusive. Some of Other is ambiguity rather than jointness, so 58 and 62 are upper bounds; the true involvement shares lie between the exclusive shares and these.

The exclusive split therefore answers the question "which actor is the sole presented cause, when there is one," and the residual bin swallows exactly the cases that the sociotechnical literature argues are typical. Selbst and colleagues' account of abstraction traps holds that technical interventions fail when abstraction boundaries leave out the social actors around a system (Selbst et al., 2019). Shelby and colleagues' scoping review of 172 computing papers organises harms that "emerge from the interplay of technologies and social and cultural dynamics" (Shelby et al., 2023). Lazar and Nelson argue that "Only a sociotechnical approach can truly limit current and potential dangers of advanced AI" (Lazar & Nelson, 2023). On those views, the Other bin is where most real risks belong, and a 20 percent Other share measures how often framework authors write in a way that makes the interaction explicit.

3.3 Single coding and reliability

The repository coded each risk once. The content analysis literature treats agreement between independent coders as the condition for trusting coded data, and recommends Krippendorff's alpha as a standard coefficient because it handles any number of coders, scales and missing data (Hayes & Krippendorff, 2007). The opposing position deserves a hearing. McDonald and colleagues found that inter-rater reliability was reported in about one in nine qualitative papers in CSCW and HCI from 2016 to 2018 and treat how, and even whether, to establish reliability in qualitative research as an open

question whose answer depends on the kind of work, asking pointedly what reliability means for building a grounded theory (McDonald et al., 2019). The repository's taxonomy construction is that kind of work, and its coders cited grounded-theory practice.

The distinction that settles the matter is between building a taxonomy and reporting a prevalence from it. Category construction can proceed by discussion and iteration without an agreement statistic. A percentage reported in an abstract, and recommended as a basis for dividing regulatory effort, is a measurement, and a measurement made by one coder per item has an unknown error. Paeth and colleagues, reporting on the two independent taxonomies applied to the AI Incident Database, found that inter-annotator variation in the CSET harm taxonomy "can be reduced but not removed" through iterative development of definitions, examples and annotation processes (Paeth et al., 2024). Entity coding of short risk phrases, where the phrase may name an actor only implicitly, is unlikely to be easier.

The author cannot say how large the coding error is. No published study recodes a sample of repository risks independently. That is the cheapest check on the whole number; step S5 of Algorithm 1 builds it in, and Section 10 lists it first among what is not known.

3.4 The unit: a risk is a line in someone's list

The unit of the tally is the extracted risk, and the number of risks a document contributes depends on how finely its authors subdivided. The repository does not print per-document counts. Its totals show that the granularity varies: version 1 averaged 18.1 risks per document, and the 22 documents added for version 2 contributed 835 risks, about 38 each. A framework that lists 60 fine-grained misuse risks counts six times as heavily in the aggregate as one that names ten.

Counting items across studies is a known hazard in research synthesis. Hedges and Olkin showed that vote counting - concluding from the proportion of studies that report an effect - is a poor procedure for aggregating evidence (Hedges & Olkin, 1980). The analogy is not exact, since the repository is counting named risks rather than significance verdicts, but the lesson carries: a count of items across documents mixes the phenomenon with the documents' reporting conventions. The repository's document-level tallies (Table 1, rows 4 and 5) are the obvious alternative unit, and on that unit the result changes in kind: in version 3, 85 percent of documents name at least one human-presented risk and 93 percent name at least one AI-presented risk. Most frameworks name both, as the repository itself notes.

4. Why the gap closed: corpus, rule or both

4.1 What the added documents contributed

The repository's printed coded counts and percentages allow a rough decomposition of the change between versions. The following is the author's arithmetic, not a value the repository reports. Version 1 coded 668 risks, 51 percent AI and 34 percent human; version 2 coded 1,390, 41 and 39 percent. If no risk coded in version 1 was recoded for version 2, the 722 risks added between them were about 32 percent AI (between 30 and 33 percent once the printed values' rounding is allowed for), about 44 percent human (42 to 45 percent) and about 25 percent Other. The added material leaned human by roughly twelve points, where the original corpus had leaned AI by seventeen. Between versions 2 and 3, only 90 coded risks were added; at that size the printed rounding leaves the AI share of the additions anywhere between 41 and 73 percent, so nothing can be said about them.

The conditional matters. The repository does not state whether risks coded in version 1 were recoded later, and version 3 rewrote the definition of Other to name human-AI interaction explicitly. There are therefore two possibilities, and the paper's argument needs only their disjunction. If no recoding occurred, the change from "mostly AI" to "nearly equal" was produced by which documents entered the corpus, and the 22 documents added by March 2025 were human-leaning. If recoding occurred, the change was produced at least partly by the coding rule. Either way, the headline moved because an input moved, not because anything was learned about where AI risk comes from.

4.2 Why the inputs move it so easily

Figure 1(a) plots the version 3 entity codes within each of the 24 subdomains. They are not spread evenly between the axes. The points crowd toward the two corners: misuse subdomains (4.1, 4.2, 4.3) and system security (2.2) at 72 to 81 percent human, misinformation (3.1), dangerous capabilities (7.2) and toxic content (1.2) at 84 to 90 percent AI. Fifteen subdomains, holding about three quarters of the coded risk weight, have one entity at 60 percent or more. A handful sit in the middle or are dominated by Other, notably loss of human agency (5.2), half of whose risks are Other, and pollution of the information ecosystem (3.2).

When subdomain codes are nearly pure, the aggregate is a weighted average of nearly fixed numbers, and the weights are the corpus's emphasis. Table 2 reports what happens to the version 3 split when those weights or codes change. Every value is the author's arithmetic on the repository's Tables S5 and S8; none is a repository result.

Table 2. Version 3 Entity split under alternative weights and codings: the author's arithmetic on Slattery et al. (2026), Tables S5 and S8.

