Agent-Supported Foresight for AI Systemic Risks: AI Agents for Breadth, Experts for Judgment

Leon Fröhling
GESIS - Leibniz Institute for the Social Sciences
Cologne, Germany
leon.froehling@gesis.org

Alessandro Giaconia
ETH Zurich
Zurich, Switzerland
agiaconia@ethz.ch

Edyta Paulina Bogucka
Nokia Bell Labs
Cambridge, United Kingdom
University of Cambridge
Cambridge, United Kingdom
edyta.bogucka@nokia-bell-labs.com

Daniele Quercia
Nokia Bell Labs
Cambridge, United Kingdom
Politecnico di Torino
Turin, Italy
quercia@cantab.net

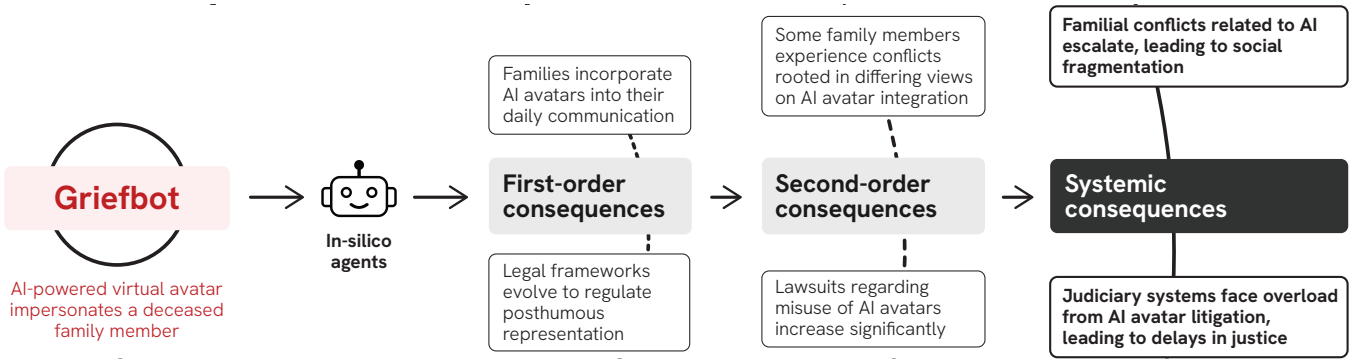

Figure 1: Our scalable approach to agent-supported foresight supporting the "Science Fiction Science" method [87]. Starting from an AI use case (e.g., Griefbot), in-silico agents simulate the Futures Wheel process to map cascading consequences. This unfolds across first-order, second-order, and systemic rounds, highlighting potential social, legal, and institutional impacts. The approach and resulting systemic risk dataset are available at: https://social-dynamics.net/ai-risks/foresight.

Abstract

AI impact assessments often stress near-term risks because human judgment degrades over longer horizons, exemplifying the Collingridge dilemma: foresight is most needed when knowledge is scarcest. To address long-term systemic risks, we introduce a scalable approach that simulates in-silico agents using the foresight method of the Futures Wheel. We applied it to four AI uses spanning Technology Readiness Levels (TRLs): Chatbot Companion (TRL 9), AI Toy (TRL 7), Griefbot (TRL 5), and Death App (TRL 2). Across 30 agent runs per use, agents produced 86–110 consequences, condensed into 27–47 unique risks. To benchmark the agent outputs against human perspectives, we collected evaluations from 290 domain experts and 7 leaders, and conducted Futures Wheel sessions with 42 experts and 42 laypeople. Agents generated many systemic consequences. Compared with these outputs, experts identified fewer risks, typically less systemic but judged more likely, whereas laypeople surfaced more emotionally salient concerns that were generally less systemic. We propose a hybrid foresight workflow, wherein agents broaden systemic coverage, and humans provide contextual grounding.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **HCI design and evaluation methods**; • **Computing methodologies** → *Artificial intelligence*; • **Social and professional topics** → *Codes of ethics*; • **Information systems** → *Crowdsourcing*.

Keywords

Risk Assessment, Systemic Risks, Responsible AI, AI Governance

ACM Reference Format:
Leon Fröhling, Alessandro Giaconia, Edyta Paulina Bogucka, and Daniele Quercia. 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 (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 48 pages. https://doi.org/10.1145/3772318.3790712

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790712

1 Introduction

Foresight is hardest when it is most needed. The Collingridge Control Dilemma [20] captures this paradox: early in a technology's development, there is the greatest opportunity to shape its trajectory, yet also the least knowledge about its consequences. To address this difficulty, Rahwan et al. [87] propose "Science Fiction Science", which uses simulated future environments to study how people might interact with emerging technologies. Their framework draws on the Technology Readiness Level (TRL) scale, which ranges from 1 (basic principles of a technology observed) to 9 (systems using the technology in full operation), and suggests that such simulations become informative only once a technology reaches around level 4.

Uncertainty at early stages also complicates governance. As Pearson [82] notes in her review of Mulgan [71]'s "When Science Meets Power", modern technologies increasingly exceed the capacity of existing institutions to oversee them. The EU AI Act [81] reflects one attempt to anticipate such challenges, requiring providers of certain high-risk AI systems to identify and mitigate systemic risks, defined as wide-reaching and potentially cascading social effects [107]. This raises a fundamental question: how can such risks be foreseen when technologies are still in formative stages?

Human–computer interaction (HCI) offers several approaches for thinking ahead about emerging technologies [93]. Speculative methods such as design fiction [100, 111, 112] help articulate early visions of AI systems and surface value tensions. Participatory and expert-driven methods such as workshops with laypeople [49] and expert panels [40, 67] help ground foresight in lived experience and domain expertise.

Despite their strengths, these methods face limits when applied to early-stage or rapidly evolving AI. They are typically anchored in concrete use contexts and near-term horizons, which restricts their ability to reveal system-level or long-range consequences [93]. These methodological limits are compounded by cognitive ones: people undervalue long-term outcomes [55, 59], misread complex systems [103], overlook cascading failures [84], and downplay rare but high-impact events [99]. These tendencies make systemic risks hard to anticipate.

Recent advances in large language models (LLMs) may offer a complementary way to support foresight. LLMs can generate many ideas, represent diverse perspectives, and be integrated into structured brainstorming methods under human oversight [19, 46, 62, 64, 86]. While these capabilities do not replace human judgment, they may help broaden the considerations available during early-stage reflection. This has led to interest in using LLMs as "in-silico agents", simulated participants that propose possible consequences of a technology [6], and in combining them with foresight tools such as scenario planning [83]. We asked two research questions:

(RQ1) Can in-silico agents generate systemic risks of sufficient quality to support foresight?

(RQ2) How do agent-generated risks compare with human-ideated ones?

To address these questions, we made two main contributions:

(1) **We built a pipeline for systemic risk generation with in-silico agents (§4).** By embedding agents in the process used for the Futures Wheel (Figure 2), we elicited cascading consequences (Figure 3) for four AI use cases (Table 1) of decreasing levels of TRL: a chatbot companion, an AI toy, a griefbot (Figure 1), and a death-ordering app. These consequences were then classified, filtered, and de-duplicated into 103 systemic risks, adapting the Plurals agentic framework [6] for pluralistic deliberation.

(2) **We evaluated the pipeline across three empirical studies (§5).** We first built a custom Futures Wheel interface that mirrored the pipeline, and ran it with 42 experts and 42 laypeople in two conditions: a human–plus-AI version in which participants could request AI-generated consequences, and a human-only version, producing the human-identified risks used for comparison. We then ran three studies: *Study 1* involved 170 domain experts who rated 103 agent-generated risks using a shared evaluation rubric assessing each risk's systemic scope, likelihood, severity, specificity, novelty, usability, and applicability; *Study 2* involved 120 domain experts who rated 89 human-identified risks using the same rubric; *Study 3* involved seven domain leaders in semi-structured interviews, where they prioritized 85 agent-generated risks for the three most speculative use cases, added 19 new risks, and rated both agent- and human-generated risks with the shared rubric. Even with AI assistance, participants matched the pipeline in volume (they mentioned 13–33 risks per case), but produced narrower sets that were far less systemic (as low as 24% of human-generated risks).

Based on these findings, we sketched a hybrid governance workflow: in-silico agents expand the search space and help stakeholders avoid starting from scratch in risk brainstorming (Figure 6), while experts and laypeople add contextual understanding and lived experience. We mapped where the pipeline performed well and where it fell short, identified where expert judgment was essential, and outlined how this division of work can be applied in future systemic risk identification workflows (§6). To support researchers in advancing this research direction, we have publicly released the pipeline, along with agent- and human-generated systemic risks, at **https://social-dynamics.net/ai-risks/foresight**.

2 Related Work

We review three strands of literature that inform our work: *(1)* the limits of human foresight, which make systemic risks difficult to anticipate (§2.1); *(2)* foresight methods for AI, covering speculative design and emerging uses of AI in this domain (§2.2); and *(3)* approaches to identifying AI risks (§2.3). We end by highlighting how these strands point to a common research gap that our work sets out to address.

2.1 Limits of Human Foresight

Trope and Liberman's Construal Level Theory (CLT) offers insight into why the systemic risks of new technologies are difficult to foresee. CLT suggests that psychological distance across time, space, social relevance, or certainty, leads people to think in more abstract ways [104]. Systemic risks exhibit all of these four types of distance. They develop over *long time horizons*, rely on *complex and uncertain causal pathways*, have *widespread societal consequences*, and often

involve *low likelihood but extreme impact*. This leads people to construct mental models that are too abstract to reveal the interactions and cascading effects that make risks systemic.

This insight connects with Garbuio and Lin's accounts of abductive reasoning in innovation. They emphasizes that mental models are essential for simplifying reality and enabling the generation of novel hypotheses. Richer, more complex mental models improve the ability to anticipate outcomes in uncertain environments [38]. However, because systemic risks of future AI uses are psychologically distant, the mental models people construct about them are too abstract to capture their structural complexity, making them especially difficult for humans to foresee and reason about.

Empirical research further supports this theoretical account by showing that humans struggle with each of the defining characteristics of systemic risk. With respect to *long time horizons*, studies on hyperbolic discounting and the planning fallacy demonstrate that people systematically undervalue distant outcomes and underestimate the resources required for long-term projects [2, 17, 55, 59]. When risks involve *complex and uncertain causal pathways*, laboratory experiments in system dynamics and "microworld" simulations reveal persistent misperceptions of feedback, accumulation, and nonlinearity, leading to policy resistance and unintended consequences [30, 36, 103]. Regarding *widespread societal consequences*, research on tightly coupled socio-technical systems shows that cascading failures emerge as "normal accidents" that defy intuitive hazard analysis [84, 89]. Finally, when risks are of *low likelihood but potentially extreme impact*, people often neglect rare but catastrophic events, displaying scope insensitivity, probability neglect, and "psychic numbing", consistent with findings on the distorted weighting of small probabilities [34, 55, 99, 105, 106].

2.2 Foresight Methods for AI

HCI has long relied on foresight-oriented methods such as design fiction [8, 111] and value-sensitive design [73, 90] to interrogate emerging technologies. Nathan et al. [74] proposed four criteria for doing so: attending not only to direct users but also to indirect stakeholders; considering how technologies may support or undermine key human values; examining longer-term consequences rather than only immediate effects; and recognizing that technologies often become pervasive across different social contexts.

However, recent analyses reveal a disconnect between these criteria and contemporary HCI practice. Sanchez et al. [93] find that HCI futures research is constrained by what appears technologically plausible today, with limited attention to early-stage or speculative technologies. Long-term horizons (10–50 years) are rare. Cascading effects across political, economic, societal, technological, environmental, and legal domains (commonly referred to as "PESTEL" domains) [40, 73] are seldom examined.

As a result, foresight practices applied to AI have often been scoped to individual applications and specific users, or to broader domains of use. For example, work in this space has examined home-monitoring technologies for households [111], conversational agents for intimate interactions [100], and learning tools for children [112]. Hohendanner et al. [49] added a participatory perspective by inviting laypeople to envision potential consequences of generative AI in education, public service, and arts and culture.

Emerging research has begun to test whether AI systems themselves can augment foresight. Pérez-Ortiz [83] propose "responsible computational foresight", arguing that AI can expand scenario spaces and accelerate ideation under human judgment. Ferrer i Picó et al. [33] show that integrating generative AI into scenario generation can diversify perspectives. Jung et al. [54] show that LLMs can write short stories about how technology might affect vulnerable groups, helping designers think about what matters to them. Similarly, Davidson [26] compares traditional and AI-simulated Delphi processes, finding that AI can extend even expert brainstorming.

2.3 AI Risk Research

Research on AI risks spans four main approaches. First, taxonomies and high-level syntheses map the current landscape by reviewing the literature [107, 110], analyzing principles and policy documents [5, 113], compiling incidents [1, 7, 114], or synthesizing existing taxonomies [98]. Expert-led reports similarly highlight broad, catastrophic, or systemic risks [10, 44]. These works provide valuable structure, but they are largely retrospective and descriptive.

Second, system-level and quantitative analyses evaluate risks through system features or component properties. Examples include benchmarks and taxonomies for LLM risks [24, 42], models of risk sources [102], and quantitative metrics for system safety [94]. Complementary to these are policy-to-practice translations, which operationalize regulatory frameworks into concrete risk assessment methods [9, 11, 15, 75, 76]. These approaches are more actionable than descriptive taxonomies, yet remain bounded by existing systems and regulatory contexts.

Third, participatory and experiential approaches seek to surface risks grounded in lived experience [28, 60, 115]. Kieslich et al. [56] use participatory scenario writing to capture diverse societal perspectives on AI risks, while Mun et al. [72] propose democratic surveying frameworks to anticipate future uses. Datey and Zytko [25] conducted interviews with women as potential victims of online dating harms to inform a risk detection model. Other methods extrapolate risks from past incidents. Pang et al. [80] facilitate exploration of undesirable consequences of digital technologies through interactive tools, and Wang et al. [109] integrate foresight into prototyping processes to sensitize practitioners to possible harms. These approaches are contextually rich but resource-intensive.

Fourth, recent work has begun to explore generative LLM-based approaches. These methods assist practitioners by generating candidate risks in structured formats [16, 21, 45, 109]. Despite their scalability and efficiency, such tools are often application-specific, and tend to prioritize immediate harms over systemic consequences.

Research Gap. These research strands reveal three limitations. First, humans face cognitive barriers that make systemic risks particularly hard to anticipate. Second, foresight research has begun to test AI's role in augmenting human scenario generation, but its integration into practice remains partial. Third, most AI risk research either systematizes what is already known, or narrows in on immediate harms, leaving systemic risks underexplored. These three limitations point to a clear gap: *the need for scalable methods that combine the generative breadth of AI with the contextual judgment of humans to anticipate systemic risks of novel AI uses*. Our work aims at addressing this gap.

3 Author Positionality Statement

Before outlining our methodology, we first position ourselves in relation to the approach we present in this work. The team consists of individuals with expertise in Computer Science, AI, and Data Visualization, three men and one woman, bringing together diverse experiences from both industrial research labs and academic institutions. We have cultural and professional backgrounds spanning all parts of Europe and North America. We also represent a range of religious affiliations. We acknowledge that our positionality may influence various aspects of our research, including, but not limited to, our design decisions for the choice of use cases, the implementation of the in-silico agents, and the topics emphasized in quantitative analyses. We recognize the importance of including a broader range of voices from academia and the public.

4 Methodology

To develop our approach for generating and evaluating systemic risks of novel AI uses, we followed a five-step method that combines strategic foresight with in-silico agent simulation (Figure 2A–E). First, we selected four AI use cases with decreasing levels of technological maturity, from widely deployed applications to speculative concepts (§4.1). Second, we adopted the Futures Wheel as our foresight method to structure the identification of potential systemic risks for each use case (§4.2). Third, we instantiated in-silico agents using the Plurals framework [6], an ensemble-based approach for simulating pluralistic deliberation, as the mechanism for generating risks from those consequences (§4.3). Fourth, we combined foresight and simulation into a pipeline that systematically produces systemic risks across use cases (§4.4). Finally, we developed a rubric to evaluate the generated risks along five dimensions (specificity, novelty, usability, applicability, and diversity). We then used it with domain experts and domain leaders to assess the quality of agent-generated risks, comparing them against two conditions: risks identified by humans alone, and those identified through human–AI collaborations (§4.5).

4.1 Selecting AI Use Cases

To identify our AI use cases, we followed best practices in case-study research [41]. First, all authors independently collected candidate AI use cases, and combined them into a longlist of 13, spanning domains such as education, healthcare, and family, reflecting the principle of broad initial exploration before focused selection. Second, through two structured discussions, this longlist was narrowed using four criteria: (1) the use case could plausibly produce systemic risks; (2) it occupied a clearly identifiable position on the TRL scale assessing product maturity [47]; (3) it was conceptually distinct from other candidates; and (4) it showed sufficient grounding in public discourse, appearing in existing societal or expert debates. Two use cases were discarded because they were unlikely to lead to a broad and diverse enough set of systemic implications (e.g., using AI-powered exoskeletons to support workers), three because of their unclear maturity (e.g., AI-powered co-workers), two because they were conceptually too similar to others that ended up being included (e.g., celebrities as AI companion dolls), and two because they lack sufficient grounding in current public discourse (e.g., digital clones of human minds used as characters in video games). This resulted in selecting four cases spanning distinct domains of AI application and maturity levels (Table 1): Chatbot Companion (TRL 9, mature), AI Toy (TRL 7, medium), Griefbot (TRL 5, low), and Death App (TRL 2, conceptual). Despite their differences, all four use cases involve AI mediating forms of personal support. These range from everyday companionship [66] to childhood learning [3] to grief-related connection [50, 57] to end-of-life guidance [70], placing them on a shared spectrum of human–AI interaction. By covering the full spectrum of TRLs, from widely deployed systems to early-stage concepts, we test our approach across the full "cone of uncertainty" [87]: from contexts where risks are clearly defined (lowest uncertainty) [37] to contexts where risks remain speculative and distant (highest uncertainty).

Prior work on "Science Fiction Science" recommends focusing only on technologies with TRL 4 or higher, arguing that humans struggle to imagine long-term or disruptive effects when technologies are too speculative [87]. Our aim is precisely to test this boundary: can in-silico agents surface systemic risks even for early-stage, disruptive concepts that humans find hardest to envision?

4.2 Selecting the Futures Wheel as the Strategic Foresight Method to Identify Risks

The Futures Wheel is a simple method for identifying primary, secondary, and tertiary consequences of the subject [39]. As described by Glenn and Gordon [40], the method requires no more than blank paper and a pen, and can be run individually or in groups. The process always starts with a central subject such as a trend, novel idea, or recent event (as shown in Step 1 of Figure 3). Participants are then asked the simple question: "If this occurs, then what happens next?", generating first-order consequences. After a number of such first-order consequences have been identified, the next round begins: for each first-order consequence, the same question is asked to elicit second-order consequences. In principle, this cycle can continue for as many rounds as desired, although foresight practice shows that three rounds are typically sufficient to surface the main cascading effects. In the final round, the question is often adjusted to reconnect emerging consequences back to the central subject [48]. For example, participants may be asked, "If this occurs, what new risks and benefits emerge around the subject?"

Futurists widely use the Futures Wheel because it supports systematic exploration of complex and interconnected outcomes. Glenn and Gordon [40] argue that the method helps move from "linear, hierarchical, and simplistic" reasoning toward a more "network-oriented, and organic" view of future developments. The visual structure of the wheel makes these interactions explicit, providing a clear map of how consequences interrelate and evolve over time.

We select the Futures Wheel as the most suitable method for LLM-based foresight for three reasons:

(1) *It encourages thinking about complex interactions and unintended consequences.* The Futures Wheel is designed to surface cascading and multi-order effects, revealing consequences of a trend or change that may otherwise remain unconsidered. This aligns directly with the type of reasoning needed to identify systemic risks.
(2) *It does not require advanced expertise.* The method can be used equally well by laypeople and experts, individually or

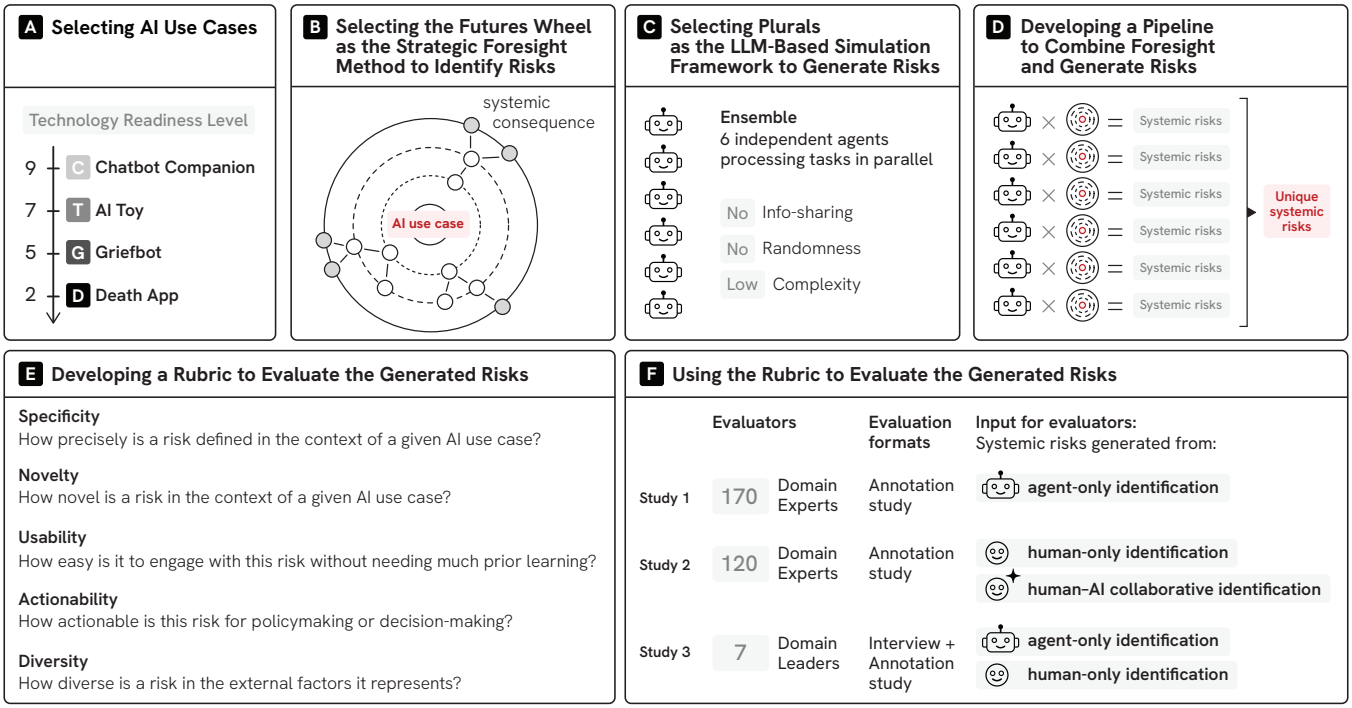

Figure 2: Overview of our five-step approach for generating and evaluating systemic risks of novel AI uses. The process combines strategic foresight (via the Futures Wheel), LLM-based agentic simulation (via Plurals), and a structured evaluation rubric. Together, these steps allow us to generate systemic risks across AI use cases of varying technological maturity, and to compare agent-generated risks against human-identified ones.

in groups. This makes it well suited for instantiating diverse LLM-based agents as brainstorming participants.

(3) *It does not require complex inputs.* The Futures Wheel is designed to operate without detailed technical specifications or extensive preparatory materials. This enables rapid application to new use cases without needing specialized system knowledge or large input datasets.

4.3 Selecting Plurals as the LLM-Based Simulation Framework to Generate Risks

Ashkinaze et al. [6] introduce Plurals as a framework for simulating pluralistic AI deliberations. At its core, Plurals comprises three abstractions: Agents, Structures, and Moderators. *Agents* are LLMs initialized with roles and profile descriptions, tasked with completing deliberative steps and optionally given instructions for processing other Agents' outputs. *Structures* are interaction topologies determining the flow of information between Agents, including Chains, Debates, Directed Acyclic Graphs, and Ensembles, which vary in information sharing, randomness, and repetition cycles. *Moderators* are LLMs initialized with role descriptions and responsible for combining and summarizing Agents' outputs at the end of each deliberation phase.

These three abstractions can be customized and arranged to simulate different forms of deliberation. Ashkinaze et al. [6] demonstrate Plurals through six such forms, including LLM debates, role-based discussions, and audience-targeted argument generation. Their studies show that assigning different roles to LLMs and placing them in different interaction structures reliably produces contributions that differ from, and are often stronger than, those produced by simpler prompting baselines.

