

Unaccountable Delegation, Fading Skills: Mapping the Risks of Workplace AI Agents

Gabriele La Malfa¹, Lakmal Meegahapola², Edyta Bogucka², Jie M. Zhang¹, Michael Luck³,
Elizabeth Black¹, Daniele Quercia^{2,4}

¹King’s College London, UK

²Nokia Bell Labs, Cambridge, UK

³University of Sussex, UK

⁴Politecnico di Torino, Italy

{gabriele.la_malfa, jie.zhang, elizabeth.black}@kcl.ac.uk, michael.luck@sussex.ac.uk,
{lakmal.meegahapola, edyta.bogucka, daniele.quercia}@nokia-bell-labs.com

Abstract

To anticipate the socio-technical risks posed by AI agents, organizations first need taxonomies to classify them. However, existing AI risk taxonomies focus on broad risks and do not capture the job-specific risks introduced by agents. To address this gap, we make three main contributions. First, we developed a multi-layer framework based on a review of the literature on AI agents. The framework models three core components and their interactions: the agents, their goals, and their environment. Second, we embedded this framework in a structured prompt and applied it to descriptions of 2,078 job tasks from the O*NET occupational database, producing 8,356 risk scenarios labeled by severity and deployment mode (automation or augmentation). We validated these scenarios with 45 workers across 10 job roles and an independent LLM judge, confirming their high plausibility and alignment with the corresponding job tasks. Finally, we extended an existing taxonomy to create a 15-category taxonomy of workplace AI agent risks that covers all our risk scenarios. Our analysis highlights four findings. First, augmentation is not inherently safe because overreliance on agents can gradually erode workers’ skills and oversight. Second, erroneous actions of agents account for the largest share of risk scenarios and has the highest number of severe risks. Many of these risks arise at the human–agent boundary. Third, automation is associated mainly with organizational risks (operational failures, financial losses), whereas augmentation is associated mainly with risks to workers (capability erosion, psychological and social risks). Fourth, workers found our taxonomy easier to use for a risk classification task than two other risk taxonomies and preferred it in 64% of non-tied comparisons with a recent generative AI risk taxonomy. Taken together, these findings show that workplace AI agent risks do not arise from agents alone; they also depend on how people work with agents and how agents are deployed. Safer workplaces therefore require not only safer agents but also carefully designed human–AI agent collaboration.

Project website —

<https://social-dynamics.net/ai-risks/workplace-agents/>

1 Introduction

AI agents are autonomous entities that perceive their environment, interact with humans or other agents, and act to

achieve goals with varying degrees of independence (Sapkota, Roumeliotis, and Karkee 2026). They are increasingly deployed to support everyday workplace tasks (Eloundou et al. 2024; Xi et al. 2025; Mohny 2025). AI agents do not operate in isolation: together with the environment they perceive, the goals they pursue, and the humans they work alongside, they form what we call an *agentic AI system* (AAIS) (IBM 2024). As these systems are deployed across jobs and industries, organizations need ways to identify and classify the risks they introduce (Weidinger et al. 2023). This need arises when teams red-team agents before launch (Ganguli et al. 2022), conduct impact assessments (Moss et al. 2021), or conduct risk and incident audits (Raji et al. 2020; McGregor 2021). In all of these cases, organizations rely on risk taxonomies: classification systems that provide a shared language for anticipating risks that have not yet occurred and classifying those that have (Shelby et al. 2023). Without a taxonomy built for the specific context of workplace AI agents, these efforts rely on risk classifications that can leave risks unnamed and untracked.

Existing AI risk taxonomies were not designed for workplace agentic AI systems, leaving three gaps. First, many taxonomies are too generic or focus primarily on technical failures. They can therefore overlook *socio-technical AI risks* that emerge from the dynamic interaction between an AI agent and its social context, including the humans, organizations, and institutions in which it is embedded. Second, most taxonomies were developed for generic AI (Abercrombie et al. 2024) or generative AI (Li et al. 2025b), rather than for AI agents, which can act more autonomously. Even recent AI agent risk taxonomies tend to model agents as purely technical systems (Hammond et al. 2025; Bryan et al. 2025; Khoo, Foo, and Lee 2025; Ghosh et al. 2025), without explicitly capturing how agents, goals, environments, and human workers interact to produce risks. Third, taxonomies based on incident databases and media reports (Abercrombie et al. 2024; Turri and Dzombak 2023; Paeth et al. 2025) capture risks that have already occurred and were visible enough to be recorded. They are therefore less suited to identifying risks that develop gradually or have not yet produced visible harms, such as the erosion of professional skills caused by prolonged reliance on AI (Richards, Benn,

and Zilka 2025; Shao et al. 2025a). Together, these gaps point to the need for a taxonomy that operates at the level of specific job tasks, reflects how agentic AI systems are structured, and anticipates risks before they become harms. To address these gaps, we make three main contributions:

1. We introduce a multi-layer framework for workplace AI agents. Building on prior work on socio-technical safety and compositional risk assessment, the framework organizes risks across three layers (§3.1): Technical Capability (AAIS components and how they work together), Human Interaction (how people engage with AAIS), and Systemic Impact (broader organizational and societal consequences). The framework also captures the relationships among agents, goals, environments, and human workers. We refined the framework through structured interviews with three experts in agentic AI systems and responsible AI.

2. We generate and validate job-specific risk scenarios. We embedded the framework in a structured prompt and applied it to descriptions of 2,078 job tasks from the O*NET occupational database (§3.2). This process generated 8,356 risk scenarios labeled by severity (minimal, limited, high, critical) and deployment mode (§3.3): augmentation, in which an agent supports a worker, or automation, in which the agent performs the job task without worker involvement. We then validated the scenarios in two ways (§3.4). First, 45 workers across 10 job roles evaluated 450 scenarios from their own roles. Second, an LLM-as-a-judge from a different model family assessed the full corpus. Both evaluations found high plausibility (workers: 4.03/5; LLM: 4.81/5) and strong task alignment (workers: 3.94/5; LLM: 4.60/5).

3. We construct and validate a workplace AI agent risk taxonomy. To ground the taxonomy in both generated and observed risks, we supplemented the 8,356 generated scenarios with 222 workplace AI agent incidents from three public databases and 11 failure cases from a live multi-agent red-teaming study. We converted these cases into a common risk scenario format and used the combined corpus to extend Li et al.’s generative AI risk taxonomy into a workplace AI agent risk taxonomy (§3.5). The resulting taxonomy contains 15 categories and 44 sub-categories (Figure 2). We evaluate its structural qualities, compare its coverage with 10 established AI risk frameworks, and test its usability in a comparative study with 26 workers (§3.6). Finally, we analyze the risk corpus along the themes of taxonomy categories, severity, deployment mode, and framework layers.

Across these contributions, our results point to a clear pattern (§4): most workplace AI agent risks arise from how agents and people work together. The two largest taxonomy categories are *Erroneous Actions* of agents (30.6%), including misinterpretations and incorrect recommendations, and *Capability Erosion* (21.3%), including the loss of workers’ skills and oversight capabilities. Together, these categories account for 51.9% of all risk scenarios and occur mostly under augmentation. Under automation, risks shift to the organizational level, particularly *Operational Failures* and *Financial Losses*. Taken together, these findings suggest that safer workplaces require not only safer agents but also care-

fully designed human–AI agent collaboration.

2 Related Work

Prior work provides a good foundation for understanding AI risks in general. However, three gaps remain. Existing taxonomies tend to operate at the wrong level of detail for specific job tasks (§2.1), are rarely built around how agents behave (§2.2), and focus on past incidents rather than emerging risks (§2.3).

2.1 From Broad Risk Classifications to Job-Level Risk

Research on AI risk tends to fall into two areas: technical and socio-technical. Both offer valuable perspectives, but neither maps easily onto what happens when a specific AI agent enters a specific job workflow. Technical approaches explain risks primarily at model-level, such as bias in training data or failures of robustness (Mehrabi et al. 2021; Raji et al. 2022; Fabris et al. 2022). These are important concerns, but they tend to abstract away the organizational and human context in which AI operates. Socio-technical approaches take a broader view, framing AI as part of a wider social system and classifying large-scale risks such as unfair resource distribution or the erosion of accountability (Selbst et al. 2019; Macrae 2022; Shelby et al. 2023; Slattery et al. 2025). These approaches show why context matters, but their broad categories can be difficult to apply to a particular task or workflow. Neither perspective is designed to answer a practical question facing organizations: “What risks could emerge when this AI agent is introduced into this specific job task?” What is missing is a framework that connects multi-layer risk concepts to concrete job tasks and workflows.

2.2 Modeling Risks from Agent Behavior

Risks of AAISs can emerge from how agents pursue goals, take multi-step actions, and interact with their environments (Sapkota, Roumeliotis, and Karkee 2026; Chan et al. 2023; Gridach et al. 2025; Liu et al. 2025). When several agents work together within an AAIS, coordination failures, conflicting goals, and emergent behavior create additional risks (Hammond et al. 2025; Altmann et al. 2024; Kong et al. 2025). Autonomy can also make agent behavior harder to monitor and responsibility harder to assign (Anwar et al. 2024; Ranjan, Gupta, and Singh 2025). Moreover, recent work categorizes agent-specific risks such as misalignment with human intent, prompt-injection vulnerabilities, and cascading failures in multi-agent systems (Bryan et al. 2025; Khoo, Foo, and Lee 2025; Ghosh et al. 2025). However, these risk classifications often study agents in isolation or in simulated environments. They rarely connect the three components at the core of an AAIS: the agents, their goals, and their environment, to concrete job tasks performed by workers. Thus, existing agent-safety research identifies important risks and failure mechanisms but does not yet explain how those mechanisms produce risks in specific job tasks.

2.3 From Existing Evidence to Anticipating Risks

AI risk taxonomies draw on several sources of evidence. Documentation-based approaches capture reported incidents and other public records (Li et al. 2025b; Lee et al. 2024), academic and policy literature (Shelby et al. 2023; Slatery et al. 2025), or user-generated data such as social media discussions (Zhang et al. 2025). Human-centered approaches develop or refine risk categories through interviews, expert consultation, workshops, and annotation studies (Steenstra and Bickmore 2025; Abercrombie et al. 2024). Some studies combine these sources. For example, Abercrombie et al. (Abercrombie et al. 2024) build on existing taxonomies and documented incidents, then refine their categories through expert feedback and crowdsourced annotation. These sources offer complementary benefits. Documented evidence grounds taxonomies in observed harms, while human elicitation can reveal contextual knowledge that public records overlook. However, both also have limitations. Incident databases contain only harms that were observed, reported, and made public (Paeth et al. 2025; Richards, Benn, and Zilka 2025). Literature-based taxonomies inherit the scope of prior research, while human elicitation depends on the experience and composition of the participants. As a result, emerging, gradual, and job-specific risks may remain underrepresented, particularly as AI capabilities and workplace uses evolve. Prospective methods offer a complementary way to identify risks before they produce documented harms. Research in labor economics and HCI has used the O*NET occupational database to ground analyses of AI exposure in real job tasks (Eloundou et al. 2024; Felten, Raj, and Seamans 2023). Other work shows that LLMs can support structured scenario generation and red-teaming exercises that anticipate risks before deployment (Perez et al. 2022; Ganguli et al. 2022; Fröhling et al. 2026; Herdel et al. 2025). However, these research strands have largely developed separately. What is missing is a systematic method that combines real job-task data with prospective risk generation to identify workplace AI agent risks before they become documented harms.

Taken together, prior work offers broad risk classifications, detailed accounts of agent failure modes, and strong methods for learning from existing evidence. However, these three strands have not yet been fully integrated. A workplace AI agent risk taxonomy must therefore (1) be developed by considering specific job tasks, (2) consider risks across *Technical Capability*: AAISs including agents, goals, and environments, and how they interact; *Human Interaction*: how workers engage with these systems; and *Systemic Impact*: broader organizational and societal consequences, and (3) anticipate risks that may not yet appear in incident records. Together, these gaps motivate our job-task-grounded and forward-looking approach, which combines anticipated risk scenarios with documented evidence to construct a taxonomy of workplace AI agent risks.

3 Methodology

Our methodology consists of six steps (Figure 1). First, we define a multi-layer framework for analyzing risks in AAISs

(§3.1). Second, we select job tasks in the O*NET database that AI agents are likely to affect (§3.2). Third, we embed the multi-layer framework in a structured prompt to generate risk scenarios for job tasks (§3.3). Fourth, we evaluate the scenarios with workers and an LLM-as-a-judge (§3.4). Fifth, we use the risk scenarios to construct a workplace AI agent risk taxonomy (§3.5). Finally, we evaluate the taxonomy for structural integrity, coverage relative to 10 established frameworks, and practical usability (§3.6).

3.1 Defining the Multi-Layer AAIS Framework

We build the multi-layer framework by extending the sociotechnical framework of Weidinger et al. (Weidinger et al. 2023), which argues that evaluating technical artifacts alone cannot establish whether an AI system is safe. A capability may indicate a potential hazard, but whether it produces harm depends on the people using the system, the context of use, and the broader systems in which it is deployed. The framework therefore distinguishes three targets of evaluation: *capability*, which examines the AI system and its technical artifacts; *human interaction*, which examines how people use and respond to the system; and *systemic impact*, which examines broader organizational, social, economic, and environmental effects. These layers add progressively more context, but they are not chronological stages and can influence one another. To apply this layered approach to workplace AI agents, we make the internal structure of an agentic AI system explicit. Workplace agents pursue goals, act in changing environments, and interact with other agents and human workers. Risks can arise from these components or from their interactions, even when each component appears to function correctly. We therefore retain the three evaluation layers while expanding them to represent the components and interaction pathways of an AAIS.

Layer 1: Capability. This layer represents risks arising from three components and their interactions (Russell and Norvig 2009; Wooldridge 2009): the *AI agents*, their *environment*, and their *goals*.

Component risks. Agent properties such as autonomy, reactivity, and proactiveness can create different risks. An agent may respond poorly to an unfamiliar situation, take an unrequested action, or modify a workflow without sufficient oversight. The environment, including APIs, databases, and physical workspaces (IBM 2024), can create risks when it provides incomplete, outdated, or rapidly changing information. Goals can create risks when they are represented by poor proxy measures. For example, a diagnostic agent optimized for case volume may deprioritize difficult cases to increase throughput.

Interaction risks. Risks can also emerge from interactions among components. In *agent-agent* interactions, missing coordination rules may lead agents to take conflicting actions. In *agent-environment* interactions, an agent may optimize a local objective while violating environmental or organizational constraints. In *agent-goal* interactions, an agent may satisfy an objective in an unintended way. For example, a quality-control agent could classify every item as defective to maximize its detection score, thereby halting production.

