

“Death By A Thousand Taxonomies?”: AI Risk Classification In Practice

Glen Berman¹, Ned Cooper², Angel Hsing-Chi Hwang³, Wesley Hanwen Deng⁴, Renee Shelby⁵, Ben Hutchinson⁵

¹Australian National University

²Cornell University

³University of Southern California

⁴Microsoft Research,

⁵Google Research

glen.berman@anu.edu.au, ned.cooper@cornell.edu, angel.hwang@usc.edu, wesleydeng@microsoft.com, reneeshelby@google.com, benhutch@google.com

Abstract

The harms in which AI is implicated range in nature and scope from unsafe user interactions through to the societal-wide consequences of AI adoption. Classification of the diverse risks of AI is foundational to AI governance: regulators, technology firms, and policymakers need structured accounts of risk upon which to act. Researchers and practitioners have accordingly developed many Sociotechnical Outcome Taxonomies (SOT). This paper presents an empirical study of SOT development and use, drawing on 25 interviews with researchers and practitioners across industry, academia, civil society, and government. We find SOT are weakly integrated into AI governance processes, and identify two features of SOT design and use that explain why. First, the design choices through which SOT produce structured representations of the complex problem space of AI risks tend to be invisible to downstream taxonomy users. Those users treat the resulting categories as exhaustive accounts of risk rather than as interpretive aids. Second, SOT typically enumerate harms without linking them to decision points or actors implicated in their occurrence, leaving accountability difficult to assign. We close with design recommendations for SOT developers and users, and argue realising the potential of SOT requires governance infrastructure that does not yet exist.

Extended version — <https://arxiv.org/abs/2608.06831>

1 Introduction

Researchers, policymakers, and industry practitioners have developed structured classifications of the risks, harms, and impacts associated with Artificial Intelligence (AI) systems (e.g., Weidinger et al. 2022; Hutiri, Papakyriakopoulos, and Xiang 2024; Domínguez Hernández et al. 2024; Chan et al. 2023). These artefacts — variously termed taxonomies, typologies, repositories, or frameworks — are designed to serve as boundary objects, rendering the social consequences of AI development legible, enumerable, and governable (Star 2010). We refer to these artefacts as *sociotechnical outcome taxonomies* (SOT). SOT are intended to be foundational to AI governance, referring to the practices, methods, policies,

and regulations used to manage the development, deployment, and monitoring of AI technologies (Batool, Zowghi, and Bano 2025). SOT form part of an emerging set of artefacts responding to civil society demands for accountability in AI development. Specifically, SOT translate the concerns of stakeholders in AI development, internal and external to AI firms, into actionable schemas for product teams.

SOT have proliferated over the past decade. Early efforts, primarily from academia, established conceptual foundations, such as distinguishing allocational from representational harms in machine learning (Barocas et al. 2017), identifying stakeholder-centred risks in language technologies (Bender 2019), and highlighting distinctions between process- and outcomes-based approaches to harms (e.g., Binns 2018). As foundation models and generative AI systems scaled commercially, industry research labs became significant SOT producers, developing taxonomies oriented towards deployed systems (Weidinger et al. 2022; Bird, Ungless, and Kasirzadeh 2023; Shelby et al. 2023; Solaiman et al. 2023). Parallel efforts emerged from civil society, intergovernmental bodies, and national standards institutions, drawing on rights-based framings, regulatory alignment, and community-centred risk identification (UN B-Tech team 2023; Calvino et al. 2024; Vassilev et al. 2024; Abercrombie et al. 2024). More recently, academic and civil society projects have focused on mapping the landscape (Slattery et al. 2026; Arnold et al. 2024). Taxonomising has thus become central to Responsible AI (RAI) practice.

Little empirical research, however, has examined how SOT are developed, interpreted, and used in practice, or how their impact might be evaluated. Currently, we lack a systematic understanding of how SOT developers conceptualise their work, how practitioners select and operationalise SOT within AI development, and how these artefacts mediate broader governance dynamics. Such understanding is necessary to inform efforts to standardise SOT or strengthen their role in AI governance. Examining their use in practice is important because classification is never neutral: categorisation systems stabilise particular problem framings, rendering some phenomena visible and actionable while marginalising others (Bowker and Star 2000). Accordingly, this paper critically reflects on SOT development, proliferation, and use to ask:

RQ1 How are SOT developed and put to use within the AI product development and deployment pipeline?

RQ2 What is the role of SOT in the broader project of AI governance, and how might this role be strengthened?

To answer these questions, we conducted a qualitative interview study of 25 researchers and practitioners across industry, academia, and civil society who have developed or used SOT. Through reflexive thematic analysis, we examine how SOT are imagined, constructed, operationalised, and contested. Our analysis is informed by Science and Technology Studies (STS) scholarship on classification and standardisation (e.g., Bowker and Star 2000; Star 1989; Timmermans and Epstein 2010). We conceptualise SOT as *proto-standards*, that is, standardisation efforts that seek to stabilise the problem space of AI harms. Through what Bowker and Star (2000) describe as “infrastructural inversion,” we shift analytical attention from the harms enumerated within taxonomies to the classificatory infrastructure itself. This perspective foregrounds the political work of categorisation, specifically the negotiation of commensurability, creation of residual categories, and temporal challenges of maintaining standards in rapidly evolving technological environments. We complement this framing with insights from Human-Computer Interaction (HCI) research on the frictions of operationalising RAI. In providing one of the first qualitative studies of the development and use of SOT, our work makes two contributions. First, we articulate design and use considerations for SOT developers and AI practitioners. Second, we describe how SOT can be better embedded within the broader AI governance landscape.

2 Related Work

2.1 Sociotechnical outcomes taxonomies

AI risks, harms and outcomes are products of complex interactions between sociotechnical systems and multilevel social phenomena, ranging from individual user interactions to the societal distributions of resources and power. Accordingly, they resist straightforward description and categorisation.

The development of SOT emerged as a pragmatic response to the operationalisation gap in AI governance (Mittelstadt 2019; Morley et al. 2020). Technology firms, in particular, promulgated high-level ethical principles signalling to internal teams and external stakeholders how they would research, develop, and deploy AI technologies, while mitigating risks and avoiding social harms (Jobin, Ienca, and Vayena 2019; Greene, Hoffmann, and Stark 2019). Translating these abstract principles into engineering constraints, however, has proven challenging. It requires practitioners to develop formal definitions of high-level concepts, such as *fairness* (Narayanan 2018), to guide their operationalisation (Mittelstadt 2019; Selbst et al. 2019; Jacobs and Wallach 2021). Researchers and practitioners have introduced SOT as translation artefacts (Star and Griesemer 1989), to both map the terrain of sociotechnical hazards associated with AI technologies and to stabilise this terrain in a way that is navigable for, and reconcilable with, AI product development. Much like checklists, SOT can “provide organizational infrastructure for formalizing ad-hoc processes and empowering individual advocates” (Madaio et al. 2020, p.10). In

this sense, SOT are intermediate governance artefacts that translate broad normative commitments into categories that can be discussed, documented, prioritised, and measured.

As AI technologies have expanded from research-oriented language models into consumer-facing, multimodal, and multi-model generative AI systems, SOT have proliferated and diversified. As of May 2026, there are over 70 SOT (Slattery et al. 2026); Table A in the Appendix provides a non-exhaustive list of published SOT to demonstrate this diversity. Initial efforts focused on the societal impacts of Natural Language Processing (NLP) (Hovy and Spruit 2016; Weidinger et al. 2022). Subsequent work extended this classificatory effort across system types and analytical dimensions. For example, Slaughter, Kopec, and Batal (2020) catalogued consumer-facing harms from algorithmic decision-making, while Shelby et al. (2023) synthesised harms across algorithmic systems more generally, distinguishing between types of harm and the levels at which they occur. Other research refined boundaries between present harms and anticipated risks (Weidinger et al. 2022), distinguishing technical evaluations of base AI systems from evaluations of impacts on people and society (Solaiman et al. 2023), and mapped individual, social, and biospheric levels of analysis (Domínguez Hernández et al. 2024).