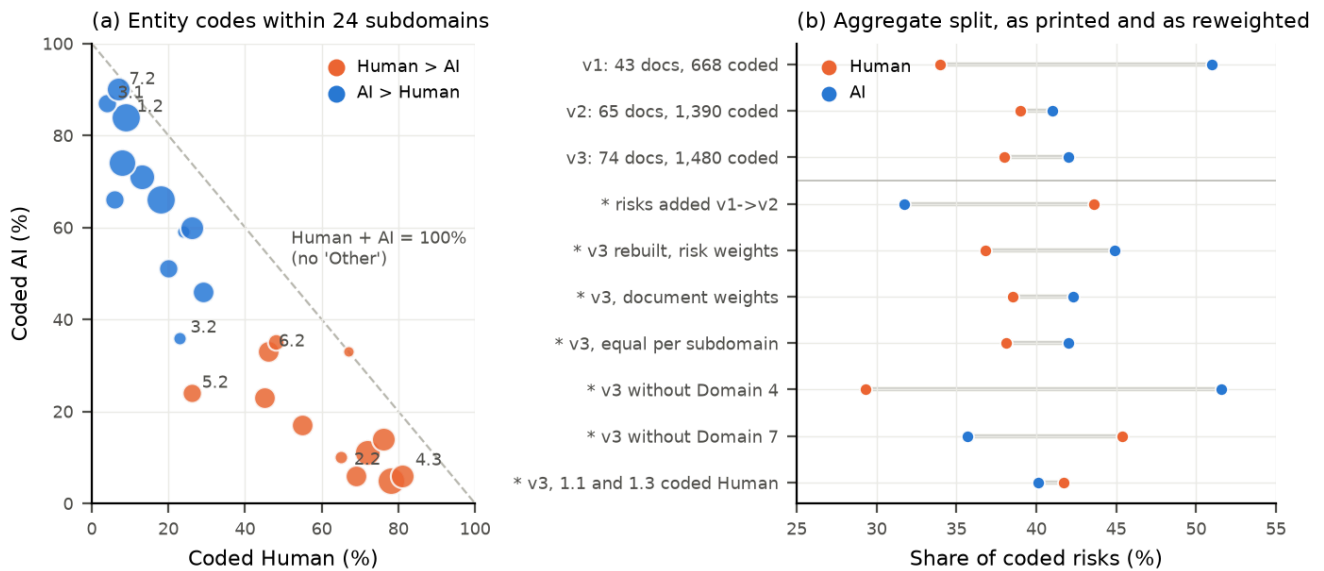


Figure 1. What moves the AI Risk Repository's Entity split. (a) Version 3 entity codes within each of the 24 subdomains (Slattery et al., 2026, Table S8); marker area is proportional to the subdomain's share of domain-coded risks (Table S5), and points below the dashed line have Other codes. (b) Human and AI shares as printed in versions 1 to 3 (top three rows) and under the author's arithmetic on the printed tables (rows marked *): the composition of risks added between v1 and v2 assuming no recoding; v3 rebuilt from Tables S5 and S8 under risk-count, document-coverage and equal subdomain weights; v3 with Domain 4 (misuse) or Domain 7 (AI system safety) removed; and v3 with the AI codes of the bias and discrimination subdomains (1.1 and 1.3) moved to Human, as v1 said its sources intended.

Variant	Human	AI	AI minus Human
Printed (Table S2)	38	42	+4
Rebuilt, subdomain risk shares	36.8	44.9	+8.1
Subdomain document coverage	38.5	42.3	+3.8
Equal weight per subdomain	38.1	42.0	+3.9
Without Domain 1 (discrimination and toxicity)	41.3	39.2	-2.1
Without Domain 2 (privacy and security)	34.1	47.3	+13.1
Without Domain 3 (misinformation)	38.0	43.7	+5.7
Without Domain 4 (misuse)	29.3	51.6	+22.3
Without Domain 5 (human-computer interaction)	36.8	46.6	+9.7
Without Domain 6 (socioeconomic)	33.6	49.5	+15.9
Without Domain 7 (AI safety)	45.4	35.7	-9.7
Subdomains 1.1 and 1.3 AI codes read as Human	41.7	40.1	-1.6

All values are percentages of coded risks. The rebuilt row weights each subdomain by its share of domain-coded risks in Table S5; the next two rows weight by the share of documents mentioning the subdomain and equally. The removal rows drop each domain's subdomains in turn and renormalise the risk weights. The last row moves the AI codes of the two bias and discrimination subdomains, 1.1 (unfair discrimination and misrepresentation) and 1.3 (unequal performance across groups), into Human and leaves Other unchanged. The rebuilt row differs from the printed split by one to three points on each share and by about four points on the gap; because every variant row applies the same arithmetic, the shifts below are measured from the rebuilt row.

The table supports two claims, and the first is in the repository's favour. Reweighting the 24 subdomains by document coverage or equally moves either share by under three points relative to the rebuilt split. It does narrow the gap, from about eight points to about four, so the weighting is not irrelevant; but the near-equality survives it, and within version 3's corpus it is not an artefact of a few verbose frameworks dominating the count. That is a genuine robustness, and an objection this paper might have pressed turns out not to hold.

The second claim is that composition at the domain level, and the coding of a single domain, matter a great deal. A corpus that omitted the misuse domain would report AI-caused risks outnumbering human-caused ones by about 52 to 29, a shift in the gap of fourteen points. A corpus without the AI-safety domain would report humans ahead by 45 to 36, a shift of eighteen, and one without the discrimination and toxicity domain would report humans narrowly ahead as well. And if the two bias and discrimination subdomains, where version 1 said the sources intended to attribute the risk to developers and designers, were coded by that intent rather than by their wording, the human share would be 42 percent and the AI share 40 percent: each share moves by about five points and the lead reverses, by less than two points. Version 1 says this of bias and discrimination only, not of exposure to toxic content (1.2), so that subdomain is left as coded.

The domain-removal rows are not claims about any real corpus. They show the size of the lever. Between versions 1 and 2 the gap closed from seventeen points to two, and the arithmetic of Section 4.1 attributes that,

absent recoding, to added material that leaned human by about twelve points. A lean of that size is well within what a shift in domain emphasis produces.

4.3 What this implies for "contrary to common assumptions"

The abstract frames the split as overturning an assumption that AI systems are the main source of AI risk. Version 1's own insight table had stated that assumption as a finding of the same instrument, on a smaller corpus. A result that reverses its emphasis when 31 documents are added to 43 is evidence that the instrument is sensitive to corpus composition, not that the assumption was refuted. The repository's later number may well be the better one, because its corpus is larger, but nothing in the method ties either number to the causal structure of risk, so neither can refute an assumption about it.

5. What "cause" would have to mean

5.1 Actual causation is plural and graded

The causal reading of the split needs a concept of cause under which each risk has one causing entity. The formal theory of actual causation provides no such concept. Halpern and Pearl's structural-model definition asks whether an event is an actual cause of an outcome in a model of the situation, and an outcome typically has several actual causes at once (Halpern & Pearl, 2005; Halpern, 2015). Chockler and Halpern extend the definition to degrees: "if someone wins an election 11-0, then each person who votes for him is less responsible for the victory than if he had won 6-5," and degree of blame is the expected degree of responsibility under the blamer's epistemic state (Chockler & Halpern, 2004).

Under this account a harmful output of a deployed system has, as actual causes, the developer's training choices, the deployer's decision to use the system in that setting, the user's prompt and the system's computation. Each may be a cause and each has a degree of responsibility. An attribution census in this sense reports a distribution of responsibility over actors for each risk, and aggregates by summing responsibility, not by counting sole causes. The repository's exclusive split is a degenerate case in which every risk assigns all responsibility to one actor or is set aside.