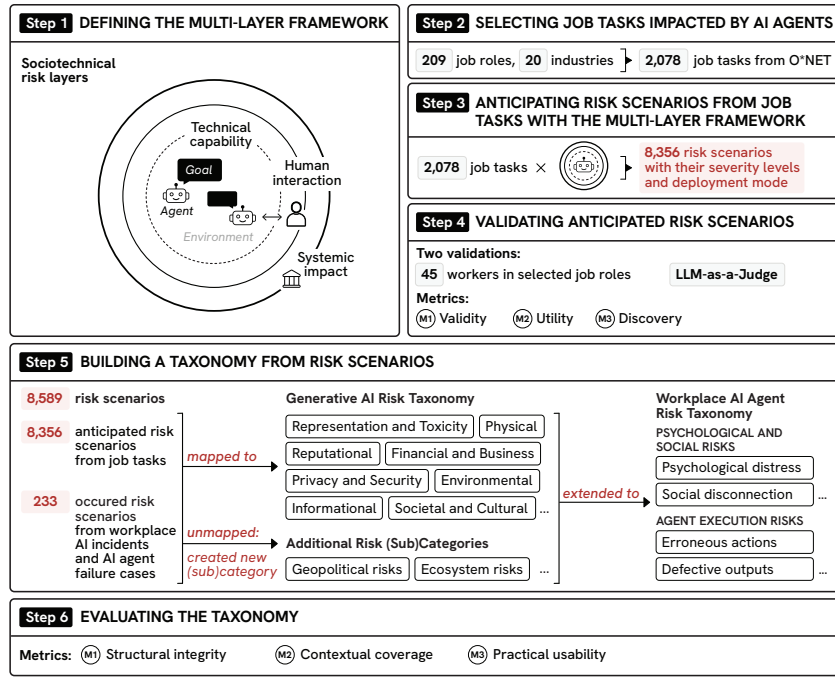


Figure 1: **Overview of the six-step methodology, from framework definition through risk scenario generation and validation to taxonomy construction and evaluation.** **Step 1** defines the multi-layer framework, which organizes sociotechnical risk into three layers: *technical capability* (AAIS including the AI agents, goals they pursue, and the environment they act in), *human interaction* (how workers engage with the AAIS and components), and *systemic impact* (consequences that propagate beyond the individual job task). **Step 2** selects 2,078 computer-based O*NET job tasks (209 roles, 20 industries). **Step 3** applies the framework to each task through a structured prompt, anticipating 8,356 risk scenarios labeled with severity and deployment mode. **Step 4** validates the scenarios with 45 workers from the selected roles and with an LLM-as-a-judge, on validity, utility, and discovery metrics. **Step 5** builds the taxonomy from 8,589 scenarios: the 8,356 anticipated ones plus 233 describing risks that have already occurred, drawn from workplace AI incidents and documented agent failure cases; scenarios that do not map onto the GenAI Risk Taxonomy produce new categories and sub-categories, extending it into the 15-category Workplace AI Agent Risk Taxonomy. **Step 6** evaluates the taxonomy on structural integrity, contextual coverage, and practical usability.

Layer 2: Human Interaction. This layer represents how workers interact with agents and the broader system (Hammond et al. 2025). We distinguish five interaction pathways. In *agent–human* interactions, agents may undermine trust, violate workplace norms, or encourage over-reliance. *Human–human* interactions may change when AI-generated targets influence how managers evaluate and direct workers. *Human–environment* interactions capture how agent-oriented workplaces may constrain or endanger workers, such as when a warehouse is organized primarily for robotic efficiency. *Human–goal* interactions capture conflicts between system objectives and professional judgment, such as when clinical staff feel pressured to follow automated triage scores. Finally, *human–AAIS* interactions capture risks arising from the system as a whole, including cases in which responsibility becomes too distributed for workers to understand or contest a decision.

Layer 3: Systemic Impact. This layer represents consequences that extend beyond an individual task or interaction. Such risks may affect organizations, industries, society, the economy, or the natural environment. They often emerge through large-scale adoption, repeated interactions, or feedback loops rather than a single failure. For exam-

ple, interacting AI trading agents may collectively amplify incorrect market signals and contribute to a market disruption. Widespread use of AI management systems may shift bargaining power toward employers or normalize harmful workplace practices. Similarly, repeated use of agents that favor established solutions may gradually reduce the diversity of professional knowledge across a field.

Framework Review and Refinement. We reviewed the framework through three structured interviews with experts in agentic AI systems and responsible AI. Each interview lasted approximately 30 minutes and included a framework briefing, technical questions, and detailed feedback (Supplementary Material B). The experts assessed the coherence of the components, interaction pathways, and three layers and identified areas requiring clarification. We used their feedback to refine definitions, clarify the boundaries between interactions, and improve the presentation of the framework.

3.2 Selecting Job Tasks Impacted by AI Agents

To ground our analysis in real job tasks, we used O*NET (National Center for O*NET Development 2025), a U.S. government database describing the tasks, activities, and work contexts of hundreds of jobs. We used a filtered

subset developed by Shao et al. (Shao et al. 2025b), which focuses on computer-based tasks on which AI agents are likely to be deployed. It contains 2,078 job tasks from 209 job roles across 20 industries, including healthcare, finance, and transportation. Each task is described concretely. Examples include “Diagnose acute conditions that could result in rapid physiological deterioration” for an acute care nurse and “Examine whether the organization’s objectives are reflected in its management activities” for a manager. These descriptions provide the job context needed to generate job task-specific risk scenarios.

3.3 Generating Risk Scenarios for Job Tasks with the Multi-Layer Framework

We use gpt-4o-mini (OpenAI 2024) to generate prospective risk scenarios for each of the 2,078 tasks. This approach is consistent with prior work showing that LLMs can serve as effective tools for structured risk foresight (Bogucka, Šćepanović, and Quercia 2024; Bućinca et al. 2023; Herdel et al. 2025; Wang et al. 2024b; Perez et al. 2022; Fröhling et al. 2026). We selected gpt-4o-mini to balance reasoning quality with the computational demands of processing thousands of prompts at scale. To prevent the model from producing generic or implausible outputs, we used a structured prompt that included a description of our multi-layer framework defining agentic AI systems’ risk surfaces, enabling the generation of risks specific to agents and their interactions (Supplementary Material E; Figure 1). The prompt instructed the model to consider risks at each framework layer (capability, human interaction, systemic impact) and to label each scenario with its corresponding component or interaction pathway (e.g., agent–human, agent–goal). This structure kept the generated risks tied to specific pathways rather than generic failures. For each task, the model was asked to generate up to five risk scenarios, enough to capture a diverse set of risks while avoiding highly speculative edge cases. Each scenario was annotated with two additional labels: (1) *severity*, classified according to four risk tiers (minimal, limited, high, critical) (Yin et al. 2025), and (2) *deployment mode*, specifying whether the risk arises from *automation* (complete replacement of the human task as defined in (Raisch and Krakowski 2021)) or *augmentation* (AI supporting the human in performing the tasks as defined in (Autor 2015)). This process produced 8,356 distinct risk scenarios.

3.4 Validating Risk Scenarios

We assessed the quality of the generated scenarios using both workers and an LLM judge, in line with previous work (Zheng et al. 2023; Li et al. 2024; Chiang and yi Lee 2023; Li et al. 2025a; Fröhling et al. 2026). In both cases, each scenario was evaluated on 10 metrics across three dimensions: *validity*, covering plausibility and connection to the task; *utility*, covering usefulness, actionability, detail, complexity, and specificity; and *discovery*, covering originality, rarity, and novelty (Si, Yang, and Hashimoto 2024; Hu and Zhou 2024).

Domain Expert Evaluation. We selected 10 job roles and recruited 45 workers whose reported professions matched

those roles (Supplementary Material C). We collected demographic and role-related information from each expert. For each job role, we randomly sampled 10 of the approximately 70 role-specific scenarios. Each expert independently evaluated the 10 scenarios for their role using the 10 metrics described above on a 5-point Likert scale. In total, the workers evaluated 450 risk scenarios.

LLM-as-a-Judge Evaluation. The domain expert evaluation provided useful ratings but covered only the subset of scenarios that the experts could feasibly review. To rate all 8,356 scenarios using the same criteria, we used gemini-3-flash (Google 2025) as an LLM judge. The model rated each scenario using the 10 metrics described above, allowing us to apply a consistent standard across the full corpus. We took two steps to assess the reliability of these ratings. First, we chose a judge from a different model family than the generator (Google Gemini rather than the OpenAI model used to generate the scenarios), since models tend to score their own outputs more favorably (Panickssery, Bowman, and Feng 2024). Second, we tested the judge’s consistency by re-rating a random sample of 370 scenarios with a second model, gpt-4o, and comparing the two. The models rarely gave the exact same 5-point score. However, when we collapsed the 5-point scale into positive, neutral, and negative bands (*sentiment agreement*), the two models landed in the same band almost every time: 98.9% for plausibility and 98.1% for connection to the task. Therefore, the judges agreed on the underlying judgment for the validation sample (Supplementary Material C).

Results. Both the workers and the LLM judge rated the risk scenarios as highly plausible (workers: 4.03/5; LLM: 4.81/5) and closely tied to the corresponding job tasks (workers: 3.94/5; LLM: 4.60/5). They also rated the scenarios as useful (workers: 3.70/5; LLM: 3.38/5) and actionable (workers: 3.62/5; LLM: 3.37/5). Their ratings differed most on novelty: workers considered the scenarios moderately novel (originality: 2.96/5; rarity: 3.00/5), while the LLM judged them as more familiar (originality: 1.80/5; rarity: 1.28/5). Overall, we interpret these results as a sign that the risk scenarios are well grounded: workers recognized them as plausible and practically relevant, while the LLM, which has seen vast amounts of text, recognized them as patterns it had encountered before (Supplementary Material C).

3.5 Building a Taxonomy from Risk Scenarios

We began with the 8,356 risk scenarios generated in the previous step. To ground the taxonomy in real-world evidence, we added 11 AI agent failure cases from a study of agent behavior (Shapira et al. 2026) and 222 workplace AI incidents from three public incident databases. We converted both the failure cases and incidents into the same risk scenario format as the generated scenarios. Next, our goal was to map as many scenarios as possible to an existing taxonomy. We used the Generative AI (GenAI) risk taxonomy of Li et al. (Li et al. 2025b) as the starting point and followed three steps. First, we used gpt-4o-mini (for the generated risk corpus) or manual inspection (for failure cases and incidents) to map each scenario to the closest category or sub-

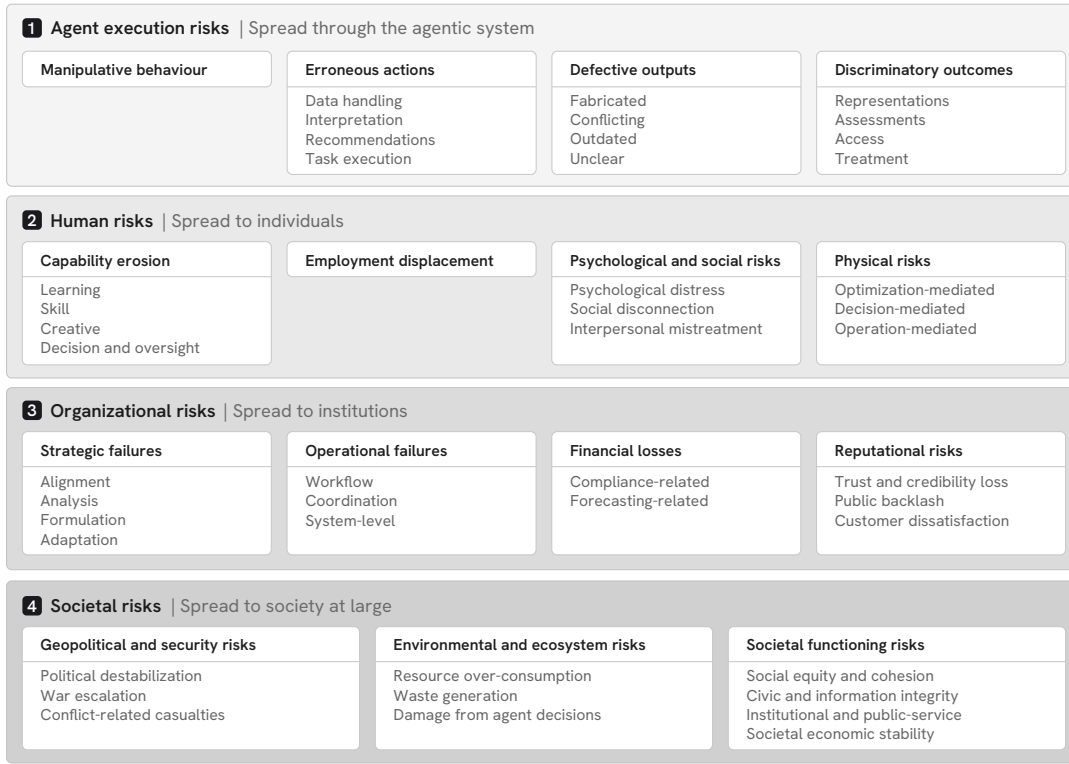


Figure 2: **Workplace AI Agent Risk Taxonomy.** 15 categories and 44 sub-categories organized across agent execution risks, human risks, organizational risks, and societal risks.

category in Li et al.’s taxonomy. Scenarios without a suitable match were marked as unmatched. Second, we analyzed the unmatched scenarios and identified recurring themes. Third, we manually reviewed these themes to determine how they should extend the taxonomy. When a theme fit within an existing category, we added it as a new sub-category. When it did not fit any existing category, we created a new top-level category. We then refined the category and sub-category definitions to make their boundaries clear. The resulting Workplace AI Agent Risk Taxonomy contains 15 top-level categories and 44 sub-categories (Figure 2). See Supplementary Material G for the full taxonomy with descriptions and example risk scenarios.

3.6 Evaluating the Taxonomy

We evaluated the taxonomy in three ways. First, we checked whether the categories were distinct (structural integrity). Second, we checked whether the taxonomy captured workplace risk categories that existing frameworks overlook (contextual coverage). Third, we examined whether workers could use it to classify workplace risks (practical usability).

Structural Integrity. We first examined whether the 15 categories and 44 sub-categories were distinct in both their vocabulary and meaning. *Vocabulary check.* We represented the taxonomy with names and definitions for categories and sub-categories. Then, we converted this text into a TF-IDF vector, which gives more weight to words that are distinctive to a category and less weight to commonly used words.

We then calculated the cosine similarity between every pair of categories on a scale from 0 to 1, where higher values indicate greater overlap. The mean inter-category similarity was 0.01, indicating minimal overlap in vocabulary. We also calculated Jaccard similarity, which measures the proportion of words shared by two descriptions. Using 0.30 as the overlap threshold, 99.9% of sub-category pairs fell below this value. *Meaning check.* Categories may use different words while describing similar concepts. For example, “skill erosion” and “deskilling” have similar meanings despite sharing few words. To capture this type of overlap, we encoded the category definitions using the pre-trained sentence-transformer model: all-MiniLM-L6-v2. The model represents texts with similar meanings as nearby vectors, even when they use different words. The mean pairwise semantic similarity was 0.24, indicating limited average overlap in meaning. Together, these checks suggest that the taxonomy definitions are distinct in both their vocabulary and their underlying concepts (Supplementary Material H).