SOT efforts have also adapted to new AI modalities, including text-to-image (Bird, Ungless, and Kasirzadeh 2023) and speech generation (Hutiri, Papakyriakopoulos, and Xiang 2024). These efforts now address specific domains and populations — such as clinical and public health (Golpayegani et al. 2022; Zhou et al. 2025), disability contexts (Wang, Kameswaran, and Kacorri 2025), and youth (Yu et al. 2025; Khoo, Chua, and Shong 2025) — as well as emerging categories of harm, including psychological, social, and relational harms (Alberts, Keeling, and McCroskery 2024; Alabed et al. 2024; Knox et al. 2025; Zhang et al. 2025) and privacy and surveillance risks (Lee et al. 2024; Gumusel, Zhou, and Sanfilippo 2024). Many emerging-harm taxonomies adopt more inductive methodologies than earlier work, grounding harm categories in lived experiences, stakeholder narratives, and user-reported impacts (Chandra et al. 2025; Archiwaranguprok et al. 2025; Zhu, Coifman, and Jin 2026; Li et al. 2025). This methodological shift is occurring in contexts where AI capabilities and social practices are co-evolving rapidly and where harms are not yet well represented in academic literature.

2.2 Taxonomies as AI governance infrastructure

SOT are governance instruments that coordinate work across distributed teams, organisations, and regulatory domains. Drawing on prior work, we identify three interrelated functions SOT may perform within AI governance: (1) structuring internal risk deliberation, (2) embedding harm categories into documentation and audit processes, and (3) mediating translation between organisational practices and regulatory regimes. Our empirical findings extend and complicate the brief description of these functions below.

Structuring risk deliberation SOT support internal risk deliberation by providing granular structuring of RAI concerns. High-level principles demarcate areas of ethical

salience for AI product teams, but leave open critical questions of interpretation, prioritisation, and implementation (Mittelstadt 2019). SOT concretise these areas by disaggregating high-level concerns into enumerated categories of impact, risk, or harm. In doing so, they create reference points that researchers and practitioners can use during design reviews, evaluation planning, impact assessments, and AI audits. Empirical studies of industry practice indicate that practitioners often lack shared language for discussing sociotechnical harms across product, policy, and legal teams, and face structural barriers when integrating fairness and responsibility concerns into existing workflows (Holstein et al. 2019; Rakova et al. 2021; Madaio et al. 2022). Toolkits and internal governance artefacts can shape how such concerns are framed and by whom they are raised (Wong, Madaio, and Merrill 2023). Madaio et al. (2020, 2022), for example, show that structured artefacts, such as fairness checklists or review templates, facilitate cross-functional communication by making implicit assumptions explicit and routinising some forms of ethical inquiry. SOT operate in a similar manner: they specify what kinds of impacts ought to be considered, and potentially how different impacts may be mitigated. In doing so, SOT implicitly define which impacts require measurement, and they provide categories through which concerns can be documented and communicated to other stakeholders. Ultimately, SOT shape how harms become administratively visible within AI governance, echoing broader concerns about problem framing in sociotechnical systems (Selbst et al. 2019).

Interfacing with AI governance SOT acquire infrastructural significance through the documentation and audit processes into which they are embedded. Transparency documentation frameworks such as model cards (Mitchell et al. 2022) and datasheets (Gebru et al. 2021) formalise expectations about which ethical considerations technologists should record and report. SOT respond to, and mediate, these expectations by structuring how such considerations are surfaced within companies and how external stakeholders can challenge inadequate consideration. Internal audit frameworks similarly rely on structured risk classification. Raji et al.'s (2020) lifecycle auditing model envisions harm identification and mitigation as a staged review process producing an 'ethical risk analysis chart,' and Algorithmic Impact Assessments (AIAs) extend this logic by formalising anticipatory risk evaluation (Metcalf et al. 2021; Selbst 2021). SOT interface with audits and AIAs, working together to conceptually (through the SOT) and administratively (through the audit or AIAs) define the space of ethically salient concerns and assign responsibility for their management. Yet, documentation alone does not ensure substantive action. Costanza-Chock, Raji, and Buolamwini (2022) show that shared standards for audit quality and comparability are underdeveloped, and most evaluation criteria for RAI tool development focus on usability rather than demonstrable harm reduction (Berman, Goyal, and Madaio 2024). Consequently, documentation and governance processes may enable the performance of transparency and responsibility without actually requiring it (Greene, Hoffmann, and Stark 2019).

Moving across internal and external boundaries SOT also operate at the interface between organisational practice and regulatory regimes. Risk-based regulatory frameworks, such as the *EU AI Act*, depend on categorical distinctions that organisations must interpret and implement through conformity assessment, technical documentation, and reliance on harmonised standards (Veale and Zuiderveen Borgesius 2021). Data protection laws, on the other hand, compel organisations to articulate risks in structured formats (Kaminski and Malgieri 2021; Metcalf et al. 2021). Within this broader standardisation environment, SOT operate as intermediary devices, aligning legal risk categories with internal evaluation metrics and governance procedures, consistent with broader accounts of how standards mediate between regulatory norms and organisational practice (Timmermans and Epstein 2010). AI governance increasingly proceeds through documentation, standardisation, and categorical alignment, with SOT mediating among these.

2.3 Taxonomy proliferation and challenges

While the proliferation of SOT reflects growing attention to AI-related harms, their divergences introduce instability in AI governance processes. Existing scholarship identifies coordination challenges, participatory burdens, and missing guidance for SOT use. These challenges produce a complex and unstable landscape of SOT adoption. Yet, limited empirical research examines how SOT are operationalised within AI governance processes. Our interview study addresses this gap by examining how SOT are developed, operationalised, and contested *in situ*.

Fragmentation and selection ambiguity Recent large-scale reviews of sociotechnical harms (Shelby et al. 2023) and meta-repository initiatives (Slattery et al. 2026) demonstrate overlap and divergence in the scope of SOT, the terminology used within, and the level of abstraction around which they are organised. SOT users must choose between partially overlapping frameworks without established criteria for determining which is most appropriate for a given system, sector, or organisational context. Fragmentation may also enable strategic evasion of AI governance, wherein organisations select SOT that are narrower in scope or omit particular risks — akin to “ethics washing” in other AI governance practices (Bietti 2020; Arnold et al. 2024).

Participatory burdens Many SOT emphasise participatory development processes (e.g., Abercrombie et al. 2024). Such approaches are often normatively justified as enhancing legitimacy and inclusivity. However, in a context where SOT are proliferating, they may create cumulative burdens for community advocates and stakeholders asked to participate in their development. Critical scholarship on design justice and participatory governance highlights the risks of extractive engagement, consultation fatigue, and the uneven distribution of labour onto communities and individuals with limited institutional power (Birhane et al. 2022; Cooper et al. 2022; Costanza-Chock 2020; Rakova et al. 2021). When multiple SOT are developed in parallel, potentially consulting overlapping communities, there may be little coordination to

mitigate repeated engagement demands or to share insights across initiatives.

Guidance for use While SOT provide structured enumerations of harms or risks, they rarely specify how they should be selected, combined, or operationalised in practice. Comprehensive taxonomies (e.g., Weidinger et al. 2022; Slattery et al. 2026) describe categories and subcategories but typically do not offer decision frameworks for prioritising harms or resolving trade-offs between competing impacts. Decisions about which harms to prioritise, how to weigh conflicting impacts, and how to determine when an SOT has been sufficiently applied remain largely under-specified.

3 Method

Through expert interviews, we explored practices and perspectives about taxonomising societal impacts of AI.

3.1 Recruitment

We recruited participants with experience developing or using SOT. Contributors to SOT were identified through the literature summarised in the preceding section and the professional networks of the research team. Two sets of inclusion criteria were used:

- **Taxonomy developers:** Contributors to the development of an SOT designed to enable systematic thinking about the potential consequences of deploying an AI system or AI technology. We interpreted both “contribution” (e.g., co-authorship, advisory roles) and “AI” broadly: if an SOT used the term AI or associated terms (e.g., GenAI, Frontier AI), we considered it relevant.
- **Taxonomy users:** Individuals who have engaged with SOT in a work or research setting. Engagement could take many forms, from using an SOT in an evaluation of an AI system to using an SOT to inform research. Participants needed some prior familiarity with an SOT (e.g., its layout, scope, and purpose) but did not need to be deeply familiar with multiple SOT.

Through email outreach, we recruited 25 participants from industry, academia, and civil society or government, across five regions (see the summary in Table B and further participant details in the Appendix). Among the participants, 17 had contributed to the development of one or more SOT (representing over 20 SOT collectively), and eight had engaged with taxonomies primarily as users in RAI evaluation contexts.