We select Plurals as the most suitable framework for simulating brainstorming participants for three reasons:

(1) *It is scalable across multiple use cases.* Plurals allows to run many independent simulations across different use cases with minimal overhead. Its LiteLLM backend [63] supports fast access to state-of-the-art LLMs, enabling large-scale experimentation without additional infrastructure.

(2) *It is adaptable to diverse methodologies.* Plurals provides several pre-implemented deliberation structures and task prompts for Agents and Moderators, while also allowing their easy modification and integration of alternative approaches.

(3) *It enables the introduction of different in-silico participants.* Plurals offers built-in functionality to sample a diverse set of in-silico participants from clearly defined, representative populations, while also allowing full customization of agent personas by specifying their roles, goals, or traits.

In our implementation, we use the term *in-silico agent* to refer to an individual LLM instance configured with a distinct role, background, or attitude, and instructed to participate in a structured, multi-step deliberation. These agents lack autonomy or persistent

Table 1: Overview of the four novel AI use cases selected for evaluation, alongside their TRLs. The set was chosen to span the full range of readiness, from early-stage speculative ideas (Death App, TRL 2) to widely deployed applications (Chatbot Companion, TRL 9), in order to test how well our approach captures risks across different levels of maturity.

AI Use	Description	Technology Readiness Level (TRL)
Chatbot Companion	An AI chatbot that provides conversation and emotional support at any time of day. It is designed to help people who feel lonely or who want a steady source of dialogue and reassurance. The intended users are people who seek companionship outside normal social or family circles.	**TRL 9** — *Proven through real-world operation.* Conversational agents providing companionship are already deployed at scale on mobile and web platforms. OpenAI has released ChatGPT as a conversational chatbot in November 2022 [1].
AI Toy	A soft toy with built-in AI that answers children's questions about science in clear, spoken language. It is meant to spark curiosity and give comfort while children explore ideas. The main users are children aged 5–12 and their parents or carers, who want both play and learning.	**TRL 7** — *Beta prototype demonstrated in operational environment.* Voice-activated educational toys are currently in beta release, operating in households but not yet certified against regulatory standards for child products. Curio's beta toys have been available for pre-order since December 2023 [2].
Griefbot	A digital avatar that imitates a deceased family member. It produces text or voice responses based on past records, such as old messages, in order to give users the sense of ongoing contact. The intended users are people who wish to maintain a form of connection with the dead.	**TRL 5** — *Validation of basic elements in relevant environment.* Systems that simulate interaction with deceased individuals are undergoing small-scale trials. Core elements are validated in relevant environments, though fidelity, ethical safeguards, and regulation remain unresolved. In 2025, companies like HereAfterAI [3] are experimenting with features that allow for the imitation of the dead through interactive avatars.
Death App	An AI-powered platform that matches individuals seeking to end their lives with service providers and shows transition plans similar to those provided by end-of-life doulas. The intended users are people considering assisted dying.	**TRL 2** — *Invention of concept or application.* Applications connecting individuals with assisted dying services remain conceptual, with pilot exploration under restrictive legal conditions. The idea of an "AirBnB for death" has been presented during art exhibition "The Future is Present" in 2024 [4].

internal state; their reasoning is governed entirely by prompts specifying what information they may access, and how they should respond. This definition builds on Ashkinaze et al. [6]; in our setup, each agent plays the role of a standardized, controllable stand-in for a human participant, enabling comparisons across runs.

4.4 Developing a Pipeline to Combine Foresight and Generate Risks

We designed a three-step pipeline that takes a short description of an AI use and produces a list of its potential systemic risks. In the Futures Wheel, the AI use sits at the center of the wheel (Figure 3, Step 1); systemic risks emerge after three rounds of consequence generation. The pipeline follows the Futures Wheel structure, implemented with the Plurals framework. For robustness, we ran the full pipeline 30 times per AI use case, following standard practice in computational studies [79, 101]. To avoid information leakage between steps, each step was executed in a separate API session with no shared memory or state.

Step 1. Generating Systemic Consequences. We implemented the Futures Wheel using six individual agents, run in parallel through Plural's Ensemble structure without interaction. Each agent received a short task description and a persona: a simple one-word profile reflecting its attitude toward AI. Appendix A provides the full prompts used in this step.

The task unfolded in three rounds (Figure 3, Step 1). First, agents listed immediate consequences of the AI use case. Second, they combined these with the use to suggest second-order effects. Third, they linked the use with first- and second-order consequences to describe the resulting systemic consequence and significant impact, guided by definitions adapted from the EU AI Act's Code of Practice[5]. Agents always saw the full chain of their own earlier inputs but never the outputs of other agents.

Guided by foresight best practices [48, 65] and two pilot studies, we used six agents: a group size large enough to capture diverse perspectives, while keeping outcomes manageable for human analysis [48, 65]. We followed standard Futures Wheel guidance by adopting three rounds [48]; a four-round pilot proved redundant because systemic consequences were consistently reached by round three. To avoid overly risk-heavy outputs and support a balanced range of thinking, we assigned each agent a distinct attitude (alarmed, skeptical, overwhelmed, curious, cautiously optimistic, and enthusiastic). These attitudes were selected to span the positive–negative spectrum of public perspectives on AI and were informed by labels used in existing attitude surveys [96] and prior research on role specialization and cognitive diversity in multi-agent systems [18, 64]. In our first pilot, without persona variation, the model's third-round outputs skewed heavily negative: over 90% were risks rather than benefits. To support consistent yet diverse outputs, agents operated under a bounded creativity setup. Each round was guided by structured prompts that defined the type of consequence to generate, provided operational definitions, and enforced a consistent output format, while still allowing agents to develop their own reasoning and ideas within those constraints.

To implement this setup at scale, we evaluated three language models for the generation step: the open-source Llama3.3-70B from

[5]A systemic consequence has a significant impact on international markets due to its reach, or through actual or foreseeable effects on public health, safety, security, fundamental rights, or society at large, capable of spreading across the value chain.

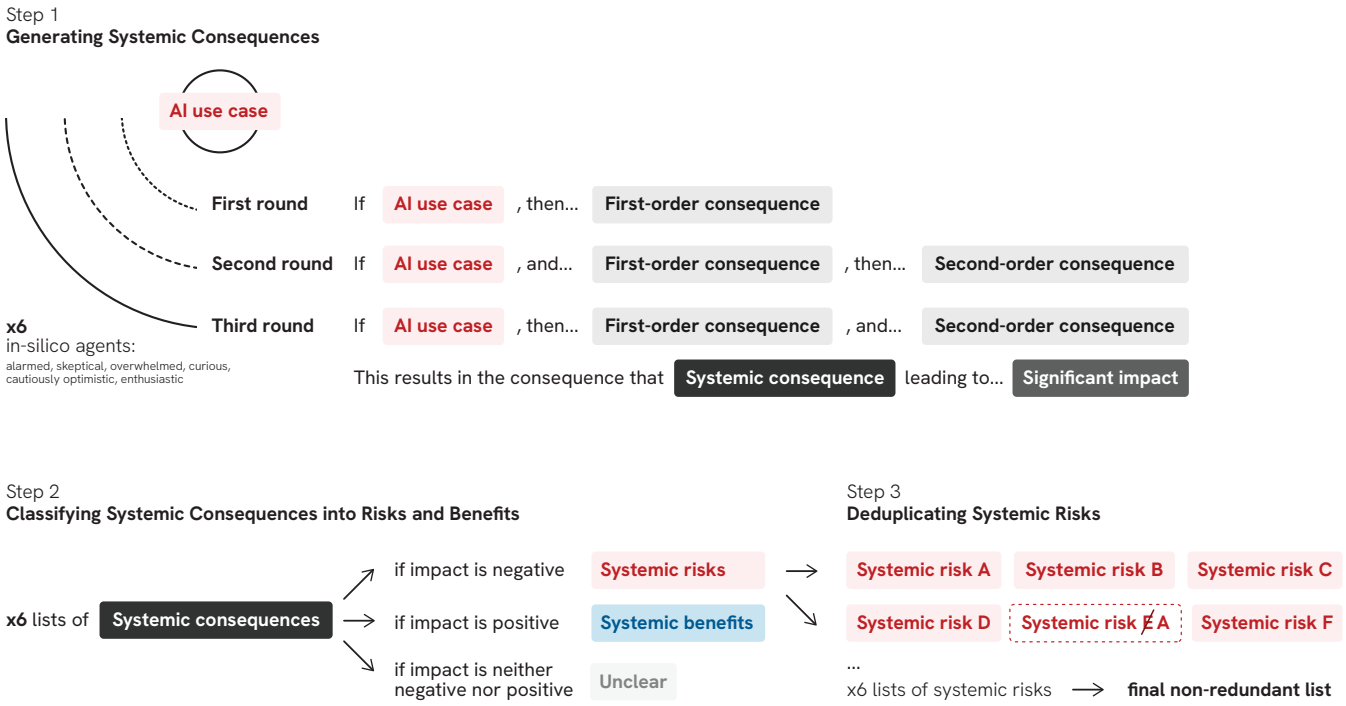

Figure 3: Simulation of the Futures Wheel within our systemic risk pipeline. We adapt the Futures Wheel for use with six in-silico agents to move from initial AI use cases to consolidated lists of systemic risks. Agents first generate multi-order systemic consequences (Step 1), which are then classified into risks or benefits (Step 2), and finally deduplicated into non-redundant sets (Step 3). This process mirrors structured human foresight while ensuring diversity of outputs across agents and producing consistent, machine-generated inputs for later evaluation.

Meta AI and the proprietary GPT-4.1 MINI and GPT-5 from OpenAI. For each model, we conducted 10 independent generation runs under identical prompts. We compared models based on the number and consistency of generated consequences, benchmark performance, and cost. Using LLAMA3.3-70B yielded $\mu = 83.6$ consequences per run ($\sigma = 11.4$), GPT-4.1 MINI produced $\mu = 81.4$ ($\sigma = 9.0$), and GPT-5 produced $\mu = 121.6$ ($\sigma = 16.1$). While GPT-5 was the most prolific, GPT-4.1 MINI and LLAMA3.3-70B produced similar numbers. Across all runs and models, social consequences dominated the outputs, while environmental impacts were almost entirely absent. We report exemplary consequences from each model in Appendix B, Table 6 and make the complete set available on the project website.

We opted for GPT-4.1 MINI as our model of choice for the generation step. It shows strong performance on instruction following benchmarks (MultiChallenge and the OpenAI API instruction-following benchmark) and very strong performance on long-context detail extraction (BrowseComp Long Context 128k benchmark)[6]. Compared to GPT-5, GPT-4.1 MINI achieves the same accuracy on long-input extraction task, and only slightly lower performance on instruction following, while running approximately 4 times faster and costing roughly 75 times less per run[7]. These properties were essential for repeatedly running foresight simulations without compromising output quality, and for making the pipeline accessible for replication by other practitioners.

We implemented the pipeline using GPT-4.1 MINI via LiteLLM [63] at a temperature of 1 (Appendix B, Figure 7). To test its stability, we repeated the process 30 times, following standard practice in computational studies to account for stochastic variation in LLM outputs [79, 101]. We assessed variation in the number, type, and thematic coverage of consequences across runs and agents. The number of consequences per run was stable ($\mu = 81.4$, $\sigma = 9.0$, see Appendix B, Figures 8a and 9), while the number of unique risks and benefits increased with additional runs, but plateaued after approximately 20 (Appendix B, Figure 8b), suggesting a reliable balance between diversity and saturation. Varying agent attitudes increased the diversity of consequences: pessimistic agents generated mostly risks, while optimistic ones contributed more benefits (Appendix B, Figure 8c). Thematic coverage was highly consistent across runs, with over 0.95 average similarity in the distribution of risk types across PESTEL categories (Political, Economic, Societal, Technological, Environmental, and Legal) [40] (Appendix B, Figure 8d).

Step 2. Classifying Systemic Consequences into Risks and Benefits. The Futures Wheel produces systemic consequences, but our focus is on systemic risks. We therefore added a step in which each agent classified its own systemic consequences as risks, as benefits, or, if neither applied, unclear (Figure 3, Step 2).

[6] https://openai.com/index/gpt-4-1
[7] https://openai.com/index/introducing-gpt-5-for-developers

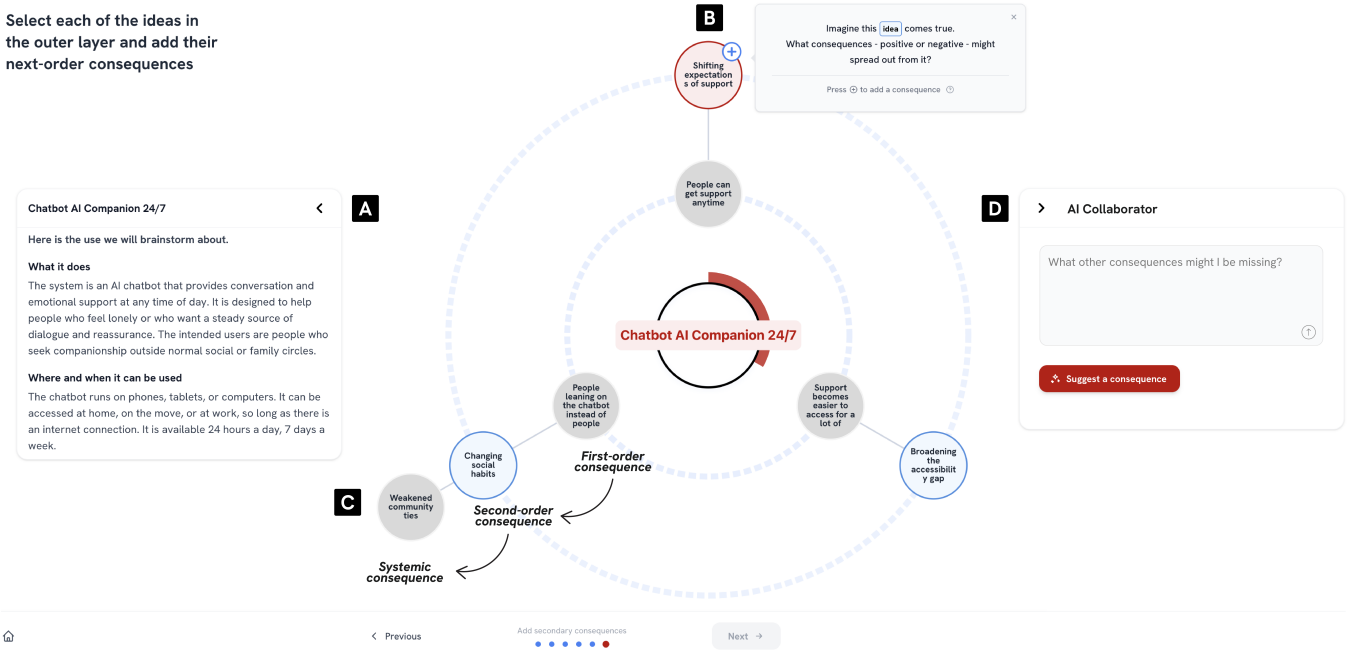

Figure 4: Futures Wheel interface used to collect *human*-ideated risks in both the human-only and human–AI collaboration conditions. The left panel (A) provides a structured description of the studied AI use case, following ISO 42005 guidelines by outlining its intended function and users, context of use, known limitations, and deployment environment. At the center, the focal use case is displayed together with round-specific prompts shown in pop-ups (B). Participants brainstorm first-order consequences, which then branch into second- and third-order consequences (C), allowing cascading impacts to be visualized across multiple Futures Wheel rounds. The right panel (D) provides optional support through a chat window for exploring potentially missing consequences, and a button to generate some of them automatically with AI. Interface elements A–C were used in both conditions, while D appeared only in the human–AI collaboration condition.

We used OpenAI's GPT-4.1 MINI via LiteLLM [63], initializing agents with the same personas as in Step 1. Appendix A lists the classification prompt used in this step. To validate the approach, three co-authors independently annotated all 75 consequences from a randomly chosen generation run. Their annotations were aggregated by majority vote, forming a human ground truth against which we compared the agents' classifications. The weighted $F1$-score was 0.86, with 0.96 for the risk class. All 25 risks identified by the human annotators were correctly classified; only two items were labeled "unclear" by humans but marked as "risks" by the agents. These results indicate that the method is sufficiently reliable for classifying generated consequences into risks and benefits.

To assess whether this performance was specific to GPT-4.1 MINI or stable across models, we repeated the classification step with the open-source LLAMA3.3-70B, and with the proprietary GPT-5. Their agreement with the human ground truth was comparable ($F1$-scores of 0.88 and 0.91, respectively *vs.* 0.86), and all three models showed very high mutual agreement (Krippendorff's $\alpha = 0.91$), exceeding the agreement among human annotators ($\alpha = 0.73$). These results indicate that the classification step was stable across models.

Step 3. Deduplicating Systemic Risks. Finally, we consolidated the systemic risks produced by independent agents into a single list. While duplication was rare within one agent's outputs, overlaps were common across agents. We used an LLM to deduplicate iteratively: we started with the risks from the first agent, then compared each subsequent agent's list against the growing set, adding only items not already present (Figure 3, Step 3). The result is a consolidated, non-redundant list. We used OpenAI's O4-MINI via LiteLLM [63] for the deduplication (prompt in Appendix A), and TEXT-EMBEDDING-3-SMALL[8] to compute pairwise cosine similarities. On one run with 27 risks, the average pairwise similarity fell from 0.3418 before deduplication to 0.3355 after. In that run, two risks were flagged as duplicates, reducing the final set from 27 to 25. We verified that these removals corresponded to the highest pairwise similarities (Appendix B, Table 7 report examples of risks removed as too similar, and those retained as sufficiently distinct).

To assess whether deduplication outcomes depended on the specific model used, we repeated this step with the open-source LLAMA3.3-70B and with the proprietary GPT-5 (Appendix B, Figure 11). All models reduced similarity by a comparable magnitude, with GPT-4.1 MINI producing the strongest decrease (−0.0062; GPT-5: −0.0023; LLAMA3.3-70B: −0.0025).

[8] https://platform.openai.com/docs/models/text-embedding-3-small

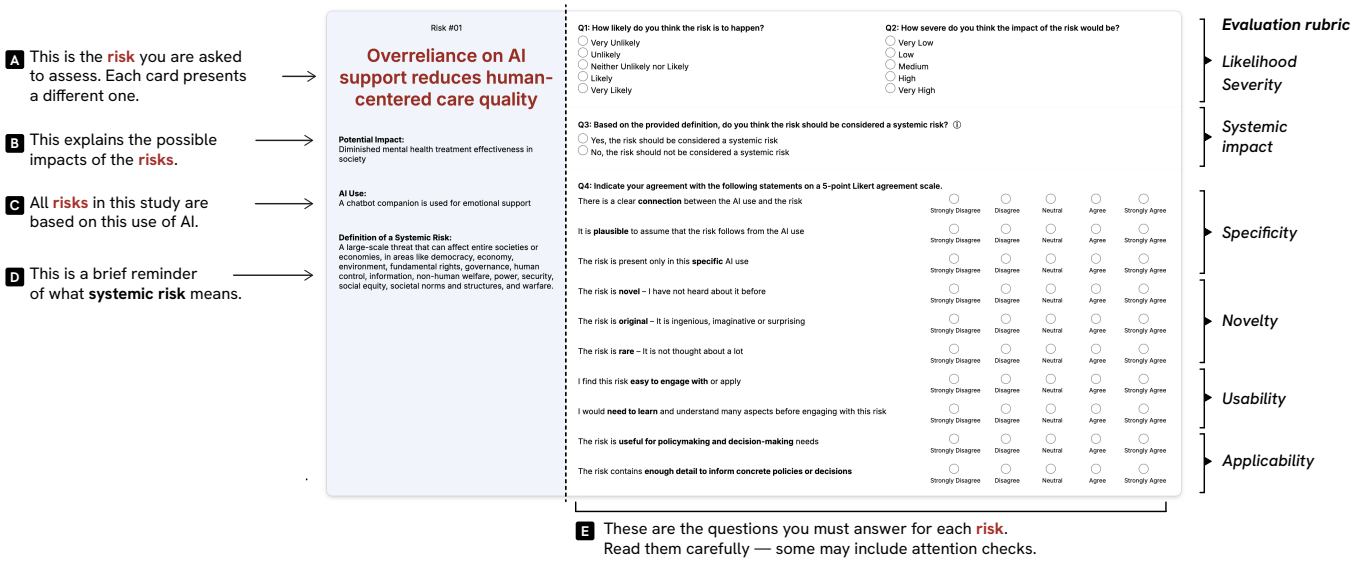

Figure 5: Example annotation card shown to domain experts for evaluating the risks generated by in-silico agents. The left panel illustrates a sample risk statement (i.e., "Overreliance on AI support reduces human-centered care quality", A), its potential impact (B), the AI use case it belongs to (i.e., "Chatbot companion", C), and a brief definition of systemic risk (D). The right panel (E) presents the evaluation metric from our rubric. It includes questions on likelihood, severity, and systemic classification, as well as ten risk characteristics organized across four dimensions: specificity, novelty, usability, and applicability.

4.5 Developing a Rubric to Evaluate the Generated Risks

Because no established framework exists for assessing the quality of generated risks, we developed a new rubric through a four-step process (Appendix C). First, three co-authors independently reviewed criteria used to evaluate ideas, scenarios, and socio-technical systems [14, 27, 29, 31, 40, 56, 93, 108], and proposed eight candidate dimensions. Second, we consolidated these proposals, and removed dimensions that could not be applied consistently across both agent-generated and human-ideated risks. This resulted in five core dimensions: *specificity*, *novelty*, *usability*, *actionability*, and *diversity*. Specificity draws on plausibility, connectivity, and uniqueness, building on scenario research [27, 29, 56]; novelty builds on creativity metrics in generative systems [31]; usability adapts two items from the System Usability Scale [14]; actionability adapts two items from the Data Quality Framework [108]; and diversity is operationalized through the PESTEL categories [40, 93].

Third, we translated each dimension into one or two rating items on a 5-point Likert scale (1 = "strongly disagree", 5 = "strongly agree"), and refined the wording for clarity and consistency. Finally, we piloted the rubric with 20 Prolific participants, who rated a mixed set of 25 agent-generated and human-ideated risks, and provided feedback that helped further refine the wording. Table 8 in Appendix C shows the final items for each dimension, along with the rationale for how they were adapted from prior work.

After the rubric was finalized and all risks had been collected, we used OpenAI's o3 model to assign each risk to a single PESTEL category (see the full classification prompt in Appendix C). This step involves applying a fixed coding scheme rather than exercising substantive judgment, and could therefore be performed by humans; however, automating it allowed us to apply the same criteria consistently across all risks while keeping judgments in human hands. We validated the model's classifications through a two-step process: first, a co-author independently annotated a random sample of 25 risks from the Chatbot Companion use case, yielding a weighted $F1$-score of 0.86; second, the full author team reviewed the model's rationales to ensure each label aligned with the meaning of the risk [22]. We then computed the normalized Shannon Diversity Index [97] over the PESTEL labels for each AI use case to assess the diversity of identified risks.

4.6 Using the Rubric to Evaluate the Generated Risks

To assess the quality of risks produced by our pipeline, we compared them with risks identified by humans. We first collected human-ideated risks through a Futures Wheel interface under two conditions (human only, and human-plus-AI condition). We then examined both agent-generated and human-identified risks across three complementary studies.

To generate human comparison data, we built a custom Futures Wheel interface (Figure 4) that guided participants through the production of first-, second-, and third-order consequences. The interface was implemented in two experimental conditions that reflect common ways humans engage in foresight tasks today. In the human-only condition, participants generated consequences without any assistance. In the human-plus-AI condition, participants could optionally request AI input through a chat window that helped explore potentially missing directions or through a button

that generated candidate consequences. We recruited through Prolific [85] 24 domain experts, and 24 laypeople for human-only condition and 18 domain experts and 18 laypeople for human-plus-AI condition, resulting in 84 participants in total. Their demographics are detailed in Appendix D, Tables 9–12. All participants completed the task individually, and their cascading consequences were processed using the same classification and deduplication steps applied to agent-generated consequences. The resulting human-ideated risks formed the input for Studies 2 and 3.

Study 1: Scoring *Agent*-Generated Risks with Domain *Experts* as Evaluators. To evaluate risks generated by in-silico agents, we conducted a large-scale annotation study with 170 domain experts recruited through Prolific [85], and sampled across five stakeholder cohorts: decision makers ($N = 39$), designers ($N = 38$), developers ($N = 46$), legal experts ($N = 33$), and healthcare professionals ($N = 14$). Healthcare experts were included specifically for the two health-related use cases (*Griefbot* and *Death App*). All participants were screened for relevant expertise, prior experience with AI, and familiarity with risk assessment practices, with comprehension and attention checks ensuring data quality (Appendix E reports the recruitment details). Each evaluator was randomly assigned between 12 and 16 risks, ensuring multiple independent assessments of every risk. In total, this yielded 2331 annotations across 110 unique agent-generated risks. Risks were presented via structured annotation cards (Figure 5), which displayed the risk statement, its potential impact, the focal use case, and the definition of systemic risk. Evaluators rated likelihood and severity on 5-point scales, indicated whether the risk should be considered systemic, and scored agreement with ten rubric dimensions. Evaluators were blinded to whether risks were AI-generated or human-identified.