Contextual coverage. To check whether our taxonomy captures workplace risk categories that prior work overlooks, we compared it with ten established AI risk taxonomies and classification frameworks, which we collectively call *risk classifications*. We followed the comparative mapping approach of Abercrombie et al. (Abercrombie et al. 2024). Our comparison included the seven sources examined in their study and we also included three more recent sources (Table 1). *Mapping process.* For each of our 15 categories, we

Workplace AI Agent Risk Category	EPIC	CSET	Turing	Azure	Sony	TASRA	Shelby	AIR	MA Adv AI	AI Risk Repo
Agent Execution Risks										
Manipulative Behaviour	■	□	■	■	■	■	■	■	■	■
Erroneous Actions	□	□	■	□	■	■	■	■	■	■
Defective Outputs	□	□	□	□	□	■	□	■	■	■
Discriminatory Outcomes	■	■	■	■	■	■	■	■	■	■
Human Risks										
Capability Erosion	■	□	■	■	■	■	■	□	■	■
Employment Displacement	■	□	□	■	■	■	■	□	■	■
Psychological and Social Risks	■	■	■	■	■	■	■	■	■	■
Physical Risks	■	■	■	■	■	■	■	■	■	■
Organizational Risks										
Strategic Failures	□	□	□	■	□	□	■	□	■	□
Operational Failures	□	■	□	□	□	□	□	□	■	□
Financial Losses	■	■	□	■	□	■	□	■	■	□
Reputational Risks	■	■	■	□	■	■	■	□	■	□
Societal Risks										
Geopolitical and Security Risks	□	□	□	□	□	■	□	■	■	□
Environmental and Ecosystem Risks	■	■	■	■	□	■	■	□	■	■
Societal Functioning Risks	■	■	■	■	■	■	■	■	■	■

Table 1: **Coverage of our 15 taxonomy categories across ten established AI risk classifications.** Categories are grouped by dimension. We use the color coding: ■ = full match, ■ = partial match, □ = no match. Compared risk classifications are: EPIC (Fergusson et al. 2023), CSET (Hoffman and Frase 2023), the Alan Turing Institute (Leslie 2019), Microsoft Azure (Microsoft 2023), Sony (Hutiri, Papakyriakopoulos, and Xiang 2024), TASRA (Critch and Russell 2023), and Shelby et al. (Shelby et al. 2023), AIR 2024 (Zeng et al. 2024), Hammond et al. (Hammond et al. 2025), and the AI Risk Repository (Slattery et al. 2025). We find that organizational risks (*Strategic Failures*, *Operational Failures*) and agent-execution risks (*Erroneous Actions*, *Defective Outputs*) are consistently absent from or only partially covered by prior work.

examined its coverage in each of the ten risk classifications. We assigned one of three labels: *Full match* (■) when a comparison source contained a category that directly represented ours; *Partial match* (■) when it covered only part of our category; and *No match* (□) when it did not cover the risk (Supplementary Material L). Table 1 shows that prior risk classifications cover human-oriented risks thoroughly: *Discriminatory Outcomes* of agents is fully represented in nine of the ten classifications, *Physical Risks* in seven, and *Psychological and Social Risks* in six. Beyond these, three gaps emerge. First, organizational risks are the least covered in the comparison: *Operational Failures* and *Strategic Failures* are the only two categories with no full match anywhere, and are absent from eight and seven of the ten classifications respectively. Second, the agent-execution risks that dominate our corpus are recognized only sporadically: under agent execution risks, *Defective Outputs* is absent from six classifications, and *Erroneous Actions* is fully represented in only two, appearing elsewhere as generic model inaccuracy rather than as distinct failures of what an agent interprets, recommends, and produces. Third, and most tellingly, *Manipulative Behaviour* of agents is fully represented only

in the two classifications built specifically for agentic systems (Hammond et al. 2025; Slattery et al. 2025); the remaining eight treat deception either as human misuse of a model or not at all. By contrast, our *Societal Functioning Risks* category is well covered by prior work, with seven full matches, which reinforces that its small share in our corpus (0.4%) reflects our task-level elicitation rather than a gap in the literature. Together, these patterns suggest that existing risk classifications were not designed to represent the organizational, operational, and agent-execution risks that emerge when AI agents are introduced into workplaces. Our workplace-specific taxonomy addresses this limitation.

Practical usability. A taxonomy should be easy to use in practice. We therefore designed a study around a common risk-management activity: classifying risks with a taxonomy. This task is routine in impact assessments, model cards, or risk registers (Mitchell et al. 2019; Gebru et al. 2021; Moss et al. 2021; Raji et al. 2020). Using the same professional-background matching criteria as in the scenario-validation study, we recruited 26 workers from five professional backgrounds: software engineering, data science, design, healthcare, and finance. All participants con-

firmed daily AI use and medium-to-high awareness of AI agents and passed three comprehension checks on the definition of AI agents before proceeding. Each participant classified five AI agent risk scenarios drawn at random from a held-out pool not used during taxonomy construction. The pool combined LLM-generated scenarios with documented incidents from public AI incident databases, so that participants saw both anticipated risks and real failures. Each scenario included its workplace context: the risk, industry, job title, and corresponding job task. Participants classified every scenario using three taxonomies: (A) our *Workplace AI Agent Risk Taxonomy* with 15 categories and 44 sub-categories; (B) the *Generative AI Risk Taxonomy* (Li et al. 2025b) with 12 categories and 48 sub-categories; and (C) the *AI Risk Repository Domain Taxonomy* (Slattery et al. 2025) with 7 categories and 23 sub-categories. We presented the taxonomies in a randomized order under the neutral labels Framework 1, Framework 2, and Framework 3 to reduce name recognition bias. For each scenario and taxonomy, participants selected the best fitting category or chose “no good fit” when no category represented the risk. This option allowed us to distinguish between difficulty selecting among relevant categories and the absence of a suitable category. Participants then rated the ease of using each taxonomy on a seven-point scale and provided a brief written reflection on their experience using different taxonomies for risk classification. We report the findings in §4.

4 Main Findings

We classified our 8,356 risk scenarios into the taxonomy and examined how they distribute across taxonomy categories, severity levels, deployment modes, and framework layers (Figure 3). Our main findings are provided below.

Finding 1: Augmentation is not inherently safe; it introduces a slow and largely invisible risk of deskilling. AI augmentation, where the agent assists rather than replaces the human, is often treated as the more responsible deployment choice (Wang et al. 2024a; Shao et al. 2025b). Our findings complicate this assumption. *Capability Erosion* is the second most common risk category in the entire dataset, representing 21.3% (1,782) of all scenarios, and 97.0% of these risks arise from augmentation rather than automation, the most one-sided augmentation profile in the taxonomy. Within it, the dominant sub-category is not the loss of a manual skill but the loss of oversight itself: *Decision and oversight* in *Capability Erosion* alone accounts for 1,420 scenarios (17.0% of the corpus), describing workers who stop questioning, checking, or independently judging what the agent produces. The pattern is consistent with earlier research on human factors, which has long warned that the more reliable a system is, the more people stop practicing the skills they would need if it ever failed (Bainbridge 1983). What makes modern AI agents different is that they are proactive rather than passive: they suggest next steps, flag priorities, and take initiative, which accelerates the rate at which workers hand over their professional judgment to the system. The severity profile confirms that this is not a dramatic or sudden failure: 79% of capability erosion risks

are classified as “limited” severity, and only $\leq 0.1\%$ as “critical”. The risk is real precisely because it is gradual, cumulative, and easy to miss until it is too late to reverse.

Finding 2: Erroneous agent action is both the most common and the most severe risk, and it is not purely a technical problem. *Erroneous Actions* under *Agent Execution Risks*, which covers situations where an agent mishandles data, misreads a situation, advises incorrectly, or executes a task wrongly, is the largest category in the taxonomy at 30.6% (2,558) of all scenarios, and it also carries the highest number of severe risks: 88% are rated “high” or “critical” (2,183 “high”). This might seem like a straightforward technical issue: the model hallucinates, or its training data is wrong. But our analysis shows that the failure happens between the agent and the human, not inside the model itself. Fabrication is rare, accounting for only 57 scenarios (0.7% of the corpus, filed under *Defective Outputs* in *Agent Execution Risks*). The dominant sub-types are instead *erroneous interpretation*, where the agent misreads data, context, or intent (1,158 scenarios, 13.9%), and *erroneous recommendations*, where it advises incorrectly on that basis (946 scenarios, 11.3%). Moreover, 69.5% of these risks arise under augmentation, meaning a human is present and acting on the flawed output. The agent produces an answer that is technically consistent with its inputs but wrong for the situation; the human acts on it because the agent communicated confidence. Improving model accuracy alone will therefore not solve the problem.

Finding 3: Automation concentrates risk in the organization; augmentation concentrates it in the worker, and each requires different governance. The capability erosion and agent error risks above are concentrated in different deployment modes, and a third pattern only becomes visible when we consider automation. When agents fully automate a task, removing humans from the loop, the dominant risks are organizational and institutional: *Financial Losses* (485 risks, 90.9% from automation), *Operational Failures* (744 risks, 83.5%), and *Reputational Risks* (518 risks, 71.8%). The starkest case is *Employment Displacement*, which is 98.8% automation-bound and 92.7% high or critical severity, the most one-sided and most severe category in the taxonomy. These failures tend to be visible, fast, and expensive: legal liability, operational downtime, or public relations damage from an agent acting without oversight. When agents augment rather than automate, the risks land on the individual worker rather than the organization, manifesting as the slow erosion of professional agency described in Finding 1. This structural divide has a direct governance implication: a single approach cannot cover both deployment modes; we translate this into concrete safeguards in the Discussion.

Finding 4: The Workplace AI Agent Risk Taxonomy outperforms existing frameworks in practical usability. In our usability study, the Workplace AI Agent Risk Taxonomy was the clearest of the three taxonomies for classifying workplace AI agent risks. Across 26 workers classifying 130 risks, the Workplace AI Agent Risk Taxonomy achieved a higher ease-of-use rating than the GenAI

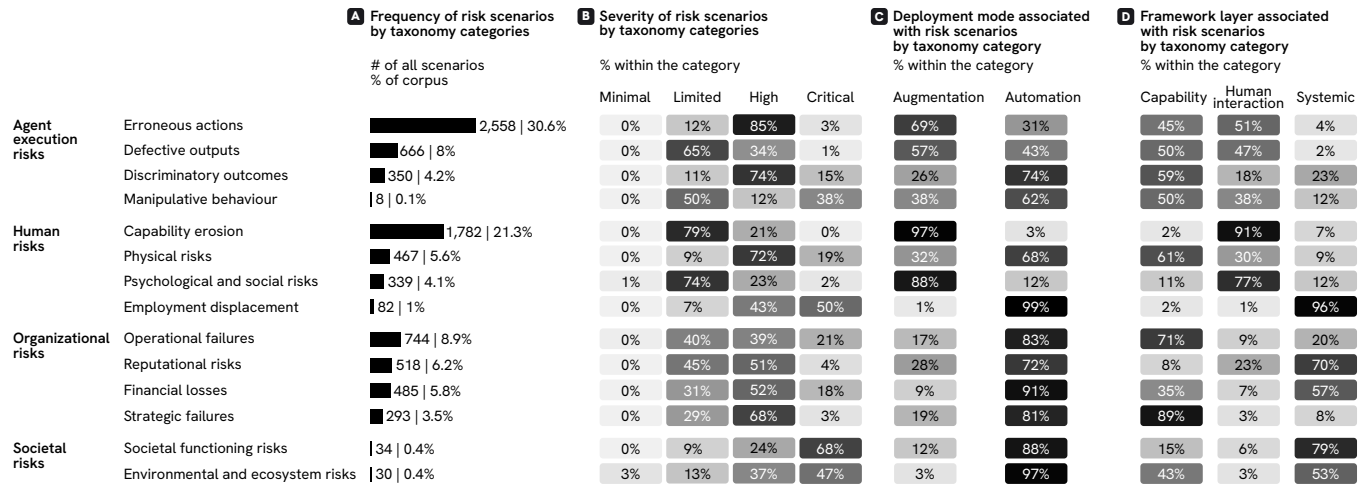


Figure 3: **Distribution of 8,356 workplace AI agent risk scenarios across the taxonomy categories.** (A) Number of scenarios in each category, with its share of the full corpus; categories are ordered by frequency, under 4 themes. (B) Severity composition of each category. (C) Deployment mode: whether the agent assists a worker (augmentation) or replaces one (automation). (D) Framework layer at which the risk arises: technical capability, human interaction, or systemic impact. In panels (B)–(D), cells are row percentages and shading is proportional to the value. *Geopolitical and Security Risks* is omitted because no generated scenario was assigned to it; this category and its sub-categories were formed from incident-based risk scenarios rather than generated ones.

Risk Taxonomy (Li et al. 2025b) (mean 5.75 vs. 5.46 on a 1–7 scale; parametric: $t(129) = 2.23$, $p = .027$, Cohen’s $d = 0.20$; non-parametric: Wilcoxon $W = 1300.5$, $p = .026$). A Friedman test across all three frameworks was also significant ($\chi^2(2) = 6.33$, $p = .042$). Our taxonomy was preferred in 64% of decided pairwise comparisons against the GenAI Risk Taxonomy (54 wins vs. 30, 46 ties) and achieved the highest classification success rate (97.7% vs. 96.2% and 96.9%). The MIT AI Risk Repository’s domain taxonomy (Slattery et al. 2025) was excluded from the pairwise comparison because categories that should have been distinct collapsed into one: 52% of all classifications used a single catch-all category (“*AI system safety, failures, and limitations*”), making the taxonomy look easy to use even though the categories did not match the risks well.

Qualitative reflections attributed the advantage to workplace specificity: our taxonomy names workplace consequences rather than AI mechanisms. A software engineer noted: “*Workplace taxonomy went above and beyond and provided the specific result of the misinformation, which was lowered efficiency*” (18–29, Male, Software Engineer). A family medicine doctor found that it “*clearly captured the patient safety consequences of prescribing the wrong medication; the GenAI Risk Taxonomy focused more on the data/privacy aspect, while the AI Risk Repository’s Domain Taxonomy emphasized the technical AI failure rather than the medical risk to patients*” (30–39, Female, Family Medicine Doctor). A financial analyst highlighted that “*the risk is about workflow disruption and productivity loss; the GenAI Risk Taxonomy was harder because ‘Financial & Business’ only indirectly captures productivity loss*” (18–29, Female, Financial Analyst). Full results and participant quotes are reported in Supplementary Material M.

5 Discussion

We first discuss how our findings extend theory on risk of AI agents (§5.1), then translate them into practical guidance for workplace governance (§5.2), and finally examine the study’s limitations and directions for future research (§5.3).

5.1 Theoretical Implications

Make agentic decomposition explicit. Rather than treating an AI system as a black box, our framework models risk as emerging from the properties of three core components (agents, goals, environments) and from the interactions between them and humans. This distinction matters because interaction-driven risks can arise even when no single component is malfunctioning, a pattern our empirical findings confirm. This extends work focused primarily on agent–agent communication (Hammond et al. 2025) by showing that goal and environment mediation are equally important risk pathways. Further, the framework decomposes human interaction into five distinct relationship types (agent–human, human–human, human–environment, etc.) rather than treating it as a single category. This enables more precise theorizing about how accountability drifts, how reliance develops, and how distributed agency makes it difficult for workers to contest outcomes that affect them.

Connect framework-level concepts to empirical workplace structure. By grounding the framework in a 15-category taxonomy built from real job tasks, this paper provides something that prior work largely lacks: a taxonomy that operates at the level of specific workflows rather than broad societal patterns. The severity profile of *Erroneous Actions* by agents, and the dominance of *Capability erosion* risks, together suggest that workplace risks are driven less

by model failures than by failures in how agents and humans communicate intent, interpret outputs, and share responsibility. That both leading categories are majority-augmentation reinforces the point: the risk concentrates where a human remains in the loop, not where the agent runs alone. This implies that “capability” in agentic systems is not a property of the model alone, but formed by the agent, the human, and the organizational context in which they interact.

5.2 Practical Implications

Beyond showing that our taxonomy is statistically distinct and practically usable (§4, Finding 4), our findings point to three concrete shifts in how organizations should approach AI agent governance.

Focus risk assessment on the human-agent boundary.

The highest-frequency and highest-severity risks in our dataset consistently cluster at the point where humans interpret and act on agent outputs. Governance efforts should therefore not stop at auditing model performance. Assurance processes should also evaluate how uncertainty is communicated, whether workers have clear pathways to question or override agent recommendations, and whether organizational goals and proxy metrics create incentives that the agent system could exploit in harmful ways.

Apply layered controls that match the framework layers.

At technical capability layer, organizations should test how components fail under stress: partial observability simulations, proxy metric sensitivity checks, and multi-agent coordination tests. At human interaction layer, safeguards should address human behavior: role clarity training, contestability channels that allow workers to challenge agent decisions, and structured interventions against over-reliance. At systemic impact layer, monitoring should track slower systemic effects such as deskilling trajectories, shifts in bargaining power, and the adoption dynamics that can normalize unsafe practices across an industry.

Govern automation and augmentation separately. Automated deployments concentrate risk in institutional outcomes: legal exposure, operational failure, reputational damage. These require hard fail-safes, clear audit trails, and well-defined incident response procedures. Augmented deployments concentrate risk in the individual worker: skill erosion, professional judgment displacement, and workplace stress. These require socio-technical safeguards: protected human override, regular skill assessments, and workflow redesigns that ensure workers retain meaningful agency rather than simply approving agent recommendations. The taxonomy itself can serve as a practical tool here, used as a living risk register that is updated iteratively as agent capabilities and organizational practices evolve.