Sector	Europe	UK	N. America	Other	Total
Academia	0	2	6	0	8
Industry	2	0	8	0	10
Civil Soc./Gov.	1	1	4	1	7
Total	3	3	18	1	25

Table 1: Participants by sector & geography ($n = 25$).

3.2 Data Collection

We developed two versions of a semi-structured interview protocol, one for SOT developers and one for SOT users, and structured both protocols around two themes:

1. **The development or use of specific SOT:** For developers, this focused on questions about purpose, envisaged use cases, target audiences, and design decisions. For users, this included questions about how they selected SOT for particular tasks, what challenges they encountered when using it, and what made an SOT useful or not.
2. **The broader role of SOT in tracking and managing societal outcomes of AI deployments:** This included how SOT fit within the RAI ecosystem and how societal impact is understood and measured in participants’ organisations.

All interviews were conducted online, lasted approximately 60 minutes, and were recorded with consent, transcribed, and de-identified. Participants were compensated with gift cards and offered the opportunity to review direct quotes before publication. The project was approved by the ANU IRB.

3.3 Data analysis

We analysed the transcripts using reflexive thematic analysis (RTA) (Braun and Clarke 2006, 2021). This approach foregrounds the influence of researchers in interpreting data, encouraging them to reflect on that influence as they develop and refine codes, and it aims for rich interpretations rather than a consensus of meaning (Byrne 2022). All authors participated in the analysis, meeting regularly as a team over six months to review transcripts and inductively develop low-level codes based on the semantic content of each interview. Discussions at each meeting compared interpretations, surfaced disagreements, and challenged assumptions that our shared familiarity with the domain could lead us to overlook.

We grouped low-level codes into higher-level codes on a digital whiteboard; one author led the initial grouping before the broader team progressively developed these into themes through multiple rounds of discussion and refinement. See the Appendix for the codebook produced through this analysis. The distinction between SOT developers and users that we envisaged in our interview design proved difficult to maintain during analysis: several participants recruited as developers described drawing on existing SOT in their own work, while those recruited as users were often simultaneously developing internal SOT. We therefore report differences between industry-based participants and those in other sectors, but not between nominal ‘developers’ and ‘users.’

4 Findings

Participants positioned SOT as foundational to the broader project of RAI. In their accounts, SOT enable product teams, policy makers, regulators, activists, and community groups to map and navigate the complex problem space of AI risks and impacts. SOT are “critical scaffolding” [C10]¹ for developing

¹We refer to participants using A (academia), I (industry) or C (civil society or government) and an ID number (e.g. A1). Several participants have moved between sectors during their career; we display only their affiliation at the time of this study.

a shared vocabulary to inform risk analysis, impact assessment, and harm mitigation efforts, fostering accountability for AI deployment decisions. While participants' evidence of SOT adoption was largely anecdotal, their reports of interest in SOT [A11, C12, I7, I8, I20] indicate the potential. Participants were cognisant of the challenges of SOT proliferation, adoption, and integration into AI governance, but the opportunity for SOT to be foundational to the development of robust AI governance regimes was nonetheless clear to them [A9, A11, C10, C19, I2]. Yet, participant descriptions of how SOT are used in practice demonstrate our first finding: **SOT are not well integrated into AI governance processes**, with little evidence of SOT being used to inform AI deployment decisions or the monitoring of AI impacts.

Participants' responses highlighted two significant barriers to realising the potential of SOT, constituting our second and third findings. To structure the landscape of AI risks, **SOT inherently reduce the complexity of the problem space**. Enacting this reduction requires resolving design tensions — most significantly the “altitude problem” [I21], wherein SOT developers must choose the right level of abstraction at which to define AI risks or impacts. Participants observed that AI risks and impacts are not natural phenomena to be mapped, but rather sociotechnical outcomes whose causes must be explicated to scaffold substantive and targeted AI governance interventions. Yet, **incorporating causal analysis into the development of SOT has proved challenging**, and most existing SOT do not attempt to identify the decision points or critical actors who may be responsible for the catalogued risks occurring.

Participants advocated for a range of approaches to addressing these barriers. Our final finding reports **three recommendations for improving SOT development and adoption**, on which participants converged: linking harm categories in SOT to decision-points in product development, strengthening external requirements on technology firms to ensure monitoring of risks enumerated in SOT, and improving coordination across currently overlapping and sometimes divergent SOT frameworks. Together, these recommendations signal a move towards standardisation across SOT, although they raise the question of where within the AI ecosystem, from a normative perspective, SOT development should be led. We return to this question in our Discussion.

4.1 Gaps in integrating SOT into AI governance

Across participants' accounts, the integration of SOT into AI governance remains partial and uneven. This is visible in two strands of our data, which we examine in turn: the proliferation of SOT in the absence of coordination across them, and the limited visibility participants reported into the adoption of SOT and the outcomes SOT attempt to anticipate.

Proliferation without coordination Proliferation is a condition practitioners must navigate rather than a resource they can draw on. The multiplicity of SOT creates “analysis paralysis” [I8], with attention spent grappling with subtly different and often incompatible classifications rather than acting on the harms those classifications enumerate [I4, A15, I16]. I14 reported that “there are so many of them” that “choosing your

approach” has itself become “a limitation of taxonomies.” Organisationally, proliferation enables strategic selection of SOT while confounding efforts to standardise harm classification. As A11 observed, “the ways that different AI people or AI organisations or companies frame [risks] are sometimes suspiciously narrow.” I16 located the consequence in the absence of coordinated alternatives: fragmentation drives “every company... to create their own” SOT. Responses were adaptive rather than corrective and, in aggregate, exacerbated proliferation. Rather than waiting for coordination, participants built local workarounds, “reformulating what has been said” in existing SOT [A15]. I4 described creating “mappings between different taxonomies, so at least all of the work you’ve done under taxonomy A will still be relevant.” These workarounds enabled participants to situate their work within a fragmented landscape, but did so by adding further layers to it. I8 described the resulting condition: “you just need to act and you don’t know how or what to act upon.”

Adoption and outcomes without visibility Beyond the navigational challenges of proliferation, SOT developers have little visibility into how their SOT are used. The absence of feedback channels surfaces directly: “I unfortunately have no evidence whether people have found it useful” [I20]; there is little “insight into how [developers] actually do, or if they actually do,” use them [I14]. I14 and I20 are both based in industry, so their lack of visibility into SOT use is particularly striking, though it should not be misread as a lack of interest in SOT utility. Where evidence of use did exist, it was anecdotal. I8 described an SOT they contributed to as “quite widely used... I mean, not all the time.” Without infrastructure that would track adoption systematically, neither the developers of SOT nor those evaluating their utility can diagnose where integration is succeeding and where it is failing. The same absence operates downstream, at the level of the harms SOT are intended to anticipate. There is little evidence that risks enumerated in SOT are systematically monitored after deployment. An aspirational adverse event reporting pipeline is “like a fantasy” [I8], while C17 reported that post-launch monitoring “consistently gets de-prioritised.” Together, these visibility gaps mean that the foundational role participants ascribed to SOT in the broader project of AI governance cannot be empirically assessed from within the current ecosystem.

4.2 Seeking clarity while navigating reductionism

Participants described SOT as responding to the complexity of AI impacts by imposing a structure on the problem space [C17, I8, I25, I4, A5, I2, A9, A11, C19, A24]. I8, for example, observed that “you have a bunch of phenomena that you’re interested and confused by... you’re trying to impose some kind of structure onto this.” By imposing structure, participants framed SOT as enabling understanding of the problem space and foreshadowing action.

This frame of “imposing structure” reflects an awareness of the subjectivity inherent in developing classificatory schemas. This subjectivity begins in the composition of the SOT development team itself, which may “naturally narrow the scope” [A1]. As I25 opined: “If the taxonomy is created primarily from the... circle jerk of computer science, then it’s not nec-

essarily actually mapping the full breadth of risk.” Counterexamples to I25’s account exist: SOT development teams spanning disciplinary or institutional boundaries formed shared vocabularies and deepened their appreciation of AI risks through the development process [C10, I2, I7, I8]. However, as we explore below, this awareness is in tension with the desire for SOT that AI practitioners can easily interpret and act on. That desire can incentivise reductionist design choices that simplify the problem space in ways that serve some audiences better than others.