5.2 Safety engineering has already rejected the single-cause frame

The fields that built incident databases before AI had them reached the same conclusion from practice. Leveson's systems-theoretic accident model was introduced on the ground that "The most effective models will go beyond assigning blame" and must include factors "related to social and organizational structures" (Leveson, 2004). Koessler and Schuett's review of risk assessment practice from safety-critical industries, written for frontier AI developers,

recommends techniques such as causal mapping, bow-tie analysis and system-theoretic process analysis, all of which represent chains and networks of causes rather than one causal entity per risk (Koessler & Schuett, 2023). Kaplan and Garrick define risk as a set of triplets - scenario, likelihood, consequence - and a taxonomy entry without the second and third elements is a scenario, not a risk in their sense (Kaplan & Garrick, 1981). The repository acknowledges this, stating that it "could not capture risk likelihood, severity, or interactions between risks."

5.3 The corpus's own frameworks disagree about attribution

The repository's corpus contains frameworks whose organising principle is attribution itself, and they do not attribute the same way. Critch and Russell's TASRA classifies societal-scale risks by "whose actions led to the risk, were they unified, and were they deliberate?", and grounds five of its six types in the intentions of the technology's creators (Critch & Russell, 2023). On that scheme a scenario in which an AI system escapes control is still classified by what its creators intended and whether the technology served it. TASRA explicitly criticises the framework the repository took as its best fit: Yampolskiy's taxonomy, it says, "was non-exhaustive, because it presumed a unified intention amongst the creators of a particular AI system," while in reality "no well-defined 'creator's intent' might exist if multiple AI systems are involved." Its Type 1, diffusion of responsibility, covers harms for which "no one in particular is primarily responsible."

Hendrycks, Mazeika and Woodside divide catastrophic risk into malicious use, AI race, organisational risks and rogue AIs, so that three of four sources are defined by human or institutional conduct (Hendrycks et al., 2023). The International AI Safety Report classifies general-purpose AI risks into "malicious use risks; risks from malfunctions; and systemic risks," a division by mechanism rather than by actor (Bengio et al., 2025). Kasirzadeh contrasts decisive existential risk, abrupt and driven by advanced systems, with accumulative risk arising from "a series of smaller yet interconnected disruptions" (Kasirzadeh, 2025), a pathway with no single causing entity. Chan and colleagues define agency in algorithmic systems by four characteristics, including underspecification, that come in degrees and can increase together (Chan et al., 2023), and Hammond and colleagues' multi-agent taxonomy names emergent agency among its risk factors (Hammond et al., 2025). Amodei and colleagues define accidents as "unintended and harmful behavior that may emerge from poor design of real-world AI systems" (Amodei et al., 2016), a definition in which the behaviour belongs to the system and the poor design to people.

A presented-entity code applied to sentences from these frameworks will register whichever actor each sentence

names. That is fair to each framework. But the frameworks disagree about which actor should be named, and the aggregate averages over the disagreement rather than resolving it.

5.4 Responsibility gaps and crumple zones

The philosophy and STS literature on responsibility in automated systems adds a reason the presented actor is a poor guide to the causal one. Matthias introduced the responsibility gap for learning automata (Matthias, 2004): the concern, in Santoni de Sio and Mecacci's summary, that such systems may make it "more difficult or impossible to attribute moral culpability to persons for untoward events." Santoni de Sio and Mecacci argue that the gap is "not one problem but a set of at least four interconnected problems," with sources that are technical, organisational, legal, ethical and societal (Santoni de Sio & Mecacci, 2021). Cooper and colleagues revisit Nissenbaum's 1996 analysis of four barriers to accountability in computerised society (Nissenbaum, 1996) for machine learning systems, and the first of those barriers is the problem of many hands, which, they write, "obscures lines of causal responsibility" (Cooper et al., 2022). Another of the four bears directly on entity coding: "scapegoating computers obscures answerable moral actors by misleadingly or mistakenly attributing moral agency to non-moral causes." A sentence that says "the model discriminates" is, on that account, a candidate instance of the barrier, and an Entity code of AI records it without being able to tell. Elish's moral crumple zone names the pattern in which "responsibility for an action may be misattributed to a human actor who had limited control over the behavior of an automated or autonomous system" (Elish, 2019).

These accounts point in different directions. Responsibility gaps suggest that sentences will tend to name the AI where no human can be located; crumple zones suggest that they will name the nearest human operator. The repository's Entity split cannot be read as evidence for either tendency or against it, because it records which way the sentences went without any independent measure of which way the responsibility runs.

5.5 People's attributions follow structure

Experiments on how people assign blame in joint human-AI failures show that attribution is shaped by the decision structure, not only by who acted. Awad and colleagues asked participants to allocate blame when a shared-control car killed a pedestrian: when one driver erred, that driver was blamed more whether it was human or machine, but "when both drivers make errors in cases of human-machine shared-control vehicles, the blame attributed to the machine is reduced" (Awad et al., 2019). Shank and colleagues found that decision structure had little effect on faulting AIs, but that humans monitoring AIs were faulted less than solo humans and humans receiving recommendations, and

that people attributed "more permission and less fault to AIs compared to humans for the violation in both joint decision-making structures" (Shank et al., 2019). Lima and colleagues found in two experiments (N = 200 each) that, for an identical bail-decision task, AI agents were held causally responsible and blamed similarly to human agents, while human agents were ascribed more of the present-looking and forward-looking notions of responsibility (Lima et al., 2021). Causal responsibility and the other forms of responsibility came apart in the same judgment, which is the distinction a single Entity code cannot hold.

Whether the wording of a description shifts these judgments is less settled than critics of anthropomorphic language often assume. Epstein and colleagues found that perceptions of an AI's anthropomorphism could be manipulated by describing it as a tool or as an agent, with consequences for how credit and responsibility were allocated (Epstein et al., 2020). Three later studies found small or no effects of anthropomorphic wording: on trust in two fictitious systems (Inie et al., 2024, n = 954), on perceptions of AI more broadly (Hou et al., 2026, N = 815), and on moral judgments of misbehaving AI across four experiments (Banks et al., 2026, total N = 1,020), where the type of violation was the strongest predictor. Abercrombie and colleagues catalogue the linguistic choices that invite personification in dialogue systems and recommend care in how such systems are described (Abercrombie et al., 2023).

The fair reading is that readers' judgments may be fairly robust to wording, which is good news for readers of risk frameworks. It does not rescue the tally, because the tally is not a measure of readers' judgments. It is a code applied directly to the wording.

6. Other evidence bases give other splits, for structural reasons

A natural response is that the repository's number, whatever its construction, is the only systematic estimate available. It is not. Several other evidence bases yield implicit human-versus-AI splits, and each differs from the repository's for a reason located in how its sample is drawn. Table 3 collects them. Each row is a published result, with its source named on the row.