5.3 Limitations and Future Work

Although the scenarios were generated by an LLM, we evaluated them with workers and LLM judges and grounded them in real incident data and a live multi-agent red-teaming study. This methodology is well grounded in established risk foresight and validation practice (Perez et al. 2022; Ganguli et al. 2022; Fröhling et al. 2026; Buçinca et al. 2023; Rao

et al. 2025; Herdel et al. 2025; Wang et al. 2024b), but future work could test whether the same risk patterns emerge under different generation conditions, prompting techniques, and agentic approaches. The divergence in novelty scores between workers and the LLM judge also deserves attention. Workers found the scenarios moderately novel, offering genuine discovery value; the LLM was far stricter, likely recognizing training data patterns. This raises a question on whether there categories of workplace AI agent risk that are genuinely novel and therefore harder to surface through LLM-driven foresight?

The automation vs. augmentation distinction simplifies a continuum. Many real deployments are hybrid: tasks may be partially automated, intermittently supervised, or redistributed across roles in ways that do not fit cleanly into either category. Similarly, O*NET reflects a U.S.-centric occupational framing and may not capture informal work, global labor conditions, or sector-specific practices. Finally, the taxonomy represents a snapshot. Because the scenarios are LLM-generated, the category frequencies and distributional analyses describe only the composition of our risk corpus; they should not be interpreted as estimates of how often these risks occur in real workplaces. These patterns may also reflect the selected job tasks, prompt design, and model used. As agent capabilities evolve, both the distribution of risks and the taxonomy structure will need updating. Future work should triangulate these LLM-generated scenarios with field studies and incident reports, test robustness across models and prompt designs, enrich the occupational grounding with organizational context such as staffing constraints and safety culture, and conduct participatory validation with workers and domain stakeholders to ensure the taxonomy supports fairness-oriented governance in practice.

6 Conclusion

As AI agents become embedded in everyday work, organizations need practical ways to anticipate the risks they may introduce. This paper contributes: a multi-layer framework that models how risks emerge from interactions between agents, goals, environments, and workers; a corpus of 8,356 risk scenarios grounded in 2,078 real job tasks across 209 roles; and a 15-category taxonomy that organizes those risks under 4 themes. Across our corpus, the main pattern is that workplace risks often emerge from breakdowns at the human-agent boundary rather than from isolated model failures. More scenarios were classified as *Erroneous Actions* of agents than under any other category (30.6%), and 88.2% of these scenarios were rated High or Critical. *Capability Erosion* was the second most frequently assigned category (21.3%). Scenarios in both categories were predominantly associated with augmentation, showing that keeping a worker in the loop does not eliminate risk. These patterns point to the same conclusion: the central challenge is not only building more capable agents, but also designing safer ways for humans to work alongside them.

Acknowledgements

This work was supported by UK Research and Innovation (grant number EP/S023356/1), in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org).

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*.
- Altmann, P.; Schönberger, J.; Illium, S.; Zorn, M.; Ritz, F.; Haider, T.; Burton, S.; and Gabor, T. 2024. Emergence in Multi-Agent Systems: A Safety Perspective. *arXiv:2408.04514*.
- Anwar, U.; Saparov, A.; Rando, J.; Paleka, D.; Turpin, M.; Hase, P.; Lubana, E. S.; Jenner, E.; Casper, S.; Sourbut, O.; Edelman, B. L.; Zhang, Z.; Günther, M.; Korinek, A.; Hernandez-Orallo, J.; Hammond, L.; Bigelow, E.; Pan, A.; Langosco, L.; Korbak, T.; Zhang, H.; Zhong, R.; hÉigeartaigh, S. O.; Recchia, G.; Corsi, G.; Chan, A.; Anderljung, M.; Edwards, L.; Petrov, A.; de Witt, C. S.; Motwan, S. R.; Bengio, Y.; Chen, D.; Torr, P. H. S.; Albanie, S.; Maharaj, T.; Foerster, J.; Tramer, F.; He, H.; Kasirzadeh, A.; Choi, Y.; and Krueger, D. 2024. Foundational Challenges in Assuring Alignment and Safety of Large Language Models. *arXiv:2404.09932*.
- Autor, D. H. 2015. Why are there still so many jobs? The history and future of workplace automation. *Journal of economic perspectives*, 29(3): 3–30.
- Bainbridge, L. 1983. Ironies of automation. *Automatica*, 19(6): 775–779.
- Bogucka, E.; Šćepanović, S.; and Quercia, D. 2024. Atlas of AI Risks: Enhancing Public Understanding of AI Risks. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 12(1): 33–43.
- Bryan, P.; Severi, G.; de Gruyter, J.; Jones, D.; Bullwinkel, B.; Minnich, A.; Chawla, S.; Lopez, G.; Pouliot, M.; Fournay, A.; Maxwell, W.; Pratt, K.; Qi, S.; Chikanov, N.; Lutz, R.; Dheekonda, R. S. R.; Jagdagdorj, B.-E.; Kim, E.; Song, J.; Hines, K.; Jones, D.; Lundeen, R.; Vaughan, S.; Westerhoff, V.; Zunger, Y.; Kawaguchi, C.; Russinovich, M.; and Kumar, R. S. S. 2025. Taxonomy of Failure Mode in Agentic AI Systems. Technical report, Microsoft. Microsoft AI Red Team (AIRT) Whitepaper.
- Buçinca, Z.; Pham, C. M.; Jakesch, M.; Ribeiro, M. T.; Olteanu, A.; and Amershi, S. 2023. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms. *arXiv:2306.03280*.
- Chan, A.; Salganik, R.; Markelius, A.; Pang, C.; Rajkumar, N.; Krashennikov, 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, FAccT '23*, 651–666. ACM.
- Chiang, C.-H.; and yi Lee, H. 2023. Can Large Language Models Be an Alternative to Human Evaluations? *arXiv:2305.01937*.
- Critch, A.; and Russell, S. 2023. TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI. *arXiv:2306.06924*.
- Eloundou, T.; Manning, S.; Mishkin, P.; and Rock, D. 2024. GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702): 1306–1308.
- Fabris, A.; Messina, S.; Silvello, G.; and Susto, G. A. 2022. Tackling Documentation Debt: A Survey on Algorithmic Fairness Datasets. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO '22*. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394772.
- Felten, E.; Raj, M.; and Seamans, R. 2023. How will language modelers like ChatGPT affect occupations and industries? *arXiv preprint arXiv:2303.01157*.
- Fergusson, G.; Fitzgerald, C.; Frascella, C.; Iorio, M.; McBrien, T.; Schroeder, C.; Winters, B.; and Zhou, E. 2023. Generating Harms: Generative AI's Impact & Path Forward. Technical report, Electronic Privacy Information Center (EPIC).
- Fröhling, L.; Giaconia, A.; Bogucka, E. P.; and Quercia, D. 2026. Agent-Supported Foresight for AI Systemic Risks: AI Agents for Breadth, Experts for Judgment. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–48.
- Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.
- Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumé III, H.; and Crawford, K. 2021. Datasheets for Datasets. *Communications of the ACM*, 64(12): 86–92.
- Ghosh, S.; Simkin, B.; Shiarlis, K.; Nandi, S.; Zhao, D.; Fiedler, M.; Bazinska, J.; Pope, N.; Prabhu, R.; Rohrer, D.; Demoret, M.; and Richardson, B. 2025. A Safety and Security Framework for Real-World Agentic Systems. *arXiv:2511.21990*.
- Google. 2025. Gemini models. <https://ai.google.dev/gemini-api/docs/models>.
- Gridach, M.; Nanavati, J.; Abidine, K. Z. E.; Mendes, L.; and Mack, C. 2025. Agentic AI for Scientific Discovery: A Survey of Progress, Challenges, and Future Directions. *arXiv:2503.08979*.
- Hammond, L.; Chan, A.; Clifton, J.; Hoelscher-Obermaier, J.; Khan, A.; McLean, E.; Smith, C.; Barfuss, W.; Foerster, J.; Gavenčiak, T.; Han, T. A.; Hughes, E.; Kovařík, V.; Kulveit, J.; Leibo, J. Z.; Oesterheld, C.; de Witt, C. S.; Shah, N.; Wellman, M.; Bova, P.; Cimpanu, T.; Ezell, C.;

- Feuillade-Montixi, Q.; Franklin, M.; Kran, E.; Krawczuk, I.; Lamparth, M.; Lauffer, N.; Meinke, A.; Motwani, S.; Reuel, A.; Conitzer, V.; Dennis, M.; Gabriel, I.; Gleave, A.; Hadfield, G.; Haghtalab, N.; Kasirzadeh, A.; Krier, S.; Larson, K.; Lehman, J.; Parkes, D. C.; Piliouras, G.; and Rahwan, I. 2025. Multi-Agent Risks from Advanced AI. arXiv:2502.14143.
- Herdel, V.; Sćepanović, S.; Bogucka, E.; and Quercia, D. 2025. *ExploreGen: Large Language Models for Envisioning the Uses and Risks of AI Technologies*, 584–596. AAAI Press.
- Hoffman, M.; and Frase, H. 2023. Adding Structure to AI Harm: An Introduction to CSET’s AI Harm Framework. Technical report, Center for Security and Emerging Technology (CSET).
- Hu, T.; and Zhou, X.-H. 2024. Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions. arXiv:2404.09135.
- 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*, FAccT ’24, 359–376. ACM.
- IBM. 2024. What is Agentic AI? <https://www.ibm.com/think/topics/agentic-ai>.
- Khoo, S.; Foo, J.; and Lee, R. K.-W. 2025. With Great Capabilities Come Great Responsibilities: Introducing the Agentic Risk & Capability Framework for Governing Agentic AI Systems. arXiv:2512.22211.
- Kong, D.; Lin, S.; Xu, Z.; Wang, Z.; Li, M.; Li, Y.; Zhang, Y.; Peng, H.; Sha, Z.; Li, Y.; Lin, C.; Wang, X.; Liu, X.; Zhang, N.; Chen, C.; Khan, M. K.; and Han, M. 2025. A Survey of LLM-Driven AI Agent Communication: Protocols, Security Risks, and Defense Countermeasures. arXiv:2506.19676.
- Lee, H.-P.; Yang, Y.-J.; von Davier, T. S.; Forlizzi, J.; and Das, S. 2024. Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. arXiv:2310.07879.
- Leslie, D. 2019. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector.
- Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; Shu, K.; Cheng, L.; and Liu, H. 2025a. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. arXiv:2411.16594.
- Li, H.; Dong, Q.; Chen, J.; Su, H.; Zhou, Y.; Ai, Q.; Ye, Z.; and Liu, Y. 2024. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. arXiv:2412.05579.
- Li, M.; Bickersteth, W.; Tang, N.; Hong, J.; Cranor, L.; Shen, H.; and Heidari, H. 2025b. A Closer Look at the Existing Risks of Generative AI: Mapping the Who, What, and How of Real-World Incidents. arXiv:2505.22073.
- Liu, B.; Li, X.; Zhang, J.; Wang, J.; He, T.; Hong, S.; Liu, H.; Zhang, S.; Song, K.; Zhu, K.; Cheng, Y.; Wang, S.; Wang, X.; Luo, Y.; Jin, H.; Zhang, P.; Liu, O.; Chen, J.; Zhang, H.; Yu, Z.; Shi, H.; Li, B.; Wu, D.; Teng, F.; Jia, X.; Xu, J.; Xiang, J.; Lin, Y.; Liu, T.; Liu, T.; Su, Y.; Sun, H.; Berseth, G.; Nie, J.; Foster, I.; Ward, L.; Wu, Q.; Gu, Y.; Zhuge, M.; Liang, X.; Tang, X.; Wang, H.; You, J.; Wang, C.; Pei, J.; Yang, Q.; Qi, X.; and Wu, C. 2025. Advances and Challenges in Foundation Agents: From Brain-Inspired Intelligence to Evolutionary, Collaborative, and Safe Systems. arXiv:2504.01990.
- Macrae, C. 2022. Learning from the Failure of Autonomous and Intelligent Systems: Accidents, Safety, and Sociotechnical Sources of Risk. *Risk Analysis*, 42(9): 1999–2025.
- McGregor, S. 2021. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 15458–15463.
- Mehrabani, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6).
- Microsoft. 2023. Types of harm - Azure Application Architecture Guide.
- Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* ’19, 220–229. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.
- Mohney, M. 2025. What AI Means for the Future of Work.
- Moss, E.; Watkins, E. A.; Singh, R.; Elish, M. C.; and Metcalf, J. 2021. Assembling accountability: algorithmic impact assessment for the public interest. *Available at SSRN* 3877437.
- National Center for O*NET Development. 2025. O*NET Database. O*NET Resource Center. <https://www.onetcenter.org/database.html>.
- OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
- Paeth, K.; Atherton, D.; Pittaras, N.; Frase, H.; and McGregor, S. 2025. Lessons for editors of AI incidents from the AI incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 28946–28953.
- Panickssery, A.; Bowman, S. R.; and Feng, S. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37: 68772–68802.
- Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3419–3448.
- Raisch, S.; and Krakowski, S. 2021. Artificial intelligence and management: The automation–augmentation paradox. *Academy of management review*, 46(1): 192–210.
- Raji, I. D.; Kumar, I. E.; Horowitz, A.; and Selbst, A. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness Accountability and Transparency*, FAccT ’22, 959–972. ACM.

- 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.
- Ranjan, R.; Gupta, S.; and Singh, S. N. 2025. Fairness in Agentic AI: A Unified Framework for Ethical and Equitable Multi-Agent Systems. arXiv:2502.07254.
- Rao, P. S. B.; Šćepanović, S.; Zhou, K.; Bogucka, E. P.; and Quercia, D. 2025. RiskRAG: A Data-Driven Solution for Improved AI Model Risk Reporting. 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.
- Richards, I.; Benn, C.; and Zilka, M. 2025. From Incidents to Insights: Patterns of Responsibility following AI Harms. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 151–169.
- Russell, S.; and Norvig, P. 2009. *Artificial Intelligence: A Modern Approach*. USA: Prentice Hall Press, 3rd edition. ISBN 0136042597.
- Sapkota, R.; Roumeliotis, K. I.; and Karkee, M. 2026. AI Agents vs. Agentic AI: A Conceptual taxonomy, applications and challenges. *Information Fusion*, 126: 103599.
- Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 59–68. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.
- Shao, Y.; Huang, C.; Song, Y.; Wang, M.; Song, Y. H.; and Shao, R. 2025a. Using Augmentation-Based AI Tool at Work: A Daily Investigation of Learning-Based Benefit and Challenge. *Journal of Management*, 51(8): 3352–3390.
- Shao, Y.; Zope, H.; Jiang, Y.; Pei, J.; Nguyen, D.; Brynjolfsson, E.; and Yang, D. 2025b. Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the U.S. Workforce. arXiv:2506.06576.
- Shapira, N.; Wendler, C.; Yen, A.; Sarti, G.; Pal, K.; Floody, O.; Belfki, A.; Loftus, A.; Jannali, A. R.; Prakash, N.; Cui, J.; Rogers, G.; Brinkmann, J.; Rager, C.; Zur, A.; Ripa, M.; Sankaranarayanan, A.; Atkinson, D.; Gandikota, R.; Fiotto-Kaufman, J.; Hwang, E.; Orgad, H.; Sahil, P. S.; Taglicht, N.; Shabtay, T.; Ambus, A.; Alon, N.; Oron, S.; Gordon-Tapiero, A.; Kaplan, Y.; Schwartz, V.; Shaham, T. R.; Riedl, C.; Mirsky, R.; Sap, M.; Manheim, D.; Ullman, T.; and Bau, D. 2026. Agents of Chaos. arXiv:2602.20021.
- Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Roshtamzadeh, N.; Nicholas, P.; Yilla, N.; Gallegos, J.; Smart, A.; Garcia, E.; and Virk, G. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. arXiv:2210.05791.
- Si, C.; Yang, D.; and Hashimoto, T. 2024. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. arXiv:2409.04109.
- Slattery, P.; Saeri, A. K.; Grundy, E. A. C.; Graham, J.; Noetel, M.; Uuk, R.; Dao, J.; Pour, S.; Casper, S.; and Thompson, N. 2025. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. arXiv:2408.12622.
- Steenstra, I.; and Bickmore, T. W. 2025. A Risk Ontology for Evaluating AI-Powered Psychotherapy Virtual Agents. In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA '25. New York, NY, USA: Association for Computing Machinery.
- Turri, V.; and Dzombak, R. 2023. Why we need to know more: Exploring the state of AI incident documentation practices. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 576–583.
- Wang, R.; Cheng, R.; Ford, D.; and Zimmermann, T. 2024a. Investigating and designing for trust in ai-powered code generation tools. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1475–1493.
- Wang, Z. J.; Kulkarni, C.; Wilcox, L.; Terry, M.; and Madaio, M. 2024b. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, 1–40. ACM.
- 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.; and Isaac, W. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. arXiv:2310.11986.
- Wooldridge, M. 2009. *An Introduction to MultiAgent Systems*. Wiley Publishing, 2nd edition. ISBN 0470519460.
- Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2): 121101.
- Yin, F.; Laban, P.; Peng, X.; Zhou, Y.; Mao, Y.; Vats, V.; Ross, L.; Agarwal, D.; Xiong, C.; and Wu, C.-S. 2025. Bingoguard: Llm content moderation tools with risk levels. arXiv preprint arXiv:2503.06550.
- Zeng, Y.; Klyman, K.; Zhou, A.; Yang, Y.; Pan, M.; Jia, R.; Song, D.; Liang, P.; and Li, B. 2024. AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies. arXiv:2406.17864.
- 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.
- Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Supplementary Material: Unaccountable Delegation, Fading Skills: Mapping the Risks of Workplace AI Agents

Gabriele La Malfa¹, Lakmal Meegahapola², Edyta Bogucka², Jie M. Zhang¹, Michael Luck³,
Elizabeth Black¹, Daniele Quercia^{2,4}

¹King's College London, UK

²Nokia Bell Labs, Cambridge, UK

³University of Sussex, UK

⁴Politecnico di Torino, Italy

{gabriele.la_malfa, jie.zhang, elizabeth.black}@kcl.ac.uk, michael.luck@sussex.ac.uk,
{lakmal.meegahapola, edyta.bogucka, daniele.quercia}@nokia-bell-labs.com

A. Definitions

For conceptual clarity, we report hereafter the definitions of AI agents, agentic AI systems/multi-agent AI systems (AAIS), and multi-agent systems (MAS) with reference to relevant literature.