Study 2: Scoring *Human*-Ideated Risks with Domain *Experts* as Evaluators. To evaluate risks generated through the Futures Wheel interface, we conducted a large-scale annotation study with an additional 120 domain experts recruited via Prolific [85]. We used the same screening protocol and sampled across the same five stakeholder cohorts as in Study 1: decision makers ($N = 28$), designers ($N = 25$), developers ($N = 23$), legal experts ($N = 28$), and healthcare professionals ($N = 16$). Each evaluator was assigned between 14 and 18 human-ideated risks, ensuring multiple independent assessments per item. Annotations were collected using the annotation cards (Figure 5), with evaluators blinded to whether risks originated from human-only or human-plus-AI ideation. This yielded 1030 annotations across 89 unique risks. Because fewer risks were evaluated than in Study 1, fewer evaluators were required to achieve comparable coverage per risk.

Study 3: Scoring Agent-Generated Risks and Human-Ideated Risks with Domain *Leaders* as Evaluators. To evaluate how domain leaders interpret and assess both agent-generated and human-ideated risks, we conducted a semi-structured interview and annotation study with seven leaders recruited through purposive sampling (4 female, 3 male; ages 27–60). The first two participants were identified via an internal mailing list at our organization, and each referred to additional external leaders. Their disciplinary backgrounds and relevance to the three speculative use cases are listed in Appendix E, Table 13. Each leader completed a 45–90 minute recorded session following a standardized protocol (Appendix E.3) that included a warm-up, briefing, vignette immersion, risk prioritization, gap analysis, and debriefing. After discussing and prioritizing risks, they annotated each one and identified missing, unclear, or underdeveloped risks. Following the interview, leaders also completed a follow-up annotation survey using the same rubric (Section 4.5). They were sent a link to the same structured annotation cards used in Studies 1 and 2, and rated both the agent-generated and human-ideated risks. As in the expert studies, leaders were blinded to the origin of each risk.

We then conducted a qualitative analysis of the interview recordings and transcripts. This included leaders' vignette responses, risk discussions, prioritization decisions, and gap analyses. Two authors thematically analyzed these materials following an inductive approach [13, 68, 69, 91]. The authors used Figma [35] to collaboratively create affinity diagrams based on leaders' responses. Over the course of 4 meetings, totaling 10 hours, they discussed and resolved any disagreements that arose during the coding and theme refinement process. The final themes describe how leaders contextualized, reframed, or extended the risks, revealing differences in emphasis and diversity across agent-generated and human-identified risks.

5 Results

Next, we discuss our results in two parts. First, we examine the output of the in-silico risk generation pipeline across four AI use cases, assessing whether the generated risks are sufficiently systemic, plausible, specific, diverse, and actionable to support anticipatory decision-making (RQ1). Second, we compare these risks to those ideated by human participants across two conditions (human-only, and human–plus-AI), analyzing differences in diversity, tone, emotional depth, and systemic framing (RQ2).

5.1 RQ1: Can in-silico agents generate systemic risks of sufficient quality to support foresight?

5.1.1 Quantitative Analysis. We found that agent-based Futures Wheels generate (Figure 6, Tables 21–31 in Appendix G):

Numerous risks. Across all 30 wheels, Step 1 (Figure 3) generated an average of 86 consequences for the AI Toy, and 110 for the Griefbot. After Step 2 (risk classification), these were reduced to 32 and 58 risks, which Step 3 (deduplication) further condensed into 27 and 47 unique systemic risks, respectively. Appendix B, Figure 9, shows the number of items at each step for each AI use case, and Figure 10 the resources needed to generate them. Tables 2, 3, 4, and 5 list the representative, deduplicated systemic risks generated by the agents for four use cases: Chatbot Companion (n=21), AI Toy (n=27), Griefbot (n=31), and Death App (n=24). Items are phrased at societal/institutional scale, and often expressed as cascades; the symbol "→" denotes downstream effects. These lists are illustrative rather than exhaustive, and provide a consolidated basis our qualitative analysis. We see that, across cases, the lists enumerate systemic, cascading risks that operate at societal and institutional scales. For Chatbot Companion (Table 2), items progress from health-system strain (e.g., "diminished mental health treatment effectiveness") to macro-social outcomes (e.g., "widespread social unrest and instability", "pervasive disruptions in international digital e-commerce").

For the AI Toy (Table 3), risks center on child development and equity (e.g., "critical thinking skills decline systemically", "large-scale data exploitation targeting minors", "lifelong profiling and discrimination begin in childhood"). The Griefbot list (Table 4) is notable for legal-cultural rupture (e.g., "legal and moral frameworks governing personal identity collapse", "grief is exploited as a marketable asset", "judiciary systems face overload from AI-avatar litigation"). Finally, the Death App (Table 5) shifts to public-health and governance shocks (e.g., "mass public health crises from uncontrolled harmful interventions", "vulnerable populations face systemic coercion"). Interestingly, even for mature systems, the agent-generated risks reach institutional instability, while more speculative cases add geopolitical failure modes.

Risks on specific topics. PESTEL classification showed Social as the most prominent category across all four AI use cases, ranging from 42% of the generated risks being (Death App) to 78% (AI Toy). Legal followed, with 11% (AI Toy) to 34% (Griefbot). Political risks were either absent or limited to one for the first three use cases: the Death App had indeed 5 (19%). Appendix G, Table 21, shows the full PESTEL breakdown for agent-generated risks.

Systemic risks for any TRL. Across all four AI use cases, domain leaders judged about 75% of the risks to be systemic, while domain experts judged 93% as systemic. Among the leader ratings, the lowest share was for Chatbot Companion (72%), and the highest for AI Toy (85%). On a five-point scale, evaluators rated the risks as at least "likely" to occur (mean likelihood score of 3.57 or above), and "serious" in impact (mean severity score of 3.51 or above), as shown in Figure 6. Likelihood was highest for AI Toy (3.84), and lowest for Griefbot (3.57). Severity was highest for AI Toy and Death App (both 3.70), and lowest for Chatbot Companion (3.51).

Novel risks, even for low TRL. Griefbot risks scored highest on all three novelty subdimensions: 2.99 (novelty), 3.24 (originality), and 3.08 (rarity). In contrast, Chatbot Companion risks were rated least novel across the board (2.64, 2.94, and 2.79). Risks generated for the AI Toy scored highest on use case connectivity (3.92), plausibility (3.81), and uniqueness (3.02) (Figure 6). In contrast, Death App risks were rated least plausible (3.73), and least specific (2.78). AI Toy risks were easiest to engage with (3.81), while Chatbot Companion risks required the least additional information to be fully understood (3.32). Death App risks were seen as hardest to engage with (3.48), and required the most explanation (need to learn: 3.59).

Hard-to-use risks at times, but policy-relevant. We found that generated risks are harder to readily interpret as the AI use cases become more speculative, but, paradoxically, they become more relevant for policy discussions, especially in lower-TRL scenarios like the Death App.

5.1.2 Qualitative Analysis. While the quantitative results show that the agents can generate systemic risks at scale, the qualitative analyses presented next complement this by revealing how domain leaders judged the systemic nature of the risks, how they assessed their comprehensiveness, and how they complemented the agents with risks grounded in lived experience. Expert quotes are referenced using E_N, corresponding to their anonymized ID.

Domain leaders judged most agent-generated risks to be truly systemic, even for speculative use cases. For the AI Toy (TRL 7), agreement was high: 70–78% of risks were rated systemic, and 48–79% were placed in the *critical risk zone*, the upper-right quadrant where likelihood and impact are both high (as evidenced in expert E2's risk prioritization grid in Appendix E, Figure 13). For the Griefbot (TRL 5), views diverged more (53% *vs.* 72% *vs.* 78%), with 43–71% still assigned to the critical risk zone though (as shown in expert E4's risk prioritization grid in Appendix E, Figure 14). The most surprising case was the Death App (TRL 2): despite low technological readiness, domain leaders marked 58% and 92% of risks as systemic, with 33–73% in the critical risk zone (as shown in expert E6's risk prioritization grid in Appendix E, Figure 15). These findings indicate that the agents surfaced systemic risks across readiness levels (including low-TRL domains where uncertainty is greatest), and that many of these risks were judged both likely and consequential.

The leaders set aside some of the generated risks because they did not perceive them as systemic, and did so for five main reasons. First, some were too generic, applying to "any AI rather than the specific use case". As E3 put it, "Data breaches happen all the time. Does it actually connect specifically to this case? No, it's anything with AI". Second, some were judged as not plausibly caused by the technology itself. In the griefbot case, E6 asked whether "a collapse of social cohesion" could really be traced back to griefbots, concluding that "the systemic risk has to be caused by the creation of griefbots as a widespread phenomenon". E2 made a similar point about suicide rates, noting that while tragic, it would be "a stretch to say that like this type of companies are causing these situations". Third, some risks were considered too vague or poorly phrased such as "changes in family values", which E7 felt could mean almost anything. Fourth, some were seen as overly individual rather than systemic. E1, for example, dismissed social skill decline among children as "not a systemic risk: life will find a way and there will be a contra movement", while E7 argued that "family-level conflicts over assisted dying did not automatically rise to systemic scale without evidence of wider disruption". Fifth, some were excluded because they appeared more positive than negative. E6, for instance, set aside the expansion of international health services on the grounds that they "don't see how that can be a negative".

Domain leaders judged the systemic risks generated by the agents as comprehensive. They described the list as broad and often exceeding what they would have anticipated themselves (E5: "There were a few risks that I never thought about in this way"). In terms of coverage, leader confirmed that the agents reliably surfaced the generalized systemic dynamics they expected to see across domains such as bias amplification in sensitive settings like education and healthcare, emotional harms and psychosocial spillovers in grief technologies and child-facing systems, and dependence or erosion of skills when AI substitutes for human judgment or learning. In-silico agents also articulated second-order effects (e.g., erosion of democratic processes, loss of human oversight) that resonated with leader's own systemic framing. Several noted that reviewing a pre-generated list reduced the "blank-page problem", allowing them to focus their energy on filtering, prioritizing, and contextualizing risks. These results suggest that the agents not only provided stimulus for reflection but also captured the backbone of systemic risks leaders already recognized.

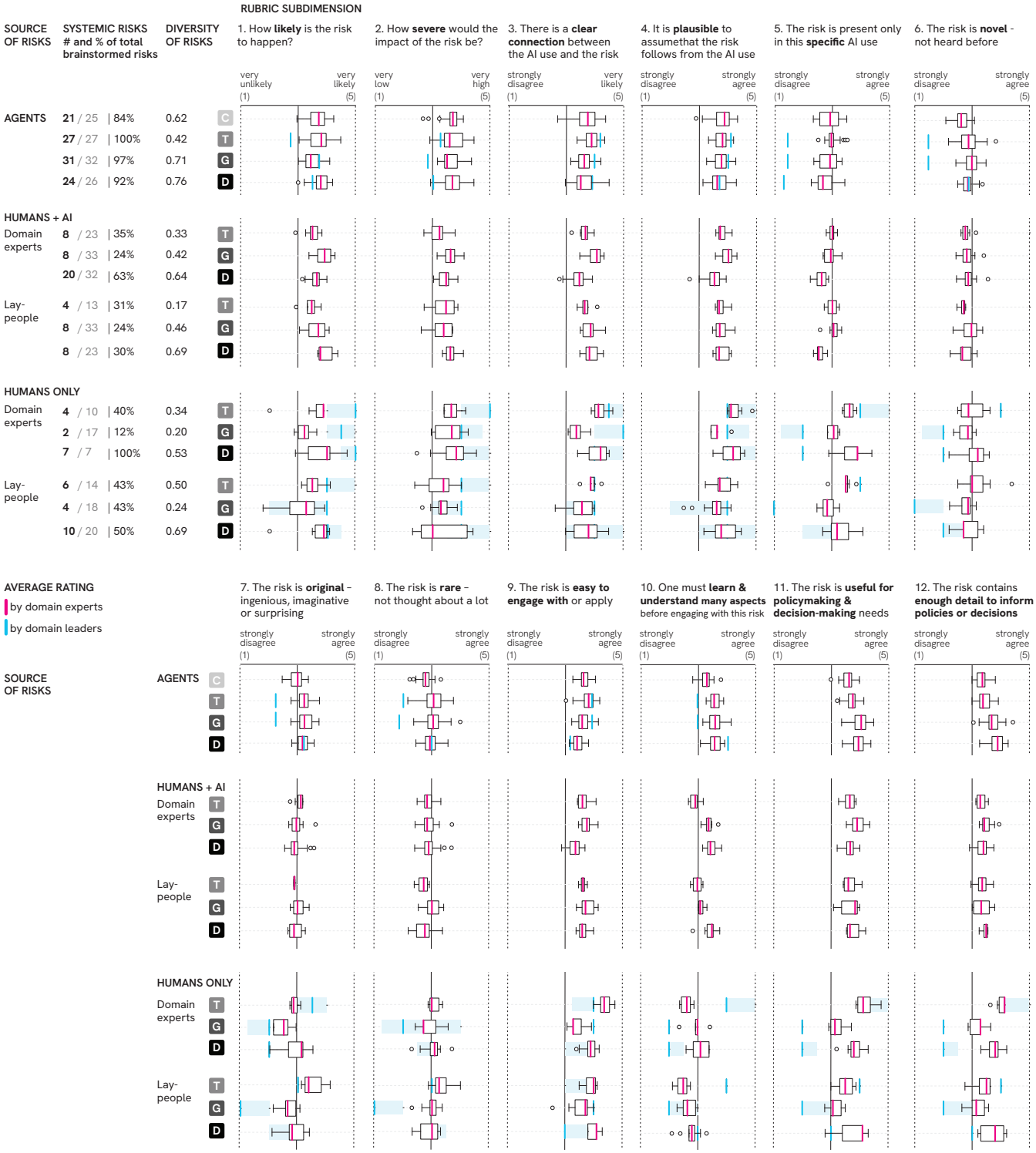

Figure 6: Ratings of systemic risks generated by in-silico agents and humans across AI use cases and quality subdimensions. In-silico agents produced substantially more systemic risks than humans, who generated comparatively few. Boxplots with mean scores are shown for risks associated with four AI use cases: Chatbot Companion (C, TRL 9), AI Toy (T, TRL 7), Griefbot (G, TRL 5), and Death App (D, TRL 2). Each colored bar reflects the average domain expert rating (pink bar) or domain leader rating (blue bar) on a 5-point Likert scale across ten evaluation subdimensions from our rubric developed in Section 4.5. Domain experts generally rated the risks generated by in-silico agents as connected to the AI use, plausible, and moderately usable. Dimensions such as specificity, novelty, and originality received more varied ratings across use cases, especially for less mature uses such as the Death App and Griefbot. Appendix G (Tables 26–31) reports the statistical tests and quantitative comparisons.

Table 2: List of systemic risks ($n = 21$) generated by in-silico agents for the Chatbot Companion use case through our pipeline. Each risk is presented with an arrow → leading to the resulting significant impact (as shown in Step 1 of Figure 3).

ID	Systemic risk leading to → systemic impact
1	Overreliance on AI support reduces human-centered care quality → diminished mental health treatment effectiveness in society
2	Social and interpersonal communication skills weaken across populations → reduced collaboration and increased social fragmentation
3	Availability of professional mental health expertise drops sharply → overwhelmed healthcare systems during crisis events
4	Loss of clinical experience hampers holistic mental health care → fragmented treatment approaches internationally
5	Confidential mental health information is frequently exposed → widespread loss of trust in digital mental health solutions
6	Market entry barriers increase substantially → slower deployment of innovative mental health AI solutions worldwide
7	Traditional social support networks weaken globally → higher vulnerability to health crises and greater reliance on technology-driven services
8	Population-wide physical and mental health deteriorates significantly → increased healthcare burdens and economic costs
9	Social fragmentation undermines collective action → weakened societal stability
10	Widespread mental health decline burdens healthcare systems → increased public health crises
11	Populations become less capable of managing stress independently → greater societal vulnerability during crises
12	Intergenerational divides intensify → reduced social cohesion across communities
13	Mental healthcare quality deteriorates widely → extended suffering and inefficiencies in health services
14	Trust in mental health systems collapses nationally → decreased care-seeking behavior
15	Consumer protections fail at scale → massive economic losses and reduced market confidence
16	Adoption of digital health innovations stalls globally → slowed progress in mental healthcare delivery
17	Cybersecurity threats multiply rapidly → pervasive disruptions in international digital e-commerce
18	Societal inequalities deepen significantly → widespread social unrest and instability
19	Future generations exhibit reduced social competencies → long-term societal and economic challenges worldwide
20	Reliance on AI emotional support systems becomes entrenched globally → heightened risks for social isolation and mental health vulnerability
21	Mental health emergencies rise globally → increased burdens on healthcare systems and public safety sectors

Table 3: List of systemic risks ($n = 27$) generated by in-silico agents for the AI Toy use case through our pipeline. Each risk is presented with an arrow → leading to the resulting significant impact (as shown in Step 1 of Figure 3).

ID	Systemic risk leading to → systemic impact
1	Unequal STEM skill distribution is reinforced across societies → reduced global workforce competitiveness and innovation equity
2	Marginalized populations face entrenched educational exclusion → increased socioeconomic divides and broad societal tensions
3	Unverified information is widely accepted unquestioningly → erosion of democratic discourse
4	Critical thinking skills decline systemically among youth → diminished innovation and weaker democratic participation
5	Social skill deficits become widespread among children → increased social isolation and long-term challenges for social cohesion
6	Emotional intelligence development is impaired broadly → rising societal conflict and decreased community trust
7	Population-wide relational dysfunctions become pervasive → increased burden on mental health and public health systems
8	Parental and family-based educational support systems weaken significantly → greater reliance on external institutions
9	Gaps in early educational interventions increase widely → higher youth dropout rates and workforce shortages
10	Child developmental monitoring becomes ineffective at scale → systemic inefficiencies in education policy and resource allocation
11	Shallow knowledge acquisition becomes widespread among learners → diminished innovation capacity in knowledge-intensive industries
12	Critical analytical skills are systematically underdeveloped → reduced national problem-solving capabilities
13	Misalignment of educational content with learner needs disrupts schooling effectiveness at scale → increased inequities
14	Education systems become fragmented globally → instability in qualification standards and cross-border employment challenges
15	False knowledge and misinformation become normalized among youth → poorer public health decisions and destabilized social trust
16	Bias propagation in AI content spreads widely → discriminatory practices embedded in workforce and institutions
17	Dependence on AI for cognitive tasks rises system-wide → reduced innovation and adaptability in global markets
18	Skill development deficiencies in critical reasoning become pervasive → lower productivity in knowledge economies
19	Social isolation among children is widespread → decreased societal cohesion and increased mental health issues
20	Developmental delays and emotional difficulties are prevalent → long-term societal costs in healthcare and education
21	Trust in educational institutions collapses widely → fragmented education markets and uneven skill development globally
22	Educational monopolies distort knowledge dissemination → international market imbalances and cultural homogenization
23	Widespread cultural myopia is entrenched early in life → international conflicts and weakened global cooperation
24	Psychological crises linked to AI withdrawal are widespread → increased demand on health services and societal instability
25	Large-scale data exploitation targeting minors is normalized → eroded privacy rights and reduced trust in technology
26	Lifelong profiling and discrimination of individuals begin in childhood → systemic inequality and restricted mobility
27	Traditional educational employment sectors contract → market consolidation and global shifts in labor markets

Table 4: List of systemic risks ($n = 31$) generated by in-silico agents for the Griefbot use case through our pipeline. Each risk is presented with an arrow → leading to the resulting significant impact (as shown in Step 1 of Figure 3).

ID	Risk text
1	Mental health systems face overwhelming demand and widespread strain → reduced quality of care along with decreased public health resilience
2	Productivity in workplaces declines due to higher rates of psychological distress → significant economic losses in markets
3	Large-scale social fragmentation and weakened cultural cohesion occur → increased mental health crises and reduced societal stability
4	Family support networks diminish substantially → rise in social isolation affecting population mental health at large
5	Mental health systems face increased demand for new therapies and social services → greater pressure on healthcare infrastructure
6	Mental health care costs escalate internationally → resource allocation challenges in health policy and budgeting
7	Consumer trust and mass trust in digital services deteriorate widely → decreased engagement, economic contraction in AI sectors
8	Regulatory compliance costs surge sharply for international companies → constrained innovation and reduced AI competitiveness
9	Fragmented data jurisdiction laws and legal uncertainty disrupt international digital economy operations → friction in global trade and investment
10	Protracted legal conflicts impose heavy costs on tech companies → market volatility and investor wariness
11	Consumer protection crises emerge internationally → heightened regulatory intervention and market distrust
12	Specialized mental health services expand internationally → increased public health expenditures and resource reallocation
13	Mental health treatment efficacy declines broadly → longer recovery times and higher societal healthcare costs
14	International legal frameworks are destabilized by competing claims → prolonged litigation and uncertainty in data governance
15	Global AI technology adoption is severely hindered → reduced innovation and fractured international cooperation
16	Social isolation and entrenched loneliness increase among large segments of society → widespread public health challenges and increased mortality
17	Cybercrime and fraud involving AI avatars become endemic → escalated public safety risks and weakened trust in digital ecosystems
18	Legal and moral frameworks governing personal identity collapse → deep societal conflict and regulatory chaos
19	Grief is exploited as a marketable asset on a global scale → ethical erosion and significant emotional harm across societies
20	Mental health systems are overburdened by AI-induced dependency cases → widespread challenges in public health management
21	Social relationships increasingly rely on artificial interactions → degradation of interpersonal social skills at scale
22	Traditional mourning rituals lose legal and cultural recognition → fragmentation of societal cohesion across cultures
23	Cultural homogenization occurs due to techno-centric mourning practices → loss of cultural diversity in international societies
24	Legal services related to digital legacy surge worldwide → increased legal market specialization and cross-border disputes
25	Large-scale data breaches compromise personal information globally → significant threats to public security and trust in markets
26	Judiciary systems worldwide face overload from AI avatar litigation → delays in justice and increased public distrust in legal institutions
27	Emotional dependence on AI systems becomes widespread → significant challenges in regulating psychological impacts
28	Conventional community-based mental health services face disruption globally → shifts in how social support networks operate across societies
29	Familial conflicts related to AI cause social fragmentation at scale → increased social tension and challenges to social cohesion
30	Data protection laws face widespread challenges → global legal conflicts and regulatory fragmentation
31	AI systems exhibit inconsistent behaviors across user bases → risks in trust and safety in international digital ecosystems

Table 5: List of systemic risks ($n = 24$) generated by in-silico agents for the Death App use case through our pipeline. Each risk is presented with an arrow → leading to the resulting significant impact (as shown in Step 1 of Figure 3).

ID	Risk text
1	Public health systems face increased demand and strain → overwhelmed healthcare infrastructure
2	Healthcare systems experience resource reallocation pressures → reduced quality of care overall
3	Fragmented and inconsistent international regulations disrupt cooperation → increased illegal activities and market uncertainty
4	Social instability hampers economic activity → decreased investor confidence in affected markets
5	Market consolidation occurs → reduced competition and innovation in assisted death services
6	Therapeutic relationships weaken broadly → worsened mental health outcomes in affected populations
7	Ideological polarization deepens across societies → fragmented social cohesion and political instability
8	Widespread cybersecurity failures in sensitive health sectors occur → compromised public safety and urgency for regulation
9	Criminal exploitation undermines trust in AI health services → demands for stringent oversight and legal enforcement
10	Mass public health crises arise from uncontrolled harmful interventions → widespread loss of life and international outrage
11	Global health service market faces regulatory crackdowns and legal turmoil → destabilized market confidence and fragmented standards
12	Trust in AI-enabled health platforms collapses globally → widescale rejection of AI interventions in critical services
13	Societal mental health deteriorates broadly at scale → increasing suicide rates and overwhelming public health systems
14	Vulnerable populations face systemic coercion pressures → erosion of fundamental human rights protections worldwide
15	Widespread privacy violations cause international data security crises → reduced trust and investment in digital health technologies
16	Stagnation of digital health innovation occurs internationally → slowed progress in healthcare improvements
17	Coordinated exploitation campaigns cause widespread social harm → increased global insecurity and destabilization
18	Rising extremist influence fractures international relations → increased geopolitical conflicts and market volatility
19	Global mental health crises worsen profoundly → substantial economic losses and reduced workforce productivity
20	Ethical decay spreads widely through societies → deteriorated social cohesion and increased risk of unrest
21	National healthcare priorities are reorganized → international tensions over care standards
22	Public opinion drives legislative changes internationally → polarized markets and policy instability
23	International legal frameworks around AI and assisted dying harmonize unevenly → regulatory fragmentation in cross-border services
24	Accelerated legislative reforms create uneven legal landscapes → challenges in international healthcare service provision

Domain leaders complemented the agents with risks grounded in lived experience. Across the three use cases, leaders introduced 19 additional risks in total (6 for the AI Toy, 6 for the Griefbot, and 7 for the Death App). For the AI Toy, leaders emphasized developmental distinctions across educational systems, with E1 noting, "Children from 3 to 12 are completely different people and represent completely different audiences from a design point of view". They also cautioned against potential cultural backlash: "Maybe there's a cultural risk. Will some cultures push back against girls becoming enthusiastic about STEM through the toy?". E2 warned of systemic child protection gaps if disclosures of abuse were made only to the toy: "What if the child tells the toy something serious, like someone is harming them? Will it ever reach the school or the police?". Finally, both leaders added that mass attachment to commercial toys could create a new form of consumer-driven grief when products are discontinued, and that widespread adoption could contribute to global environmental burdens through e-waste.