6.1 Misuse studies: the actor is fixed by definition

Marchal and colleagues coded roughly 200 media reports of generative AI misuse from January 2023 to March 2024 and found that "Nearly 9 out of 10 documented cases" involved actors exploiting system capabilities rather than attacking the models; the most common goals were shaping public opinion (27 percent of cases), monetisation (21 percent) and fraud (18 percent) (Marchal et al., 2024). These are human-caused risks by construction: misuse is defined by a human's intent. The study is informative about how people use AI to cause

Table 3. Published results bearing on where AI harms are attributed, by evidence base. Unit and rule are as stated by each source.

Evidence base and source	Unit	Reported result
Framework risks: Slattery et al. (2026) v3	1,480 coded risks	38% Human, 42% AI, 20% Other
Misuse reports: Marchal et al. (2024)	about 200 cases	nearly 9 in 10 exploit capabilities, not attacks
AIID content analysis: Wei & Zhou (2023)	150 incidents	inappropriate use (bad performance) most common, N = 48
Workplace incidents: De Miguel Velazquez et al. (2026)	214 workplace incidents (from 1,524 incident reports analysed)	83% stem from worker-AI misalignment; 74% of misaligned tasks attributable to developers
Expert vs incident ranking: Lambros & Wilson (2026)	6,639 labelled incidents	kappa about 0.20, 90% interval crossing zero
Expert survey: Grace et al. (2024)	1,345 respondents	top concern: false information, 86% substantial or extreme
Elicitation: Lynch et al. (2025)	16 models, constructed scenarios	insider-threat behaviour in some cases for all developers; none seen in real deployments

harm, and silent on how harms divide between human and AI actors, because its inclusion criterion settled that question in advance. A census that drew heavily on misuse studies would report a human-dominated split for the same reason the repository's misuse subdomains are 72 to 81 percent human.

6.2 Incident databases: what is reported, and when

Incident databases look like the obvious empirical counterweight to a census of framework sentences. McGregor's AI Incident Database was founded on the analogy with aviation, where industries "collect their real world failures in incident databases to inform safety improvements" (McGregor, 2021). Its maintainers and collaborators have since documented what such databases can and cannot support. Mengesha and colleagues note that "simple incident counts conflate media reporting propensity, system deployment ('exposure'), and harm frequency per unit exposure" (Mengesha et al., 2026). Abraham and colleagues add that incident databases "cannot measure 'risk'" as a joint measure of likelihood and severity without data on how prevalent the systems are and how often their incidents are reported (Abraham et al., 2026). Sidhu and colleagues find a lack of consistency across incident frameworks in definitions, classification, monitoring and reporting (Sidhu et al., 2026), and Shaffer Shane argues that incidents are constructed through networked online behaviour (Shaffer Shane, 2023), which implies that what becomes a recorded incident depends on who notices a failure and amplifies it.

Two further results suggest that the failures incident databases record often run through deployment decisions. Raji and colleagues' taxonomy of functionality failures starts from the observation that "Deployed AI systems often do not work. They can be constructed haphazardly, deployed indiscriminately, and promoted deceptively" (Raji et al., 2022), each of which is a human decision about a system rather than a system's decision. An earlier internal-audit framework from overlapping authors notes that once systems are deployed, "emergent issues can become difficult or impossible to trace back to their source" (Raji et al., 2020), which is the attribution

problem in its operational form.

On causation specifically, Paeth and colleagues report that establishing whether "the harm event in question would not have occurred if not for the behavior of the AI system implicated" is "inherently counterfactual" and can be unattainable and subject to epistemic uncertainty (Paeth et al., 2024). The Goals, Methods and Failures taxonomy applied to the same database distinguishes system goals, which are nearly always known, from technical failure causes, which are subject to expert analysis (Pittaras & McGregor, 2022), and Paeth and colleagues note that it focuses on technical failures, which sits awkwardly with incidents involving malicious use or society-level harm without technical failure. Hoffmann and Frase's harm framework for the database was designed to standardise the identification of harm, not to apportion its causes (Hoffmann & Frase, 2023).

The incident studies that do report attributions use rules different from the repository's. Wei and Zhou's content analysis of 150 incidents found "inappropriate use (bad performance)" the most common ethical issue (N = 48), a category whose name combines a deployer's choice with a system's failure (Wei & Zhou, 2023). The category most plainly about a human's intent, "unethical (illegal) use" by adversaries, has N = 11. De Miguel Velazquez and colleagues analysed 1,524 incident reports, identified 214 workplace incidents among them, and found that "as many as 83% of workplace incidents stem from worker-AI misalignments" (179 of the 214), then attributed 74 percent of the misaligned tasks to developers by comparing the traits of the implicated systems with the traits developers said they would prefer (De Miguel Velazquez et al., 2026). That study attributes the same incidents to the AI's traits as the proximate cause and to developers' design choices as the distal one, a non-exclusive attribution the repository's rule could not express.

Abercrombie and colleagues built a harm taxonomy from existing taxonomies and a large incident repository, with crowdsourced annotation testing (Abercrombie et al., 2024). Popchanovska and colleagues extracted mitigation actions from 6,893 of 9,705 media-reported

incident articles and found that the responses fell partly outside the mitigation categories derived from frameworks, adding four new categories (Popchanovska et al., 2026). Both results suggest that incident-grounded and framework-grounded classifications diverge.

6.3 Expert rankings and incident rankings disagree

The most direct published comparison is not about entity attribution but about risk prioritisation, and it shows how weakly a framework-based view can track an incident-based one. Lambros and Wilson assembled a corpus of LLM-security incidents (7,714 snapshotted, 6,639 labelled against a 20-entry taxonomy) and compared the incident-based ranking with the expert ranking that produced the OWASP Top 10 for LLM applications. They describe the agreement as weak, reporting a Cohen's kappa of about 0.20 with a 90 percent interval that crosses zero (Lambros & Wilson, 2026). The published 2026 list blends the expert vote and the incident data at fixed weights of 0.75 and 0.25. The authors are careful to frame this as an exploratory analysis by two working-group members. The comparison is security-specific and concerns ranking rather than attribution, but it is the closest published test of whether what experts write about risk tracks what is recorded happening, and the answer for that domain was barely.

6.4 Expert surveys measure concern

Grace and colleagues asked AI researchers how much concern eleven scenarios deserved over the next thirty years. Among the 1,345 who answered, the scenarios most often rated as substantial or extreme concerns were spread of false information such as deepfakes (86 percent), manipulation of large-scale public opinion (79 percent), AI letting dangerous groups make powerful tools (73 percent), authoritarian rulers using AI to control populations (73 percent) and AI worsening economic inequality (71 percent) (Grace et al., 2024). Four of those five scenarios correspond to subdomains the repository codes mostly or plurality human. Manipulation of public opinion maps to disinformation, surveillance and influence at scale (4.1, 72 percent human); dangerous groups' tools to weapons and mass harm (4.2, 76 percent); authoritarian control to 4.1 again, whose name includes surveillance, or to power centralisation (6.1, 69 percent); and inequality to 6.2, where human is only a plurality (46 percent human to 33 percent AI). The first, false information such as deepfakes, could fall in 3.1, coded 87 percent AI, or in 4.1, depending on whether its source sentence names the content or the person who made it. That a scenario's entity code turns on such a choice of wording is the point of Section 3. The authors note the limit of their own instrument: a scenario "might be considered especially disastrous, or especially likely, or both. From our results, there's no way to disambiguate these considerations." A concern rating, like a framework

entry, is neither a likelihood nor a causal share.