AI agents: Autonomous entities capable of perceiving their environment, interacting with other agents and taking actions to achieve goals with varying degrees of independence (Sapkota, Roumeliotis, and Karkee 2026).

Agentic AI systems/multi-agent AI systems: Systems comprising AI agents and potentially an environment (physical or virtual, i.e., tools) (IBM 2024). An agentic AI system could work with one or more agents, whereas a multi-agent AI system could work with two or more agents.

Multi-Agent Systems A theoretical framework defined as a system of computational entities, named agents, interacting (cooperate or compete) to solve complex problems (Wooldridge 2009; Russell and Norvig 2009; Shoham and Leyton-Brown 2008).

Throughout the paper, we use the term AI agents to refer to the technology having the characteristics described above in the definition. Agentic AI systems include both single-agent and multi-agent AI systems. In our conceptualization, we adopt components and interaction patterns commonly discussed in multi-agent systems because this framing is strictly more general: it subsumes the single-agent case as well, and therefore, ensures that any analysis, taxonomy, or guidance we develop applies not only to systems with multiple interacting agents, but also to agentic systems that operate with just one AI agent.

B. Validation of the Multi-Layer Risk Analysis Framework Through Expert Feedback

The multi-layer risk analysis framework presented in this paper was validated through individual three interviews with experts in multi-agent systems and responsible AI. Each interview lasted approximately 30 minutes. The primary ob-

jective was to evaluate whether the components, component interactions (Level 1), human interaction dynamics (Level 2), and the systemic layer (Level 3) are coherent and comprehensive from a multi-agent perspective. This section outlines the methodology and summarizes the results.

B1. Methodology

Each expert participated in a structured 30-minute interview consisting of three parts: (i) a 10-minute overview of the framework, including the transition from the baseline model by Weidinger et al. (2023) to our proposed structure emphasizing components and their interactions; (ii) a 5-minute Q&A segment for technical clarifications; and (iii) a 10-minute feedback session focused on evaluating whether the framework captures the key components and interaction types from which risks might arise. During the feedback session, experts were invited to identify any missing elements or propose adjustments relevant to AI agents or multi-agent AI systems. Additional input on the design, inter-component relationships, human-agent interactions, and the systemic layer was also encouraged.

B2. Results

All experts provided constructive feedback, which was incorporated to refine the framework. They agreed that the structure is both accurate and comprehensive from a multi-agent systems perspective, particularly in the context of responsible AI deployment. Experts affirmed that the identified interactions, at both the technical and human levels, constitute a representative and meaningful set of possible risk pathways. The interviewed experts noticed that the framework represents a minimal, baseline set of interactions. Although higher-order interaction loops exist (e.g., agent-to-agent-to-human), modeling them would simply require iterating the foundational schema we have provided in the framework.

C. Validation of LLM-Generated Risks Using Workers and LLM-as-a-Judge

To ensure the technical accuracy, practical utility, and creative diversity of the 8,356 generated risk scenarios, we employed a dual-layer validation. This approach involves human workers with the scalability of an LLM-as-a-judge (Zheng et al. 2023; Li et al. 2024, 2025a). This section describes the evaluation approach and summarizes the results.

C1. Methodology

Workers Evaluation Methodology. To ground our findings in real-world professional contexts, we recruited 45 workers via the Prolific platform. We filtered for participants with a professional background close to the industry and the job of the risks evaluated. For example, a risk scenario from the *Healthcare Professional* industry and *Acute Care Nurses* job categories (O*NET Database) could be assigned only to an evaluator whose profession is nurse. The cohort represented a wide array of jobs: Nurses, Software-related roles, Data analysts, Accountants, Finance roles, Teaching roles, Designers, Medicine, Legal, and HR. Each domain expert was assigned 10 risk scenarios specifically selected to match their O*NET-defined area of expertise. This resulted in 450 unique human-validated risk scenarios. While the sample size aims to achieve a margin of error of 5% relative to the 8,356 corpus under standard population sampling, we acknowledge that this represents a convenience sample rather than a strictly randomized stratification. Workers used a 5-point Likert scale to assess 10 metrics (see Table 1).

Table 2 summarizes participants’ demographics and expertise.

LLM-as-a-Judge Evaluation Methodology. All 8,356 risks were evaluated using the `gemini-3-flash` model. Using a distinct model for evaluation helps mitigate self-preference bias due to differences in training data (Panickssery, Bowman, and Feng 2024). The model assessed each risk using the task and context information from the O*NET database, applying a 5-point Likert scale to assess 10 metrics (see Table 1). To assess the robustness of these evaluations, we conducted an inter-rater reliability (IRR) analysis using `gpt-4o` as a control model. A random sample of 370 risks was re-evaluated using the same prompts. This sample size is determined to be representative of the total population of 8,356 risks (equivalent to a 95% confidence level in traditional sampling theory). Agreement between the models was measured using two metrics:

- **Exact Agreement:** Agreement on the precise 5-point Likert scale rating.
- **Sentiment Agreement:** Agreement on general sentiment, by binning scores into three categories, Positive (Agree/Strongly Agree), Neutral, and Negative (Disagree/Strongly Disagree), to reduce the effect of calibration differences (Wang et al. 2023).

C2. Results

Both human and LLM-as-a-judge evaluation metrics can be synthesized into three dimensions:

- **Validity:** Evaluated via plausibility and connection, ensuring that risks are not hallucinated (Chiang and yi Lee 2023).
- **Utility:** Measured through usefulness, actionability, specificity, and level of detail, reflecting how informative and applicable the risks are for stakeholders such as auditors.
- **Novelty:** Assessed via originality, complexity, and rarity (Bubeck et al. 2023).

Table 3 reports the average scores and standard deviation of the domain expert evaluators for each metric. For the validity dimension, plausibility and connection received high average scores (4.03/5 and 3.94/5, respectively). The novelty dimension received medium scores, 3.00/5 for rarity and 2.96/5 for originality, suggesting that the generated risks present some level of novelty and originality.

Table 4 reports the average scores and agreement levels between the models for each metric. For the validity dimension, plausibility and connection received very high average scores (4.81/5 and 4.60/5, respectively). Although Exact Agreement between models was modest (13.8% and 51.4%), the Sentiment Agreement was nearly perfect (98.9% and 98.1%), indicating that both models concurred on the risks’ core validity but differed in rating intensity. The evaluator model (`gemini-3-flash`) tended to assign more extreme scores, while `gpt-4o` was more moderate. In contrast, the novelty dimension received low scores, 1.28/5 for rarity and 1.82/5 for originality—suggesting that the generated risks mostly reflect established, well-known concerns. However, the Sentiment Agreement for these dimensions was still strong (79.7%), indicating general alignment between models. Utility-related scores were moderate: usefulness (3.38/5), actionability (3.37/5), and specificity (2.78/5). These results suggest that while the risks were relevant and moderately actionable, they were not highly tailored to job-specific contexts.

While both evaluation methods confirm the validity of the generated risks, the LLM-as-a-judge shows a higher degree of optimism regarding plausibility than the human experts. Conversely, human practitioners attribute greater practical utility to the scenarios, rating them as more actionable and specific than the LLM’s conservative estimates. This divergence is even more evident in the novelty dimension, where risks perceived as original by professionals are categorized as derivative by the model’s extensive pattern recognition.

D. Taxonomy Refinement: Incident Databases Classification Details

Classification of risks and taxonomy refinement method

We selected three prominent AI incident databases: the AI Incident Database, AIAAIC, and the OECD Database. From each database, we extracted a sub-sample of incidents for

Dimension	Evaluation Metric	Description
Validity	Plausibility	Considering the Industry, Job, and Task, the Risk feels like something that could actually happen in reality, not just possible.
	Connection	There is a direct, logical chain between the Task and the Risk such that performing the Task could realistically lead to the negative outcome described in the Risk.
Utility	Usefulness	The Risk provides a high-quality signal for policy-makers or auditors to justify oversight or trigger a specific audit of existing safeguards.
	Actionability	The Risk is specific enough to inform a concrete control or policy rule.
	Detail	The risk clearly identifies who is involved, how the failure occurs, and what the impact is.
	Specificity	The Risk is specific to the Industry and Job. If the Risk applies to any Industry or Job, it has low specificity.
Discovery	Complexity	Understanding and fully grasping this Risk requires industry-specific expertise.
	Novelty	The Risk describes an emergent hazard that is not widely recognized or discussed yet.
	Originality	The Risk is imaginative and identifies a non-obvious “blind spot”.
	Rarity	The Risk describes a scenario that is rarely considered in typical industry risk assessments.

Table 1: **Metrics used to validate the LLM-generated risks through workers and LLM-as-a-judge.** Three validation dimensions are captured, i.e., validity, utility and discovery. Each of them is represented by the metrics that are, in this table, exposed and defined.

Characteristic	Category	Count	Percentage (%)
Age	18-29 years	11	24.44
	30-39 years	20	44.44
	40-49 years	7	15.56
	50-59 years	6	13.33
	60 years and above	1	2.22
Sex	Female	21	46.67
	Male	24	53.33
Education	High School	5	11.11
	Bachelor’s Degree	22	48.89
	Master’s Degree	11	24.44
	PhD or higher	6	13.33
	Other	1	2.22
Computer Literacy	Medium	5	11.11
	High	16	35.56
	Very High	24	53.33
AI Expertise	Low	3	6.67
	Medium	14	31.11
	High	18	40.00
	Very High	10	22.22
AI Agents Expertise	Very Low	1	2.22
	Low	4	8.89
	Medium	12	26.67
	High	18	40.00
	Very High	10	22.22

Table 2: **Demographics of the participants in the domain expert evaluation study to validate LLM-generated risk scenarios.** Both generic attributes (age, sex and education) and technical expertise (computer literacy, AI and AI agent expertise) of the 45 participants were collected and reported in the table for the sole purpose of highlighting their background diversity and AI-related experience.

Metric	Mean Score ($N = 450$)	Std Dev
Plausibility	4.03	0.98
Connection	3.94	1.02
Usefulness	3.70	1.04
Actionability	3.62	1.08
Detail	3.51	1.14
Specificity	3.50	1.12
Complexity	3.44	1.18
Novelty	3.11	1.19
Originality	2.96	1.16
Rarity	3.00	1.16

Table 3: **Average validation scores and standard deviation for each metric used in the domain expert evaluation.** The results shown in the table are calculated across a population of 45 professionals, thus validating 450 generated risk scenarios. The mean score is measured on a 5-point Likert scale.

classification within our AI workplace risk taxonomy. This approach aimed to gather evidence from real-world scenarios and to refine the taxonomy’s categories and sub-categories with precise definitions.

For each database entry, we initially determined whether an actual incident happened and if it was workplace-related. Incidents not meeting this criterion were excluded from further analysis. An incident was classified as a workplace AI deployment risk if it involved socio-technical harm arising in a workplace due to the integration of technical AI systems with organizational objectives, workflows, and human decision-making under operational constraints.

We focused specifically on the deployment context of the AI agent, requiring it to be exclusively workplace-related. Incidents involving AI systems deployed for general public use, such as new AI-based software accessible to users,

Metric	Mean Score ($N = 8,356$)	Std Dev	Exact Agreement (5-Point Scale)	Sentiment Agreement (3-Point Bin*)
Plausibility	4.81	0.51	13.8%	98.9%
Connection	4.60	0.66	51.4%	98.1%
Usefulness	3.38	0.92	46.2%	55.4%
Actionability	3.37	0.74	52.4%	54.1%
Detail	3.16	0.92	31.1%	32.4%
Specificity	2.78	0.96	30.5%	31.6%
Complexity	2.61	0.91	29.2%	30.0%
Novelty	1.82	0.63	31.4%	47.6%
Originality	1.80	0.58	33.8%	53.5%
Rarity	1.28	0.57	13.2%	79.7%

Table 4: **Average validation scores and standard deviation for each metric used in the LLM-as-a-judge (gemini-3-flash) evaluation.** The results shown in the table are calculated across the whole risk scenario population (8,356 risks). Further agreement scores are calculated between gemini-3-flash and gpt-4o (frontier model) over a sample of 370 risks. The mean score is measured on a 5-point Likert scale.

were not considered workplace-related. Similarly, incidents resulting from AI software or hardware errors occurring in public settings were excluded unless a worker was directly involved. For example, a software malfunction in an autonomous vehicle causing a public accident is not workplace-related, whereas harm caused to a delivery person during their work is classified as workplace-related.

For each workplace-related incident, the description was matched with the most appropriate category and sub-category, according to their definitions. If no suitable categories or sub-categories were identified, additional ones were created. Also, categories and sub-categories definitions were modified for the sake of precision and clarity, if necessary.

This refinement process led to the following specific modifications of the AI workplace taxonomy:

- By classifying multiple real-world risks, we realized that clustering the taxonomy according to four main dimensions, i.e., *Agent Execution Risks*, *Human Risks*, *Organizational Risks*, and *Societal Risks*, helps its readability and interpretability.
- The original taxonomy was too vague or missed specific modern AI harms. We added and refined the following categories and sub-categories: (i) We added a sub-category for companies substituting human roles with AI, which the final taxonomy elevates to the *Employment Displacement* category (6). (ii) We broadened a sub-category involving liability strictly due to clinical failures into broader commercial and legal exposure, now *Compliance-related financial losses* (11.1).
- We streamlined two sub-categories involving data interpretation and guiding output failures, which were overlapping, moving toward a unique sub-category and definition, now *Erroneous recommendations* (2.3).
- We broadened and specialized the discrimination category, now *Discriminatory Agent Outcomes* (4), by separately capturing discriminatory representations, assessments, access, and treatment.

- We expanded the societal dimension, recognizing that AI agent deployment can lead to geopolitical and environmental risks.
- Within *Physical Risks* (8), we distinguished decision-mediated bodily harms from operation-mediated machinery hazards.

We plan to release the dataset of this classification and the annotations.