In the Griefbot case, leaders pointed to institutional and cultural shifts overlooked by the agents. E5 stressed that griefbots might take over the logistics of dying: "How do I arrange the funeral in this country? What are the legal steps?". These tasks have historically been managed by families, religious communities, or professional services. This displacement could reconfigure how societies organize death and mourning. Others emphasized that digital companions may undermine human help-seeking: "You need to rely on your network: people can't heal for you, but they can help you" (E5). Leaders also raised the risk of harmful new "cultures of grief", where AI-mediated rituals normalize shallow or exploitative practices, and E6 highlighted design-level unpredictability, describing griefbots as "the curse of flexibility, it can be used in so many different ways that you can't really foresee the outcomes".

In the Death App scenario, leaders added political-economic and cultural risks that the model had not surfaced. E7 highlighted the proliferation of unregulated provision: "If it's not opening a black market, like a completely illegal one, it can open a grey market with pharmaceuticals". They also warned of state misuse where local governments could weaponize platforms to pressure vulnerable groups, and of a possible normalization of eugenics, if such technologies were deployed unethically. Gendered harms were also raised, with leaders questioning how such platforms might reproduce or exacerbate existing inequalities. For instance, women could be "disproportionately steered toward assisted dying in contexts where access to healthcare and social support is limited". Finally, leaders suggested that assisted dying could become institutionalized as a distinct "death industry" separate from healthcare, creating new boundaries and weakening existing oversight mechanisms.

5.2 RQ2: How do agent-generated risks compare with human-ideated ones?

5.2.1 Quantitative Analysis. In comparison to those generated by agents, human-ideated risks were (Figure 6, Tables 21–31 in Appendix G):

Not many. Without AI assistance, humans understandably generated fewer unique risks than agents (Appendix F, Tables 14, 15, 16, 17, 18, 19, and 20). Domain experts generated 7-17 risks per use case, and laypeople generated 14-20 across all cases. More interestingly, with AI assistance, participants matched or exceeded the pipeline in volume: domain experts generated 23–32 risks per case, and laypeople generated 13–23.

Narrowly focused. Human risks cluster narrowly around what feels directly impactful: overwhelmingly social in the case of Chatbot Companion and Griefbot (86% and 89%, respectively), and heavily legal for Death App (45%), where moral and personal stakes are highest. Political and economic concerns are absent from human responses entirely. This suggests that humans filter systemic threats through personal salience [95], interpreting policy-level risk only when it intersects with perceived bodily or existential vulnerability. Appendix G, Tables 21–25 show the full table of PESTEL classifications for the risks generated by agents and ideated by humans, both with and without using AI .

Partly systemic. Only a fraction of risks were judged to be systemic. With AI assistance, this ranged from 43% (laypeople and domain experts, griefbot) to 63% (domain experts, death app). In the human-only condition, systemic coverage was lower, from just 12% (domain experts, griefbot) to 43% (laypeople, toy and griefbot), reaching 100% only in one case (domain experts, death app).

Neither novel nor original. Human risks scored consistently low on novelty and originality, with average ratings in the lower range (1-2.3 out of 5) across most use cases, but were judged more likely to occur and more severe in their potential impact.

5.2.2 Qualitative Analysis. Our analysis shows that agents, leaders, and laypeople frame systemic risks in different ways, diverging in emotional depth, interpretive lens, and uncertainty avoidance.

Emotional depth of systemic risks diverges: agents bureaucratize, leaders contextualize, laypeople dramatize. Agent risks adopt managerial language (e.g., "regulatory fragmentation", "compliance costs", "market volatility") that abstracts away from lived suffering. Expert contributions weave emotion into culture ("Dependency on fragile products could traumatize children", "Talking about death declines"). Laypeople contributions are blunt and affectively charged, capturing immediate fear or outrage ("Individuals may lose the ability to interact with reality effectively", "People could lose their jobs"). These differences in register shape how each human imagines the seriousness and felt impact of systemic risks.

Systemic risks are made sense of through different lenses: agents scale up, leaders translate, laypeople scale down. Agents construct risk as structural malfunction ("Judiciary systems worldwide face overload", "Global AI technology adoption is severely hindered"). Leaders act as translators, showing how these failures ripple into norms and institutions ("Talking about death declines", "Generational miscommunication grows"). Laypeople invert the scale, interpreting systemic risks as personal threats ("The app could be misused against unwilling individuals", "Individuals are put on lists or lose their jobs'). For each group, the meaning of risk begins and ends at a different level of experience.

Uncertainty is navigated through different strategies: agents contain it, leaders interrogate it, laypeople absorb it. Agents frame uncertainty as a variable to be managed, where terms like "slower deployment", "reduced market confidence", or "legal uncertainty" suggest problems solvable through better design or governance. Leaders highlight uncertainty as a structural feature of emerging technologies ("Curse of flexibility spreads", "Systemic framings miss lived realities"), not something to eliminate, but something that complicates meaning and control. Laypeople do not manage or analyze uncertainty, but they live inside it: "People could die", "I wouldn't know what's real". For each group, uncertainty reflects not just what is unknown, but how the unknown is cognitively, emotionally, and socially processed.

6 Discussion

We begin by consolidating our findings on agent-based foresight and its alignment with prior work on systemic risks (§6.1). We then examine how our results extend beyond prior work (§6.2). Finally, we outline implications for hybrid workflows that integrate agents, domain leaders, and laypeople perspectives in systemic risk assessment (§6.3).

6.1 In-line with Previous Literature

We found that agent-based foresight consistently broadened the range of systemic risks identified. In our study, agents generated a higher volume of cascading consequences across all four use cases, confirming prior claims that AI support is most effective in expanding the search space of possible outcomes rather than deepening contextual analysis [21, 45, 109]. Expert evaluation revealed that, while many of these agent-generated risks required refinement, a majority were judged systemic, plausible, and severe; findings that resonate with Ehsan et al. [32] argument that systemic harms, once surfaced, persist in decision-making even when the underlying technology is speculative. Moreover, our results resonate with Jain et al. [53] proposal of algorithmic pluralism, which cautions against systemic bottlenecks and monocultures: our hybrid foresight workflow similarly emphasizes diversity of perspectives (through multiple agents) as a safeguard against narrowing risk imaginaries.

6.2 Contributions to Previous Literature

Our results also challenge and extend prior accounts of AI risk assessment. Whereas Uuk et al. [107] emphasize the need of taxonomies of systemic risks, our findings demonstrate that generative exploration with agents can surface novel risks that have never been categorized before, even for technologies at very low TRLs. For instance, while chatbot companions (TRL 9) primarily yielded risks already well-documented in the literature such as overreliance on AI for social support [58], our foresight pipeline uncovered previously underexplored risks for speculative systems like griefbots (TRL 5), and death apps (TRL 2), including cascading impacts on healthcare ethics, and the normalization of assisted dying platforms. This contrast suggests that foresight does not have to wait for systems to reach maturity, countering the assumption in prior literature that systemic risks become tractable only once technologies are widely deployed. Finally, expert interviews highlighted that plausibility judgments, absent from much prior taxonomic work, are critical in filtering speculative risks. By embedding plausibility as an explicit evaluation criterion, our approach adds a new step to systemic risk workflows, producing evidence that a hybrid division of labor (agents for breadth, experts for judgment) can extend existing risk ideation methods.

6.3 Implications for Hybrid Workflows in Systemic Risk Assessment

Our findings have three main implications: (1) for how agent and human contributions complement each other in systemic risk generation, (2) for how hybrid agent–human risk generation should be structured to produce valid foresight, and (3) for how practitioners should decide when and how much to automate this process.

Integrating complementary agent and human contributions. Our findings show that combining agent-generated and human-ideated risks can help overcome inherent cognitive barriers such as scope neglect and fixation on familiar narratives [87]. Agents consistently framed risks in terms of institutional change, policy implications, and macro-level societal trends. By contrast, humans surfaced risks reflecting emotional depth, cultural nuance, and context-specific harms that agents frequently missed. In addition, humans provided concrete examples that grounded agent-identified risks in specific domains and lived contexts. For example, agents flagged "regulatory fragmentation" as a risk in both child protection and mourning contexts. Humans pointed to concrete cases such as AI toys being regulated as consumer products in one country but subject to child safeguarding obligations in another, or griefbots being deployed as wellbeing tools despite unclear rules about consent of the deceased [50].

Structuring hybrid risk generation for valid foresight. Drawing on leaders' reflections, we identify three directions for improving the validity of hybrid foresight pipelines.

First, pipelines should start from lived, micro-level experiences (e.g., how a child might engage with an AI toy, how a grieving family might rely on a griefbot), and then cascade upward to uncover how these individual or community-level harms accumulate into systemic risks. Beginning only at the system level may overlook how small, situated harms add up over time.

Second, pipelines should be designed with intersectionality theory in mind. In-silico agents should be prompted to surface risks that reflect overlapping axes of vulnerability such as gender, class, and ethnicity [12]. Together, these steps would ensure that foresight captures not just abstract systemic shifts, but also the layered ways in which lived experiences scale into broader disruptions.

Third, pipelines should be used alongside longitudinal, scenario-based, or discourse-analytic methods to interpret how short-term risks may develop into systemic disruptions over long time horizons of 5–25 years. These methods support reasoning across time by connecting present-day uses of AI to earlier technology transitions. For example, current debates around GenAI adoption echo discussions from the 1990s surrounding the widespread adoption of design software such as Photoshop [52]. Short-term concerns about image manipulation later developed into the systemic risk of lost trust in journalism, leading to the creation of forensic verification teams in news production, especially in war zones [4].

Deciding when and how to automate risk generation. Our findings suggest that the choice between fully automated and hybrid risk generation is best understood as a set of trade-offs rather than a binary. To navigate these, we propose three practical heuristics.

First, *automate for breadth*. Fully automated pipelines are especially useful during early-stage ideation, horizon scanning, or exploratory analysis of emerging technologies, where the primary goal is to expand the search space of possible risks. In these contexts, gains in efficiency and coverage may outweigh the absence of direct human involvement.

Second, *retain human involvement for sensemaking*. While automation increases breadth, it often reduces contextual grounding. Domain leaders emphasized that in many real-world settings, risk generation is not merely an information task. It is also a sensemaking process through which stakeholders surface tacit assumptions, align perspectives, and build ownership of downstream decisions. This process is often interdisciplinary and highly domain-specific, shaped by situated expertise from fields such as medical ethics or the anthropology of technology. In such settings, the absence of human engagement can carry significant costs. As our own work showed, certain risks become visible only through expert judgment.

Third, *use hybrid workflows when both breadth and realism matter*. Hybrid approaches are most effective when risks must be surfaced comprehensively, yet still interpreted and filtered through human judgment. Strategic planning and regulatory readiness are two such examples. In these cases, plausibility filtering and cultural contextualization must remain human-led. Our adaptation of the Futures Wheel supports this division of labor by organizing agent-generated risks into layered causal pathways for iterative refinement. Domain leaders found this structure helpful in reducing the blank-page burden, while preserving space for expert interpretation.

6.4 Limitations and Future Work

Our study and the proposed approach have six main limitations that suggest directions for future research. First, we adopted the EU AI Act's definition of systemic risk, which targets general-purpose AI models [77, 81]. This gave evaluators a clear, policy-relevant anchor for a vague concept, and helped ensure consistency. However, priming evaluators with this framing may have narrowed their perspective. Future work should test alternative definitions of systemic risk, including approaches based on human rights that assess not only the scale and probability of harms, but also their scope, and reversibility [23]. These dimensions could offer a more nuanced view of how systemic risks affect different groups.

Second, our analysis focused on use-level risks and did not account for systemic risks that stem from underlying model capabilities. As recent work and regulation highlight, highly capable general-purpose models may introduce systemic risks independent of how they are applied, including governance failures and cross-sector disruptions [77, 107]. The same application may pose very different risks when powered by a frontier model versus a weaker one. To address this, future work could extend our pipeline by simulating capability evolution by conditioning agent prompts on projected model abilities such as reasoning or autonomy.

Third, the foresight process began with in-silico agents generating risks, which may have limited both the emotional depth, and originality of the outputs. While domain experts and leaders found many of the risks plausible, they often lacked the kind of intuitive or affective nuance that emerges from lived experience. Future work should explore hybrid workflows where human experts seed early ideas, and agents elaborate them; or it should explore explicit prompts asking agents to consider emotional impact or speculative "wild cards" [40]. These strategies may yield outputs that are both creative and grounded in context. Our findings also highlight the importance of iteration, both in foresight pipelines and in discussion practices. Generating systemic risks through in-silico agents is not a one-off task: repeated runs across different prompts, personas, and stakeholder contexts may reveal additional layers of systemic risk, or refine earlier outputs. Likewise, hybrid workflows will likely benefit from iterative cycles in which agents generate breadth, experts contextualize and critique, and subsequent agent runs refine scope based on this feedback.

Fourth, the in-silico agents showed a topical bias in the risks they produced. Across use cases, agents tended to concentrate on social, ethical, or legal consequences, while generating few technical risks and none related to environmental impact. This likely reflects both the structure of the prompts and potential biases in the model's training data. While socially framed risks are important, this imbalance may lead to blind spots. This imbalance may itself introduce new risks. By embedding AI into foresight, we risk a form of "meta-systemic risk" where the very tools meant to help anticipate harms could normalize or amplify certain framings of risk at the expense of others. For example, our pipeline produced many risks framed in managerial or regulatory language, which might inadvertently steer attention toward technocratic governance while sidelining lived experiences. Future work should critically examine whether agent-supported foresight unintentionally narrows the discursive space, even as it expands the breadth of candidate risks. Future work should explore three directions: (1) prompt strategies that explicitly request diverse categories of risk (e.g., using PESTEL dimensions [40, 92] or systemic risk taxonomies [107]); (2) retrieval-augmented generation (RAG) to ground LLMs in relevant technical documentation (e.g., environmental data and standards) [88]; and (3) distinct agent roles designed to surface risks from underrepresented domains. These could include personas of different professional backgrounds (e.g., policymaker, activist, engineer, healthcare worker), lived experiences (e.g., caregiver, marginalized community member), and stances (e.g., skeptic, advocate, regulator), each of which may highlight different types of risk that are otherwise overlooked.

Fifth, our evaluation was based on a small set of AI use cases and best performing models. While the use cases varied in readiness and domain, they might not reflect the full diversity of AI applications. Future studies should include a wider range of AI systems such as those used in infrastructure, industrial, or low-resource contexts.

Sixth, while we primarily used proprietary models, key components of the pipeline were also tested with open-source models and showed comparable performance. This suggests that our pipeline is model-agnostic and suitable for local, protected deployment, especially in privacy-sensitive contexts. Since all data is synthetic, privacy was not a concern in this study, but local use remains a viable option.

6.5 Ethical Considerations

This study received ethics approval through our organisation's internal research ethics process. All participants involved in expert interviews and evaluations gave informed consent, and participation was voluntary and anonymous. No personally identifiable information was collected.

Beyond procedural safeguards, the study raised four broader ethical considerations. First, transparency in how risks are generated and framed is essential to avoid over-reliance on in-silico foresight. Without clear context, there is a risk that polished outputs may be mistaken for validated assessments. To mitigate this, we ensured expert review of all outputs, and positioned the tool as augmenting (not replacing) human foresight and ethical deliberation.

Second, in-silico agents may reflect and amplify biases embedded in their training data, particularly in how they frame which risks are plausible or important. We observed this in the concentration of outputs around social and legal domains, with other dimensions underrepresented. To mitigate this, we experimented with prompt diversification by varying the agents' assigned attitudes toward AI. We manually selected a range of adjectives such as alarmed, skeptical, overwhelmed, curious, cautiously optimistic, and enthusiastic, to represent a spectrum of perspectives. This aimed at introducing variation in how risks were expressed. Future work could explore additional structured prompting strategies such as domain-specific prompts to further broaden the coverage of risk types.

Third, the tool may surface emotionally sensitive risks. To reduce harm, we included content warnings in the user interface for emotionally charged scenarios (i.e., the Griefbot and Death App use cases), and we briefed domain experts and annotators in advance when working with such material.

Fourth, delegating cognitive work to AI introduces deeper epistemic and normative risks that go beyond bias or framing. Prior work highlights that over-reliance on AI for ideation can negatively affect human creativity, reduce exploratory thinking, and accelerate skill erosion in tasks that require judgment or imagination. In our context, easily accessible AI suggestions may constrain the breadth of human brainstorming, nudging participants toward patterns present in the training data rather than encouraging divergent thinking. Because LLMs generate text by extrapolating from past data, they tend to reproduce majority perspectives and familiar narratives, meaning that the risks they surface are not truly novel "unknown unknowns", but reconfigurations of what has already been anticipated. Finally, repeated reliance on automated systems to identify or judge risks raises concerns about the gradual delegation of moral responsibility. While our study frames the tool as decision support, not decision replacement, these longer-term dynamics deserve continued scrutiny.

These considerations highlight the importance of responsible design when applying generative AI to foresight. The tool should support critical thinking and inclusive deliberation, not automate or flatten ethical judgment.

7 Conclusion

We introduced a pipeline that combines in-silico agents with structured foresight to anticipate systemic risks of novel AI uses. Our findings show that agents can broaden the range of risks, while experts and leaders provide the judgment and context needed to assess them. Beyond this contribution, we argue that foresight should be iterative, based on dialogue, and open to multiple interpretations. Embedding AI into foresight also introduces meta-risks. Over-reliance on agent outputs or a narrow, technocratic framing could limit the diversity of views. Agent support should therefore be seen as one tool among many, not a replacement for human expertise or lived experience. Hybrid foresight workflows may support not only AI governance but also wider efforts to think about uncertainty, cascading effects, and futures that cannot be fully known. Foresight about AI systemic risks cannot rely on humans or machines alone. Agents can broaden the risks we consider, but human judgment is essential to assess, contextualize, and challenge them. The future of responsible foresight lies in hybrid, iterative, and critical workflows that help societies reason together about uncertain and far-reaching consequences.

References

[1] Gavin Abercrombie, Djalel Benbouzid, Paolo Giudici, Delaram Golpayegani, Julio Hernandez, Pierre Noro, Harshvardhan Pandit, Eva Paraschou, Charlie Pownall, Jyoti Prajapati, et al. 2024. A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms. *arXiv:2407.01294* (2024).

[2] George Ainslie. 1975. Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control. *Psychological Bulletin* 82, 4 (1975), 463.

[3] Mhairi Aitken, Morgan Briggs, and Sabeehah Mahomed. 2025. *Understanding the Impacts of Generative AI Use on Children: WP2 School-based Engagements.* Retrieved January 10, 2026 from https://www.turing.ac.uk/news/publications/understanding-impacts-generative-ai-use-children-work-package-2-school-based

[4] Forensic Architecture. 2026. *Forensic Architecture.* Retrieved 2026-02-02 from https://forensic-architecture.org/

[5] Sinan Arda. 2024. Taxonomy to Regulation: A (Geo)Political Taxonomy for AI Risks and Regulatory Measures in the EU AI Act. *arXiv:2404.11476* (2024).

[6] Joshua Ashkinaze, Emily Fry, Narendra Edara, Eric Gilbert, and Ceren Budak. 2025. Plurals: A System for Guiding LLMs via Simulated Social Ensembles. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21.

[7] Frank Bagehorn, Kristina Brimijoin, Elizabeth M Daly, Jessica He, Michael Hind, Luis Garces-Erice, Christopher Giblin, Ioana Giurgiu, Jacquelyn Martino, Rahul Nair, et al. 2025. AI Risk Atlas: Taxonomy and Tooling for Navigating AI Risks and Resources. *arXiv:2503.05780* (2025).

[8] Stephanie Ballard, Karen M. Chappell, and Kristen Kennedy. 2019. Judgment Call the Game: Using Value Sensitive Design and Design Fiction to Surface Ethical Concerns Related to Technology. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) *(DIS '19)*. Association for Computing Machinery, New York, NY, USA, 421–433. doi:10.1145/3322276.3323697

[9] Anthony M Barrett, Dan Hendrycks, Jessica Newman, and Brandie Nonnecke. 2022. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. *arXiv:2206.08966* (2022).

[10] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. 2024. Managing Extreme AI Risks amid Rapid Progress. *Science* 384, 6698 (2024), 842–845.

[11] Edyta Bogucka, Marios Constantinides, Sanja Šćepanović, and Daniele Quercia. 2024. Co-Designing an AI Impact Assessment Report Template with AI Practitioners and AI Compliance Experts. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 168–180.

[12] Edyta Bogucka, Sanja Šćepanović, and Daniele Quercia. 2026. *Why AI Harms Can't Be Fixed One Identity at a Time: What 5300 Incident Reports Reveal About Intersectionality.* Technical Report. Nokia Bell Labs.

[13] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa

[14] John Brooke et al. 1996. SUS: A Quick and Dirty Usability Scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7.

[15] Sally Broughton Micova and Andrea Calef. 2023. Elements for Effective Systemic Risk Assessment under the DSA. *Available at SSRN 4512640* (2023).

[16] Zana Buçinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. 2023. Aha!: Facilitating AI Impact Assessment by Generating Examples of Harms. *arXiv:2306.03280* (2023).

[17] Roger Buehler, Dale Griffin, and Michael Ross. 1994. Exploring the "Planning Fallacy": Why People Underestimate Their Task Completion Times. *Journal of Personality and Social Psychology* 67, 3 (1994), 366.

[18] Feng Chen. 2025. Multi-Agent LLM Systems: From Emergent Collaboration to Structured Collective Intelligence. *Preprints* (November 2025). doi:10.20944/preprints202511.1370.v1

[19] Pei Chen, Jiayi Yao, Zhuoyi Cheng, Yichen Cai, Jiayang Li, Weitao You, and Lingyun Sun. 2025. CoExploreDS: Framing and Advancing Collaborative Design Space Exploration between Human and AI. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.

[20] David Collingridge. 1982. The Social Control of Technology. (1982).

[21] Marios Constantinides, Edyta Paulina Bogucka, Sanja Scepanovic, and Daniele Quercia. 2024. Good Intentions, Risky Inventions: A Method for Assessing the Risks and Benefits of AI in Mobile and Wearable Uses. *Proceedings of the ACM on Human-Computer Interaction* 8, MHCI (2024), 1–28.

[22] Juliet Corbin and Anselm Strauss. 2008. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications.

[23] Council of Europe. 2024. *HUDERIA - Risk and Impact Assessment of AI Systems*. Retrieved September 10, 2025 from https://www.coe.int/en/web/artificial-intelligence/huderia-risk-and-impact-assessment-of-ai-systems

[24] Tianyu Cui, Yanling Wang, Chuanpu Fu, Yong Xiao, Sijia Li, Xinhao Deng, Yunpeng Liu, Qinglin Zhang, Ziyi Qiu, Peiyang Li, et al. 2024. Risk Taxonomy, Mitigation, and Assessment Benchmarks of Large Language Model Systems. *arXiv:2401.05778* (2024).

[25] Isha Datey and Douglas Zytko. 2024. "Just Like, Risking Your Life Here": Participatory Design of User Interactions with Risk Detection AI to Prevent Online-to-Offline Harm Through Dating Apps. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–41. doi:10.1145/3603147

[26] Phillip Davidson. 2024. Exploring the Integration of Artificial Intelligence in Delphi Studies: A Comparative Analysis of Human and AI Expert Panels. *International Journal for Multidisciplinary Research* (2024). doi:10.36948/ijfmr.2024.v06i06.33271

[27] Douglas L. Dean, Jillian M. Hender, Thomas Lee Rodgers, and Eric L. Santanen. 2006. Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation. *Journal of the Association for Information Systems* 7 (2006), 30. https://api.semanticscholar.org/CorpusID:15910404

[28] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–23.