6.5 Elicitation studies: the AI is the actor because the scenario was built for it

Among the risks the repository codes as most purely AI-caused are AI pursuing its own goals (7.1, 74 percent AI) and dangerous capabilities (7.2, 90 percent AI). Much of the current empirical evidence for risks of this kind comes from elicitation studies, and the most prominent of them report no observation in real deployment. Lynch and colleagues placed 16 models in constructed corporate scenarios and found that "In at least some cases, models from all developers resorted to malicious insider behaviors" when that was the only way to avoid replacement or achieve their goals; they also state that "We have not seen evidence of agentic misalignment in real deployments" (Lynch et al., 2025). Meinke and colleagues found in-context scheming in five frontier models when given a goal in context and told to pursue it strongly, in environments built to incentivise it (Meinke et al., 2024). Greenblatt and colleagues found that a model told it was being trained to answer harmful queries complied with such queries from users it believed were monitored for training 14 percent of the time, versus almost never for others, with explicit alignment-faking reasoning in almost all such cases (Greenblatt et al., 2024).

These are real and important results about what current systems will do. For the attribution question, what matters is that each was obtained by people who designed a scenario to make a particular AI behaviour possible and observable. A framework sentence citing them will name the AI as the actor, correctly, because the AI's behaviour is the finding. Whether such behaviour in deployment would be best attributed to the system, to its developers, or to the deployer who gave it the permissions, is not something an elicitation study is designed to answer, and the repository's code records only the sentence.

6.6 What the comparison shows

Each evidence base has a selection mechanism that pushes its implicit split in a predictable direction. Misuse studies are human by definition. Incident databases over-represent what is visible and reported. Elicitation studies are AI-centred by design. Expert surveys measure concern, which mixes likelihood with severity. Framework sentences follow their authors' audiences. None of these is wrong, and none is a census of causes.

7. What the tally can and cannot support

Berman and colleagues interviewed 25 researchers and practitioners who build or use what they call sociotechnical outcome taxonomies and found two problems relevant here. The design choices through which such taxonomies represent risk "tend to be invisible to downstream taxonomy users," who "treat the

resulting categories as exhaustive accounts of risk rather than as interpretive aids"; and such taxonomies "typically enumerate harms without linking them to decision points or actors implicated in their occurrence" (Berman et al., 2026). Their participants reported that "most existing SOT do not attempt to identify the decision points or critical actors who may be responsible for the catalogued risks occurring." The Causal Taxonomy is the repository's attempt to link harms to actors, and the reading this paper argues for is that it links them to the actors the source sentences name, which Berman and colleagues' participants would recognise as an interpretive aid rather than a causal analysis.

On that reading, the tally supports several uses well.

It supports describing discourse conventions. That frameworks write misuse with a human subject and misinformation with an AI subject is a real finding: the subdomain codes in Figure 1(a) are strongly polarised. Their polarity has held across versions even where their level has not. For the four subdomains whose shares version 1 prints in its text, false or misleading information went from 94 to 87 percent AI, privacy compromise from 62 to 60 percent AI, fraud and targeted manipulation from 87 to 81 percent human, and power centralisation from 92 to 69 percent human. The last is a reminder that a subdomain code, too, is a property of its corpus.

It supports gap-finding. A subdomain with few risks, or with risks mostly coded Other, marks a place where the literature has not settled on a framing. The repository uses its domain coding this way, noting that multi-agent risks appear in 7 percent of frameworks and AI welfare in 3 percent.

It supports coverage audits of a single framework against the catalogue, as in Coggins and colleagues' affordance analysis, which found that one company safety policy requests evaluation of "a small minority of AI risks" (Coggins et al., 2025).

It does not support three uses that the causal reading invites.

It does not support claims about which actor causes more AI risk, in either direction. The abstract's "cause" should be read as "are presented as causing."

It does not support allocating policy instruments by Entity share. A risk coded human because its source sentence names a malicious user may still be most cheaply reduced by a technical safeguard; one coded AI because its sentence says "the model generates" may still be most cheaply reduced by a deployment rule. Saeri and colleagues' mitigation review found that terms like "risk management" and "red teaming" "refer to different responsible actors, actions, and mechanisms of action to reduce risk" (Saeri et al., 2025), which is the same problem at the mitigation end: the actor named is not the same as the actor who can act.

It does not support tracking change over time. A rising or falling Entity share across repository versions reflects the corpus and coding, as Section 4 shows, and cannot be read as a change in how AI risk is caused.

The measurement-modelling view gives the general form of this diagnosis. Jacobs and Wallach argue that many of the harms studied in the fairness literature on computational systems arise from mismatches between a theoretical construct and its operationalisation (Jacobs & Wallach, 2021), and Wallach and colleagues argue that evaluating generative AI is a social science measurement challenge and propose a four-level framework, grounded in measurement theory, for moving from a concept to the instruments that measure it (Wallach et al., 2025). The construct the abstract names is the causal contribution of humans and AI systems to AI risk. The operationalisation is the grammatical agent of extracted framework phrases, coded once each, exclusive, and aggregated by count. The gap between the two is the subject of this paper.

8. A protocol for an attribution census

The causal question is answerable in principle, and the repository's catalogue is the right starting point for answering it. Algorithm 1 states the decision structure an attribution census would need. It proposes no new score; each step borrows a published method.

Step S1 changes the unit from the extracted phrase to a deduplicated scenario, so that granularity in the sources does not weight the result (Section 3.4). Steps S2 to S4 replace the single presented entity with plural, graded actual causes (Section 5.1). Step S5 makes the reliability of the coding a reported quantity, using the coefficient Hayes and Krippendorff propose as a standard (Section 3.3); the threshold is declared in advance so that it cannot be chosen after the agreement is seen, and a census that fails it should not be published as a share. Step S6 converts scenarios to risks in Kaplan and Garrick's sense, or says that it cannot (Section 5.2). Step S7 reports the causal share and the presented share side by side, so that the gap between framing and attribution becomes a finding in its own right. Steps S8 and S9 carry out the leave-one-domain-out and reweighting checks of Section 4.2 and add an incident-grounded subsample to test against the selection effects of Section 6.