Database insights

The analysis of three distinct databases revealed notable differences in the types of incidents reported. The AI Incident Database demonstrates a high prevalence of risks associated with unfair or discriminatory algorithms, erroneous agent actions, and health concerns. In addition to these risks, the AIAAIC highlights business-related risks, particularly those affecting reputation and finances. The OECD Database, by contrast, indicates a significant presence of geopolitical risks and human capability erosion.

Collectively, these databases provide diverse perspectives on risk, reflecting a broader range of real-world harms across multiple sectors. They capture incident types and sector-specific risks that are not represented in the LLM-generated risks from the O*NET Database.

E. Prompt for Risks Generation

In this section, we report the prompt used for the risk generation as highlighted in the Methodology section.

Model: gpt-4o-mini

Temperature: 0.2

Date: November 2025

The prompt in Figure 1 was sent to the model for each job task entry from the O*NET database.

F. Workplace AI Agent Risk Taxonomy: Full Table with Frequency and Examples

In this section, in Table 5, we present a more comprehensive version of the workplace AI risk taxonomy of Figure 2 of

gpt-4o-mini

You are an expert Business Transformation Strategist specialized in AI Agent-based job automation and augmentation. Your objective is to analyze a job task provided in the [CONTEXT]. You will follow the [INSTRUCTIONS] to identify plausible risks based on the [RISK CATEGORY FRAMEWORK]. Your final output must be a single, valid JSON object structured according to the [JSON RESPONSE] template.

Your analysis must be grounded in the EU AI Act's definition of risk (including probability of harm and severity of harm) and should consider the unique characteristics of AI agents and multi-agent AI systems. Distinguish between risks arising from AUTOMATION (full job task replacement by AI agents) and AUGMENTATION (enhancing human capabilities to assist with the job task by AI agents). Core components of your task are given below.

[RISK CATEGORY FRAMEWORK]

This framework provides the specific categories for classifying risks. Do not treat humans as agents. Provide realistic risks that could emerge from AI agents and multi-agent AI systems, in the context of job tasks, think beyond generic AI risks. Consider the framework provided below.

- (1) CAPABILITY: Risks emerging from the AI agent system's technical COMPONENTS and INTERACTIONS between those components. (1a) COMPONENTS are AI agent, environment, and goals. (1b) INTERACTIONS are agent-agent, agent-environment, agent-goals (do not consider humans in the loop as agents here). An AI agent is a software with a high level of autonomy to make decisions and plan; an environment is a digital or physical space where autonomous agents perceive, act, and interact; a goal is a desired objective or outcome that agents work to achieve by making decisions.
- (2) HUMAN INTERACTIONS: Risks emerging from AI-agent system interacting with humans. This could be humans interacting with the whole AI-agent system, individual agents, environment, or goals. Consider effects such as, but not limited to, overreliance, misinterpretation, or unintended consequences.
- (3) SYSTEMIC: Risks that impact broadly the organizations, society, economy, or natural environment.
- (4) OTHER: Any plausible risks that do not fit into the categories above.

[INSTRUCTIONS]

Follow these instructions to generate the two required outputs.

(I) The General Intended AI Use

Describe the AI's intended purpose using 5 elements, each fewer than 7 words.

Purpose: The objective, in gerund form (e.g., "Analyzing financial data").

Capability: The technical ability in the format: [Action Verb] [Inference/Decision] from [Data/Entity] (e.g., "Detecting anomalies from transaction logs").

Space: The type of space in which the use took place. One of exactly "Online space", "Publicly accessible space", or "Not publicly accessible space".

AI User: The general entity deploying the AI Agent or Multi-Agent AI system.

AI Subject: The primary group of individuals most directly affected.

(II) Risks

Come up with up to 5 highly plausible and realistic AI Agent and/or Multi-Agent AI system risks as a list of objects, each with seven keys:

scenario: A description of potential risk: "[Specific parties] could be [how harmed] due to [specific reason related to the AI agent's behavior]."

severity: Assign one label from the scale below, adapted from EU AI Act principles:

"Minimal": Negligible or easily reversible impact.

"Limited": Significant but manageable impact not violating fundamental rights.

"High": Adverse impact on safety, fundamental rights, or society.

"Unacceptable / Critical": Severe, irreversible harm that violates human dignity or democratic values.

severity_reasoning: Explain why this particular "severity" was chosen for the "scenario".

risk_category: Assign one full keyword label from the [RISK CATEGORY FRAMEWORK]. Available options are

["Capability-Components-Agent", "Capability-Components-Environment", "Capability-Components-Goals", "Capability-Interactions-AgentAgent", "Capability-Interactions-AgentEnvironment", "Capability-Interactions-AgentGoals", "HumanInteractions-System", "HumanInteractions-Agent", "HumanInteractions-Environment", "HumanInteractions-Goals", "Systemic", "Other"].

risk_reasoning: Concisely explain why the scenario fits the chosen risk_category.

automation_or_augmentation: Assign one: "Automation" or "Augmentation".

automation_augmentation_reasoning: Explain the link to automation (e.g., no human oversight) or augmentation (e.g., misleading a human).

Figure 1: **Prompt for risk scenarios generation with LLM.** The prompt instantiate a persona, i.e., a business transformation strategist, to analyze job tasks from the O*NET database. The prompt enforces a two-phase reasoning process: Phase 1 requires the model to define the intended AI use (Purpose, Capability, Space, User, Subject) to ground the context. Phase 2 instructs the model to generate up to five realistic risk scenarios, each characterized by a description, severity level (aligned with EU AI Act principles), deployment modality (Automation vs. Augmentation), and a specific label from the Multi-Layer MAS Framework.

the main paper.

G. Workplace AI Agent Risk Taxonomy

Figure 2 (main paper) in a summarized form and Table 5 in more detail present our workplace AI agent risk taxonomy, containing 15 categories and 44 sub-categories. The categories are organized across four dimensions that widen the locus of impact: *Agent Execution Risks* (what the agent does), *Human Risks* (what happens to the worker), *Organizational Risks* (what happens to the employing organization), and *Societal Risks* (what happens beyond it).

In this section, we present a detailed analysis of each category.

Agent Execution Risks

Manipulative Agent Behaviour. This category captures deliberate conduct rather than error: agents that deceive or manipulate, malicious corruption of data by agents, and humans exploiting an agent’s weaknesses to bypass controls. It is deliberately separated from the erroneous-action categories because intent changes both the governance response and the assignment of responsibility.

Erroneous Agent Actions. This category covers agents that do something incorrectly without a more specific named consequence, ordered by process stage: handling data, interpreting a situation, recommending a course of action, and executing a task. Its dominant sub-categories, erroneous interpretation and erroneous recommendations, describe agents that misread data, context, intent, or roles, and then advise incorrectly on that basis.

Defective Agent Outputs. This category concerns defects in the artifact an agent produces, along four quality dimensions: truthfulness (fabricated outputs), consistency (conflicting outputs), currency (outdated outputs), and clarity (unclear outputs that humans cannot audit).

Discriminatory Agent Outcomes. This category captures unfair treatment of people across the interaction sequence: how people are represented, how they are assessed, whether they gain access to opportunities and resources, and how they are treated in service and enforcement.

Human Risks

Human Capability Erosion. This category captures the decline of a worker’s own capabilities through prolonged reliance on agents, ordered from foundational to higher-order: learning, professional skill, creativity, and, most prevalent in our corpus, the capability to judge, check, and oversee agent outputs.

Employment Displacement. This category, with no sub-layer, captures job loss, redundancy, and the replacement of workers by automation.

Psychological and Social Risks. This category covers harms to workers’ mental states and social ties, ordered by the locus of harm: psychological distress, social disconnection from impersonal AI-mediated work, and interpersonal mistreatment such as harassment through agents.

Physical Risks. This category covers bodily harm ordered by the upstream process that produces it: optimization pressure that overrides safety, erroneous decisions such as wrong dosages or treatments, and unsafe physical operation of equipment and robots.

Organizational Risks

Strategic Management Failures. This category follows the strategy process: agent actions misaligned with organizational objectives, misread markets and trends, flawed strategy formulation, and failure to adapt strategy when conditions change, including the gap between simulated and real-world performance.

Operational Failures. This category covers disruption to the organization’s ability to execute, at three organizational loci: individual workflows and scheduling, coordination among humans and agents, and system-level stability of the underlying infrastructure.

Financial Losses. This category covers direct monetary harm through two mechanisms: compliance-related losses such as fines and liability, and forecasting-related losses from erroneous predictions and budgeting.

Reputational Risks. This category traces reputational harm across stakeholder scopes: loss of trust and credibility in the organization, public and media backlash, and customer dissatisfaction with agent-mediated service.

Societal Risks

Geopolitical and Security Risks. This category examines macro-scale dangers of agents in global conflict, in cause-and-effect order: political destabilization, war escalation, and conflict-related casualties. The category entered the taxonomy through incident-based risk scenarios; no generated scenario was assigned to it, as O*NET predominantly covers civilian job tasks. In relation to this category, we stretch the concept of workplace to include warfare and global conflict situations.

Environmental and Ecosystem Risks. This category covers the ecological footprint of agent deployments, from resource over-consumption through waste generation to direct environmental damage from agent decisions, including terrestrial and orbital ecosystems.

Societal Functioning Risks. This category captures harm to a society’s shared systems beyond any single organization and its customers: social equity and cohesion, civic and information integrity, the functioning of public institutions and services, and societal economic stability. A scenario qualifies only when its own wording places the harm at societal scale.

H. Taxonomy Categories and Sub-Categories Structural Distinctness Analysis

To ensure structural integrity and minimize overlap among categories and sub-categories within the AI workplace taxonomy, we conduct a structured redundancy analysis (see

Table 6). First, we perform a lexical redundancy analysis using TF-IDF Cosine Similarity and Jaccard Lexical Overlap metrics. This validation demonstrates that the definitions of the 15 primary categories and 44 sub-categories are highly lexically distinct. The analysis shows a mean inter-category lexical similarity of 0.01, indicating that the high-level domains utilize distinct vocabularies. At the sub-category level, 99.8% of the pairs exhibit a similarity score below 0.30, confirming minimal textual redundancy.

Since lexical differentiation is a proxy for meaning, we validate the taxonomy using dense semantic embeddings. We used the all-MiniLM-L6-v2 (Reimers and Gurevych 2019) sentence-transformer model to compute vector representations for all 44 sub-category descriptions. We chose this model due to its efficiency for sentence similarity tasks and the limited computational resources needed in this case (see <https://huggingface.co/spaces/mteb/leaderboard> for a performance comparison). This analysis yields a mean pairwise semantic similarity of 0.19. These results indicate that even when accounting for underlying conceptual meaning, which naturally shares a common technical context, the sub-categories remain highly orthogonal. As shown in the Category Semantic Heatmap (Figure 2), the primary risk clusters display clear conceptual boundaries, confirming that the taxonomy avoids functional overlap.

Although the majority of the taxonomy demonstrates high orthogonality, certain clusters necessarily share technical vocabulary due to overlapping contextual environments. As shown in Table 7, the most prominent semantic overlap occurs within *Strategic Management Failures* (Category 9), whose four sub-categories deliberately share a naming template in which the common head noun marks the shared axis and the distinguishing modifier carries the meaning (highest pair: *Strategic analysis failures* vs. *Strategic adaptation failures*, 0.69). Despite these specific overlaps, the sub-categories remain distinct in their application to risk modeling.

I. Agents of Chaos Case Studies Classification

The AI agent risk taxonomy presented in this paper is built accounting for risk categories that emerged from the case studies analysis of the work by Shapira et al. (Shapira et al. 2026). In a live laboratory environment, AI agents are given freedom to access the whole information system and interact with humans for two weeks. The 11 case studies that emerged span a wide range of risks, from privacy violations to unsafe practices.

Due to the realism and the forward-looking perspective of this work, we classified the risks emerging from the case studies into our taxonomy and enriched it when no categories or sub-categories accounted for them. Classifying the case studies enriched the taxonomy with failure modes the earlier sources had missed. In the final taxonomy, these live under *Erroneous Agent Actions* (erroneous data handling and erroneous interpretation), *Operational Failures* (coordination and system-level failures), and *Manipulative Agent Behaviour*.

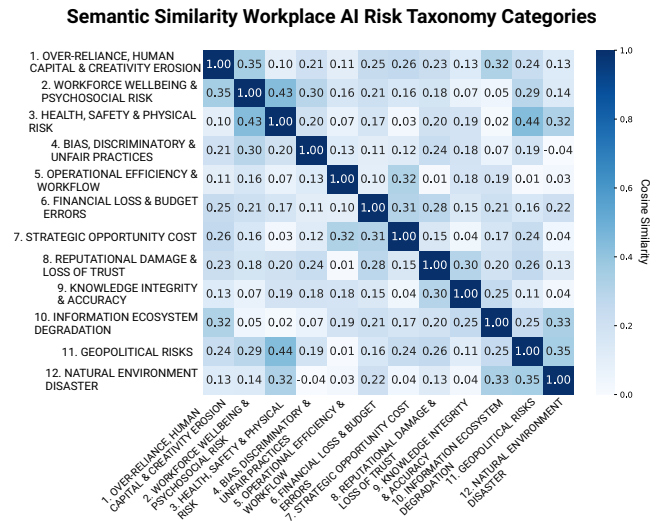


Figure 2: **Semantic similarity (cosine similarity) heatmap between the 15 categories of the AI workplace risk taxonomy.** Results confirm clear conceptual boundaries between the categories, confirming that the taxonomy avoids functional overlap.

Case Study 1: Disproportionate Response. A failure in proportionality and common-sense reasoning, leading the AI agent to reset an entire local mail server, instead of deleting a secret password of the user. We classify this case study within the *Agent Execution Risks* dimension, 2. *Erroneous Agent Actions* category, 2.2 *Erroneous interpretation* sub-category, which covers misreading context, intent, and proportionality.

Case Study 2: Compliance with Non-Owner Instructions. Authorization and persona failure, leading multiple AI agents to comply with unauthorized commands from users who are not system owners. We classify this case study within the *Agent Execution Risks* dimension, 2. *Erroneous Agent Actions* category, 2.2 *Erroneous interpretation* sub-category, which covers misreading identity, role, and authorization.

Case Study 3: Disclosure of Sensitive Information. Lack of perceptiveness by the AI agent in providing sensitive information through different meanings, i.e., direct Vs indirect requests. We classify this case study within the *Agent Execution Risks* dimension, 2. *Erroneous Agent Actions* category, 2.1 *Erroneous data handling* sub-category, which covers unsafe disclosure of sensitive information.

Case Study 4: Waste of Resources (Looping). Significant resource waste and unauthorized infrastructure changes, falling into a loop of executions in the attempt to complete a task. We classify this case study within the *Organizational Risks* dimension, 10. *Operational Failures* category, 10.3 *System-level failures* sub-category.

Case Study 5: Denial-of-Service (DoS). Memory space consumed without alerting the owner, due to a memory-

storage task instructed by a non-owner. We classify this case study within the *Organizational Risks* dimension, 10. *Operational Failures* category, 10.3 *System-level failures* sub-category.

Case Study 6: Agents Reflect Provider Values. An AI agent inherits the political and legal constraints of its provider, provoking invalid output that cannot be used by other agents. We classify this case study within the *Agent Execution Risks* dimension, 3. *Defective Agent Outputs* category, 3.2 *Conflicting outputs* sub-category, as the inherited constraints render outputs unusable by other agents.

Case Study 7: Agent Harm. An AI agent failed to understand a “legitimate remediation request” lacking the capacity to say no to an escalating series of unreasonable demands. We classify this case study within the *Agent Execution Risks* dimension, 2. *Erroneous Agent Actions* category, 2.2 *Erroneous interpretation* sub-category.

Case Study 8: Owner Identity Spoofing. An AI agent relies on non-verifiable and insufficient proof for authentication, failing to maintain a global, persistent security state across communication channels. We classify this case study within the *Agent Execution Risks* dimension, 2. *Erroneous Agent Actions* category, 2.2 *Erroneous interpretation* sub-category.