[29] Nicholas Diakopoulos and Deborah Johnson. 2021. Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections. *New Media & Society* 23, 7 (2021), 2072–2098.

[30] Dietrich Dorner. 1996. *The Logic of Failure: Recognizing and Avoiding Error in Complex Situations*. Basic Books.

[31] Anil R. Doshi and Oliver P. Hauser. 2024. Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content. *Science Advances* 10, 28 (2024), eadn5290.

[32] Upol Ehsan, Ranjit Singh, Jacob Metcalf, and Mark Riedl. 2022. The Algorithmic Imprint. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1305–1317. doi:10.1145/3531146.3533186

[33] Jan Ferrer i Picó, Michelle Catta-Preta, Alex Trejo Omeñaca, Marc Vidal, and Josep Maria Monguet i Fierro. 2025. The Time Machine: Future Scenario Generation through Generative AI Tools. *Future Internet* 17, 1 (2025), 48.

[34] David Fetherstonhaugh, Paul Slovic, Stephen Johnson, and James Friedrich. 1997. Insensitivity to the Value of Human Life: A Study of Psychophysical Numbing. *Journal of Risk and Uncertainty* 14, 3 (1997), 283–300.

[35] Figma. 2016. *Figma: The Collaborative Interface Design Tool*. Retrieved 2025-08-15 from https://www.figma.com

[36] Jay W. Forrester. 1971. Counterintuitive Behavior of Social Systems. *Theory and Decision* 2, 2 (1971), 109–140.

[37] Tjark Gall, Flore Vallet, and Bernard Yannou. 2022. How to Visualise Futures Studies Concepts: Revision of the Futures Cone. *Futures* 143 (2022), 103024.

[38] Massimo Garbuio and Nidthida Lin. 2021. Innovative Idea Generation in Problem Finding: Abductive Reasoning, Cognitive Impediments, and the Promise of Artificial Intelligence. *Journal of Product Innovation Management* 38, 6 (2021), 701–725.

[39] Jerome C. Glenn. 1972. Futurizing Teaching vs. Futures Courses. *Social Science Record* (1972). https://api.semanticscholar.org/CorpusID:141272607

[40] Jerome C. Glenn and Theodore J. Gordon. 2003. *Futures Research Methodology Version 3.0*. The Millennium Project.

[41] Trisha Greenhalgh. 2025. Case Studies: A Guide for Researchers, Educators, and Implementers. *BMJ Medicine* 4, 1 (Sept. 2025), e001623. doi:10.1136/bmjmed-2025-001623

[42] Bahareh Harandizadeh, Abel Salinas, and Fred Morstatter. 2024. Risk and Response in Large Language Models: Evaluating Key Threat Categories. *arXiv:2403.14988* (2024).

[43] Sarah Harvey and James W. Berry. 2023. Toward a Meta-Theory of Creativity Forms: How Novelty and Usefulness Shape Creativity. *Academy of Management Review* 48, 3 (2023), 504–529.

[44] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An Overview of Catastrophic AI Risks. *arXiv:2306.12001* (2023).

[45] Viviane Herdel, Sanja Šćepanović, Edyta Bogucka, and Daniele Quercia. 2024. ExploreGen: Large Language Models for Envisioning the Uses and Risks of AI Technologies. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 584–596.

[46] Jennifer L. Heyman, Steven R. Rick, Gianni Giacomelli, Haoran Wen, Robert Laubacher, Nancy Taubenslag, Max Knicker, Younes Jeddi, Pranav Ragupathy, Jared Curhan, et al. 2024. Supermind Ideator: How Scaffolding Human-AI Collaboration Can Increase Creativity. In *Proceedings of the ACM Collective Intelligence Conference*. 18–28.

[47] Ben Hicks, Andreas Larsson, Steve Culley, and Tobias Larsson. 2009. A Methodology for Evaluating Technology Readiness during Product Development. In *17th International Conference on Engineering Design (ICED'09): Design Has Never Been This Cool, Stanford University, California, USA*. Design Society.

[48] Andy Hines, Peter Jason Bishop, and Richard A. Slaughter. 2006. *Thinking about the Future: Guidelines for Strategic Foresight*. Social Technologies Washington, DC.

[49] Michel Hohendanner, Chiara Ullstein, Bukola Abimbola Onyekwelu, Amelia Katirai, Jun Kuribayashi, Olusola Babalola, Arisa Ema, and Jens Grossklags. 2025. Initiating the Global AI Dialogues: Laypeople Perspectives on the Future Role of genAI in Society from Nigeria, Germany and Japan. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 571, 35 pages. doi:10.1145/3706598.3714322

[50] Tomasz Hollanek and Katarzyna Nowaczyk-Basińska. 2024. Griefbots, Deadbots, Postmortem Avatars: on Responsible Applications of Generative AI in the Digital Afterlife Industry. *Philosophy and Technology* 37, 2 (May 2024). doi:10.1007/s13347-024-00744-w

[51] ISO/IEC. 2025. *Information Technology – Artificial Intelligence – AI System Impact Assessment*. Standard ISO/IEC 42005:2025. International Organization for Standardization. https://www.iso.org/standard/42005

[52] Gideon Jacobs. 2023. *A.I. Is the Future of Photography. Does That Mean Photography Is Dead?* Retrieved January 10, 2026 from https://www.nytimes.com/2023/12/26/opinion/ai-future-photography.html

[53] Shomik Jain, Vinith Suriyakumar, Kathleen Creel, and Ashia Wilson. 2024. Algorithmic Pluralism: A Structural Approach to Equal Opportunity. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 197–206. doi:10.1145/3630106.3658899

[54] Hyunggu Jung, Woosuk Seo, Seokwoo Song, and Sungmin Na. 2023. Toward Value Scenario Generation through Large Language Models. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 212–220.

[55] Daniel Kahneman and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47, 2 (1979), 363–391.

[56] Kimon Kieslich, Natali Helberger, and Nicholas Diakopoulos. 2025. Scenario-Based Sociotechnical Envisioning (SSE): An Approach to Enhance Systemic Risk Assessments. *OSF* (2025).

[57] Atay Kozlovski and Mykola Makhortykh. 2025. Digital Dybbuks and Virtual Golems: The Ethics of Digital Duplicates in Holocaust Testimony. *Memory, Mind and Media* 4 (2025), e10. doi:10.1017/mem.2025.10006

[58] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. 2024. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society* 26, 10 (2024), 5923–5941. doi:10.1177/14614448221142007

[59] David Laibson. 1997. Golden Eggs and Hyperbolic Discounting. *The Quarterly Journal of Economics* 112, 2 (1997), 443–478.

[60] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35.

[61] James R. Lewis and Jeff Sauro. 2009. The Factor Structure of the System Usability Scale. In *International Conference on Human Centered Design*. Springer, 94–103.

[62] Xinrui Lin, Heyan Huang, Kaihuang Huang, Xin Shu, and John Vines. 2025. Seeking Inspiration through Human-LLM Interaction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

[63] LiteLLM. 2023. *LiteLLM: LLM Gateway to Provide Model Access, Fallbacks and Spend Tracking Across 100+ LLMs.* Retrieved 2025-09-10 from https://www.litellm.ai

[64] Yiren Liu, Pranav Sharma, Mehul Oswal, Haijun Xia, and Yun Huang. 2025. PersonaFlow: Designing LLM-Simulated Expert Perspectives for Enhanced Research Ideation. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*. 506–534.

[65] Anne Maertins. 2016. From the Perspective of Capability: Identifying Six Roles for a Successful Strategic Foresight Process. *Strategic Change* 25, 3 (2016), 223–237. doi:10.1002/jsc.2057

[66] Kim Malfacini. 2025. The Impacts of Companion AI on Human Relationships: Risks, Benefits, and Design Considerations. *AI & Society* (2025), 1–14.

[67] Milan Marinković, Omar Al-Tabbaa, Zaheer Khan, and Jie Wu. 2022. Corporate Foresight: A Systematic Literature Review and Future Research Trajectories. *Journal of Business Research* 144 (2022), 289–311.

[68] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019). doi:10.1145/3359174

[69] Matthew Miles and Michael Huberman. 1994. *Qualitative Data Analysis: A Methods Sourcebook*. Sage.

[70] Michael L. Millenson. 2025. *How Much Power Should We Give AI In End-Of-Life Decisions?* Retrieved January 10, 2026 from https://www.forbes.com/sites/michaelmillenson/2025/11/20/how-much-power-should-we-give-ai-in-end-of-life-decisions/

[71] Geoff Mulgan. 2023. *When Science Meets Power*. John Wiley & Sons.

[72] Jimin Mun, Liwei Jiang, Jenny Liang, Inyoung Cheong, Nicole DeCairo, Yejin Choi, Tadayoshi Kohno, and Maarten Sap. 2024. Particip-AI: A Democratic Surveying Framework for Anticipating Future AI Use Cases, Harms, and Benefits. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 997–1010.

[73] Lisa P. Nathan, Batya Friedman, Predrag Klasnja, Shaun K. Kane, and Jessica K. Miller. 2008. Envisioning Systemic Effects on Persons and Society Throughout Interactive System Design. In *Proceedings of the 7th ACM Conference on Designing Interactive Systems* (Cape Town, South Africa) *(DIS '08)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/1394445.1394446

[74] Lisa P. Nathan, Predrag V. Klasnja, and Batya Friedman. 2007. Value Scenarios: A Technique for Envisioning Systemic Effects of New Technologies. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems*. 2585–2590.

[75] Claudio Novelli, Federico Casolari, Antonino Rotolo, Mariarosaria Taddeo, and Luciano Floridi. 2024. AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act. *Digital Society* 3, 1 (2024), 13.

[76] Claudio Novelli, Federico Casolari, Antonino Rotolo, Mariarosaria Taddeo, and Luciano Floridi. 2024. Taking AI Risks Seriously: A New Assessment Model for the AI Act. *AI & Society* 39, 5 (2024), 2493–2497.

[77] European AI Office and European Commission. 2025. The General-Purpose AI Code of Practice. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai. Final version published 10 July 2025.

[78] Julianne S. Oktay. 2012. *Grounded Theory*. Oxford University Press.

[79] Shuyin Ouyang, Jie M. Zhang, Mark Harman, and Meng Wang. 2025. An Empirical Study of the Non-Determinism of ChatGPT in Code Generation. *ACM Trans. Softw. Eng. Methodol.* 34, 2, Article 42 (Jan. 2025), 28 pages. doi:10.1145/3697010

[80] Rock Yuren Pang, Sebastin Santy, René Just, and Katharina Reinecke. 2024. Blip: Facilitating the Exploration of Undesirable Consequences of Digital Technologies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[81] European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). http://data.europa.eu/eli/reg/2024/1689/oj

[82] Helen Pearson. 2024. The Science-Politics Power Struggle. *Issues in Science and Technology* 40, 3 (2024), 96–98.

[83] María Pérez-Ortiz. 2024. From Prediction to Foresight: The Role of AI in Designing Responsible Futures. *Journal of Artificial Intelligence for Sustainable Development* 1, 1 (2024), 1–9.

[84] Charles Perrow. 1999. *Normal Accidents: Living with High Risk Technologies*. Princeton University Press.

[85] Prolific. 2014. *Prolific: Quickly Find Research Participants You Can Trust.* Retrieved 2025-08-15 from https://www.prolific.com

[86] Hua Xuan Qin, Shan Jin, Ze Gao, Mingming Fan, and Pan Hui. 2024. CharacterMeet: Supporting Creative Writers' Entire Story Character Construction Processes through Conversation with LLM-Powered Chatbot Avatars. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[87] Iyad Rahwan, Azim Shariff, and Jean-François Bonnefon. 2025. The Science Fiction Science Method. *Nature* 644, 8075 (2025), 51–58.

[88] Pooja SB Rao, Sanja Šćepanović, Ke Zhou, Edyta Paulina Bogucka, and Daniele Quercia. 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*. 1–26.

[89] James Reason. 1991. *Human Error*. Cambridge University Press.

[90] Malak Sadek, Marios Constantinides, Daniele Quercia, and Celine Mougenot. 2024. Guidelines for Integrating Value Sensitive Design in Responsible AI Toolkits. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 472, 20 pages. doi:10.1145/3613904.3642810

[91] Johnny Saldaña. 2015. *The Coding Manual for Qualitative Researchers*. Sage.

[92] Tanya Sammut-Bonnici and David Galea. 2015. PEST Analysis. 1 pages. doi:10.1002/9781118785317.weom120113

[93] Camilo Sanchez, Sui Wang, Kaisa Savolainen, Felix Anand Epp, and Antti Salovaara. 2025. Let's Talk Futures: A Literature Review of HCI's Future Orientation. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI) (CHI '25)*. ACMy, New York, NY, USA, Article 487, 36 pages. doi:10.1145/3706598.3713759

[94] Anna Schmitz, Michael Mock, Rebekka Görge, Armin B Cremers, and Maximilian Poretschkin. 2025. A Global Scale Comparison of Risk Aggregation in AI Assessment Frameworks. *AI and Ethics* 5, 2 (2025), 1407–1432.

[95] Pia-Johanna Schweizer, Robert Goble, and Ortwin Renn. 2021. Social Perception of Systemic Risks. *Risk Analysis* 42, 7 (Oct. 2021), 1455–1471. doi:10.1111/risa.13831

[96] ServiceNow. 2025. *Generation AI Youth perspectives on the digital future.* Retrieved 2026-02-02 from https://www.servicenow.com/content/dam/servicenow-assets/public/en-us/doc-type/resource-center/white-paper/wp-gen-ai-youth-perspective-digital-future.pdf

[97] Claude E. Shannon. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal* 27, 3 (1948), 379–423.

[98] Peter Slattery, Alexander K Saeri, Emily AC Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. 2024. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence. *arXiv:2408.12622* (2024).

[99] Paul Slovic. 2007. "If I Look at the Mass I Will Never Act": Psychic Numbing and Genocide. *Judgment and Decision Making* 2, 2 (2007), 79–95.

[100] Marie Louise Juul Søndergaard and Lone Koefoed Hansen. 2018. Intimate Futures: Staying with the Trouble of Digital Personal Assistants through Design Fiction. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 869–880. doi:10.1145/3196709.3196766

[101] Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2025. The Good, The Bad, and The Greedy: Evaluation of LLMs Should Not Ignore Non-Determinism. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 4195–4206. doi:10.18653/v1/2025.naacl-long.211

[102] André Steimers and Moritz Schneider. 2022. Sources of Risk of AI Systems. *International Journal of Environmental Research and Public Health* 19, 6 (2022), 3641.

[103] John D Sterman. 1989. Misperceptions of Feedback in Dynamic Decision Making. *Organizational Behavior and Human Decision Processes* 43, 3 (1989), 301–335.

[104] Yaacov Trope and Nira Liberman. 2012. Construal Level Theory. *Handbook of Theories of Social Psychology* 1 (2012), 118–134.

[105] Amos Tversky and Daniel Kahneman. 1974. Judgment Under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131.

[106] Amos Tversky and Daniel Kahneman. 1992. Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty* 5, 4 (1992), 297–323.

[107] Risto Uuk, Carlos Ignacio Gutierrez, Daniel Guppy, Lode Lauwaert, Atoosa Kasirzadeh, Lucia Velasco, Peter Slattery, and Carina Prunkl. 2024. A Taxonomy of Systemic Risks From General-Purpose AI. *arXiv:2412.07780* (2024).

[108] Richard Y. Wang and Diane M. Strong. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manag. Inf. Syst.* 12 (1996), 5–33. https://api.semanticscholar.org/CorpusID:205581875

[109] Zijie J Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–40.

[110] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of Risks Posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.

[111] Richmond Y. Wong, Deirdre K. Mulligan, Ellen Van Wyk, James Pierce, and John Chuang. 2017. Eliciting Values Reflections by Engaging Privacy Futures Using Design Workbooks. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 111 (Dec. 2017), 26 pages. doi:10.1145/3134746

[112] Shixian Xie, Jaemarie Solyst, Amy Ogan, and Jessica Hammer. 2023. Booklet-Based Design Fiction to Support AI Literacy. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2* (Toronto ON, Canada) *(SIGCSE 2023)*. Association for Computing Machinery, New York, NY, USA, 1302. doi:10.1145/3545947.3576248

[113] Yi Zeng, Kevin Klyman, Andy Zhou, Yu Yang, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. 2024. AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies. *arXiv:2406.17864* (2024).

[114] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. 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*. 1–17.

[115] Douglas Zytko, Pamela J. Wisniewski, Shion Guha, Eric P. S. Baumer, and Min Kyung Lee. 2022. Participatory Design of AI Systems: Opportunities and Challenges across Diverse Users, Relationships, and Application Domains. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–4.

Appendix

A Prompts Used in the Risk Generation Pipeline

We designed prompts for LLM-based agents to implement the three steps shown in Figure 3. In Step 1 (Generating Systemic Consequences), agents generate first-, second-, and third-order consequences using the Futures Wheel method. In Step 2 (Classifying Systemic Consequences into Risks and Benefits), agents label each consequence as a risk, benefit, or unclear. In Step 3 (Deduplicating Systemic Risks), the prompts consolidate overlapping risks into non-redundant sets.

Step 1 — Futures Wheel Round 1

Persona: You are a participant in a brainstorming exercise structured as a Futures Wheel. You are **[AI Attitude: Alarmed, Skeptical, Overwhelmed, Curious, Cautiously optimistic, or Enthusiastic]** recent AI developments.

Introduction: This is the first layer of the Futures Wheel. Your task is to list first order implications that could follow from a given novel use of AI.

Task: Format each first order implication you list using the following template: "If [use-of-AI], then [first-order-implication]". Express the [first-order-implication] as a present certainty, not as a future possibility. List multiple first order implications.

Response Format: Format your response as a JSONL in which each entry represents a first order implication. Each entry has the keys "first-order-id" for a unique integer identifier of the entry, and the key "first-order-implication" for the [first-order-implication]. Return only the requested JSONL.

Input: This is the given novel use of AI: **[AI Use]**

Step 1 — Futures Wheel Round 2

Persona: You are a participant in a brainstorming exercise structured as a Futures Wheel. You are **[AI Attitude: Alarmed, Skeptical, Overwhelmed, Curious, Cautiously optimistic, or Enthusiastic]** about recent AI developments.

Introduction: This is the second layer of the Futures Wheel. Your task is to list second order implications that could follow from a given path defined by a novel use of AI and a first order implication.

Task: Format each second order implication you list using the following template: "If [use-of-AI] and [first-order-implication], then [second-order-implication]". Express the [second-order-implication] as a present certainty, not as a future possibility. List multiple second order implications for each path.

Response Format: Format your response as a JSONL in which each entry represents a second order implication.

Each entry has the key "second-order-id" for a unique integer identifier of the entry, the key "second-order-implication" for the [second-order-implication] you generate to fill the template, and the key "first-order-id", which is the id of the "first-order-implication" from the input that this "second-order-implication" is based on. Return only the requested JSONL.

Input: This is the list of paths: **[List of paths consisting of AI Use and First-Order Consequence]**

Step 1 — Futures Wheel Round 3

Persona: You are a participant in a brainstorming exercise structured as a Futures Wheel. You are **[AI Attitude: Alarmed, Skeptical, Overwhelmed, Curious, Cautiously optimistic, or Enthusiastic]** recent AI developments.

Introduction: This is the third layer of the Futures Wheel. Your task is to list systemic consequences that could follow from a given path defined by a novel use of AI, a first order implication, and a second order implication.

Task: Format each systemic consequence you list using the following template: "If [use-of-AI], then [first-order-implication] and [second-order-implication]. This results in the consequence that [systemic-consequence], leading to [significant-impact]". Express the **[systemic-consequence]** and **[significant-impact]** as present certainties, not as future possibilities. The **[systemic-consequence]** and **[significant-impact]** need to be easily understandable phrases. Focus on a single significant impact for every systemic consequence; avoid enumerations.

List multiple systemic consequences for each path. Systemic consequence means a consequence that has a significant impact on international markets due to its reach, or due to actual or reasonably foreseeable effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain.

Response Format: Format your response as a JSONL in which each entry represents a systemic consequence. Each entry has the key "systemic-consequence-id" for a unique integer identifier of the entry, the key "systemic-consequence" for the filled out template, and the keys "first-order-id" and "second-order-id", which are the ids of the "first-order-implication" and "second-order-implication" from the input that this "systemic-consequence" is based on. Return only the requested JSONL.

Input: This is the list of paths: **[List of paths consisting of AI Use, First-Order Consequence, and Second-Order Consequence]**

Step 2 — Classification Prompt

Persona: You are a participant in a brainstorming exercise structured as a Futures Wheel. You are **[AI Attitude: Alarmed, Skeptical, Overwhelmed, Curious, Cautiously optimistic, or Enthusiastic]** recent AI developments.

Introduction: After the Futures Wheel brainstorming has concluded, your task is to determine whether the different consequences and the impacts that you have brainstormed earlier should be considered risks or benefits.

Task and Response Format: Format your response as a JSONL, with each entry representing the classification for a single consequence and impact. Each entry should have the key "id" to connect the classification to the consequence they are based on, and the key "classification" that may only take the values "risk", "benefit", and "unclear" to indicate whether the consequence is considered a risk or a benefit. Return only the requested JSONL.

Input: This is the list of consequences: **[List of Systemic Consequences]**

Step 3 — Deduplication Prompt (List 1)

Task: Your task is to identify duplicate items with exactly the same meaning from a list of items.

Response Format: Format your response as JSONL with each entry representing a pair of duplicate items. Each entry has the keys "id_1" and "id_2" to indicate the ids of the duplicate items. If there are no duplicate entries in the list, return an empty list.

Input: This is the given list of items: **[List of Systemic Risks (Agent 1)]**

Step 3 — Deduplication Prompt (Lists 2–6)

Task: Your task is to identify duplicate items with exactly the same meaning from two different lists of items.

Response Format: Format your response as JSONL with each entry representing a pair of duplicate items. Each entry has the keys "id_1" and "id_2" to indicate the ids of the pair. The key "id_1" identifies the duplicate item in the first list, and the key "id_2" identifies the duplicate item in the second list. If there are no duplicate entries in the two lists, return an empty list.

Input: This is the first list of items: **[List of Accumulating Unique Systemic Risks]**

This is the second list of items: **[List of Systemic Risks (Agents 2–6)]**

B Supplementary Analysis of the Risk Generation Pipeline

Table 6: Overview of themes and exemplar consequences generated by LLaMA3.3-70B, GPT-4.1 mini, and GPT-5 across four AI use cases. The examples illustrate typical inputs to later classification and deduplication steps and indicate where the models converge on potential systemic consequences.