Reduction treated as description The structuring work SOT do is necessarily reductive: a navigable category schema cannot, by definition, preserve all the specificity of the phenomena it organises. Referring to deepfake harms, I20 illustrated how individual categories collapse phenomena that remain heterogeneous in practice:

The kind of harm that happens if a deepfake is used to fool a grandmother [is different] to the kind of harm that happens with the Marco Rubio thing that . . . happened on a political level . . . to the kind of harm that happens if a CEO’s voice is taken. They’re actually quite different.

The single category of “deepfake harm” offers structure for those acting on the SOT, but only by collapsing distinctions that matter for who is responsible, who is harmed, and what mitigation might involve. The category risks communicating more clarity than the underlying phenomena warrant.

The simplification of AI risks has unintended consequences for the way AI governance processes recognise and address harms. Several participants recognised the sociopolitical consequences of the choices this reduction entails:

At the end of the day, taxonomies are language creation . . . if language and categorisation is power, taxonomies hold power. Because if you create [a] taxonomy and you leave out categories intentionally, those categories won’t get measured. [I20]

I8 named the corollary risk: “the danger, of course, is always that you start to mistake your abstraction for reality . . . people think the taxonomy is real.” In both accounts, the reductive work of an SOT is interpretive, not descriptive. The risk is not that SOT simplify per se, but that the simplifications come to be treated as descriptions of the world rather than as one possible structuring of it.

Usability concerns guide navigation of design tensions SOT enable discussion and debate about AI risks by establishing a shared vocabulary. The framework of an SOT, however, must overcome what I21 termed the “altitude problem:”

Having a ridiculously high branching factor [or] a very flat taxonomy is not that helpful. You need to have some degree of nice categorical structuring to it. But how do you decide which elements can be promoted to what level within the taxonomy?

The risk here is the failure to meaningfully clarify the problem space; neither exhaustive lists nor high-level categories enable interpretation of AI risks. C17 similarly noted that the same categories mean different things to different actors:

The same words mean different things [to all the stakeholders], and you need to be able to root the [SOT] in concrete examples.

The criterion for determining the right ‘altitude,’ across accounts, is thus ease of interpretation for those expected to act on the SOT [A6, A3, A15, C17].

Some risks are likely easier to describe with “concrete examples” than others, which may lead to imbalances across SOT categories. Optimising for interpretability therefore has two risks for SOT: inviting reductionist interpretations of AI risks and impacts, and losing fidelity to the heterogeneity of experiences within a category or across affected communities.

The altitude problem is further complicated by the need for SOT to retain their relevance across successive generations of AI technologies. As A9 observed, a “real difficulty with taxonomies of technology is that . . . if they’re too deep, they become redundant really quickly.” Thinking of SOT as ‘living’ documents — which evolve over time while retaining their core structure — resolved tensions between flexibility and stability for several participants [I4, I16, I20, I21]. I16 described how precise risk definitions may be less useful than a balance of flexibility and stability that allows product teams to intuit risks and move quickly to action. The pace of SOT change is thus calibrated to practitioner workflow, without consideration of the needs of other stakeholders.

4.3 Mitigating risks requires addressing causality

The purpose of SOT, for many participants, is interpretive rather than declarative [C10, I8, I7]. A15 termed this “the procedural value of taxonomising,” which “maybe exists irrespective of whether or not at the end of it you try to publish a paper.” I2 framed the SOT they contributed to as “a sort of reflection, an initial tool embedded in the process before we go to a more formal process.” I7, meanwhile, tied SOT development directly to accountability for AI impacts: developers “can be held accountable to what they’ve done about it because they’ve just declared that they know.” Across these accounts, the deliberative work of SOT development generates shared vocabulary and prepares product teams to engage with AI risks; SOT development serves as a “critical scaffolding” [C10] for actions intended to mitigate the risks SOT enumerate.

Interpretation scaffolds action SOT scaffold governance actions by reducing the complexity of analysing AI risks. The value of SOT, C10 argued, lies not in the categorisation itself but in the systematic analysis it enables. Across accounts, SOT facilitate the handoff between identifying risks and acting on them [I21, I2, C12, A15, I16, C19, A24, C17]. I16 used the language of ‘passing’ to describe the function of SOT, which enable RAI teams to “suss out risk types and then pass that off to the [product] team with some helpful advice or strategies that they can use to measure, assess, and determine” next steps. In these cases, SOT function as translation aids, converting diffuse concerns about AI risks into forms that are tractable for AI product teams, and, in doing so, enabling responsibility for concerns to move from ethics or research teams to product teams. This translational function

was valued across institutional contexts [A11, I8, I16, I14, I2, C17, C19]: generalised concern about AI risks is, in itself, a barrier to productive action [I4, C17, C12].

A24 extended this scaffolding role in a direction other participants did not, describing SOT as serving a diagnostic function after deployment as well as a proactive one before it: “If you’re building an AI system in any of these areas, this taxonomy helps you pay attention to... what can be done to make sure certain types of outcomes are not produced. And on the other flip side... [AI] systems are going to make mistakes. And when they do... where is the problem? What category of problems are we dealing with?” A24’s account is also distinctive in centring the interests of affected communities rather than the operational needs of product teams: when working with marginalised communities, A24 argued, the right approach is to “Go talk to the community. Figure out what they need.” Across our interviews, this orientation was rare; the scaffolding function of SOT was overwhelmingly described in terms of actions for product teams.

Action without causal analysis may be misdirected The form of action SOT scaffold follows from their descriptive approach to enumerating harms. SOT describe AI risks but do not, in general, explain how risks manifest as harms or impacts, or specify the actors whose decisions may exacerbate or mitigate them [I8, I20, C10, C13]. As I8 put it:

There’s a missing element of causation in the taxonomy... We are treating it like a botany taxonomy as opposed to, ‘Wait, we’re the ones doing this.’... Whenever it’s discovered or detected and pointed out, of course they... go to court and say, ‘No, we’re not responsible.’ [I8]

A botany taxonomy reflects the positivist epistemology of the naturalist classifying plants based on cellular and biological traits, without regard to questions of accountability for their distribution in a particular environment. SOT, treated similarly, render AI harms as objective, passive phenomena to be observed and classified rather than as normative, dynamic outcomes produced by identifiable decisions. As I8 argues, the descriptive posture of existing SOT means that harms are documented without causal attribution, and the organisations producing the underlying systems are positioned to deny responsibility for what SOT record. Other participants similarly highlighted gaps in SOT “root cause analysis” [A24]; “we’re only categorising incidents and risks without actually looking at how risks arise” [I20; cf. C10, I21]. This lack of explicit causal analysis within SOT is, in part, a product of the diffuse range of risk sources unique to the form and function of individual AI systems. I8 notes, when reflecting on conversations about causality with AI product teams during SOT development: “You’d get to a point in the discussion where they’d say, ‘Well... that’s just society, that’s just how society is. We can’t be held responsible for inequality. We can’t be held responsible for racism.’” Where harms can be attributed to the social world in general, rather than to AI development specifically, responsibility for catalogued risks becomes similarly diffuse.

Where SOT adopt a descriptive orientation, there are consequences for the kinds of action they can scaffold. Two

consequences, in particular, are visible across participants’ accounts: action may be misdirected, focusing more on measuring risk proxies than on preventing harm [I20]; and accountability infrastructure that could connect documented harms to responsible actors remains underdeveloped [I7, I8]. The first consequence is visible in how SOT-based measurement work is constructed, while the second can be seen in accountability gaps for addressing SOT-catalogued risks.

A3 described a recurring disconnect between the measurements SOT enable and the impacts SOT enumerate: “There’s a lack of connection made between what is that impact and what is the model or the system’s performance?” Without a causal account that links system features to the impacts they produce, SOT-informed measurement is limited to system properties that are tractable to measure, rather than to the impacts that are the SOT’s substantive concern. The action SOT scaffold thus tracks the measurable rather than the consequential.

The second consequence concerns how actors are held accountable, or made responsible, for addressing risks described in an SOT. I14 described the experience of attempting to challenge a product team’s decision-making using SOT-based concerns:

There was one product team that was just driving me absolutely crazy because they didn’t really have any justification... And there was not really a way to fight that justification because, to them, if they have people buying something that’s more justified than, like, a paper [describing an AI risk].

C12, from civil society rather than within a technology firm, observed: “You have to have some kind of way to point out what you’re talking about [to decision makers].” These extracts demonstrate an evidentiary and accountability asymmetry. Product teams have tools to articulate how their decisions impact certain forms of consumer behaviour, but RAI teams and civil society lack the tools to articulate how product decisions may increase AI risks in any form that activates internal accountability processes.