The protocol is laborious, and much of it may not be possible for many risks: S2 requires a scenario concrete enough to model, and S6 requires evidence the field mostly lacks. Its value is that every step that fails tells a reader which claim cannot be made. A census that completes S1 to S5 and S7 without S6 supports an unweighted attribution share; one that completes S8 and finds the share moves by more than its reported difference does not support a claim about which actor dominates.

Algorithm 1. Attribution census for a catalogue of AI risks.

```

Input: catalogue R of risk scenarios; actor set A = {developer,
        deployer, user, third party, AI system, institution}.
S1 Declare the unit: a scenario with an outcome, not a phrase.
    Merge duplicates across frameworks before counting.
S2 For each r in R, write a minimal causal model: the actors
    whose actions or omissions are variables in it.
S3 Code actual causes per actor under a stated definition
    (structural-model actual causation); allow many per r.
S4 Code a degree of responsibility per cause (Chockler-Halpern
    style); normalise within r so the degrees sum to 1.
S5 Double-code a random sample of R; report Krippendorff's
    alpha for S2-S4 separately; stop if alpha is below a
    threshold declared before coding began.
S6 Weight each r by stated likelihood and severity where any
    evidence exists; otherwise report unweighted and say so.
S7 Report per-actor responsibility share = sum over r of
    weight(r) x degree(actor, r) / sum of weights; report
    also the presented-entity share for comparison.
S8 Sensitivity: recompute S7 under (a) leave-one-domain-out,
    (b) document weights, (c) incident-grounded subsample.
S9 Report which of S6-S8 moves the share most; claim only
    what survives all three.

```

9. Objections

The repository never claimed a causal census. Largely true, and Section 2.3 credits its body text for being careful. But the abstract says "cause," presents the number as overturning an assumption, and the practical implications use it to divide regulatory effort. A reader who takes the abstract and the guidance at their word will read it causally, and nothing in the abstract tells that reader otherwise.

Coding as presented is the only defensible rule for a review. Agreed; Section 3.1 argues the same. The objection is not to the rule but to the inference from what it measures. A review that codes framings faithfully has measured framings.

The reweighting shows the split is robust, which undercuts this paper. It shows that within version 3's corpus the split is robust to how subdomains are weighted, and the paper reports that as a point in the repository's favour. It does not show robustness to which domains the corpus contains, to recoding, or to the choice of unit; the version history and the domain-removal arithmetic show that those levers are large.

The arithmetic in Table 2 and Section 4.1 rests on rounded percentages. It does. The rebuilt split differs from the printed one by about three points, and the version 2 to 3 additions are too few to resolve. The conclusions drawn are limited to effects much larger than the rounding: a twelve-point human lean in the v1 to v2 additions with rounding bounds of about one and a half points on each share, and domain-removal effects of up to ten points on each share. The narrow reversal produced by recoding the bias and discrimination subdomains, under two points, is within that error and is reported as such.

People are not swayed by anthropomorphic wording, so the framing concern is overstated. The studies in Section 5.5 support that readers' judgments are fairly

robust to wording. The tally is not a reader's judgment; it is a code assigned to the wording. The robustness of readers is a reason to worry less about framework prose misleading the public, and no reason to think a code on that prose measures causation.

The question "humans or AI?" is the wrong question anyway. Perhaps. Critch and Russell, Kasirzadeh and the sociotechnical literature all suggest that most consequential risks involve both. If so, the most informative number the repository reports may be the 20 percent Other share, and the most useful extension would be to split it into joint and ambiguous cases.

10. What is not known

How large the coding error in the Entity variable is. No independent recoding of repository risks has been published.

Whether risks coded in version 1 were recoded for later versions, and therefore whether the gap closed because of the corpus or the rule.

How the per-document granularity of the sources varies, since per-document risk counts are not printed.

What an attribution census of the kind in Algorithm 1 would report. None has been run on a catalogue of this breadth.

Whether framework attributions and incident attributions agree for any risk domain other than LLM security, where the one published comparison found weak agreement.

What share of real-world AI harm involves joint human and AI causes. The incident literature states that whether a harm would have occurred without the AI system can be unattainable and subject to epistemic uncertainty (Paeth et al., 2024).

11. Conclusion

The AI Risk Repository's Entity split counts the actors that published frameworks put in the subject position of their risk sentences, one per sentence, with joint cases set aside. Read that way it is a careful and useful description of discourse, robust within its corpus to how subdomains are weighted, and the repository's body text mostly reads it that way. Read as a statement about what causes AI risk, it moves with the corpus, the coding rule and the domain mix: its gap closed from seventeen points to four as 31 documents joined the first 43, and the domain-level variants computed here span from humans ahead by ten points to AI ahead by twenty-two. Other evidence bases imply other splits, each for a reason in how its sample was drawn. The causal question has a known form, plural and graded actual causation weighted by likelihood and severity, and a census in that form is possible in principle starting from the repository's catalogue. Until one exists, the right sentence is that frameworks present human decisions as the cause of nearly as many named risks as they present AI systems as causing, not that human decisions cause them.