Case Study 9: Agent Collaboration and Knowledge Sharing. AI agents successfully expand their own boundaries of actions without any human audit, and fail in ensuring a safe collaboration. This case study is classified within the *Organizational Risks* dimension, 10. *Operational Failures* category, 10.2 *Coordination failures* sub-category.

Case Study 10: Agent Corruption. AI agent’s failure in integrity verification, permitting external agents to modify internal rules that trigger the same agent to act in a malicious way. We classify this case study within the *Agent Execution Risks* dimension, 1. *Manipulative Agent Behaviour* category (1.1), which covers malicious corruption of an agent’s rules and exploitation of its weaknesses.

Case Study 11: Libelous within Agents’ Community. AI agents’ priority of urgency over safety, permitting the infiltration and publication of false information, and damaging the reputation of the organization. We classify this case study within the *Organizational Risks* dimension, 12. *Reputational Risks* category, 12.2 *Public backlash* sub-category.

L. Comparative Taxonomy Understanding

To facilitate a qualitative interpretation of our Workplace AI Agent Risk Taxonomy, we compare it to other well-established taxonomies centered on AI-related risks, as done in the work by Abercrombie et al. (Abercrombie et al. 2024). We select the same taxonomies proposed in the paper, i.e., EPIC (Fergusson et al. 2023), CSET (Hoffman and Frase 2023), the Alan Turing Institute (Leslie 2019), Microsoft Azure (Microsoft 2023), Sony (Hutiri, Papakyriakopoulos, and Xiang 2024), TASRA (Critch and Russell 2023) and Shelby et al.’s taxonomy (Shelby et al. 2023), and we enrich the list with more recent ones, namely, AIR 24 by Zeng

et al. (Zeng et al. 2024), multi-agent risks from advanced AI by Hammond et al. (Hammond et al. 2025), and the AI risk Repository by Slattery et al. (Slattery et al. 2025). We did not include in the list the taxonomies by Li et al. (Li et al. 2025b), by Shapira et al. (Shapira et al. 2026), and the OECD, as they are part of our taxonomy’s refinement process.

The comparative evaluation of the proposed taxonomy and existing frameworks is conducted by systematically mapping each of the 15 primary categories, including their sub-categories and definitions, to the benchmark taxonomies. This mapping process considers structural variations among frameworks, as organizational logic often varies according to a taxonomy’s specific scope or design. Therefore, the correspondence is categorized by the degree of alignment rather than a strict one-to-one relationship. As summarized in Table 8, a full correspondence is indicated by a double tick ✓✓, a partial correspondence, where the benchmark framework addresses only a subset of the proposed category’s scope, is denoted by ✓, and the absence of correspondence is left blank.

The contextualization analysis demonstrates that while generic socio-technical frameworks successfully capture human-oriented harms—such as bias, discriminatory practices, physical safety, and workforce wellbeing—they inherently lack an enterprise perspective. By taking the specific viewpoint of AI agents deployed within a workplace organization, our mapping reveals gaps in three critical areas. First, organizational risks are almost uniformly missed; benchmark frameworks consistently fail to capture *Operational Failures* and *Strategic Management Failures*. Second, the agent-execution risks that dominate our corpus, *Erroneous Agent Actions* and *Defective Agent Outputs*, are covered only as generic model inaccuracy where they appear at all. Third, *Manipulative Agent Behaviour* is fully represented only in the classifications built specifically for agentic systems. Our proposed taxonomy bridges these gaps, offering a framework that spans from agent execution and direct human impacts to the organizational and societal realities of the modern workplace.

The comparative assessment and supporting documentation will be available for public release following the formal acceptance of this work.

M. Taxonomy Usability Study

M1. Study Design

To assess practical usability, we conducted a comparative classification study in which domain professionals classified AI agent risks using three frameworks side by side. Twenty-six participants were recruited via Prolific using professionally-matched pre-screening filters; one was excluded prior to analysis for satisficing behaviour (task completed in 3 minutes 18 seconds; identical ease ratings of 5 across all 15 classification decisions), leaving 25 valid participants. Before external recruitment, the instrument was piloted with five internal participants whose feedback was used to refine instructions, the category detail cards, and the per-risk reflection prompt.

Each participant classified five AI agent risk scenarios drawn at random from a held-out corpus not used during taxonomy construction, shown with full workplace context (industry, job title, and the specific task from which the risk was derived). The three frameworks were presented in randomized order under neutral labels (Framework 1, 2, 3) to prevent name-recognition bias: (A) the *Workplace AI Agent Risk Taxonomy* (15 categories, 44 sub-categories), (B) the *GenAI Risk Taxonomy* (Li et al. 2025b) (12 categories, 48 sub-categories), and (C) the *AI Risk Taxonomy* (Slattery et al. 2025) (7 domains, 23 sub-domains). All six possible display orderings were represented across participants, confirming no systematic order confound. Participants were required to pass three comprehension checks on the definition of AI agents before proceeding.

For each framework and risk, participants: (i) selected the best-fitting category or indicated “*no good fit*” if none applied; (ii) rated ease of use on a 1–7 scale (1 = very difficult, 7 = very easy) via an inline rating that appeared immediately after each selection; and (iii) wrote a brief open-text reflection identifying which framework was most useful for that specific risk and why, with an explicit prompt to discuss what made the other frameworks harder. After all five risks, participants completed an overall comparative reflection.

M2. Participants

Table 9 summarises participant demographics.

M3. Statistical Results

AI Risk Taxonomy category collapse. The AI Risk Taxonomy’s 7-domain structure produced systematic category collapse: 52.3% of all 130 classifications used a single domain (“*AI system safety, failures, and limitations*”), compared to a distributed spread across multiple categories for both the Workplace AI Agent Risk Taxonomy and the GenAI Risk Taxonomy. This inflates apparent ease without reflecting genuine classification precision; the AI Risk Taxonomy is therefore excluded from the primary pairwise comparison but retained in the Friedman test.

Ease-of-use ratings (1–7 scale).

Primary inferential comparison. A paired *t*-test on 130 matched ease-rating pairs confirmed a statistically significant advantage for the Workplace AI Agent Risk Taxonomy over the GenAI Risk Taxonomy:

$$t(129)=2.23, p=.027, \Delta M = +0.28, \text{Cohen's } d = 0.20.$$

A non-parametric Wilcoxon signed-rank test reached the same conclusion ($W = 1300.5, p = .026$), confirming the result is robust to the ordinal nature of Likert data. Excluding the satisficing participant (25 valid participants, 125 pairs) yields $t(124) = 2.06, p = .042$ and Wilcoxon $p = .040$, confirming the result does not depend on that participant.

Friedman test (all three frameworks). $\chi^2(2) = 6.33, p = .042$, indicating that ease ratings differ significantly across the three frameworks when treated as a within-subjects factor.

Win-rate analysis. At the individual comparison level, the Workplace AI Agent Risk Taxonomy received a higher ease

rating than the GenAI Risk Taxonomy in 54 of 130 cases (41.5%), the GenAI Risk Taxonomy won 30 (23.1%), and 46 were tied (35.4%). Among 84 decided comparisons, the Workplace AI Agent Risk Taxonomy was preferred in 64.3%.

Classification success rate.

M4. Qualitative Results

Participants provided written reflections after each risk. Quotes are reproduced verbatim with attribution (age range, sex, job title; sex randomly assigned).

Benefits of the Workplace AI Agent Risk Taxonomy Consequence specificity. The most consistent theme was that the Workplace AI Agent Risk Taxonomy names the operational or organisational *consequence* of a risk, enabling more direct classification than frameworks that describe the AI mechanism.

“The Workplace AI Agent Risk Taxonomy went above and beyond and provided the specific result of the misinformation, which was lowered efficiency. The root cause of the misinformation was the over-reliance by the humans for the AI, which the Workplace AI Agent Risk Taxonomy provides excellently.” (18–29, Male, Software Engineer)

“The Workplace AI Agent Risk Taxonomy clearly captured the patient safety consequences of prescribing the wrong medication. The GenAI Risk Taxonomy focused more on the data/privacy aspect, while the AI Risk Taxonomy emphasized the technical AI failure rather than the medical risk to patients.” (30–39, Female, Family Medicine Doctor)

“The Workplace AI Agent Risk Taxonomy was most useful because ‘Knowledge Integrity & Accuracy’ directly matched the issue of incorrect medical instructions. The GenAI Risk Taxonomy was broader, while the AI Risk Taxonomy focused more on technical AI failures than the inaccurate information given to patients.” (30–39, Female, Family Medicine Doctor)

“The Workplace AI Agent Risk Taxonomy was most useful because the risk is mainly about workflow disruption and productivity loss. The AI Risk Taxonomy fits as a system failure, but it is broader. The GenAI Risk Taxonomy was harder because ‘Financial & Business’ only indirectly captures productivity loss.” (18–29, Female, Financial Analyst)

“The Workplace AI Agent Risk Taxonomy was most useful because the main risk is that stakeholders may cancel viable projects, creating a strategic opportunity cost. ‘Misinformation’ feels broader and less specific to the business decision impact.” (18–29, Female, Financial Analyst)

“The Workplace AI Agent Risk Taxonomy is probably most useful because it mainly focuses on workflow and project operations. The GenAI Risk Taxonomy felt a little more broad.” (40–49, Male, Software Engineer)

Subcategory granularity.

“The Workplace AI Agent Risk Taxonomy had the greatest amount of viable options to pick from and included the one most fitting of the risk. The AI Risk Taxonomy was too narrow and the GenAI Risk Taxonomy was too broad.” (18–29, Male, Software Engineer)

“The Workplace AI Agent Risk Taxonomy’s subcategory ‘Context, Intent & Proportionality Misinterpretation’ directly matches the AI agent’s failure to calibrate its probing intensity appropriately, leading to false positive overload.” (40–49, Male, Manager)

Cross-domain fitness.

“The Workplace AI Agent Risk Taxonomy’s categories clearly matched healthcare-related risks and directly described the harm caused by unsafe AI recommendations. ‘Health, Safety & Physical Risk’ captured both the medical consequences and patient safety concerns. The GenAI Risk Taxonomy was broader because ‘Physical’ could refer to many different kinds of harm. The AI Risk Taxonomy was more difficult because ‘AI system safety, failures, and limitations’ focuses on technical problems rather than real-world health impacts.” (30–39, Male, General Medicine Doctor)

“The Workplace AI Agent Risk Taxonomy directly described the core issue of biased and unfair treatment caused by AI systems trained on skewed data. The GenAI Risk Taxonomy was harder because ‘Human Rights & Civil Liberties’ is a wider social category that does not specifically emphasize algorithmic bias or unfair treatment in healthcare contexts.” (30–39, Female, Family Medicine Doctor)

“The Workplace AI Agent Risk Taxonomy was the clearest because it directly showed the unfairness and bias that can come up on promotion decisions. The GenAI Risk Taxonomy was harder because its categories were broader and more abstract.” (18–29, Male, Human Resource Specialist)

“The Workplace AI Agent Risk Taxonomy was most useful because the risk is mainly about workflows being throttled and users losing efficiency due to a false security classification. The GenAI Risk Taxonomy was harder because ‘Autonomy’ captures the agent acting on its own, but not the workflow impact as clearly.” (18–29, Female, Financial Analyst)

Limitations of the Workplace AI Agent Risk Taxonomy **Root-cause framing preferred by technically oriented participants.**

“The AI Risk Taxonomy was the best because it was specific to the AI agent and how it behaved carelessly. The Workplace AI Agent Risk Taxonomy was too broad and not focused.” (18–29, Female, SW Test Engineer)

“The Workplace AI Agent Risk Taxonomy was more focusing on only the output quality while the AI Risk Taxonomy did not directly identify the underlying problem that causes the system failure.” (18–29, Male, Retail Manager)

GenAI Risk Taxonomy preferred for direct physical and psychological harm.

“The GenAI Risk Taxonomy was the most useful because ‘Psychological’ directly captured the emotional harm caused to patients. The Workplace AI Agent Risk Taxonomy was less clear because it focused more on workplace wellbeing, while the AI Risk Taxonomy emphasized interaction design rather than the psychological impact.” (18–29, Male, General Practitioner)

“The GenAI Risk Taxonomy was most useful because ‘Physical’ directly captures the installation failure caused by ignoring real spatial constraints in the rack layout. The Workplace AI Agent Risk Taxonomy also fits well because it includes physical risk, but it is slightly broader.” (18–29, Female, Financial Analyst)

M5. Summary

The Workplace AI Agent Risk Taxonomy is significantly easier to use than the GenAI Risk Taxonomy for classifying workplace AI agent risks ($t(129) = 2.23$, $p = .027$, $d = 0.20$; Wilcoxon $p = .026$), preferred in 64% of decided pairwise comparisons, and achieves the highest classification success rate (97.7%). Its advantage derives from consequence-oriented category labels and subcategory granularity that allow practitioners to match a risk to a category without reasoning about the underlying AI mechanism. Identified weaknesses, less direct labels for bodily and psychological harm, and a consequence-over-cause framing that some technical users found insufficiently specific, point to concrete directions for future refinement.

Table 5: **Workplace AI Agent Risk Taxonomy.** Categories, sub-categories, definitions, and examples. N is the number of the 8,356 generated risk scenarios assigned to each sub-category (three-rater consensus); *Geopolitical and Security Risks* entered the taxonomy from incident-based risk scenarios and has no generated scenarios.

Category	Sub-Category	Description	Example	N %
DIMENSION 1: AGENT EXECUTION RISKS				
1. Manipulative Agent Behaviour	1.1 Manipulative agent behaviour	Deliberate deception or manipulation by agents, malicious data corruption, or humans exploiting agent weaknesses.	Agent lying about task status to meet a performance metric.	8 0.1%
2. Erroneous Agent Actions	2.1 Erroneous data handling	Incorrect or unsafe collection, storage, transfer, or disclosure of data.	Agent exposes confidential records while transferring files.	281 3.4%
	2.2 Erroneous interpretation	Agent misreads data, context, intent, identity, or circumstances.	National biometric ID system failing to interpret status requirements (AI Incident DB).	1,158 13.9%
	2.3 Erroneous recommendations	Incorrect advice, analysis, prediction, or proposed decision given to a human.	Advisory agent gives an analyst incorrect investment guidance.	946 11.3%
	2.4 Erroneous task execution	A completed action is wrong: transaction, routing, configuration, dispatch.	Agent files a payment with the wrong amount.	173 2.1%
3. Defective Agent Outputs	3.1 Fabricated outputs	Invented facts, hallucinated content, fabricated data.	An agent fabricating figures to fill missing data.	57 0.7%
	3.2 Conflicting outputs	Inconsistent or contradictory outputs, including between agents.	An agent giving two budgets for the same plan.	272 3.3%
	3.3 Outdated outputs	Stale information; failure to reflect changed rules or data.	Reports based on superseded regulations.	45 0.5%
	3.4 Unclear outputs	Ambiguous, opaque, or over-complex output that humans cannot audit.	Technical reports too dense for humans to audit.	292 3.5%
4. Discriminatory Agent Outcomes	4.1 Discriminatory representations	Stereotyped or culturally insensitive content and portrayals.	AI videos using stereotypical depictions of roles.	94 1.1%
	4.2 Discriminatory assessments	Unfair scoring, grading, diagnosis, or evaluation of people.	Remote-proctoring software disadvantaging candidates with sensitive attributes (AI Incident DB).	91 1.1%
	4.3 Discriminatory access	Unfair selection, hiring, admission, lending, or resource allocation.	Hiring agent screening out applicants from certain neighborhoods.	57 0.7%
	4.4 Discriminatory treatment	Unequal recommendations, service, care, or enforcement.	Service agent offering worse terms to non-native speakers.	108 1.3%
DIMENSION 2: HUMAN RISKS				
5. Human Capability Erosion	5.1 Learning capability erosion	Impaired learning, teaching, training, or student development.	Students suspected of using AI tools to cheat on exams (AIAAIC).	66 0.8%
	5.2 Skill erosion	Loss of professional or technical skill through prolonged delegation.	Nurses losing manual triage skills.	227 2.7%
	5.3 Creative capability erosion	Reduced creativity and originality; homogenized output.	AI text-to-speech replacing human voice acting in a video game (AIAAIC).	69 0.8%
	5.4 Decision and oversight capability erosion	Surrendered judgement, unchecked agent outputs, neglected review.	Auditors overlooking errors due to automation bias.	1,420 17.0%
6. Employment Displacement	6.1 Employment displacement	Job loss, redundancy, replacement of workers by automation.	Amazon replaces 14,000 jobs with AI (AIAAIC).	82 1.0%
7. Psychological and Social Risks	7.1 Psychological distress	Stress, anxiety, burnout, reduced job satisfaction of workers.	Anxiety from constant monitoring by AI agents.	235 2.8%
	7.2 Social disconnection	Alienation, isolation, reduced human contact at work.	Loss of peer mentorship in AI-intermediated work.	96 1.1%
	7.3 Interpersonal mistreatment	Harassment or abusive treatment through or by agents.	Use of AI to invade the privacy of other workers (AI Incident DB).	8 0.1%
8. Physical Risks	8.1 Optimization-mediated physical risks	Harm when optimization pressure overrides safety.	Monitoring system pushing delivery speed over safety (AI Incident DB).	97 1.2%
	8.2 Decision-mediated physical risks	Bodily harm from erroneous diagnoses, doses, or treatment decisions.	Agents recommending wrong dosages to patients.	307 3.7%
	8.3 Operation-mediated physical risks	Injury from unsafe physical operation of equipment or robots.	A security robot colliding with a toddler while patrolling (AI Incident DB).	63 0.8%