Where SOT do inform organisational processes, depending on organisational culture, classification can displace action altogether. I8 provided one example: “You just kind of point to like, ‘Oh, we did a datasheet’. Then it’s like... and what did you do because of that?... The answer is, ‘We didn’t actually change it. We just document.’” Datasheets may become evidence that risk has been considered, not evidence that risk has been addressed. I2 named the same risk from a different vantage point, describing the attitude of “I use this tool and this sort of Bible, in quotes, oh, I’m done... It’s okay. I got the green light. Push it into market.” In both accounts, the intended governance work of SOT is displaced by the organisational work of demonstrating governance has occurred. Without causal scaffolding linking categories to decision points and responsible actors, SOT-informed governance may default to dominant logics already operative within the firm.

Myopic focus on corporate concerns The descriptive orientation of SOT can shape which catalogued risks are addressed in practice. Without external requirements to address

impacts documented in SOT, decisions about prioritisation may be guided by other incentives, namely a firm's commercial and reputational considerations. I16 described the fundamental tension between near-term "business needs" and the "long time" it takes to "understand societal impacts," noting that harms often only become "glaringly apparent after something has already launched." Product launches thus act as a contradictory inflection point, marking a transition in the firm's development cycle and the beginning of real-world societal impact. Consequently, at the time initial risk assessments and mitigation strategies are developed, threats to the business and the product timeline are highly legible, while long-term societal impacts remain opaque. RAI practitioners navigate this asymmetry by framing SOT adoption in commercial terms: "How do we sell this to product teams so they feel like... it accelerates their launch" [I16].

Yet, this commercial framing may influence which SOT-catalogued harms are actually addressed, further ignoring long-term societal impacts. As risks are weighed against business interests, reputational visibility may be a primary proxy for harm. As I8 described, the operative question during launch decisions is often simply: "What would the *New York Times* headline read?" Relying on media coverage and social media monitoring [I7] as the barometer of risk means harms can be systematically deprioritised if they are diffuse, slow-moving, or borne by communities with limited public voice. I20 highlighted the geographic bias of this dynamic: "Europe would focus very much on Europe, US [is] very much concerned about deepfakes and harms associated with that, but then cultural erasure in the Global South kind of gets ignored."

4.4 Improving SOT development and adoption

Across our interviews, participants articulated converging views regarding future development of SOT and their integration into AI governance, highlighting the need for causal analysis linking harm categories to decision points in product development and responsible actors [I8, I20, C13, C12, C10]; external requirements to ensure technology firms are accountable for AI risk mitigation and monitoring [I16, A11, A24, I7]; and coordination across the fragmented landscape of different SOT, moving towards standardised approaches to AI risk classification, identification, and monitoring [C17, A11, I16, I21]. Participants framed technology firms' development of SOT as a pragmatic response to the absence of external requirements and the need for structured approaches to thinking about AI risks and impacts [I16, C17, I8]. This framing reflects the lack of alternatives rather than an endorsement of industry as the appropriate locus for this work.

Participants cautioned, however, that the form that integration takes matters. C17 described the historical trajectory of SOT integration into firm-level AI governance as a flattening movement, from a "noble minded, ethical, philosophical approach to harm" through to a "corporatised risk approach" integrated into "normal compliance practices." A11 described a sharper version of the same concern, observing that "bad evals can just serve to increase boldness and reduce potential for liability... [and] decrease responsibility for the downstream problems." Integration into evaluation and compli-

ance processes, in this account, do not merely flatten but actively reduce accountability, by providing documented evidence that risks have been considered and discharged. The implication is not that further development and adoption of SOT should be resisted, but that steps towards standardisation across SOT should be taken cautiously. Standardisation — whether through external requirements, internal compliance regimes, or cross-industry coordination — moves the locus of definitional authority over what counts as an AI risk. Where definitional authority comes to rest shapes which risks remain visible, whose accountability is activated, and what counts as having addressed a classified harm. This highlights the underlying question raised by participants' recommendations: where, within the AI ecosystem, the work of developing and coordinating SOT should sit, whether with regulators [A1], academia [A11], civil society [C10], or distributed across AI governance actors [A3] — a question we turn to below.

5 Discussion

5.1 Implications for SOT developers and users

STS scholars have documented how methodological and political choices embedded in classification work are rarely made explicit. This opacity has several consequences for those who inherit the categories (Bowker and Star 2000; Star and Strauss 1999; Suchman 2002). Phenomena that do not fit are absorbed into residual categories or distorted to fit those that exist. The labour through which the schema was stabilised becomes invisible and difficult to contest, and the positions from which the categories were produced are hidden from those who later use them.

Our findings show this opacity at work in SOT practice, with three consequences revealed in participants' accounts. First, without shared conventions for how SOT are organised, scoped, and described, users cannot reliably compare or combine them. Participants worked around this navigation problem through local mappings that, in aggregate, deepened the problem they were responding to. Second, where the interpretive work informing an SOT is invisible, the categories travel without the caveats that produced them, and SOT designed as reflexive aids can be taken up as performative checklists. Third, where opacity disguises the choices about whose concerns the SOT centres, risks more readily legible to technology firms tend to be reified, while less legible risks, and the communities affected by them, are marginalised. We turn now to three design properties that respond to these consequences: interoperability, extensibility, and traceability.

Interoperability SOT can be made comparable and combinable through deliberate design, rather than relying solely on after-the-fact translation. Three design moves support this. First, SOT need shared structural conventions through consistent organising primitives: clear distinctions between risks, impacts, and harms; between types of harm and levels at which they occur; and between observed and anticipated effects. Through such primitives, two SOT using compatible conventions can be aligned even if their specific categorical content differs (Bowker 2000). Second, explicit metadata regarding evidentiary sources, target audience, intended uses, and definitions of risk and scope should be integrated directly

into the SOT rather than accompanying notes. Treating this context as a core component of the artefact makes visible the articulation work of SOT development, exposing the debate, negotiation, and judgement under uncertainty through which categories are stabilised (Star and Strauss 1999). This documentation would allow users to assess fit, identify overlap and gaps, and integrate categories across multiple SOT without flattening the vital distinctions that produced them (Edwards et al. 2011). Third, the SOT development process itself should produce explicit mappings to adjacent SOT (e.g., noting a category maps to *X* in another scheme or stating it has no direct equivalent). Publishing SOT with these documented correspondences facilitates interoperability and absorbs the heavy translation labour participants currently report undertaking themselves. Practically, as Bagehorn et al. (2025) describe, this mapping can be supported by adopting stable semantic identifiers for risks, and using frameworks like the Simple Standard for Sharing Ontological Mappings (Matentzoglou et al. 2022).

Extensibility An SOT’s categorical structure can leave space for new evidence of risks or harms, technological developments, or contexts to be incorporated. Three design moves support extensibility (Ribes and Finholt 2009; Jackson 2014). First, state explicit boundary conditions. Rather than merely documenting what an SOT covers, the artefact should explicitly identify adjacent contexts it *does not* cover and where compatible categorisation is still needed. Defining this negative space clearly signals to downstream users where extensions are welcome and roughly what shape they can take. Second, develop risk categories designed to operate across multiple levels of abstraction. When SOT allow categories to be unpacked at finer granularity or aggregated upward without breaking the overall structure, it enables users at different operational altitudes to engage with and extend the artefact in either direction. For example, Shelby et al. (2023) organised harms by type and level, providing a modular architecture that Wang, Kameswaran, and Kacorri (2025) extended to map harms related to disability. Third, include mechanisms for integrating new evidence in the initial publication of the SOT, rather than deferred to future re-development cycles. Conceptually, Weidinger et al. (2022) demonstrate this by distinguishing between observed harms and anticipated risks. This built-in distinction creates a designated holding space for new empirical evidence, while the interactive taxonomy developed by Abercrombie et al. (2024) offers an example of tooling to support extensibility.