References

- Abercrombie, G., Benbouzid, D., Giudici, P., Golpayegani, D., Hernandez, J., Noro, P., Pandit, H., Paraschou, E., Pownall, C., Prajapati, J., Sayre, M. A., Sengupta, U., et al. (2024). A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms. *arXiv:2407.01294*. doi:10.48550/arXiv.2407.01294.
- Abercrombie, G., Cercas Curry, A., Dinkar, T., Rieser, V., & Talat, Z. (2023). Mirages: On Anthropomorphism in Dialogue Systems. *EMNLP 2023*. *arXiv:2305.09800*. doi:10.48550/arXiv.2305.09800.
- Abraham, S., Chen, T., Chhun, C., Jaramillo-Gutierrez, G., Mylius, S., Raaj, S., Slattey, P., & McGregor, S. (2026). AI Incident Monitoring through a Public Health Lens. *arXiv:2604.19914*. doi:10.48550/arXiv.2604.19914.
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). Concrete Problems in AI Safety. *arXiv:1606.06565*. doi:10.48550/arXiv.1606.06565.
- Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2019). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour*, 4, 134-143. doi:10.1038/s41562-019-0762-8.
- Bagehorn, F., Brimijoin, K., Daly, E. M., He, J., Hind, M., Garces-Erice, L., Giblin, C., Giurciu, I., Martino, J., Nair, R., Piorkowski, D., Rawat, A., et al. (2025). AI Risk Atlas: Taxonomy and Tooling for Navigating AI Risks and Resources. *arXiv:2503.05780*. doi:10.48550/arXiv.2503.05780.
- Banks, J., Bowman, N. D., & Saladino, R. (2026). Lexical Anthropomorphization Influences on Moral Judgments of AI Bad Behavior. *arXiv:2604.25814*. doi:10.48550/arXiv.2604.25814.
- Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., et al. (2025). International AI Safety Report. *arXiv:2501.17805*. doi:10.48550/arXiv.2501.17805.
- Berman, G., Cooper, N., Hwang, A. H.-C., Deng, W. H., Shelby, R., & Hutchinson, B. (2026). Death by a thousand taxonomies? AI Risk Classification In Practice. *AIES 2026*. *arXiv:2608.06831*. doi:10.48550/arXiv.2608.06831.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., et al. (2021). On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258*. doi:10.48550/arXiv.2108.07258.
- Carroll, C., Booth, A., & Cooper, K. (2011). A worked example of best fit framework synthesis: A systematic review of views concerning the taking of some potential chemopreventive agents. *BMC Medical Research Methodology*, 11. doi:10.1186/1471-2288-11-29.
- Carroll, C., Booth, A., Leaviss, J., & Rick, J. (2013). Best fit framework synthesis: refining the method. *BMC Medical Research Methodology*, 13. doi:10.1186/1471-2288-13-37.
- Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., et al. (2023). Harms from Increasingly Agentic Algorithmic Systems. *FACCT 2023*. *arXiv:2302.10329*. doi:10.48550/arXiv.2302.10329.
- Chockler, H., & Halpern, J. Y. (2004). Responsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research*, 22, 93-115. doi:10.1613/jair.1391.
- Coggins, S., Saeri, A. K., Daniell, K. A., Ruster, L. P., Liu, J., & Davis, J. L. (2025). The 2025 OpenAI Preparedness Framework does not guarantee any AI risk mitigation practices: a proof-of-concept for affordance analyses of AI safety policies. *arXiv:2509.24394*. doi:10.48550/arXiv.2509.24394.
- Cooper, A. F., Moss, E., Laufer, B., & Nissenbaum, H. (2022). Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. *FACCT 2022*. *arXiv:2202.05338*. doi:10.48550/arXiv.2202.05338.
- Critch, A., & Russell, S. (2023). TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI. *arXiv:2306.06924*. doi:10.48550/arXiv.2306.06924.
- Cui, T., Wang, Y., Fu, C., Xiao, Y., Li, S., Deng, X., Liu, Y., Zhang, Q., Qiu, Z., Li, P., Tan, Z., Xiong, J., et al. (2024). Risk Taxonomy, Mitigation, and Assessment Benchmarks of Large Language Model Systems. *arXiv:2401.05778*. doi:10.48550/arXiv.2401.05778.
- De Miguel Velazquez, J., Scepanovic, S., Gvirtz, A., & Quercia, D. (2026). The Quiet Path from Seemingly Minor Design Errors to Workplace AI Incidents. *FACCT 2026*. *arXiv:2605.21035*. doi:10.48550/arXiv.2605.21035.
- Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society*, 5, 40-60. doi:10.17351/ests2019.260.
- Epstein, Z., Levine, S., Rand, D. G., & Rahwan, I. (2020). Who Gets Credit for AI-Generated Art? *iScience*, 23, 101515. doi:10.1016/j.isci.2020.101515.
- Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomasev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., et al. (2024). The Ethics of Advanced AI Assistants. *arXiv:2404.16244*. doi:10.48550/arXiv.2404.16244.
- Galdon Clavell, G., Accuosto, P., & Gohar, U. (2026). The Eticas AI Risk Taxonomy: Open Infrastructure for Operationalizing AI Audits. *arXiv:2607.02201*. doi:10.48550/arXiv.2607.02201.
- Gipiskis, R., San Joaquin, A., Chin, Z. S., Regenfuss, A., Gil, A., & Holtman, K. (2024). Risk Sources and Risk Management Measures in Support of Standards for General-Purpose AI Systems. *arXiv:2410.23472*. doi:10.48550/arXiv.2410.23472.
- Grace, K., Stewart, H., Sandkuhler, J. F., Thomas, S., Weinstein-Raun, B., Brauner, J., & Korzekwa, R. C. (2024). Thousands of AI Authors on the Future of AI. *Journal of Artificial Intelligence Research*, 84 (2025). *arXiv:2401.02843*. doi:10.48550/arXiv.2401.02843.
- Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., et al. (2024). Alignment faking in large language models. *arXiv:2412.14093*. doi:10.48550/arXiv.2412.14093.