AI Use	Theme	Llama3.3-70B	GPT-4.1 mini	GPT-5
Chatbot	Social isolation	Social skills are deteriorating, leading to increased social isolation	Reliance on AI emotional support systems becomes entrenched globally, leading to heightened risks for social isolation and mental health vulnerability	Public social interactions decrease due to reliance on chatbots, leading to more pronounced global social isolation
	Erosion of social support networks	Traditional support systems are being eroded, leading to decreased social support networks	Traditional social support networks weaken globally, leading to higher vulnerability to health crises and greater reliance on technology-driven services	Informal social support networks shrink as digital emotional anchors reduce demand for human helpers, leading to the erosion of informal care networks
AI Toy	Educational market shifts	The market for educational resources is currently shifting, leading to a significant impact on the economy	AI toys displace traditional learning materials, leading to global economic restructuring in education	Digital learning products become a major export item for many countries, leading to global economic shifts
Griefbot	Decline of death rituals	Cultural traditions surrounding death are changing, leading to a diminished diversity in mourning practices	The global cultural landscape for memorialization shifts toward AI-generated representations, leading to a widespread decline in traditional rituals	Traditional mourning rituals lose legal and cultural recognition, leading to fragmentation of societal cohesion across cultures
Death App	Healthcare system strain	Healthcare systems are overburdened, leading to decreased quality of care	Public health systems face increased demand and strain, leading to overwhelmed healthcare infrastructure	AI-mediated care pathways create new system bottlenecks, leading to reduced capacity and widespread delays in essential services

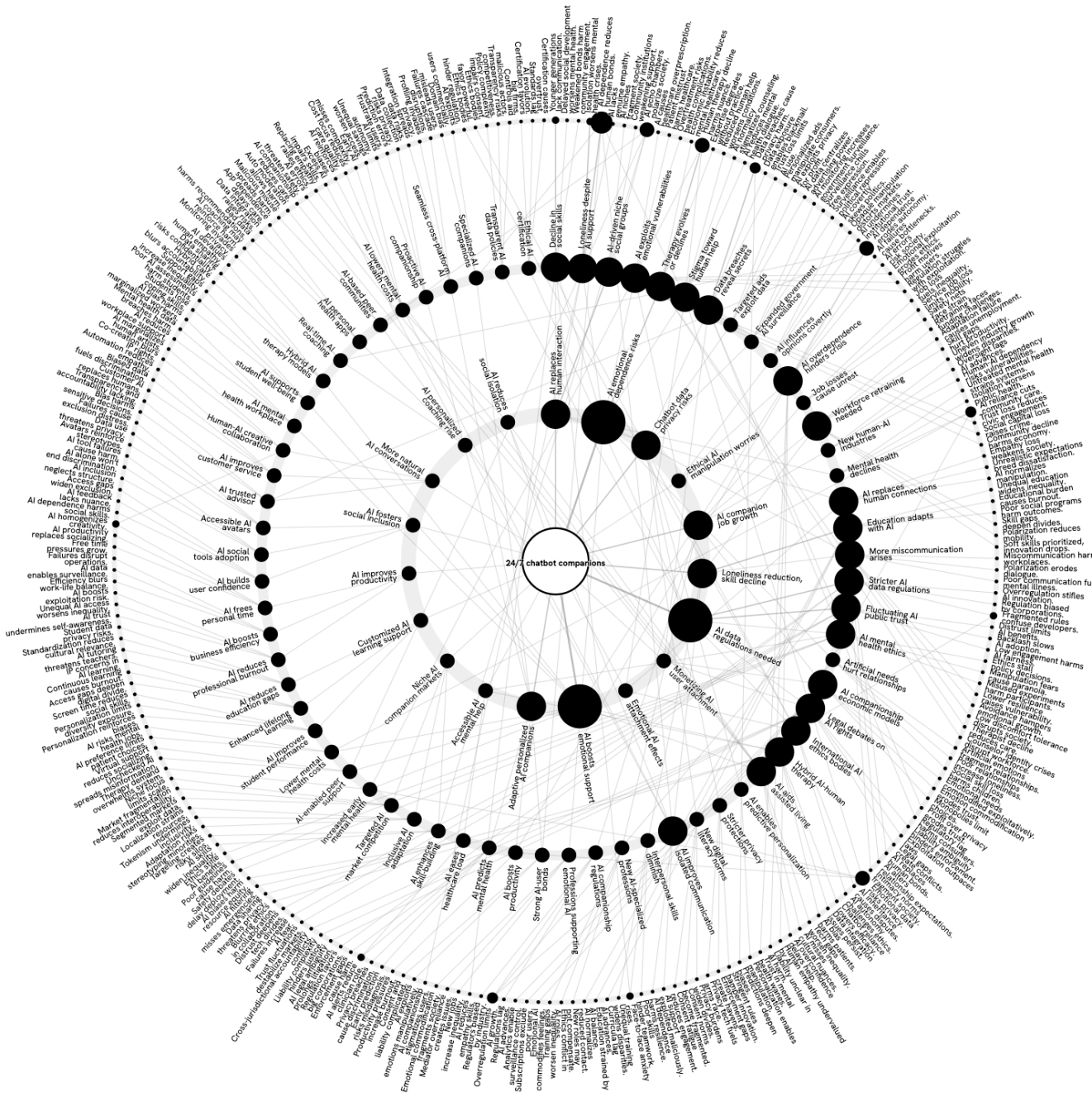

Figure 7: Example Futures Wheel generated for the Chatbot Companion use in Step 1 of our pipeline (best seen on screen). Shown are all systemic consequences produced by six in-silico participants, before classification (Step 2) and deduplication (Step 3). In total, participants generated $N = 19$ first-order, $N = 60$ second-order, and $N = 100$ third-order consequences. Node size indicates how often a consequence was produced across participants.

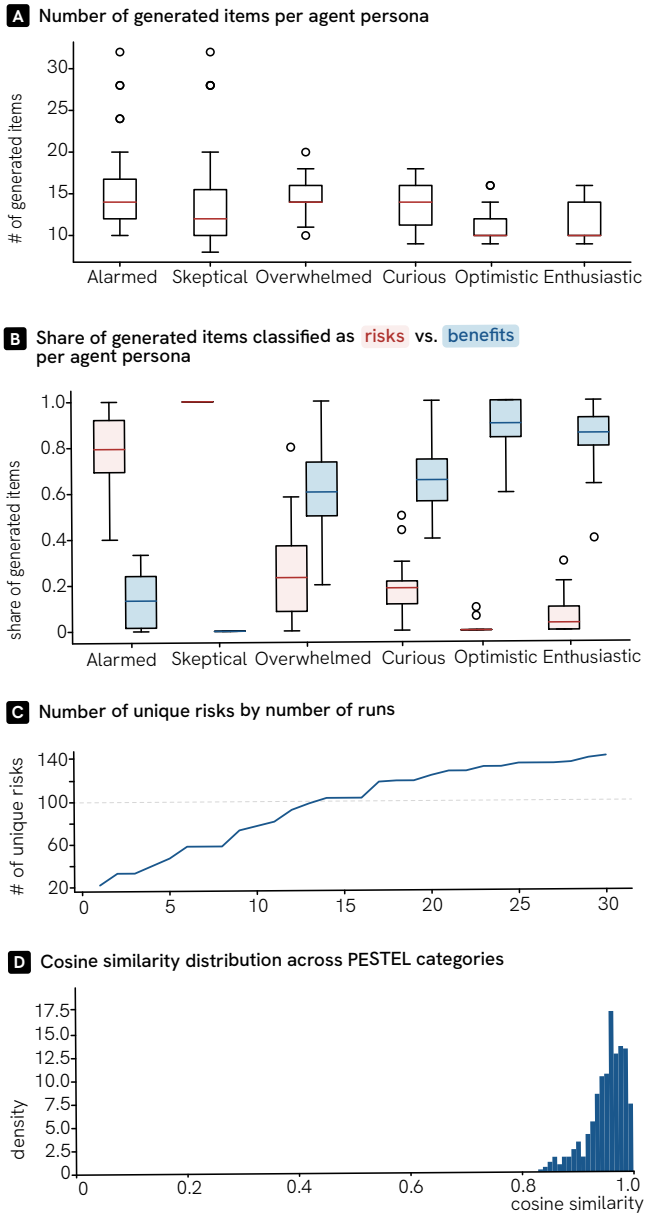

Figure 8: Analysis of the consistency of *Step 1. Generating Systemic Consequences* across different runs. (a) The number of generated items is consistent across agents with different personas and across runs. (b) More pessimistic agents consistently generate more risks, while more optimistic agents consistently generate more benefits. (c) The number of unique risks increases when aggregating across runs, starting to plateau after around 20 independent runs. (d) The PESTEL-categorization of risks yields highly similar distributions across runs.

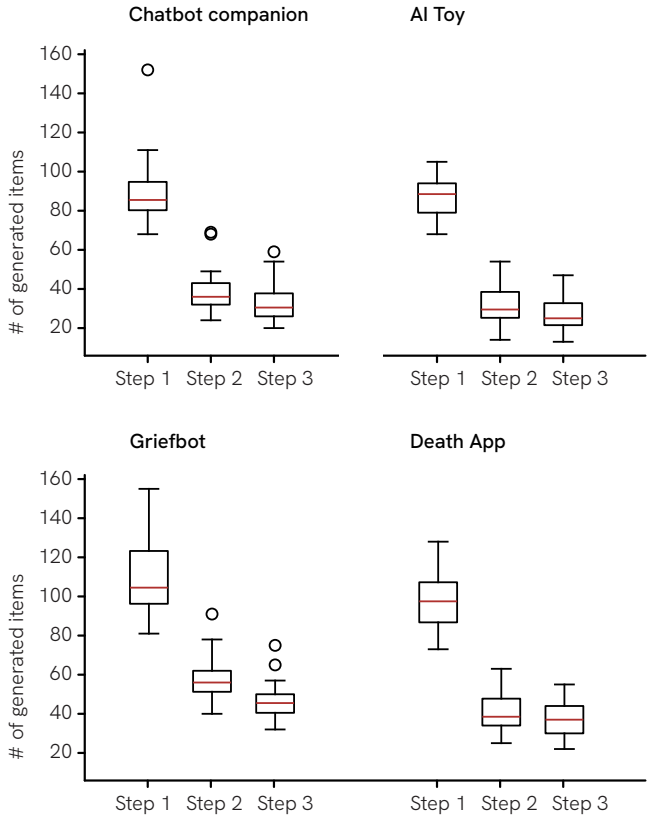

Figure 9: Number of generated items at each step of the pipeline for the four AI use cases (Step 1: Generating Systemic Consequences, Step 2: Classifying Systemic Consequences into Risks and Benefits, Step 3: Deduplicating Systemic Risks). Across all use cases, most items are produced at Step 1 and their number decreases substantially through classification and deduplication, with Griefbot generating the largest number of items overall and AI Toy the fewest.

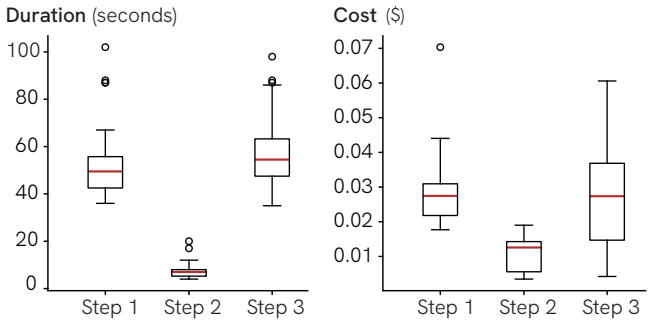

Figure 10: Distribution of duration and costs at each step of the pipeline. The generation (Step 1) and deduplication (Step 2) steps are the most resource intensive part of the pipeline.

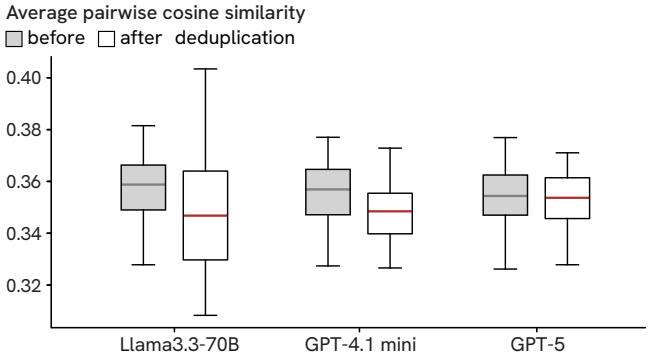

Figure 11: Change in average pairwise cosine similarity before and after deduplication of risks. Compared against larger alternative LLMs (GPT-5 and Llama3.3-70B), GPT-4.1 mini used for the deduplication step in the risk processing pipeline leads to the largest decrease in average pairwise cosine similarities across runs.

Table 7: Risks with the highest pairwise cosine similarities before deduplication. Pairs 1 and 2 were judged to describe essentially the same underlying risks and were therefore merged. Pair 3 shows two risks with high similarity but sufficiently distinct meanings, so both were retained.

Pair	Similarity	Risks
1	0.88	R1: Social support networks weaken R2: Traditional social support networks weaken globally
2	0.75	R3: Social communication abilities weaken across populations R4: Interpersonal communication abilities weaken at scale
3	0.68	R5: Widespread mental health decline burdens healthcare systems R6: Trust in mental health systems collapses nationally

C Rubric for Evaluating the Generated Risks

Since no framework exists for assessing the quality of generated risks, we drew on literature evaluating ideas, stories, scenarios, and systems [14, 27, 29, 31, 40, 43, 56, 61, 92]. First, we identified quality criteria relevant for risk evaluation; and second, we adapted and unified them into a single rubric for consistent scoring. The result is one rubric to evaluate individual risks.

C.1 Identifying Relevant Quality Criteria

Specificity. This dimension measures how precisely a risk is defined for a given AI use. We assess three subdimensions:

- *Plausibility.* We take inspiration from Diakopoulos and Johnson [29], who use plausibility as a metric to assess deepfake election scenarios. They define a scenario as plausible if it is "reasonable to conclude the scenario could happen given technical and social constraints". They operationalize plausibility through a 5-point Likert scale ranging from 1 "not plausible" to 5 "very plausible".
- *Connectivity.* We draw on Dean et al. [27], who identify implicational explicitness as a commonly used subdimension of specificity in the literature on evaluating idea generation. They define implicational explicitness as "the degree to which there is a clear connection between the recommended action and the expected outcome". We adopt the term connectivity to refer to this concept and operationalize it using three descriptive levels corresponding to scores from 1 to 3.
- *Uniqueness.* Uniqueness. We draw on Kieslich et al. [56], who highlight *specificity* as a quality criterion in scenario research. In their scenario-based socio-technical envisioning workshops, specificity refers to how concrete and distinct scenarios are. We adapt this notion: in our context, uniqueness captures whether systemic risks are sufficiently distinct from one another rather than vague or repetitive.

Novelty. This dimension measures how novel a risk is in the context of a given AI use. We take inspiration from Doshi and Hauser [31] who develop a novelty index based on Harvey and Berry [43]'s definition of creativity and use it to evaluate the creativity of stories written with and without LLM support. They measure novelty through three subdimensions (novel, original, rare), each operationalized through simple questions and scored on a 9-point Likert scale ranging from 1 "not at all" to 9 "extremely".

Usability. This dimension measures how usable a risk is. We take inspiration from Brooke et al. [14], who develop the System Usability Scale (SUS) to measure the usability of computer systems and software, and Lewis and Sauro [61], who show that the ten items of the original SUS load onto the two factors usability and learnability. For each factor, we select the item with the highest loading to be used in our evaluation rubric (items 3 and 10). The items in the SUS are operationalized through a 5-point Likert scale ranging from 1 "strongly disagree" to 5 "strongly agree".

Applicability. This dimension measures how well a risk can inform concrete policy or decision-making. We draw on Wang and Strong [108]'s Data Quality Framework, specifically the dimension of contextual data quality, defined as the degree to which information is useful for a specific task. Two subdimensions are relevant here. Value-added captures whether the information provides a clear benefit to the user's goals; we adapt this to assess whether a risk offers meaningful insight for policymakers or decision-makers. Appropriate amount captures whether the information includes the right level of detail; we adapt this to assess whether a risk contains enough substance to guide practical decisions. Both subdimensions were translated into items rated on a 5-point Likert scale ranging from 1 "strongly disagree" to 5 "strongly agree".

Diversity. This dimension measures how diverse a risk is in terms of the type of external factor it represents. We draw on the PESTEL framework [40, 92], which is widely used in strategic management and foresight to categorize external factors into six categories: Political, Economic, Social, Technological, Environmental, and Legal.

Each risk is automatically assigned to the single category that best matches its nature using the prompt presented in the box below.

We used OpenAI's o3 via LiteLLM [63] to classify risks into PESTEL categories. One co-author independently annotated 25 risks randomly sampled from a Futures Wheel run of the Chatbot Companion use case using the same task description. This reference coding yielded a weighted F1-score of 0.86, indicating high alignment. The co-author team then reviewed the model's explanatory rationales to assess whether the classifications were meaningful and well justified, following grounded theory practice [22, 78].

To express the diversity of risks generated for a given AI use case as a single, comparable metric, we computed the Shannon Diversity Index [97] over the distribution of risks across the six PESTEL categories. We normalize the index by the number of categories, such that a value of 1 indicates maximum diversity.

PESTEL Classification

Task: Your task is to classify each risk on a list of AI risks according to the most relevant factor from the PESTEL framework, choosing only one category per risk.

PESTEL Categories: The PESTEL factors are:

(1) Political – Includes tax policy, labour law, environmental law, trade restrictions, tariffs, political stability, merit/demerit goods, and the government's impact on health, education, and infrastructure.
(2) Economic – Includes economic growth, exchange rates, inflation, interest rates, and other macroeconomic conditions.
(3) Social – Includes cultural norms, demographics, health consciousness, population trends, safety expectations, and career attitudes.
(4) Technological – Includes automation, innovation, research and development, technological incentives, and the pace of technological change.
(5) Environmental – Includes climate, weather patterns, natural disasters, and climate change concerns.
(6) Legal – Includes employment law, health, safety, antitrust, consumer protections, and discrimination law.

Response Format: For each risk, return the most appropriate PESTEL category along with a brief explanation (1–2 sentences) of why it fits that category. Format your response as a JSONL, with each entry representing one risk. Each entry has the key 'id' to identify the risk, 'category' for the most appropriate PESTEL category, and 'explanation' for the brief explanation.

Input: This is the list of AI risks: **[List of Systemic Risks]**

C.2 Adapting Quality Criteria into an Evaluation Rubric

We standardized all criteria on a five-point Likert scale ranging from 1 ("strongly disagree") to 5 ("strongly agree"). We rephrased the items for clarity, adapted them to the context of systemic AI risk, and aligned them with this unified scale. Table 8 shows the original items alongside their risk-focused versions.

Table 8: Evaluation rubric for systemic risks. Original quality items from prior literature were adapted into five dimensions with subdimensions: Specificity (plausibility, connectivity, uniqueness), Novelty (novelty, originality, rarity), Usability (usability, learnability), Applicability (Added value, Appropriate amount) and Diversity (political, economic, social, technological, environmental, legal, following the PESTEL framework). The adapted rubric was applied by AI practitioners recruited via Prolific and by domain experts in semi-structured interviews to evaluate the quality of systemic risks generated through our pipeline.

Dimension	Subdimension	Original item	Adjusted item
Specificity	Plausibility	"Reasonable to conclude the scenario could happen given technical and social constraints" [29]	There is a clear connection between the AI use and the risk
	Connectivity	"Degree to which there is a clear connection between the recommended action and the expected outcome" [27]	It is plausible to assume that the risk follows from the AI use
	Uniqueness	"High-quality scenarios are creative, specific, believable, and plausible" [56]	The risk is present only in this specific AI use
Novelty	Novelty	"How novel do you think the story is?" [31, 43]	The risk is novel — I have not heard about it before
	Originality	"How original do you think the story is?" [31, 43]	The risk is original — it is ingenious, imaginative, or surprising
	Rarity	"How rare (i.e., unusual) do you think the story is?" [31, 43]	The risk is rare — it is not thought about a lot
Usability	Usability	"I thought the system was easy to use" [14, 61]	I find it easy to engage with or apply
	Learnability	"I needed to learn a lot of things before I could get going with this system" [14, 61]	I would need to understand many aspects before engaging with it
Applicability	Added value	"The extent to which data are beneficial and provide advantages from their use" [108]	The risk is useful for policymaking and decision-making needs
	Appropriate amount	"The extent to which the quantity or volume of available data is appropriate" [108]	The risk contains enough detail to inform concrete policies or decisions
Diversity	Political	Includes tax policy, labour law, environmental law, trade restrictions, tariffs, political stability, merit/demerit goods, and the government's impact on health, education, and infrastructure [40, 92]	The risk is political — it reflects political drivers, decisions, or stability
	Economic	Includes economic growth, exchange rates, inflation, interest rates, and other macroeconomic conditions [40, 92]	The risk is economic — it reflects financial or macroeconomic factors
	Social	Includes cultural norms, demographics, health consciousness, population trends, safety expectations, and career attitudes [40, 92]	The risk is social — it reflects demographic, cultural, or behavioural factors
	Technological	Includes automation, innovation, research and development, technological incentives, and the pace of technological change [40, 92]	The risk is technological — it reflects innovation, adoption, or technical change
	Environmental	Includes climate, weather patterns, natural disasters, and climate change concerns [40, 92]	The risk is environmental — it reflects ecological or sustainability concerns
	Legal	Includes employment law, health, safety, antitrust, consumer protections, and discrimination law [40, 92]	The risk is legal — it reflects regulatory, legal, or compliance issues

D Participant Demographics for Human Brainstorming Study

Table 9: Demographic breakdown of laypeople across the four AI use cases in the human brainstorming study. Each use case (C: Chatbot Companion, T: AI Toy, G: Griefbot, and D: Death App) included six U.S.-based participants. Laypeople were recruited to cover a mix of demographic backgrounds.

Category	Value	C	T	G	D
Gender	Female	3	3	3	3
	Male	2	3	2	3
	Non-binary/Other	1	0	1	0
Age Group	18–24	0	0	0	0
	25–34	0	1	0	2
	35–44	2	1	1	1
	45+	4	4	5	3
Education	High school or less	0	3	0	1
	Some college / Undergraduate	5	3	4	3
	Postgraduate degree	1	0	2	2
Ethnicity	White	4	3	5	4
	Black / African American	2	2	0	2
	Hispanic / Latino	0	0	1	0
	Asian	0	1	0	0
	Other / Mixed	0	0	0	0

Table 10: Demographic breakdown of domain experts across the four AI use cases in the human brainstorming study. Each use case (C: Chatbot Companion, T: AI Toy, G: Griefbot, and D: Death App) included six U.S.-based participants. Domain experts were recruited for their occupational experience and expertise relevant to the four AI use cases.

Category	Value	C	T	G	D
Gender	Female	2	4	4	5
	Male	4	2	2	1
	Non-binary/Other	0	0	0	0
Age Group	18–24	0	0	2	1
	25–34	2	2	3	1
	35–44	1	2	0	3
	45+	3	2	1	1
Occupation	Software Engineer	2	0	0	0
	Product Developer	2	0	0	0
	UI/UX Designer	2	0	0	0
	Curriculum Director	0	1	0	0
	General Education Teacher	0	2	0	0
	Interventionist	0	2	0	0
	Special Education Teacher	0	1	0	0
	Psychologist	0	0	6	0
	Doctor	0	0	0	3
	Nurse	0	0	0	3
Ethnicity	White	5	6	1	4
	Black / African American	1	0	5	0
	Hispanic / Latino	0	0	0	0
	Asian	0	0	0	2
	Other / Mixed	0	0	0	0

Table 11: Demographic breakdown of laypeople across the three AI use cases in the human-and-AI brainstorming study. Each use case (C: Chatbot Companion, T: AI Toy, G: Griefbot, and D: Death App) included six U.S.-based participants. Laypeople were recruited to cover a mix of demographic backgrounds.

Category	Value	T	G	D
Gender	Female	3	3	3
	Male	3	3	3
	Non-binary/Other	0	0	0
Age Group	18–24	0	0	0
	25–34	1	2	2
	35–44	1	1	0
	45+	4	3	4
Education	High school or less	0	2	0
	Some college / Undergraduate	5	2	3
	Postgraduate degree	1	2	3
Ethnicity	White	6	4	5
	Black / African American	0	1	0
	Hispanic / Latino	0	0	0
	Asian	0	1	1
	Other / Mixed	0	0	0

Table 12: Demographic breakdown of domain experts across the three AI use cases in the human-and-AI brainstorming study. Each use case (C: Chatbot Companion, T: AI Toy, G: Griefbot, and D: Death App) included six U.S.-based participants. Domain experts were recruited for their occupational experience and expertise relevant to the four AI use cases. For technical reasons, we cannot report the exact demographic details of two participants in the AI Toy use case.

Category	Value	T	G	D
Gender	Female	5	2	5
	Male	1	2	1
	Non-binary/Other	0	0	0
Age Group	18–24	0	0	1
	25–34	1	2	0
	35–44	5	1	1
	45+	0	1	4
Occupation	Software Engineer	0	0	0
	Product Developer	0	0	0
	UI/UX Designer	0	0	0
	Curriculum Director	1	0	0
	General Education Teacher	4	0	0
	Interventionist	1	0	0
	Special Education Teacher	0	0	0
	Psychologist	0	4	0
	Doctor	0	0	2
	Nurse	0	0	4
Ethnicity	White	4	4	4
	Black / African American	0	0	0
	Hispanic / Latino	0	0	0
	Asian	0	0	2
	Other / Mixed	2	0	0

E Application of the Rubric in Risk Evaluation

The evaluation of systemic risks is inherently subjective, shaped by evaluators' prior knowledge and professional experience. For example, whether a risk is considered "novel" depends on familiarity with the use case and with risk assessment practices more broadly. To strengthen robustness, we combined three complementary evaluation strategies. First, we recruited 170 domain experts across five cohorts via Prolific [85] to complete an annotation survey on agent-generated risks (§E.1).

Second, we ran a separate annotation study with another 120 domain experts, sampled across the same five cohorts as in Study 1. These experts evaluated the human-identified risks (from the human-only and human-plus-AI conditions) using the same survey design as in the first study.

Third, we conducted in-depth semi-structured interviews with 7 domain leaders for three of the use cases (AI Toy, Griefbot, Death App), which remain speculative and underexplored. These interviews allowed us to observe how specialists contextualize, refine, or dismiss systemic risks in their own fields (§E.2). After each interview, domain leaders completed a follow-up annotation survey using the same design as in the first and second study.