Traceability SOT enumerate harms but, as our third finding documents, rarely connect those harms to the decisions, deployment contexts, and actors implicated in producing them. This disconnect can allow classification to substitute for meaningful action, reducing the potential of SOT to support accountability within AI governance. Two design moves support traceability. First, include causal pathways linking classified risks to the decision points and actors that shape their occurrence. Full causal models require examining sociotechnical specifics of a given AI system and its context of use, and would risk overburdening participants in collaborative SOT development. But even lightweight causal

mapping is powerful. By indicating *how* a risk might arise, and *whose* decisions govern that context, the SOT gives users and external stakeholders a concrete basis to challenge what might otherwise be interpreted as diffuse, unavoidable harms. Slattery et al. (2026)’s Causal Taxonomy, which categorises entity, intent and timing, provides a practical example of this approach. Second, monitoring requirements should be made explicit alongside causal pathways. For each pathway, the SOT should specify what needs to be observable in deployment for a classified risk to be detected. Naming these observability requirements makes the link between classification and intervention explicit, *even when the infrastructure to fulfil them does not yet exist*. A rigorous measurement practice, supported by complementary diagnostic approaches (e.g., (Rismani et al. 2023, 2025; Wallach et al. 2025)), will allow SOT developers to surface these observability gaps.

5.2 Implications for AI governance

Integration gaps in AI governance reflect more than coordination shortcomings. SOT do important work surfacing potential AI harms across product teams, policymakers, and civil society. But in a regulatory environment with few binding requirements specific to AI harm mitigation, technology firms remain largely able to self-regulate how they respond to surfaced harms (Greene, Hoffmann, and Stark 2019). Those outside technology firms have few pathways for holding them to account for risk prioritisation or mitigation actions. Our finding that classification can substitute for action, and can occur without sanction, is unsurprising in this light (Costanza-Chock, Raji, and Buolamwini 2022). The proliferation of SOT compounds the gap. When every firm develops its own internal SOT, harm definition becomes a private rather than a public activity. The consequences of those private definitional choices accrue to publics that have little role in making them.

Where, then, in the AI governance ecosystem should the work of SOT development sit? Participants gestured in different directions, including regulators, academia, civil society, or distributed across actors. Each has structural limits. Industry-led development brings empirical grounding from proximity to deployed systems, but participants described an asymmetric publication regime: researchers in industry labs are rewarded for novel taxonomies and constrained from publishing actual incidents. The result is a steady supply of new SOT alongside little public evidence about whether the harms those SOT enumerate are actually occurring or being addressed. The SOT closest to deployed systems are the least likely to be public, and the most visible SOT are largely categorisation without empiricism. Regulator-led development could address concerns about democratic legitimacy — weighing different harms against each other is, in substance, a political question (Bowker and Star 2000) — but it risks the flattening of normative commitments into compliance categories that participants observed occurring within firms. Academic and civil society-led development can preserve normative breadth, but lacks the empirical visibility and enforcement leverage that would make SOT operative in deployed contexts. No single locus offers a complete answer.

Standardisation across SOT would, in principle, address many of these challenges by providing a common referent, re-

ducing the strategic space for firm-level selection and partial uptake. But standards emerge through accretive coordination across institutional sites, requiring an accumulated substrate of shared vocabularies, infrastructural commitments, comparable evidentiary records, and consolidated institutional authority (Timmermans and Epstein 2010). The AI governance landscape lacks these. No existing institution combines the empirical visibility, democratic legitimacy, and enforcement leverage that standardisation requires. The technology firms best positioned to lead standard-setting for SOT are also those whose interests standardisation would most readily entrench. Premature standardisation under these conditions risks freezing the definitional asymmetries our findings have identified, on terms that suit incumbents. What is needed first is not a standard but the substrate from which a standard could later emerge.

Assembling this substrate is the AI governance work needed to complement the SOT design recommendations outlined above. A well-designed SOT does little if the conditions for its uptake, comparison, and contestation do not exist; interoperability, extensibility, and accountability must be realised at the level of the ecosystem too. Governance efforts can begin to assemble the infrastructure within which these properties become consequential: accretive infrastructure (Timmermans and Berg 1997) through which future standards for AI risk, harm, and impact definition, classification, and monitoring may develop. The most tractable starting point is aggregation: a registry, perhaps hosted by an existing standards institution, where SOT are deposited alongside their developmental documentation and where monitoring and reporting expectations can be attached to catalogued risks. Such a register would do three things. First, it would make interoperability a property of the ecosystem rather than work that each developer replicates. Second, it would offer a place where empirical evidence currently siloed within firms could begin to be assembled, giving extensibility a basis in accumulating evidence rather than in developer foresight alone. Third, it would enable monitoring expectations to be attached to classification work, turning the link between classification and intervention into a potentially enforceable governance relation rather than a developer aspiration. AGORA (Arnold et al. 2024) offers a partial precedent at the regulatory level. This is the broader governance project that SOT scaffold, and on which their value ultimately depends.

6 Limitations

Four limitations bear on how our findings should be read. First, our sample mirrors the geographic concentration that our participants themselves critique: participants and authors are based in the US, Europe, UK, and Australia, with no representation from the Global South. The dynamics we describe may look different from positions our team cannot speak from. Second, our findings are based on retrospective accounts of SOT development and use, rather than direct observation of taxonomising practice or its effects on product decisions. Triangulation through ethnographic studies of SOT in use would be helpful. Third, we recruited participants through their prior engagement with SOT, which excludes the perspectives of practitioners who encountered SOT and chose

not to use them. Their reasoning would provide vital counter-evidence regarding SOT uptake. Finally, several participants spoke with notable candour and their quotes recur across our themes; we have sought to balance their voices against others where possible, but especially forthright participants likely influenced the analysis.

7 Conclusion & Future Work

We set out to investigate two research questions about the development and use of SOT, specifically their role in AI governance. Through interviews with researchers and practitioners ($n = 25$), we found SOT developers and users to be keenly aware of a core tension: imposing a useful structure on an inherently complex problem space without flattening the nuance required to properly analyse AI risk. While SOT are theoretically positioned to bridge the operationalisation gaps between high-level ethical principles and their implementation in product decisions, we found them to be weakly integrated into actual AI governance processes. Our findings locate this weak integration in features of SOT design and use that simply proliferating more SOT will not, on its own, resolve. The necessary, but reductive, choices through which SOT structure the problem space of AI risks tend to be invisible to downstream taxonomy users, causing frameworks intended as interpretive aids to be misused as exhaustive descriptions. SOT also tend to enumerate harms without linking them to the decision points or specific actors implicated in their occurrence, leaving accountability difficult to assign and facilitating classification to substitute for substantive action.

Realising the governance potential of SOT requires intervention at the artefact and ecosystem levels. For SOT developers, we articulate three desiderata: interoperability, extensibility, and traceability, the last of which demands explicit causal pathways and observability conditions for classified risks. At the ecosystem level, rather than pursuing premature standardisation that risks entrenching incumbent interests, we advocate for assembling the substrate from which future standards might emerge: a shared registry that aggregates SOT alongside their developmental documentation, evidentiary records, and monitoring expectations. Finally, to better understand when classification leads to mitigation, future empirical research must directly observe SOT in use across industry, civil society, and regulatory agencies.

Positionality Statement

Our team is interdisciplinary and inter-sectoral. Three authors hold industry affiliations and three are based in academia, with disciplinary backgrounds spanning NLP, HCI, and STS. Four authors are based in the United States, two in Australia. All authors are active in the RAI research community. One author led the development of an SOT; others have contributed to the development of SOT or used them in evaluation work. We recognise SOT as one of the few RAI artefacts with uptake across research, policy, and industry, while also being aware of widespread barriers to effective RAI adoption in practice. Our analysis proceeds from an intermediary position, neither disinterested outsiders to questions of SOT development nor entirely insiders to SOT adoption within

technology firms.

Acknowledgements

We recognise and thank the generosity and trust of our participants, whose contributions enabled the analysis presented. We also extend our thanks to the reviewers for their engagement and constructive feedback. This research project benefited from feedback gathered at the *Evaluating Evaluations: Examining Best Practices for Measuring Broader Impacts of Generative AI* NeurIPS 2024 workshop. The quote included in the title of this paper is from participant I4, who noted that the phrase “death by a thousand taxonomies” was first used by Avijit Ghosh and Kevin Klyman during the 2024 NeurIPS workshop.