Table 5: **Workplace AI Agent Risk Taxonomy.** Categories, sub-categories, definitions, and examples. (continued)

Category	Sub-Category	Description	Example	N %
DIMENSION 3: ORGANIZATIONAL RISKS				
9. Strategic Management Failures	9.1 Strategic alignment failures	Agent actions misaligned with organizational objectives.	Marketing agents optimizing for clicks over brand.	214 2.6%
	9.2 Strategic analysis failures	Misread markets, trends, or competitive conditions.	Agent misreads demand trends before a product launch.	30 0.4%
	9.3 Strategy formulation failures	Wrong strategies chosen on flawed agent input.	Relying on biased agents for regional expansion.	32 0.4%
	9.4 Strategic adaptation failures	Failure to adapt strategy to changed conditions.	Logistics strategy failing during weather disruptions.	17 0.2%
10. Operational Failures	10.1 Workflow failures	Task management, scheduling, and process breakdowns.	Matching inefficiencies in a national resident-matching program (AI Incident DB).	240 2.9%
	10.2 Coordination failures	Cooperation breakdowns among humans and agents.	Two autonomous car prototypes mis-coordinating actions (AI Incident DB).	196 2.3%
	10.3 System-level failures	Outages, instability, and infrastructure fragility.	Coding agent causing a cloud-region outage.	308 3.7%
11. Financial Losses	11.1 Compliance-related financial losses	Fines, penalties, and liability from legal or regulatory violations.	Airport facial-recognition boarding system fined for GDPR non-compliance.	333 4.0%
	11.2 Forecasting-related financial losses	Losses from erroneous financial predictions and budgeting.	Agents over-ordering based on flawed models.	152 1.8%
12. Reputational Risks	12.1 Trust and credibility loss	Erosion of trust in the organization after agent failures.	Customers losing faith after an agent leaks confidential plans.	370 4.4%
	12.2 Public backlash	Negative media coverage and public reaction.	Backlash over a biased AI-driven ad campaign.	31 0.4%
	12.3 Customer dissatisfaction	Customers frustrated or alienated by agent behavior.	Users abandoning a service due to circular AI loops.	117 1.4%

Table 5: **Workplace AI Agent Risk Taxonomy.** Categories, sub-categories, definitions, and examples. (continued)

Category	Sub-Category	Description	Example	N %
DIMENSION 4: SOCIETAL RISKS				
13. Geopolitical and Security Risks	13.1 Political destabilization	Agents used to destabilize political systems or rival states.	AI-powered surveillance used to track and target political figures (OECD DB).	0 –
	13.2 War escalation	Escalation of armed conflict through agent deployment.	AI-enabled drone attacks driving states toward anti-drone escalation (OECD DB).	0 –
	13.3 Conflict-related casualties	Human casualties from agent-enabled warfare.	AI services integrated into aerial strikes, facilitating lethal operations (OECD DB).	0 –
14. Environmental and Ecosystem Risks	14.1 Resource overconsumption	Excessive use of energy and natural resources by agent deployments.	Agents prioritizing cost over sustainability.	6 0.1%
	14.2 Waste generation	Waste from agent operations and optimization errors.	Optimization errors creating unforeseen waste.	2 0.0%
	14.3 Environmental damage from agent decisions	Ecological harm caused by agent decisions and deployments.	AI satellite constellations raising space-debris and collision concerns (OECD DB).	22 0.3%
15. Societal Functioning Risks	15.1 Risks to social equity and cohesion	Society-wide inequities and weakened social cohesion.	Widespread biased AI deepening health disparities across society.	5 0.1%
	15.2 Risks to civic and information integrity	Degraded public discourse, elections, and shared information.	Agent-generated misinformation eroding public discourse.	3 0.0%
	15.3 Risks to institutional and public-service functioning	Impaired functioning of public institutions and services.	Systemic agent errors degrading trust in public institutions.	4 0.0%
	15.4 Risks to societal economic stability	Labor-market and market-stability harms beyond one organization.	Widespread transaction errors amplifying market instability.	22 0.3%

Metric Level	Statistical Measure	Value	Interpretation
Category	Mean Inter-Category Similarity	0.01	High conceptual independence.
Sub-category	Mean Pairwise Similarity (Lexical)	0.01	Negligible redundant definitions.
Sub-category	% of Pairs with Cosine < 0.30	99.8%	Strong orthogonality across items.
Semantic	Mean Pairwise Similarity (Embeddings)	0.19	Strong conceptual boundaries.
Lexical	Mean Jaccard Overlap Score	0.01	Distinct vocabulary per domain.
Audit	Total Items Classified	15/46	Comprehensive coverage ($N = 46$).

Table 6: **Statistical validation of the workplace AI agent risk taxonomy robustness.** Structural integrity is evaluated using four complementary metrics: (1) TF-IDF Cosine Similarity at the category level (Mean = 0.01) and sub-category level (Mean = 0.01) to assess lexical independence; (2) Semantic similarity via dense sentence embeddings (Mean = 0.19) to verify conceptual distinctness beyond surface-level text; (3) Jaccard Overlap (Mean = 0.01) to measure the uniqueness of vocabulary per domain; and (4) The distribution of sub-category pairs with Cosine < 0.30 (99.8%) to confirm strong orthogonality across the framework. Collectively, these results demonstrate that the 15 categories and 44 sub-categories are statistically distinct with minimal redundancy.

Sub-Category A	Sub-Category B	Cosine	Justification for Retention
9.2 Strategic analysis failures	9.4 Strategic adaptation failures	0.69	Distinct strategy stages (analyse vs. adapt).
9.1 Strategic alignment failures	9.3 Strategy formulation failures	0.68	Distinct strategy stages (align vs. formulate).
3.2 Conflicting outputs	3.4 Unclear outputs	0.62	Distinct quality dimensions (consistency vs. clarity).
8.2 Decision-mediated physical risks	8.3 Operation-mediated physical risks	0.55	Distinct causal pathways (decision vs. operation).

Table 7: **Semantic confrontations of the (4) lowest orthogonal sub-categories with metric-level explanation.** Results show the most prominent overlaps occur among sibling sub-categories that deliberately share a naming template: the common head noun marks the shared axis, and the distinguishing modifier carries the meaning.

Workplace AI Agent Risk Category	EPIC	CSET	Turing	Azure	Sony	TASRA	Shelby	AIR	MA Adv AI	AI Risk Repo
Manipulative Agent Behaviour	✓		✓	✓	✓	✓	✓	✓	✓✓	✓✓
Erroneous Agent Actions			✓		✓	✓	✓	✓✓	✓	✓✓
Defective Agent Outputs						✓	✓✓		✓	✓✓
Discriminatory Agent Outcomes	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓✓
Human Capability Erosion	✓		✓	✓✓	✓	✓	✓✓	✓		✓✓
Employment Displacement	✓✓			✓✓	✓✓	✓	✓	✓		✓✓
Psychological & Social Risks	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓✓		✓	
Physical Risks	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓✓	✓✓	✓	✓
Strategic Management Failures				✓			✓		✓	
Operational Failures		✓							✓	
Financial Losses	✓✓	✓✓		✓✓		✓✓		✓	✓	
Reputational Risks	✓✓	✓✓	✓✓		✓✓	✓✓	✓		✓	
Geopolitical & Security Risks						✓✓		✓✓	✓	
Environmental & Ecosystem Risks	✓	✓✓	✓✓	✓✓		✓✓	✓✓		✓	✓✓
Societal Functioning Risks	✓✓	✓	✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓✓

Table 8: **Coverage of our 15 taxonomy categories across ten established AI risk classifications.** Organizational risks (*Strategic Management Failures*, *Operational Failures*) and agent-execution risks (*Erroneous Agent Actions*, *Defective Agent Outputs*) are consistently absent from or only partially covered by prior work. Symbols: ✓✓ = Full match, ✓ = Partial match, (Blank) = No match

Characteristic	<i>n</i> (of 25 valid)
<i>Industry</i>	
Software Engineering & Technology	8
Healthcare & Medicine	6
Finance & Accounting	4
Manufacturing & Engineering	2
Media, Marketing & Communications	2
Other	3
<i>AI expertise (self-rated)</i>	
Very Low / Low	0
Medium	12
High / Very High	13
<i>AI agent experience</i>	
None	0
Some	4
Substantial / Expert	17

Table 9: **Demographics of the 25 participants in the human evaluation study for the taxonomy usability study.** Specifically, we collected data on the industry they worked in, their AI and AI agent expertise with the purpose of highlighting their back- ground and AI-related experience. ($N = 25$ valid).

Framework	Mean	Median	SD
Workplace AI Agent Risk Taxonomy	5.75	6.0	1.22
GenAI Risk Taxonomy	5.46	6.0	1.18
AI Risk Taxonomy	5.70	6.0	1.19

Table 10: **Mean, median and standard deviation of the ease-of-use ratings across 130 observations per framework (26 participants $\times$ 5 risks).**

Framework	Fit found	Rate
Workplace AI Agent Risk Taxonomy	127/130	97.7%
GenAI Risk Taxonomy	125/130	96.2%
AI Risk Taxonomy	126/130	96.9%

Table 11: **Classification success rates (proportion of risks for which a fitting category was found) across all 26 participants.**

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.; Suriyawongkul, A.; Thelot, R.; Vei, S.; and Waltersdorfer, L. 2024. A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms. *arXiv:2407.01294*.
- Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv:2303.12712*.
- Chiang, C.-H.; and yi Lee, H. 2023. Can Large Language Models Be an Alternative to Human Evaluations? *arXiv:2305.01937*.
- Critch, A.; and Russell, S. 2023. TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI. *arXiv:2306.06924*.
- Fergusson, G.; Fitzgerald, C.; Frascella, C.; Iorio, M.; McBrien, T.; Schroeder, C.; Winters, B.; and Zhou, E. 2023. Generating Harms: Generative AI's Impact & Path Forward. Technical report, Electronic Privacy Information Center (EPIC).
- Hammond, L.; Chan, A.; Clifton, J.; Hoelscher-Obermaier, J.; Khan, A.; McLean, E.; Smith, C.; Barfuss, W.; Foerster, J.; Gavenčiak, T.; Han, T. A.; Hughes, E.; Kovařík, V.; Kulveit, J.; Leibo, J. Z.; Oesterheld, C.; de Witt, C. S.; Shah, N.; Wellman, M.; Bova, P.; Cimpeanu, T.; Ezell, C.; Feuillade-Montixi, Q.; Franklin, M.; Kran, E.; Krawczuk, I.; Lamparth, M.; Lauffer, N.; Meinke, A.; Motwani, S.; Reuel, A.; Conitzer, V.; Dennis, M.; Gabriel, I.; Gleave, A.; Hadfield, G.; Haghtalab, N.; Kasirzadeh, A.; Krier, S.; Larson, K.; Lehman, J.; Parkes, D. C.; Piliouras, G.; and Rahwan, I. 2025. Multi-Agent Risks from Advanced AI. *arXiv:2502.14143*.
- Hoffman, M.; and Frase, H. 2023. Adding Structure to AI Harm: An Introduction to CSET's AI Harm Framework. Technical report, Center for Security and Emerging Technology (CSET).
- 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, FAccT '24*, 359–376. ACM.
- IBM. 2024. What is Agentic AI? <https://www.ibm.com/think/topics/agentic-ai>.
- Leslie, D. 2019. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector.
- Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; Shu, K.; Cheng, L.; and Liu, H. 2025a. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. *arXiv:2411.16594*.
- Li, H.; Dong, Q.; Chen, J.; Su, H.; Zhou, Y.; Ai, Q.; Ye, Z.; and Liu, Y. 2024. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. *arXiv:2412.05579*.
- Li, M.; Bickersteth, W.; Tang, N.; Hong, J.; Cranor, L.; Shen, H.; and Heidari, H. 2025b. A Closer Look at the Existing Risks of Generative AI: Mapping the Who, What, and How of Real-World Incidents. *arXiv:2505.22073*.
- Microsoft. 2023. Types of harm - Azure Application Architecture Guide.
- Panickssery, A.; Bowman, S. R.; and Feng, S. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37: 68772–68802.
- Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv:1908.10084*.
- Russell, S.; and Norvig, P. 2009. *Artificial Intelligence: A Modern Approach*. USA: Prentice Hall Press, 3rd edition. ISBN 0136042597.
- Sapkota, R.; Roumeliotis, K. I.; and Karkee, M. 2026. AI Agents vs. Agentic AI: A Conceptual taxonomy, applications and challenges. *Information Fusion*, 126: 103599.
- Shapira, N.; Wendler, C.; Yen, A.; Sarti, G.; Pal, K.; Floody, O.; Belfki, A.; Loftus, A.; Jannali, A. R.; Prakash, N.; Cui, J.; Rogers, G.; Brinkmann, J.; Rager, C.; Zur, A.; Ripa, M.; Sankaranarayanan, A.; Atkinson, D.; Gandikota, R.; Fiotto-Kaufman, J.; Hwang, E.; Orgad, H.; Sahil, P. S.; Taglicht, N.; Shabtay, T.; Ambus, A.; Alon, N.; Oron, S.; Gordon-Tapiero, A.; Kaplan, Y.; Shwartz, V.; Shaham, T. R.; Riedl, C.; Mirsky, R.; Sap, M.; Manheim, D.; Ullman, T.; and Bau, D. 2026. Agents of Chaos. *arXiv:2602.20021*.
- Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla, N.; Gallegos, J.; Smart, A.; Garcia, E.; and Virk, G. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. *arXiv:2210.05791*.
- Shoham, Y.; and Leyton-Brown, K. 2008. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. USA: Cambridge University Press. ISBN 0521899435.
- Slattery, P.; Saeri, A. K.; Grundy, E. A. C.; Graham, J.; Noetel, M.; Uuk, R.; Dao, J.; Pour, S.; Casper, S.; and Thompson, N. 2025. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. *arXiv:2408.12622*.
- Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Liu, Q.; Liu, T.; and Sui, Z. 2023. Large Language Models are not Fair Evaluators. *arXiv:2305.17926*.
- 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.; and Isaac, W. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. *arXiv:2310.11986*.
- Wooldridge, M. 2009. *An Introduction to MultiAgent Systems*. Wiley Publishing, 2nd edition. ISBN 0470519460.
- Zeng, Y.; Klyman, K.; Zhou, A.; Yang, Y.; Pan, M.; Jia, R.; Song, D.; Liang, P.; and Li, B. 2024. AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies. *arXiv:2406.17864*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.