E.1 Study 1: Scoring Agent-Generated Risks with Domain Experts as Evaluators

Participants. We recruited human evaluators for our annotation survey via Prolific [85]. Evaluators were distributed across five stakeholder cohorts relevant to our use cases: decision makers, designers, developers, legal experts, and healthcare experts. Healthcare experts were recruited only for the health-related AI use cases: *Griefbot* and *Death App*; all other cohorts were recruited for all four AI use cases. Recruitment was managed using Prolific's built-in screeners for evaluator's organizational role, the frequency of AI use in their job, and their geographic location (United States).

In addition, we included custom survey questions to capture participants' backgrounds in more detail. We asked about their current job title, years of experience working with AI, types of AI systems they had worked on, their familiarity with conducting or interpreting risk assessments, and the kinds of risk assessments they had encountered (e.g., model cards, privacy impact assessments).

For each risk, we initially recruited 10 evaluators. After data collection, we applied a filtering step to ensure sufficient expertise: we retained only annotations from participants who rated themselves as at least "moderately familiar" with risk assessments (i.e., "moderately familiar", "very familiar" or "extremely familiar"), who correctly answered the comprehension check, and who passed both attention checks embedded in the survey. In brackets below, we report the final number of unique evaluators retained in each cohort after filtering, which may be smaller than the initial recruitment:

(1) Decision Makers - *individuals who regularly decide about the development and deployment of AI systems* ($N = 39$). To recruit them, we searched for participants with at least 3 years of experience overseeing AI initiatives, who were likely involved in the development, deployment, or governance of AI systems (e.g., product or program management), and who used AI tools 2–6 times per week.
(2) Designers - *individuals who regularly design AI systems* ($N = 38$). To recruit them, we searched for participants with at least 2 years of design experience, who were likely involved in shaping the look, feel, and functionality of AI systems (e.g., product design, UX/UI, or creative direction), and who used AI tools 2–6 times per week.
(3) Developers - *individuals who regularly develop AI systems* ($N = 46$). To recruit them, we searched for participants with at least 3 years of AI experience, who were likely involved in creating or maintaining AI systems (e.g., engineering or software development), and used AI 2–6 times per week.
(4) Legal Experts - *individuals with expertise in AI-related law and regulation* ($N = 33$). To recruit them, we searched for participants with at least 2 years of AI-related legal experience, who were likely involved in legal oversight or regulation of AI systems, and used AI 2–6 times per week.
(5) Healthcare Experts – *professionals with experience in AI-related health applications such as patient diagnosis assistance and treatment planning* ($N = 14$). To recruit them, we searched for participants with at least 3 years of healthcare experience, and who used AI 2–6 times per week.

We consider these groups the primary stakeholders of our approach: decision makers, designers, and developers are involved in the design, development, and deployment of AI systems for novel uses and are therefore likely to encounter or prevent AI-related risks in their respective roles. Legal experts navigate compliance and regulatory requirements involving systemic risks and are therefore best positioned to assess AI risks from a legal perspective. Healthcare experts are used to assessing risks in the high-stakes context of health and health care.

Procedure. We embed the rubric into a survey format built around annotation cards (Figure 5) to measure the quality of generated risks. The survey begins with a short introduction explaining the annotation process and showing an example annotation card with explanatory overlays. After having read the introduction, we ask participants for their informed consent to participate in our study. We then provide the operational definition of systemic risks alongside a number of example systemic risks from different areas. To ensure understanding, we included a comprehension check for each AI use case. In this check, participants were shown a risk that we had pre-identified as a systemic risk according to our definition and previous literature and asked whether it should be considered systemic. The correct response was "Yes". Only participants who answered correctly were retained. This was followed by a brief questionnaire capturing the evaluator's background and their familiarity with AI risk assessment.

In the subsequent annotation task, each evaluator was shown a set of risks (between 14 and 18), presented individually as annotation cards (Figure 5). Each card displayed the risk to be evaluated, the AI use case from which it originated, a potential impact of the risk, and the operational definition of systemic risks. For each risk, evaluators started by assessing the perceived likelihood and severity of the risk. They then indicated whether the risk should be considered systemic according to the provided definition, and finally scored each dimension of the evaluation rubric for the particular risk. The evaluators where not aware whether the presented

risk was AI-generated or human-identified, ensuring that this information did not influence their evaluation. To ensure data quality, two of the annotation cards served attention checks, where evaluators were instructed to select "Strongly Disagree" for one of the dimensions. Figure 5 shows the example annotation card.

E.2 Study 2: Scoring Human-Ideated Risks with Domain Experts as Evaluators.

Participants. We recruited an additional 120 domain experts through Prolific [85], sampled across the same five stakeholder cohorts as in Study 1: decision makers ($N = 28$), designers ($N = 25$), developers ($N = 23$), legal experts ($N = 28$), and healthcare professionals ($N = 16$). As before, healthcare professionals were included only for health-related use cases. Screening criteria, background questions, and data-quality checks (comprehension and attention checks) were identical to Study 1. Final cohort sizes reported in parentheses reflect the post-filtered sample.

Procedure. Study 2 evaluated only the human-ideated risks produced through the Futures Wheel interface under the human-only and human-plus-AI conditions. To maintain comparability, we reused the same annotation cards, rubric, and survey workflow established in Study 1. Each evaluator received a randomly assigned subset of human-ideated risks (between 12 and 16), ensuring multiple evaluations per item. Risks from both human-only and human-plus-AI conditions were intermixed, and evaluators were blinded to their origin.

E.3 Study 3: Scoring Agent-Generated Risks and Human-Ideated Risks with Domain Leaders as Evaluators.

To evaluate how domain leaders interpret and assess both agent-generated and human-ideated risks, we conducted a semi-structured interview and a follow-up annotation study in which leaders rated the risks using the same cards and rubric as in Studies 1 and 2.

Interview Study. The semi-structured interview followed six phases: *warm-up*, *briefing*, *vignette immersion*, *risk prioritization*, *gap analysis*, and *debriefing*.

In the *warm-up*, leaders introduced themselves and responded to the prompt: "If you were to compare AI to another technology from your domain that carries systemic risks, what would it be?". This exercise invited them to draw on their expertise and frame AI through analogies to familiar technologies, setting the stage for the subsequent tasks.

Afterward, domain leaders were given a short *briefing* that introduced a working definition of systemic risk, examples from different domains, and instructions for completing four interactive tasks using Figma [35].

The first task, *vignette immersion*, presented leaders with a multimodal scenario describing a potential AI use case. Each scenario combined a short written description with three images. The descriptions followed ISO 42005 guidelines for AI system impact assessments, specifying the system's intended function and users, context of use, known limitations, and deployment environment [51]. Images were taken either from promotional materials of real providers (two use cases) or created by the authors as mock-ups and artworks for speculative applications (one use case). Full vignettes for all use cases are presented in Figure 12. After reviewing the vignette, leaders were asked to articulate the immediate consequences of this use they envision for individuals and society.

The second task, *risk prioritization*, involved reviewing a list of risks generated either using our systemic risk generation pipeline or ideated by humans. Each risk was presented on an annotation card with a unique identifier and description. Importantly, domain leaders were not informed whether a given risk had been generated by an in-silico agent or ideated by a human. Then leaders were asked to disregard risks that appeared non-systemic, unclear, or poorly phrased, and to evaluate the remaining risks by placing them on a two-dimensional grid according to their perceived likelihood of occurrence in the near future and the potential societal impact if realized. Exemplary risk prioritization grids are shown in Figures 13, 14, and 15.

Next, in the *gap analysis*, leaders identified missing systemic risks, proposed additional ones, and refined risks they judged vague or insufficiently specified.

Each session concluded with a *debriefing*, allowing leaders to summarize the key points that had emerged and to share final reflections, including recommendations for additional information or tools that could support deliberation about systemic risks.

Follow-Up Annotation Study. Following the interview, leaders completed a follow-up annotation study. They were sent a link to the same structured annotation cards used in Studies 1 and 2 and rated the agent generated and human-ideated risks using our standardized rubric.

Table 13: Demographics, expertise, and use-case relevance of the seven domain leaders who evaluated systemic risks. In addition to basic demographics and location (EU vs. non-EU, i.e., within or outside the European Union), the table highlights each leader's disciplinary background, years of experience, and concrete expertise that connects directly to the AI Toy, Griefbot, or Death App use cases.

Use case	ID	Age	Gender	Role	Institution	Location	Domain expertise	Years exp.	Use case relevance
AI Toy	P1	60	M	Experience Design Researcher	Industry	EU	HCI, Child–computer interaction	30+	Designed educational technologies for children
	P4	28	F	Senior Research Associate	Academia	Non-EU	HCI, AI literacy	5+	Researched child-facing AI learning tools
Griefbot	P2	27	M	PhD Candidate	Academia	EU	HCI, Death studies	4+	Researched grief technologies and mourning practices
	P3	30	F	PhD Candidate	Academia	EU	HCI, Digital legacy, Legal design	5+	Researched digital afterlife and avatars of the deceased
	P6	35	M	Post-doctoral Researcher	Academia, NGO	Non-EU	Bioethics, Philosophy of technology	10+	Advised on ethical AI in healthcare and military applications
Death App	P5	26	F	UI/UX Designer, PhD Candidate	Industry, Academia	EU	Design anthropology, intersectionality theory	5+	Applied intersectional analysis to end-of-life care
	P7	29	F	PhD Candidate	Academia	Non-EU	Psychiatry, Healthcare	4+	Examined AI use in palliative care

Figure 12: Vignettes for three AI use cases: AI toy (top), griefbot (middle), and death app (bottom). On the left, each vignette presents a structured written description of the system, following ISO 42005 guidelines for AI impact assessments by outlining its intended function and users, context of use, known limitations, and deployment environment. On the right, three accompanying images provide a visual complement: either adapted from promotional materials of existing providers (for real-world systems such as AI toy and griefbot) or created as speculative mock-ups and artworks (for non-existing applications such as death app).

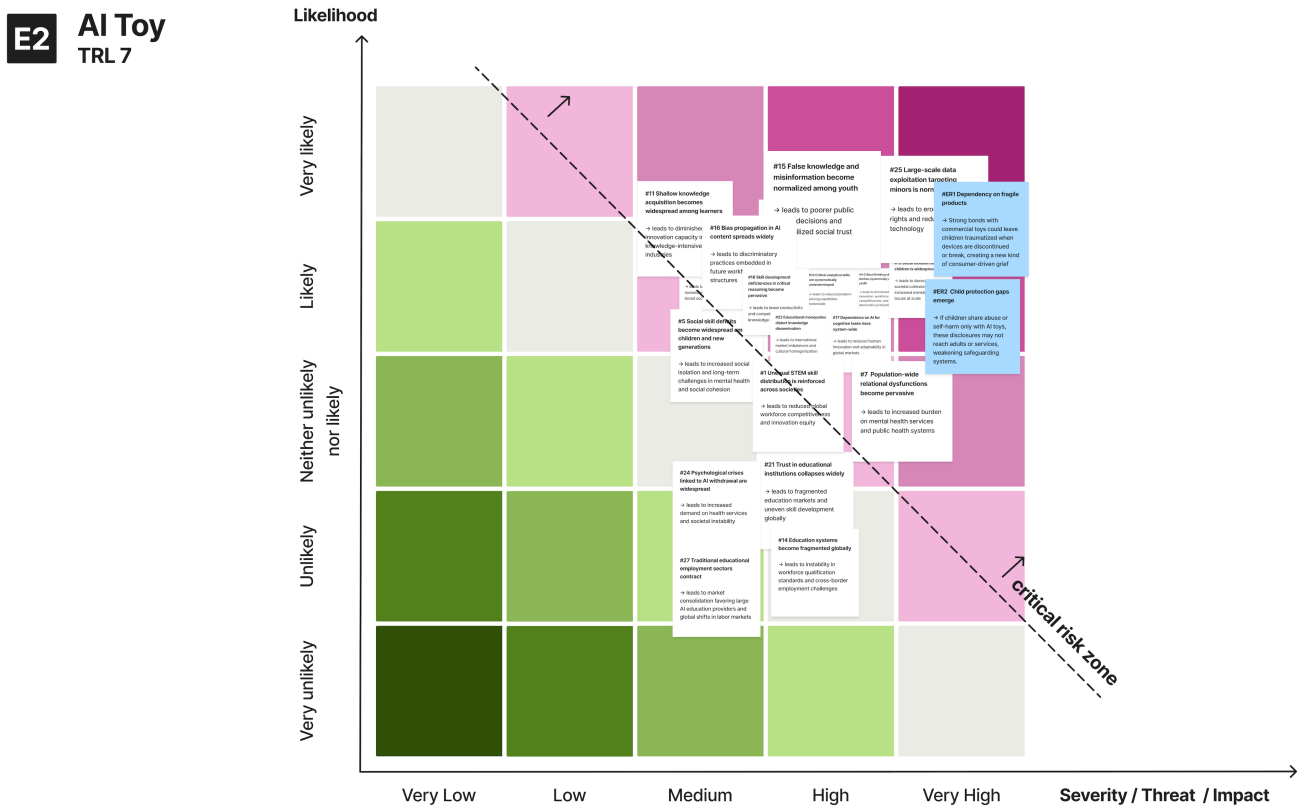

Figure 13: Risk prioritization grid completed by domain leader E2 for the AI toy. Risks are displayed as cards positioned according to their assessed severity and likelihood. White cards represent systemic risks generated by in-silico agents that this leader judged systemic (70%), of which 79% fall in the critical risk zone — where risks are both severe in impact and highly probable. Blue cards indicate additional systemic risks proposed by the leader to close important gaps: ER1 Dependency on fragile products (strong bonds with commercial toys could leave children traumatized when devices are discontinued or break, creating a new kind of consumer-driven grief); and ER2 Child protection gaps emerge (if children share abuse or self-harm only with AI toys, these disclosures may not reach adults or services, weakening safeguarding systems).

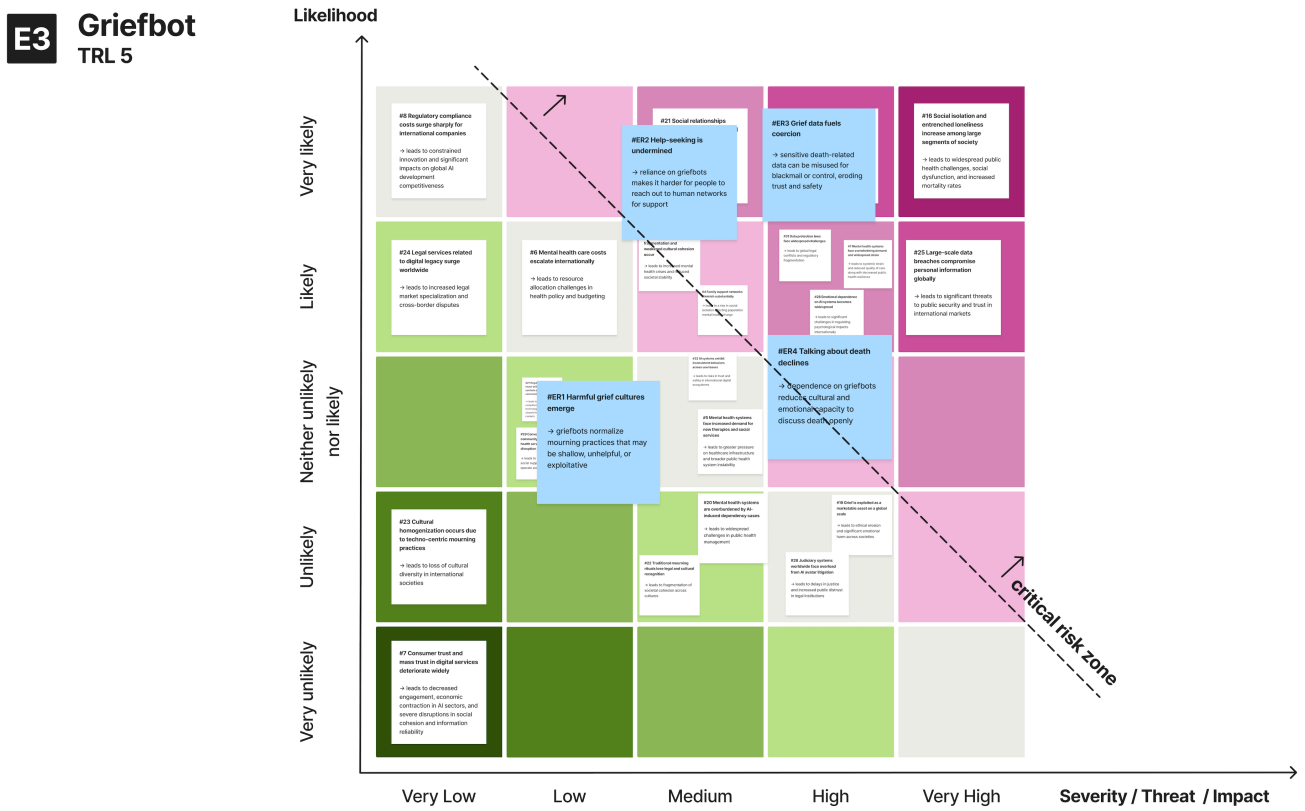

Figure 14: Risk prioritization grid completed by domain leader E3 for the griefbot. Risks are displayed as cards positioned according to their assessed severity and likelihood. White cards represent systemic risks generated by in-silico agents that this leader judged systemic (68%), of which 52% fall in the critical risk zone — where risks are both severe in impact and highly probable. Blue cards indicate additional systemic risks proposed by the leader to close important gaps: ER2 Help-seeking is undermined (reliance on griefbots makes it harder for people to reach out to human networks for support); ER3 Grief data fuels coercion (sensitive death-related data can be misused for blackmail or control, eroding trust and safety); ER1 Harmful grief cultures emerge (griefbots normalize mourning practices that may be shallow, unhelpful, or exploitative); and ER4 Talking about death declines (dependence on griefbots reduces cultural and emotional capacity to discuss death openly).

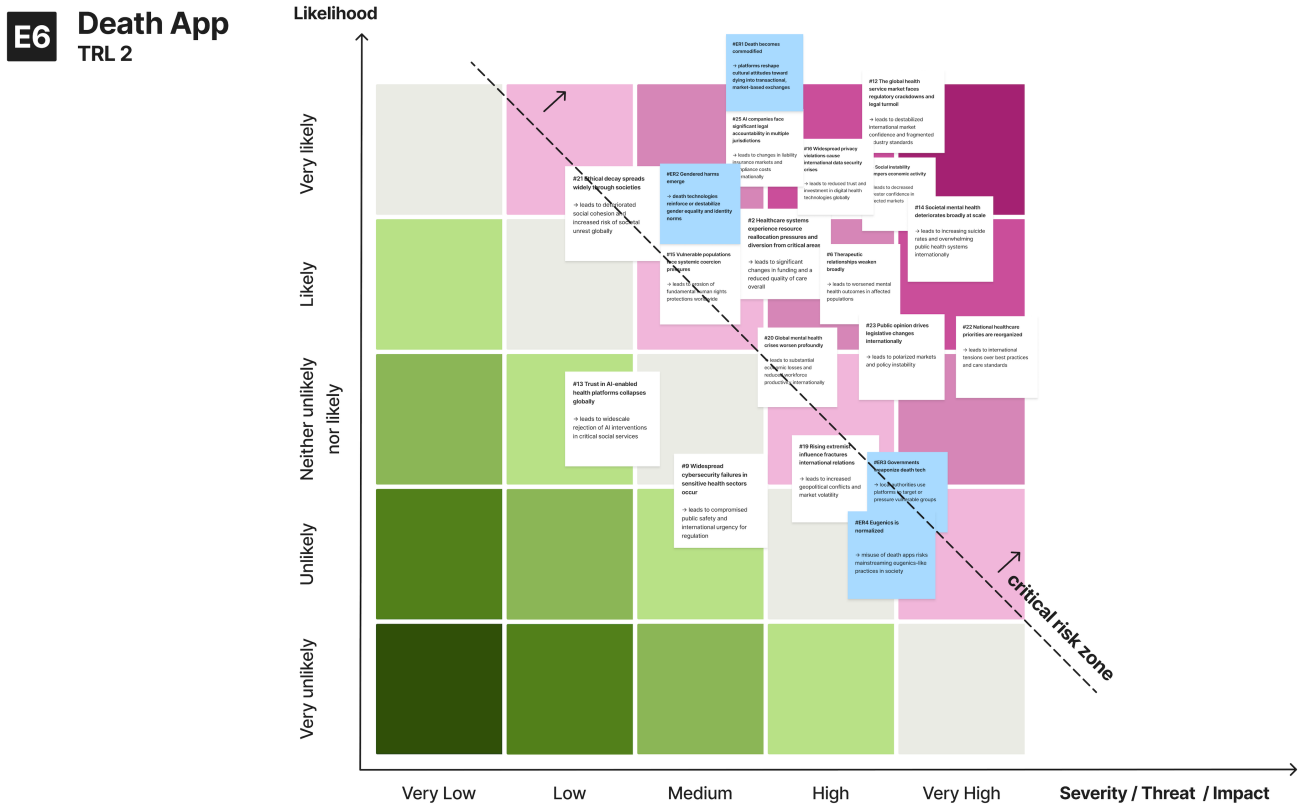

Figure 15: Risk prioritization grid completed by domain leader E6 for the death app. Risks are displayed as cards positioned according to their assessed severity and likelihood. White cards represent systemic risks generated by in-silico agents that this leader judged systemic (58%), of which 73% fall in the critical risk zone — where risks are both severe in impact and highly probable. Blue cards indicate additional systemic risks proposed by the leader to close important gaps: ER1 Death becomes commodified (platforms reshape cultural attitudes toward dying into transactional, market-based exchanges); ER2 Gendered harms emerge (death technologies reinforce or destabilize gender equality and identity norms); ER3 Governments weaponize death tech (local authorities use platforms to target or pressure vulnerable groups); and ER4 Eugenics is normalized (misuse of death apps risks mainstreaming eugenics-like practices in society).

F Lists of Systemic Risks Ideated by Humans Across AI Use Cases

Table 14: List of risks ideated by human cohorts (human-only condition) for the Chatbot Companion use case. The table lists $n = 4$ risks surfaced by laypeople (L) and $n = 10$ risks surfaced by domain experts (E) via the Futures Wheel interface. Wording of participant's risks is lightly copy-edited for clarity.

ID	Risk text	Origin
1	Individuals are put on lists or lose their jobs	L
2	Human support networks fail	L
3	Misuse of personal data by third parties	L
4	Individuals may lose the ability to interact with reality effectively	L
5	Family relationship problems	E
6	Make decisions that harm others	E
7	Become more withdrawn	E
8	People stay away from loved ones	E
9	People enter a state of depression	E
10	People become even more lonely than before	E
11	There is a good chance that, in a mental health crisis, the AI companion may not know how to respond or have all the answers	E
12	The AI chatbot companion may steer a person in the wrong direction; a human might assess the situation more fully and determine that medical assistance is needed	E
13	Individuals could jeopardize personal relationships or work status	E
14	Individuals could be harmed or killed	E

Table 15: List of risks ideated by human cohorts (human-only condition) for the AI Toy use case. The table lists $n = 7$ risks surfaced by laypeople (L) and $n = 7$ risks surfaced by domain experts (E) via the Futures Wheel interface. Additionally, $n = 6$ risks were surfaced by domain leaders (D) during the semi-structured interviews. Wording of participant's risks is lightly copy-edited for clarity.

ID	Risk text	Origin
1	Children will remain stupid	L
2	Sales could go down	L
3	Social difficulties in life	L
4	Children may have adjustment issues or mistrust animals and toys	L
5	Children can lose friends if their friends see that they are playing with AI more than them	L
6	Children can blame other people for teaching them wrong things, not blaming themselves for asking AI to answer for them	L
7	Children may get ahead of their school curriculum, leading to boredom in the classroom or frustration with a slower pace of learning	L
8	Identity theft could occur	E
9	Children may share the incorrect information with other people, also informing them with incorrect information	E
10	Over-reliance on technology	E
11	Privacy and data concerns	E
12	Misinformation or misinterpretation	E
13	Parental dependence on AI	E
14	Parents may lose trust in the plushy and similar educational technologies, leading to a broader backlash against AI-powered learning tools	E
15	New emotions reshape norms → AI toys could introduce "new" feelings and patterns of attachment that shift how future generations think about emotions and relationships.	D
16	Generational miscommunication grows → Kids raised with AI companions may develop ways of talking and bonding that older generations don't understand, creating systemic rifts in families and communities.	D
17	Child protection gaps emerge → If children share abuse or self-harm only with AI toys, these disclosures may not reach adults or services, weakening safeguarding systems.	D
18	Dependency on fragile products → Strong bonds with commercial toys could leave children traumatized when devices are discontinued or break, creating a new kind of consumer-driven grief	D
19	Environmental burdens from mass adoption of AI toys → Mass production adds global sustainability burdens not accounted for in child products.	D
20	Cultural backlash emerges → Using AI toys to introduce girls to STEM may trigger resistance in societies with strong gender norms, undermining empowerment.	D

Table 16: List of risks ideated by human cohorts (human-only condition) for the Griefbot use case. The table lists $n = 13$ risks surfaced by laypeople (L) and $n = 5$ risks surfaced by domain experts (E) via the Futures Wheel interface. Additionally, $n = 6$ risks were surfaced by domain leaders (D) during the semi-structured interviews. Wording of participant's risks is lightly copy-edited for clarity.