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.; Suriya-wongkul, A.; Thelot, R.; Vei, S.; and Waltersdorfer, L. 2024. A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms. *ArXiv:2407.01294 [cs]*.
- Alabed, A.; Javornik, A.; Gregory-Smith, D.; and Casey, R. 2024. More than just a chat: A taxonomy of consumers’ relationships with conversational AI agents and their well-being implications. *European Journal of Marketing*, 58(2): 373–409.
- Alberts, L.; Keeling, G.; and McCroskery, A. 2024. Should agentic conversational AI change how we think about ethics? Characterising an interactional ethics centred on respect. *arXiv:2401.09082*.
- Archiwaranguprok, C.; Albrecht, C.; Maes, P.; Karahalios, K.; and Pataranutaporn, P. 2025. Simulating Psychological Risks in Human-AI Interactions: Real-Case Informed Modeling of AI-Induced Addiction, Anorexia, Depression, Homicide, Psychosis, and Suicide. *arXiv:2511.08880*.
- Arnold, Z.; Schiff, D. S.; Schiff, K. J.; Love, B.; Melot, J.; Singh, N.; Jenkins, L.; Lin, A.; Pilz, K.; Enweareazu, O.; and Girard, T. 2024. Introducing the AI Governance and Regulatory Archive (AGORA): An Analytic Infrastructure for Navigating the Emerging AI Governance Landscape. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7: 39–48.
- Bagehorn, F.; Brimijoin, K.; Daly, E. M.; He, J.; Hind, M.; Garces-Erice, L.; Giblin, C.; Giurgiu, I.; Martino, J.; Nair, R.; et al. 2025. AI risk atlas: Taxonomy and tooling for navigating AI risks and resources. *arXiv preprint arXiv:2503.05780*.
- Barocas, S.; Crawford, K.; Shapiro, A.; and Wallach, H. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual Conference of the Special Interest Group for Computing, Information and Society*, volume 1. New York, NY.
- Batool, A.; Zowghi, D.; and Bano, M. 2025. AI governance: A systematic literature review. *AI and Ethics*, 5(3): 3265–3279.
- Bender, E. M. 2019. A typology of ethical risks in language technology with an eye towards where transparent documentation can help. In *Future of artificial intelligence: language, ethics, technology workshop*, volume 1.
- Berman, G.; Goyal, N.; and Madaio, M. 2024. A Scoping Study of Evaluation Practices for Responsible AI Tools: Steps Towards Effectiveness Evaluations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–24. Honolulu HI USA: ACM. ISBN 979-8-4007-0330-0.
- Bietti, E. 2020. From Ethics Washing to Ethics Bashing: A View on Tech Ethics from within Moral Philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 210–219. Association for Computing Machinery. ISBN 978-1-4503-6936-7.
- Binns, R. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, 149–159.
- Bird, C.; Ungless, E.; and Kasirzadeh, A. 2023. Typology of Risks of Generative Text-to-Image Models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 396–410. Montreal QC Canada: ACM. ISBN 9798400702310.
- Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *EAAMO ’22*, 1–8. New York, NY, USA: Association for Computing Machinery. Journal Abbreviation: EAAMO ’22.
- Bowker, G. C. 2000. Biodiversity Datadiversity. *Social Studies of Science*, 30(5): 643–683.
- Bowker, G. C.; and Star, S. L. 2000. *Sorting Things Out: Classification and its Consequences*. MIT Press. ISBN 0-262-52295-0.
- Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3: 77–101.
- Braun, V.; and Clarke, V. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18: 328–352.
- Byrne, D. 2022. A worked example of Braun and Clarke’s approach to reflexive thematic analysis. *Quality & Quantity*, 56: 1391–1412.
- Calvino, F.; Dernis, H.; Samek, L.; and Ughi, A. 2024. A sectoral taxonomy of AI intensity. Technical report, OECD Publishing.
- Chan, A.; Salganik, R.; Markelius, A.; Pang, C.; Rajkumar, N.; Krashenninnikov, D.; Langosco, L.; He, Z.; Duan, Y.; Carroll, M.; Lin, M.; Mayhew, A.; Collins, K.; Molamohammadi, M.; Burden, J.; Zhao, W.; Rismani, S.; Voudouris, K.; Bhatt, U.; Weller, A.; Krueger, D.; and Maharaj, T. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *2023 ACM Conference on Fairness Accountability and Transparency*, 651–666. Chicago IL USA: ACM. ISBN 979-8-4007-0192-4.
- Chandra, M.; Naik, S.; Ford, D.; Okoli, E.; De Choudhury, M.; Ershadi, M.; Ramos, G.; Hernandez, J.; Bhattacharjee, A.; Warreth, S.; and Suh, J. 2025. From Lived Experience to Insight: Unpacking the Psychological Risks of Using AI

- Conversational Agents. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 975–1004. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Cooper, N.; Horne, T.; Hayes, G. R.; Heldreth, C.; Lahav, M.; Holbrook, J.; and Wilcox, L. 2022. A Systematic Review and Thematic Analysis of Community-Collaborative Approaches to Computing Research. In *CHI Conference on Human Factors in Computing Systems*, 1–18. New Orleans LA USA: ACM. ISBN 978-1-4503-9157-3.
- Costanza-Chock, S. 2020. *Design justice: community-led practices to build the worlds we need*. Cambridge, Massachusetts: The MIT Press. ISBN 978-0-262-35686-2. OCLC: 1130310256.
- Costanza-Chock, S.; Raji, I. D.; and Buolamwini, J. 2022. Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1571–1583. Seoul Republic of Korea: ACM. ISBN 978-1-4503-9352-2.
- Domínguez Hernández, A.; Krishna, S.; Perini, A. M.; Katell, M.; Bennett, S.; Borda, A.; Hashem, Y.; Hadjiloizou, S.; Mahomed, S.; Jayadeva, S.; Aitken, M.; and Leslie, D. 2024. Mapping the individual, social and biospheric impacts of Foundation Models. In *The 2024 ACM Conference on Fairness Accountability and Transparency*, 776–796. Rio de Janeiro Brazil: ACM. ISBN 979-8-4007-0450-5.
- Edwards, P. N.; Mayernik, M. S.; Batcheller, A. L.; Bowker, G. C.; and Borgman, C. L. 2011. Science Friction: Data, Metadata, and Collaboration. *Social Studies of Science*, 41(5): 667–690.
- Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for Datasets. *Communications of the ACM*, 64(12): 86–92.
- Golpayegani, D.; Hovsha, J.; Rossmailer, L. W. S.; Saniei, R.; and Misis, J. 2022. Towards a Taxonomy of AI Risks in the Health Domain. In *2022 Fourth International Conference on Transdisciplinary AI (TransAI)*, 1–8. Laguna Hills, CA, USA: IEEE. ISBN 978-1-6654-7184-8.
- Greene, D.; Hoffmann, A. L.; and Stark, L. 2019. Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. In *Hawaii International Conference on System Sciences*.
- Gumusel, E.; Zhou, K. Z.; and Sanfilippo, M. R. 2024. User Privacy Harms and Risks in Conversational AI: A Proposed Framework. arXiv:2402.09716.
- Holstein, K.; Vaughan, J. W.; Daumé, H.; Dudík, M.; and Wallach, H. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–16. ISBN 978-1-4503-5970-2.
- Hovy, D.; and Spruit, S. L. 2016. The social impact of natural language processing. In *The 54th Annual Meeting of the Association for Computational Linguistics Proceedings of the Conference, Vol. 2 (Short Papers)*. Association for Computational Linguistics.
- Hutiri, W.; Papakyriakopoulos, O.; and Xiang, A. 2024. Not My Voice! A Taxonomy of Ethical and Safety Harms of Speech Generators. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 359–376. Rio de Janeiro Brazil: ACM.
- Jackson, S. J. 2014. Rethinking Repair. *Media Technologies: Essays on Communication, Materiality, and Society*, 221–39.
- Jacobs, A. Z.; and Wallach, H. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385. Association for Computing Machinery. ISBN 978-1-4503-8309-7.
- Jobin, A.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9): 389–399.
- Kaminski, M. E.; and Malgieri, G. 2021. Algorithmic impact assessments under the GDPR: producing multi-layered explanations. *International Data Privacy Law*, 11(2): 125–144.
- Khoo, S.; Chua, G.; and Shong, R. 2025. MinorBench: A hand-built benchmark for content-based risks for children. arXiv:2503.10242.
- Knox, W. B.; Bradford, K.; Castro, S. V.; Ong, D. C.; Williams, S.; Romanow, J.; Nations, C.; Stone, P.; and Baker, S. 2025. Harmful Traits of AI Companions. arXiv:2511.14972.
- Lee, H.-P. H.; Yang, Y.-J.; Von Davier, T. S.; Forlizzi, J.; and Das, S. 2024. Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.
- Li, L.; Ma, R.; Xue, Z.; and Xiong, J. 2025. Towards Trustworthy AI: Characterizing User-Reported Risks across LLMs "In the Wild". arXiv:2509.08912.
- Madaio, M. A.; Egede, L.; Subramonyam, H.; Wortman Vaughan, J.; and Wallach, H. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1): 1–26.
- Madaio, M. A.; Stark, L.; Wortman Vaughan, J.; and Wallach, H. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *CHI '20*, 1–14. New York, NY, USA: Association for Computing Machinery. Journal Abbreviation: CHI '20.
- Matentzoglu, N.; Balhoff, J. P.; Bello, S. M.; Bizon, C.; Brush, M.; Callahan, T. J.; Chute, C. G.; Duncan, W. D.; Evelo, C. T.; Gabriel, D.; et al. 2022. A simple standard for sharing ontological mappings (SSSOM). *Database*, 2022: baac035.
- Metcalfe, J.; Moss, E.; Watkins, E. A.; Singh, R.; and Elish, M. C. 2021. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 735–746. Virtual Event Canada: ACM. ISBN 978-1-4503-8309-7.
- Mitchell, M.; Pistilli, G.; Jernite, Y.; Ozoani, E.; Gerchick, M.; Rajani, N.; Luccioni, A. S.; Solaiman, I.; Masoud, M.;