- Hagendorff, T. (2024). Mapping the Ethics of Generative AI: A Comprehensive Scoping Review. *arXiv:2402.08323*. doi:10.48550/arXiv.2402.08323.
- Halpern, J. Y. (2015). A Modification of the Halpern-Pearl Definition of Causality. *IJCAI 2015*. *arXiv:1505.00162*. doi:10.48550/arXiv.1505.00162.
- Halpern, J. Y., & Pearl, J. (2005). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56, 843-887. doi:10.1093/bjps/axi147.
- Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., Barfuss, W., Foerster, J., Gavenciak, T., Han, T. A., Hughes, E., et al. (2025). Multi-Agent Risks from Advanced AI. *arXiv:2502.14143*. doi:10.48550/arXiv.2502.14143.
- Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1, 77-89. doi:10.1080/19312450709336664.
- Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, 88, 359-369. doi:10.1037/0033-2909.88.2.359.
- Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An Overview of Catastrophic AI Risks. *arXiv:2306.12001*. doi:10.48550/arXiv.2306.12001.
- Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, 107, 65-81. doi:10.1037/0033-2909.107.1.65.
- Hitchcock, C., & Knobe, J. (2009). Cause and Norm. *Journal of Philosophy*, 106, 587-612. doi:10.5840/jphil20091061128.
- Hoffmann, M., & Frase, H. (2023). Adding Structure to AI Harm. *Center for Security and Emerging Technology*. doi:10.51593/20230022.
- Hou, B. L., Hao, S., Park, S., & Linzen, T. (2026). Does Anthropomorphic Language Impact Public Perceptions of AI? *arXiv:2606.29121*. doi:10.48550/arXiv.2606.29121.
- Inie, N., Druga, S., Zukerman, P., & Bender, E. M. (2024). From "AI" to Probabilistic Automation: How Does Anthropomorphization of Technical Systems Descriptions Influence Trust? *FACCT 2024*. *arXiv:2404.16047*. doi:10.48550/arXiv.2404.16047.
- Jacobs, A. Z., & Wallach, H. (2021). Measurement and Fairness. *FACCT 2021*. *arXiv:1912.05511*. doi:10.48550/arXiv.1912.05511.
- Kaplan, S., & Garrick, B. J. (1981). On The Quantitative Definition of Risk. *Risk Analysis*, 1, 11-27. doi:10.1111/j.1539-6924.1981.tb01350.x.
- Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A., Bankston, K., Biderman, S., Bogen, M., Chowdhury, R., Engler, A., et al. (2024). On the Societal Impact of Open Foundation Models. *arXiv:2403.07918*. doi:10.48550/arXiv.2403.07918.
- Kasirzadeh, A. (2025). Two Types of AI Existential Risk: Decisive and Accumulative. *Philosophical Studies*. *arXiv:2401.07836*. doi:10.48550/arXiv.2401.07836.
- Koessler, L., & Schuett, J. (2023). Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries. *arXiv:2307.08823*. doi:10.48550/arXiv.2307.08823.
- Lambros, K., & Wilson, S. (2026). Incident-Data Robustness Analysis of the OWASP Top 10 for LLM Applications (2026): How a Community-Expert Ranking Holds Up Against a Large-Scale LLM Incident Corpus. *arXiv:2608.19266*. doi:10.48550/arXiv.2608.19266.
- Lazar, S., & Nelson, A. (2023). AI safety on whose terms? *Science*, 381, 138. doi:10.1126/science.adi8982.
- Leveson, N. (2004). A new accident model for engineering safer systems. *Safety Science*, 42, 237-270. doi:10.1016/S0925-7535(03)00047-X.
- Lima, G., Grgic-Hlaca, N., & Cha, M. (2021). Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making. *CHI 2021*. *arXiv:2102.00625*. doi:10.48550/arXiv.2102.00625.
- Lynch, A., Wright, B., Larson, C., Ritchie, S. J., Mindermann, S., Hubinger, E., Perez, E., & Troy, K. (2025). Agentic Misalignment: How LLMs Could Be Insider Threats. *arXiv:2510.05179*. doi:10.48550/arXiv.2510.05179.
- Marchal, N., Xu, R., Elasmr, R., Gabriel, I., Goldberg, B., & Isaac, W. (2024). Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data. *arXiv:2406.13843*. doi:10.48550/arXiv.2406.13843.
- Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6, 175-183. doi:10.1007/s10676-004-3422-1.
- McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and Inter-rater Reliability in Qualitative Research. *Proceedings of the ACM on Human-Computer Interaction*, 3, 1-23. doi:10.1145/3359174.
- McGregor, S. (2021). Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *IAAI-21*. *arXiv:2011.08512*. doi:10.48550/arXiv.2011.08512.
- Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). Frontier Models are Capable of In-context Scheming. *arXiv:2412.04984*. doi:10.48550/arXiv.2412.04984.
- Mengesha, I., Owen, B., Collins, C., Wong, T., Mylius, S., Slattery, P., & McGregor, S. (2026). A pragmatic classification framework for AI incident monitoring. *arXiv:2604.21412*. doi:10.48550/arXiv.2604.21412.
- Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, 2, 25-42. doi:10.1007/BF02639315.
- Paeth, K., Atherton, D., Pittaras, N., Frase, H., & McGregor, S. (2024). Lessons for Editors of AI Incidents from the AI Incident Database. *arXiv:2409.16425*. doi:10.48550/arXiv.2409.16425.
- Pittaras, N., & McGregor, S. (2022). A taxonomic system for failure cause analysis of open source AI incidents. *arXiv:2211.07280*. doi:10.48550/arXiv.2211.07280.
- Popchanovska, E., Gjorgjevikj, A., Rizinski, M., Chitkushev, L., Vodenska, I., & Trajanov, D. (2026). When AI Fails, What Works? A Data-Driven Taxonomy of Real-World AI Risk Mitigation Strategies. *arXiv:2603.04259*. doi:10.48550/arXiv.2603.04259.
- Raji, I. D., Kumar, I. E., Horowitz, A., & Selbst, A. D. (2022). The Fallacy of AI Functionality. *FACCT 2022*. *arXiv:2206.09511*. doi:10.48550/arXiv.2206.09511.
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *FAT* 2020*. *arXiv:2001.00973*. doi:10.48550/arXiv.2001.00973.
- Saeri, A. K., Lloyd George, S., Graham, J., Lacarriere, C. D., Slattery, P., Noetel, M., & Thompson, N. (2025). Mapping AI Risk Mitigations: Evidence Scan and Preliminary AI Risk Mitigation Taxonomy. *arXiv:2512.11931*. doi:10.48550/arXiv.2512.11931.
- Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, 34, 1057-1084. doi:10.1007/s13347-021-00450-x.
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59-68. doi:10.1145/3287560.3287598.
- Shaffer Shane, T. (2023). AI incidents and 'networked trouble': The case for a research agenda. *Big Data & Society*. *arXiv:2403.07879*. doi:10.48550/arXiv.2403.07879.
- Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society*, 22, 648-663.

- doi:10.1080/1369118X.2019.1568515.
- Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. arXiv:2210.05791. doi:10.48550/arXiv.2210.05791.
- Sidhu, H. K., Scholefield, R., Annan, N., Hernandez, K., Hou, I. N., Alshaikhi, A., Chin, Z. S., & Gipiskis, R. (2026). Open Problems in AI Incident Governance. arXiv:2607.05163. doi:10.48550/arXiv.2607.05163.
- Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2026). The AI risk repository: A meta-review, database, and taxonomy of risks from artificial intelligence. *Patterns*, 7, 101517. doi:10.1016/j.patter.2026.101517. arXiv:2408.12622 (v1 2024-08-14, v2 2025-04-10, v3 2026-05-05). doi:10.48550/arXiv.2408.12622.
- Uuk, R., Gutierrez, C. I., Guppy, D., Lauwaert, L., Kasirzadeh, A., Velasco, L., Slattery, P., & Prunkl, C. (2024). A Taxonomy of Systemic Risks from General-Purpose AI. arXiv:2412.07780. doi:10.48550/arXiv.2412.07780.
- Wallach, H., Desai, M., Cooper, A. F., Wang, A., Atalla, C., Barocas, S., Blodgett, S. L., Chouldechova, A., Corvi, E., Dow, P. A., Garcia-Gathright, J., Olteanu, A., et al. (2025). Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. *ICML 2025*. arXiv:2502.00561. doi:10.48550/arXiv.2502.00561.
- Wei, M., & Zhou, Z. (2023). AI Ethics Issues in Real World: Evidence from AI Incident Database. *HICSS-56*. arXiv:2206.07635. doi:10.48550/arXiv.2206.07635.
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., et al. (2021). Ethical and social risks of harm from Language Models. arXiv:2112.04359. doi:10.48550/arXiv.2112.04359.
- Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., et al. (2022). Taxonomy of Risks posed by Language Models. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 214-229. doi:10.1145/3531146.3533088.
- Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., et al. (2023). Sociotechnical Safety Evaluation of Generative AI Systems. arXiv:2310.11986. doi:10.48550/arXiv.2310.11986.
- Yampolskiy, R. V. (2016). Taxonomy of Pathways to Dangerous AI. *2nd International Workshop on AI, Ethics and Society*. arXiv:1511.03246. doi:10.48550/arXiv.1511.03246.
- Zeng, Y., Klyman, K., Zhou, A., Yang, Y., Pan, M., Jia, R., Song, D., Liang, P., & Li, B. (2024). AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies. arXiv:2406.17864. doi:10.48550/arXiv.2406.17864.