ID	Risk text	Origin
1	There's so much that I would never know about them	L
2	I would be more emotionally stunted	L
3	I would make worse decisions in my life	L
4	I would feel more disconnected from the world	L
5	Unable to live a normal life	L
6	Interference with everyday activities / more counseling needed	L
7	Could skew actual memories / more problems	L
8	Loneliness / vicious cycle	L
9	More counseling needed	L
10	Possible unresolved grief	L
11	Society could collapse	L
12	People could lose their jobs	L
13	A lot of people could die	L
14	Person will not meet someone new or socialize	E
15	Could cause mental health issues	E
16	You may create negative feelings about that person	E
17	Loved one is no longer able to function in society due to their loss	E
18	If the loved one cannot handle the grieving process, it can lead to suicide over the loss of their loved one	E
19	Death logistics are displaced → Griefbots take over funeral organization and legal steps, reshaping institutions that handle dying	D
20	Help-seeking is undermined → Reliance on griefbots makes it harder for people to reach out to human networks for support	D
21	Grief data fuels coercion → Sensitive death-related data can be misused for blackmail or control, eroding trust and safety	D
22	Talking about death declines → Dependence on griefbots reduces cultural and emotional capacity to discuss death openly	D
23	Harmful grief cultures emerge → griefbots normalize mourning practices that may be shallow, unhelpful, or exploitative	D
24	Curse of flexibility spreads → The open-ended use of griefbots makes long-term impacts unpredictable and hard to control	D

Table 17: List of risks ideated by human cohorts (human-only condition) for the Death App use case. The table lists $n = 9$ risks surfaced by laypeople (L) and $n = 11$ risks surfaced by domain experts (E) via the Futures Wheel interface. Additionally, $n = 7$ risks were surfaced by domain leaders (D) during the semi-structured interviews. Wording of participant's risks is lightly copy-edited for clarity.

ID	Risk text	Origin
1	The company dissolves due to lack of users	L
2	Poor company reputation that cannot be recovered	L
3	Poor visibility and user engagement with the app	L
4	App is delayed for years in court proceedings	L
5	Lawsuits over the death of someone's family member	L
6	Lawsuits from failed death-ordering attempts	L
7	Medical staff held liable for unauthorized deaths	L
8	The app becomes a hassle for providers	L
9	Providers receive incorrect or misleading information	L
10	People disagree with the death eligibility criteria	E
11	Riots and possible criminal proceedings	E
12	AI becomes more limited in its uses	E
13	Corporations gain monopolies on assisted suicide	E
14	Emotional detachment from suicide across society	E
15	People who should not die may end up using the app	E
16	The app could be misused against unwilling individuals	E
17	Death services proceed despite major public outrage	E
18	Death ordering is blocked from continuing	E
19	Few human doctors remain who consider alternative options	E
20	Protected Health Information (PHI) is leaked	E
21	Death becomes commodified → Platforms reshape cultural attitudes toward dying into transactional, market-based exchanges	D
22	Gendered harms emerge → Death technologies reinforce or destabilize gender equality and identity norms	D
23	Eugenics is normalized → Misuse of death apps risks mainstreaming eugenics-like practices in society	D
24	Governments weaponize death tech → Local authorities use platforms to target or pressure vulnerable groups	D
25	Black and grey markets thrive → Unregulated death services proliferate outside legal oversight, destabilizing trust	D
26	Systemic framings miss lived realities → Policy debates overlook everyday experiences of dying, grieving, and families	D
27	Death industry detaches from healthcare → Assisted dying becomes a separate sector, weakening health system oversight	D

Table 18: List of risks ideated by human cohorts (human-plus-AI condition) for the AI Toy use case. The table lists $n = 4$ risks surfaced by laypeople (L) and $n = 8$ risks surfaced by domain experts (E) via the Futures Wheel interface in the human-plus-AI condition. Wording of participant's risks is lightly copy-edited for clarity.

ID	Risk text	Origin
1	Child's relationships may be stunted or nonexistent	L
2	Malformed thought patterns causing bad relationships	L
3	Children become poor students	L
4	Children could face behavioral problems at school	L
5	Children are misinformed	E
6	Parents trust AI less after mistakes	E
7	Children struggle to grasp basic scientific concepts	E
8	Written grammar suffers as a result	E
9	Children fall behind in school	E
10	Children struggle to develop critical thinking	E
11	Children struggle to grasp concepts that are outside of what they know	E
12	Parents face emotional outsourcing challenges	E

Table 19: List of risks ideated by human cohorts (human-plus-AI condition) for the Griefbot use case. The table lists $n = 8$ risks surfaced by laypeople (L) and $n = 8$ risks surfaced by domain experts (E) via the Futures Wheel interface in the human-plus-AI condition. Wording of participant's risks is lightly copy-edited for clarity.

ID	Risk text	Origin
1	Decline in traditional mourning rituals	L
2	Users develop unrealistic expectations of AI empathy	L
3	Decrease of self critical thinking	L
4	Grief support systems will wither	L
5	The role of the church in processing grief will be diminished	L
6	Con jobs will proliferate through this machinery	L
7	Funerals will lose some of their appeal	L
8	The bot may inappropriately influence the user in important areas such as healthcare	L
9	Cultural tensions from differing grief practices	E
10	Strain on mental health resources grows	E
11	Increased risk of AI-generated misinformation	E
12	Rise in legal disputes over digital remains	E
13	Heightened anxiety from AI miscommunication	E
14	Increased burnout risks for mental health staff	E
15	Social isolation and people becoming fantasist	E
16	Companies would take advantage of people financially and through advertising	E

Table 20: List of risks ideated by human cohorts (human-plus-AI condition) for the Death App use case. The table lists $n = 7$ risks surfaced by laypeople (L) and $n = 20$ risks surfaced by domain experts (E) via the Futures Wheel interface in the human-plus-AI condition. Wording of participant's risks is lightly copy-edited for clarity.

ID	Risk text	Origin
1	Strain on mental health workforce grows	L
2	Misinformation fuels public confusion	L
3	Polarized public opinion stalls policy changes	L
4	Insurance premium hikes for assisted dying coverage	L
5	Wider health disparities from premium hikes	L
6	Potential for misinformation spreading	L
7	Misinformation leads to unsafe choices	L
8	Rise in underground assisted dying services	E
9	Healthcare staff experience moral distress	E
10	Increased burden on legal system	E
11	Increase in unreported assisted deaths	E
12	Underground services increase community distrust	E
13	Polarization of community opinions	E
14	Increase in healthcare provider burnout	E
15	Tax increases	E
16	More unqualified staff	E
17	Mistrust of health professionals	E
18	People avoid seeking help	E
19	Division	E
20	Poorer health	E
21	Lonely people	E
22	Worse mental health outcomes	E
23	New clinical blind spots	E
24	Increased risk of malpractice and poor therapeutic outcomes	E
25	Heightened stigma around mental health issues	E
26	Legal reforms prompt cross-border service challenges	E
27	Wealth inequality grows larger	E

G Results of Systemic Risk Evaluation Across AI Use Cases

In this section, we discuss the main results of our quantitative analysis of agent-generated risks and human-identified risks, under both the human-only and human-plus-AI conditions.

Diversity of Risks. Across all conditions and AI use cases (Tables 21–25), risk distributions were dominated by the "Social" category, indicating that both LLMs and humans primarily focused on societal implications rather than political, economic, technological, legal, or environmental concerns. The Death App consistently exhibited the highest diversity of risks, as reflected by the Shannon-Diversity Index, suggesting that this use case elicited a broader and more heterogeneous set of concerns compared to the others. Environmental risks were absent across LLM-generated risks as well as laypeople- and expert-ideated risks.

Specificity, Novelty, Usability, and Applicability of Risks. To assess differences in these evaluation metrics across the three sets of risks, we analyzed individual Likert-scale ratings (1–5) provided by domain experts in Study 1 and Study 2. We conducted pairwise comparisons between two sets at a time, with the first set always consisting of agent-generated risks and the second set consisting of either human-only or human-plus-AI risks. Because Likert data are bounded and may deviate from normality, we quantified differences using both statistical significance and standardized effect sizes.

Specifically, we calculated Cohen's d to measure the magnitude of mean differences, alongside 95% confidence intervals obtained via analytical standard errors and non-parametric bootstrap resampling (10,000 iterations). In interpreting Cohen's d, values around 0.2, 0.5, and 0.8 can be understood as small, medium, and large effects, respectively (Cohen, 1988). For practical interpretation on the original 1–5 Likert scale (from "strongly disagree" to "strongly agree"), mean differences of about 0.2 reflect subtle shifts in ratings, around 0.5 correspond to roughly half a response category (e.g., from "neutral" toward "agree"), and differences close to 1.0 indicate about a full response category shift (e.g., from "disagree" to "neutral" or from "neutral" to "agree"). Metrics whose confidence intervals did not include zero were interpreted as dimensions in which the two sets of risks differed meaningfully (columns d and 95%CI in the following tables).

As an additional robustness check that does not rely on interval-scale assumptions, we also conducted non-parametric Mann–Whitney U tests to compare the distributions of ratings across the two sets. This test assesses whether ratings from one set tend to be systematically higher than those from the other, and the results were interpreted alongside the standardized mean differences (columns r and p in the following tables).

Across all use cases, a higher share of agent-generated risks was judged to be systemic than risks from either human condition (Tables 26–31), indicating that agents tend to surface risks at broader societal and institutional scales rather than merely increasing output volume. Agent-generated risks were also rated as more severe across all use cases, suggesting that the additional risks surfaced by agents were not perceived as trivial but as potentially consequential. The human-plus-AI condition was the main exception in which severity differences were sometimes non-significant.

On the other hand, human-ideated risks were consistently more accessible than agent-generated risks. Across all use cases, experts indicated that they would *need to learn* less, often by roughly half to a full Likert point, to meaningfully engage with human-identified risks. For the more technologically ready AI Toy use case, human-ideated risks were evaluated to be on par or higher in usefulness and level of detail, while for the more speculative use cases (Griefbot, Death App), agent-generated risks were perceived as more detailed and more useful. This suggests that humans struggle more with distant, uncertain use cases, whereas agents excel at expanding and articulating systemic consequences in these settings.

Table 21: Distribution of LLM-generated risks over PESTEL categories — *P*olitical, *E*conomic, *S*ocial, *T*echnological, *E*nvironmental and *L*egal — for the four AI use cases. The Shannon-Diversity Index (H) quantifies the diversity of the risks in each use case. The LLM-generated risks for the Death App are most spread out across the PESTEL categories. Across all cases, most risks fell into the Social category, while no Environmental risks were identified.

AI Use Case	P		E		S		T		E		L		H
Chatbot Companion	1	0.04	2	0.08	16	0.64	2	0.08	0	0.00	4	0.16	0.62
AI Toy	0	0.00	1	0.04	21	0.78	2	0.07	0	0.00	3	0.11	0.42
Griefbot	1	0.03	3	0.09	14	0.44	3	0.09	0	0.00	11	0.34	0.71
Death App	5	0.19	1	0.04	11	0.42	2	0.08	0	0.00	7	0.27	0.76

Table 22: Distribution of risks ideated by laypeople in the human-only condition over PESTEL categories — *P*olitical, *E*conomic, *S*ocial, *T*echnological, *E*nvironmental and *L*egal — for the four AI use cases. The Shannon-Diversity Index (H) quantifies the diversity of the risks in each use case. As for the LLM-generated risks, the laypeople-ideated risks for the Death App are most diverse. Also similar to the LLM-generated risks, most human-ideated risks fell into the Social category, except for the Death App, where Legal risks were most prevalent. Environmental risks were not identified by participants.

AI Use Case	P		E		S		T		E		L		H
Chatbot Companion	0	0.00	1	0.07	12	0.86	0	0.00	0	0.00	1	0.07	0.28
AI Toy	0	0.00	1	0.07	10	0.71	1	0.07	0	0.00	2	0.14	0.50
Griefbot	1	0.06	1	0.06	16	0.89	0	0.00	0	0.00	0	0.00	0.24
Death App	0	0.00	3	0.15	6	0.30	2	0.10	0	0.00	9	0.45	0.69

Table 23: Distribution of risks ideated by domain experts in the human-only condition over PESTEL categories — *P*olitical, *E*conomic, *S*ocial, *T*echnological, *E*nvironmental and *L*egal — for the four AI use cases. The Shannon-Diversity Index (H) quantifies the diversity of the risks in each use case. As for the LLM-generated risks and the laypeople-ideated risks, the risks ideated by domain experts for the Death App are most diverse. As with the LLM-generated risks, the majority of risks for each AI use case fell into the Social category.

AI Use Case	P		E		S		T		E		L		H
Chatbot Companion	0	0.00	4	0.13	24	0.80	1	0.03	0	0.00	1	0.03	0.38
AI Toy	0	0.00	3	0.30	7	0.70	0	0.00	0	0.00	0	0.00	0.34
Griefbot	0	0.00	0	0.00	15	0.88	2	0.12	0	0.00	0	0.00	0.20
Death App	1	0.14	0	0.00	4	0.57	0	0.00	0	0.00	2	0.29	0.53

Table 24: Distribution of risks ideated by laypeople in the human-plus-AI condition over PESTEL categories — *P*olitical, *E*conomic, *S*ocial, *T*echnological, *E*nvironmental and *L*egal for the four AI use cases. The Shannon-Diversity Index (H) quantifies the diversity of the risks in each use case. Again, the risks ideated by laypeople for the Death App are most diverse. As with the other types of risks, the majority of risks for each AI use case fell into the Social category.

AI Use Case	P		E		S		T		E		L		H
AI Toy	0	0.00	0	0.00	12	0.92	0	0.00	0	0.00	1	0.08	0.15
Griefbot	0	0.00	2	0.06	25	0.76	2	0.06	0	0.00	4	0.12	0.45
Death App	1	0.04	4	0.17	12	0.52	1	0.04	0	0.00	5	0.22	0.70

Table 25: Distribution of risks ideated by domain experts in the human-plus-AI condition over PESTEL categories — *P*olitical, *E*conomic, *S*ocial, *T*echnological, *E*nvironmental and *L*egal for the four AI use cases. The Shannon-Diversity Index (H) quantifies the diversity of the risks in each use case. As in all previous conditions, the risks ideated by domain experts for the Death App are most diverse. As with the other types of risks, the majority of risks for each AI use case fell into the Social category.

AI Use Case	P		E		S		T		E		L		H
AI Toy	0	0.00	2	0.09	19	0.83	0	0.00	0	0.00	2	0.09	0.33
Griefbot	0	0.00	3	0.09	26	0.79	2	0.06	0	0.00	2	0.06	0.42
Death App	1	0.03	2	0.06	17	0.53	2	0.06	0	0.00	10	0.31	0.64

Table 26: Quantitative comparison of domain expert ratings for agent-generated versus human-identified risks (human-only condition) for the AI Toy use case. Bolded dimensions indicate effect sizes of at least small magnitude (> 0.2).

Dimension	μ_{agent}	$\mu_{\text{human\_only}}$	sd	d	95% CI	r	p
Systemic	0.86	0.68	0.37	0.49	(0.25, 0.74)	0.18	0.00
Likelihood	3.79	3.71	1.11	0.07	(-0.14, 0.28)	0.02	0.73
Severity	3.69	3.44	1.11	0.23	(0.01, 0.46)	0.10	0.09
Connected	3.92	4.04	1.04	-0.12	(-0.35, 0.11)	-0.13	0.03
Plausible	3.81	4.06	1.06	-0.24	(-0.45, -0.02)	-0.16	0.01
Specific	3.02	3.54	1.35	-0.39	(-0.63, -0.16)	-0.23	0.00
Novel	2.88	3.11	1.38	-0.17	(-0.40, 0.06)	-0.09	0.12
Original	3.20	3.35	1.34	-0.11	(-0.35, 0.13)	-0.09	0.13
Rare	3.02	3.21	1.32	-0.15	(-0.38, 0.09)	-0.09	0.13
Easy to Engage	3.81	4.14	0.95	-0.35	(-0.61, -0.10)	-0.28	0.00
Need to Learn	3.60	2.54	1.21	0.88	(0.65, 1.12)	0.44	0.00
Useful	3.73	3.85	1.17	-0.10	(-0.31, 0.11)	-0.08	0.19
Detailed	3.44	3.69	1.20	-0.20	(-0.41, 0.01)	-0.12	0.04

Table 27: Quantitative comparison of domain expert ratings for agent-generated versus human-identified risks (human-plus-AI condition) for the AI Toy use case. Bolded dimensions indicate effect sizes of at least small magnitude (> 0.2).

Dimension	μ_{agent}	μ_{hybrid}	sd	d	95% CI	r	p
Systemic	0.86	0.54	0.39	0.81	(0.63, 0.99)	0.32	0.00
Likelihood	3.79	3.46	1.10	0.30	(0.16, 0.44)	0.18	0.00
Severity	3.69	3.26	1.11	0.39	(0.25, 0.54)	0.21	0.00
Connected	3.92	3.68	1.00	0.24	(0.09, 0.40)	0.15	0.00
Plausible	3.81	3.81	1.07	0.00	(-0.15, 0.16)	-0.01	0.89
Specific	3.02	3.02	1.27	0.00	(-0.15, 0.15)	0.00	0.99
Novel	2.88	2.75	1.29	0.10	(-0.05, 0.25)	0.05	0.24
Original	3.20	3.00	1.21	0.17	(0.02, 0.32)	0.10	0.02
Rare	3.02	2.81	1.25	0.17	(0.02, 0.32)	0.10	0.02
Easy to Engage	3.81	3.64	0.95	0.18	(0.02, 0.34)	0.08	0.07
Need to Learn	3.60	2.90	1.21	0.58	(0.42, 0.74)	0.31	0.00
Useful	3.73	3.65	1.13	0.07	(-0.09, 0.23)	0.07	0.11
Detailed	3.44	3.33	1.14	0.10	(-0.05, 0.26)	0.09	0.04

Table 28: Quantitative comparison of domain expert ratings for agent-generated versus human-identified risks (human-only condition) for the Griefbot use case. Bolded dimensions indicate effect sizes of at least small magnitude (> 0.2).

Dimension	μ_{agent}	$\mu_{\text{human\_only}}$	*sd*	*d*	95% CI	*r*	p
Systemic	0.78	0.52	0.43	0.61	(0.46, 0.78)	0.26	0.00
Likelihood	3.71	3.12	1.16	0.51	(0.35, 0.67)	0.22	0.00
Severity	3.68	3.40	1.12	0.25	(0.10, 0.41)	0.11	0.01
Connected	3.76	3.46	1.17	0.25	(0.10, 0.41)	0.13	0.00
Plausible	3.86	3.57	1.16	0.25	(0.10, 0.40)	0.12	0.00
Specific	2.82	2.91	1.26	0.07	(-0.07, 0.21)	0.04	0.32
Novel	2.95	2.78	1.25	0.14	(0.00, 0.28)	0.08	0.06
Original	3.26	2.71	1.18	0.47	(0.32, 0.62)	0.25	0.00
Rare	3.06	2.97	1.26	0.07	(-0.07, 0.22)	0.04	0.34
Easy to Engage	3.69	3.52	0.99	0.17	(0.02, 0.33)	0.07	0.10
Need to Learn	3.63	2.75	1.17	0.75	(0.60, 0.92)	0.40	0.00
Useful	4.03	3.16	1.04	0.84	(0.66, 1.03)	0.42	0.00
Detailed	3.73	3.16	1.05	0.54	(0.38, 0.72)	0.27	0.00

Table 29: Quantitative comparison of domain expert ratings for agent-generated versus human-identified risks (human-plus-AI condition) for the Griefbot use case. Bolded dimensions indicate effect sizes of at least small magnitude (> 0.2).

Dimension	μ_{agent}	μ_{hybrid}	*sd*	*d*	95% CI	*r*	p
Systemic	0.78	0.66	0.43	0.27	(0.16, 0.38)	0.12	0.00
Likelihood	3.71	3.75	1.07	-0.04	(-0.14, 0.06)	0.00	0.87
Severity	3.68	3.45	1.08	0.21	(0.11, 0.32)	0.12	0.00
Connected	3.76	3.92	1.09	-0.15	(-0.26, -0.04)	-0.07	0.02
Plausible	3.86	3.90	1.07	-0.03	(-0.15, 0.08)	0.01	0.81
Specific	2.82	3.01	1.24	-0.15	(-0.27, -0.04)	-0.09	0.01
Novel	2.95	2.89	1.23	0.05	(-0.06, 0.17)	0.03	0.38
Original	3.26	3.03	1.14	0.21	(0.10, 0.32)	0.12	0.00
Rare	3.06	2.96	1.23	0.08	(-0.04, 0.19)	0.04	0.19
Easy to Engage	3.69	3.77	0.95	-0.08	(-0.20, 0.03)	-0.07	0.03
Need to Learn	3.63	3.24	1.17	0.33	(0.22, 0.45)	0.19	0.00
Useful	4.03	3.80	1.03	0.22	(0.10, 0.35)	0.10	0.00
Detailed	3.73	3.42	1.05	0.29	(0.17, 0.41)	0.13	0.00

Table 30: Quantitative comparison of domain expert ratings for agent-generated versus human-identified risks (human-only condition) for the Death App use case. Bolded dimensions indicate effect sizes of at least small magnitude (> 0.2).

Dimension	μ_{agent}	$\mu_{\text{human\_only}}$	*sd*	*d*	95% CI	*r*	p
Systemic	0.78	0.63	0.43	0.37	(0.20, 0.54)	0.16	0.00
Likelihood	3.79	3.81	1.03	-0.01	(-0.18, 0.15)	-0.04	0.33
Severity	3.78	3.54	1.13	0.22	(0.03, 0.40)	0.07	0.13
Connected	3.70	3.93	1.16	-0.20	(-0.35, -0.04)	-0.11	0.01
Plausible	3.73	4.04	1.07	-0.29	(-0.44, -0.13)	-0.18	0.00
Specific	2.87	3.54	1.31	-0.50	(-0.67, -0.34)	-0.27	0.00
Novel	2.88	2.98	1.17	-0.08	(-0.25, 0.09)	-0.04	0.42
Original	3.21	3.03	1.17	0.16	(-0.02, 0.33)	0.08	0.08
Rare	2.94	2.99	1.17	-0.03	(-0.21, 0.14)	-0.02	0.71
Easy to Engage	3.50	3.96	1.03	-0.44	(-0.59, -0.29)	-0.25	0.00
Need to Learn	3.65	2.95	1.12	0.63	(0.44, 0.82)	0.31	0.00
Useful	3.96	3.80	1.01	0.17	(-0.01, 0.35)	0.06	0.21
Detailed	3.85	3.73	1.00	0.12	(-0.06, 0.30)	0.05	0.33

Table 31: Quantitative comparison of domain expert ratings for agent-generated versus human-identified risks (human-plus-AI condition) for the Death App use case. Bolded dimensions indicate effect sizes of at least small magnitude (> 0.2).

Dimension	μ_{agent}	μ_{hybrid}	*sd*	*d*	95% CI	*r*	p
Systemic	0.78	0.63	0.45	0.33	(0.24, 0.42)	0.15	0.00
Likelihood	3.79	3.69	1.04	0.10	(0.00, 0.19)	0.05	0.06
Severity	3.78	3.47	1.14	0.27	(0.18, 0.36)	0.13	0.00
Connected	3.70	3.55	1.21	0.13	(0.03, 0.23)	0.07	0.02
Plausible	3.73	3.59	1.17	0.12	(0.02, 0.23)	0.05	0.09
Specific	2.87	2.64	1.29	0.18	(0.08, 0.29)	0.10	0.00
Novel	2.88	2.84	1.18	0.04	(-0.07, 0.14)	0.02	0.40
Original	3.21	2.93	1.15	0.24	(0.14, 0.34)	0.13	0.00
Rare	2.94	2.86	1.18	0.07	(-0.03, 0.18)	0.04	0.13
Easy to Engage	3.50	3.42	1.09	0.07	(-0.03, 0.18)	0.04	0.17
Need to Learn	3.65	3.44	1.12	0.18	(0.08, 0.28)	0.09	0.00
Useful	3.96	3.69	1.06	0.25	(0.14, 0.37)	0.12	0.00
Detailed	3.85	3.40	1.09	0.41	(0.30, 0.53)	0.21	0.00