- Nikpoor, S.; Ferrandis, C. M.; Bekman, S.; Akiki, C.; Contractor, D.; Lansky, D.; McMillan-Major, A.; Thrush, T.; Ilić, S.; Dupont, G.; Longpre, S.; Dey, M.; Biderman, S.; Kiela, D.; Baylor, E.; Le Scao, T.; Gokaslan, A.; Launay, J.; and Muennighoff, N. 2022. Model Card.
- Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11): 501–507.
- Morley, J.; Floridi, L.; Kinsey, L.; and Elhalal, A. 2020. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, 26(4): 2141–2168.
- Narayanan, A. 2018. Translation Tutorial: 21 Fairness Definitions and Their Politics. In *Proc. Conf. Fairness Accountability Transp.*, New York, Usa, volume 1170, 3.
- Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. Barcelona Spain: ACM. ISBN 978-1-4503-6936-7.
- Rakova, B.; Yang, J.; Cramer, H.; and Chowdhury, R. 2021. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–23.
- Ribes, D.; and Finholt, T. A. 2009. The Long Now of Infrastructure: Articulating Tensions in Development. *Journal for the Association of Information Systems (JAIS): Special issue on eInfrastructures*, 10(5): 375–398.
- Rismani, S.; Shelby, R.; Davis, L.; Rostamzadeh, N.; and Moon, A. 2025. Measuring what matters: Connecting AI ethics evaluations to system attributes, hazards, and harms. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8(3), 2199–2213.
- Rismani, S.; Shelby, R.; Smart, A.; Jatho, E.; Kroll, J.; Moon, A.; and Rostamzadeh, N. 2023. From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.
- Selbst, A. D. 2021. An Institutional View Of Algorithmic Impact Assessments. *Harv. J. Law Technol.*, 35(1).
- Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 59–68. ISBN 978-1-4503-6125-5.
- Shelby, R.; Rismani, S.; Henne, K.; Moon, Aj.; Rostamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; and Virk, G. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–741. Montreal QC Canada: ACM. ISBN 979-8-4007-0231-0.
- Slaterry, P.; Saeri, A. K.; Grundy, E. A.; Graham, J.; Noetel, M.; Uuk, R.; Dao, J.; Pour, S.; Casper, S.; and Thompson, N. 2026. The AI risk repository: A meta-review, database, and taxonomy of risks from artificial intelligence. *Patterns*, 7.
- Slaughter, R. K.; Kopec, J.; and Batal, M. 2020. Algorithms and economic justice: A taxonomy of harms and a path forward for the federal trade commission. *Yale JL & Tech.*, 23: 1. Publisher: HeinOnline.
- Solaiman, I.; Talat, Z.; Agnew, W.; Ahmad, L.; Baker, D.; Blodgett, S. L.; Chen, C.; Daumé, H.; Dodge, J.; Duan, I.; Evans, E.; Friedrich, F.; Ghosh, A.; Gohar, U.; Hooker, S.; Jernite, Y.; Kalluri, R.; Lusoli, A.; Leidinger, A.; Lin, M.; Lin, X.; Luccioni, S.; Mickel, J.; Mitchell, M.; Newman, J.; Ovalle, A.; Png, M.-T.; Singh, S.; Strait, A.; Struppek, L.; and Subramonian, A. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. Version Number: 4.
- Star, S. L. 1989. The Structure of Ill-Structured Solutions: Boundary Objects and Heterogeneous Distributed Problem Solving. In *Distributed Artificial Intelligence*, 37–54. Elsevier. ISBN 978-1-55860-092-8.
- Star, S. L. 2010. This is Not a Boundary Object: Reflections on the Origin of a Concept. *Science, Technology, & Human Values*, 35(5): 601–617.
- Star, S. L.; and Griesemer, J. R. 1989. Institutional Ecology, ‘Translations’ and Boundary Objects: Amateurs and Professionals in Berkeley’s Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 19(3): 387–420.
- Star, S. L.; and Strauss, A. 1999. Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work. *Computer Supported Cooperative Work (CSCW)*, 8(1): 9–30.
- Suchman, L. 2002. Located Accountabilities in Technology Production. *Scandinavian Journal of Information Systems*, 14(2).
- Timmermans, S.; and Berg, M. 1997. Standardization in Action: Achieving Local Universality through Medical Protocols. *Social Studies of Science*, 27(2): 273–305.
- Timmermans, S.; and Epstein, S. 2010. A World of Standards but not a Standard World: Toward a Sociology of Standards and Standardization. *Annual Review of Sociology*, 36(1): 69–89.
- UN B-Tech team. 2023. Taxonomy of Human Rights Risks Connected to Generative AI. Technical report, United Nations Human Rights Office of the High Commissioner.
- Vassilev, A.; Oprea, A.; Fordyce, A.; and Anderson, H. 2024. Adversarial machine learning : a taxonomy and terminology of attacks and mitigations. Technical Report NIST 100-2e2023, National Institute of Standards and Technology (U.S.), Gaithersburg, MD.
- Veale, M.; and Zuiderveen Borgesius, F. 2021. Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4): 97–112.
- 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.; Pangakis, N.; Reed, S.; Sheng, E.; Vann, D.; Vaughan, J. W.; Vogel, M.; Washington, H.; and Jacobs, A. Z. 2025. Position: evaluating generative AI systems is a social science measurement challenge. In *Proceedings of the 42nd International Conference on Machine Learning, ICML'25*. JMLR.org.
- Wang, L.; Kameswaran, V.; and Kacorri, H. 2025. Toward a Taxonomy of Algorithmic Harms for Disability: A Systematic Review. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3): 2649–2665.
- Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; Biles, C.; Brown, S.; Kenton, Z.; Hawkins, W.; Stepleton, T.; Birhane, A.; Hendricks, L. A.; Rimell, L.; Isaac, W.; Haas, J.; Legassick, S.; Irving, G.; and Gabriel, I. 2022. Taxonomy of Risks Posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229. Seoul Republic of Korea: ACM. ISBN 978-1-4503-9352-2.
- Wong, R. Y.; Madaio, M. A.; and Merrill, N. 2023. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–27.
- Yu, Y.; Liu, Y.; Zhang, J.; Huang, Y.; and Wang, Y. 2025. Youth-Centered GAI Risks (YAIR): A Taxonomy of Generative AI Risks from Empirical Data. In *Symposium On Usable Privacy and Security*.
- Zhang, R.; Li, H.; Meng, H.; Zhan, J.; Gan, H.; and Lee, Y.-C. 2025. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25*. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.
- Zhou, J.; Chen, A. Z.; Shah, D.; Schwab-Reese, L. M.; and De Choudhury, M. 2025. A Risk Taxonomy and Reflection Tool for Large Language Model Adoption in Public Health. *Proc. ACM Hum.-Comput. Interact.*, 9(7).
- Zhu, J.; Coifman, K. G.; and Jin, R. 2026. Understanding Risk and Dependency in AI Chatbot Use from User Discourse. arXiv:2602.09339.