News-Informed Probabilistic Models for AI Risk Analysis

Mattia Fumagalli[1], Stefano M. Nicoletti[2], Diego Calvanese[1], and Giancarlo Guizzardi[3]

[1] KRDB Research Centre on Knowledge and Data, Free University of Bozen-Bolzan, Bolzano, Italy
[2] Formal Methods and Tools (FMT), University of Twente, Enschede, The Netherlands
[3] Semantics, Cybersecurity, & Services (SCS), University of Twente, Enschede, The Netherlands

Abstract. The growing adoption of *Artificial intelligence (AI)* has heightened concerns about the need to raise awareness of AI's potential risks. Although several studies have explored AI-related risks, a lack of practical tools to support comprehensive and accessible risk assessment remains. To address this gap, we present an approach that assists in developing a practical solution. The proposed method can be employed to build probabilistic models derived from news reports on incidents involving AI technologies. It aligns with key requirements identified in the literature on AI risk assessment and enables efficient data retrieval and analysis. These capabilities can then be used to support quantitative risk assessment. The feasibility and effectiveness of the approach are validated through a proof-of-concept implementation.

Keywords: AI risk · AI risk assessment · AI risk modeling

1 Introduction

Technologies classified as *Artificial Intelligence (AI)* are increasingly present in people's daily lives. They are typically used to support decision-making processes and/or to generate content across a variety of domains, ranging from medicine to finance. The pervasive diffusion of these artefacts, combined with their often opaque functioning and their accessibility beyond expert circles, has led to growing interest in the risks associated with AI and in its governance [7].

The efforts undertaken by projects aimed at defining regulatory frameworks and analyzing categories of AI-related risks are clear examples of this concern. Among these, one of the most significant initiatives is the development of the *EU Artificial Intelligence Act (AIA)*,[4] which defines and discusses four risk categories: *unacceptable*, *high*, *limited*, and *minimal*. Accordingly, studies have sought to further explore the notion of risk within the specific context of AI technologies.

[4] `https://artificialintelligenceact.eu/`

The works in [13,3,27] and the archive of articles accompanied by a structured taxonomy of AI-related hazards, released by the MIT,[5] are pivotal examples in this direction.

All these initiatives must confront a series of challenges arising from the numerous variables involved in discussions about AI. Chief among them is the difficulty of reaching a shared definition of what constitutes AI, and consequently, of determining the specific operations such systems are expected to perform. Following this, attention must be given to the range of *hazardous events* involving these technologies, the statistical data concerning such events, and the various *subjects involved*, both as *those responsible* for developing the technologies and as *those potentially exposed* to their risks. Moreover, when it comes to control mechanisms, there remains the issue of designing and implementing appropriate mitigation measures to address the identified risks.

For those working in the field of risk assessment applications, these challenges are long-standing and well-known [19,32]. In this regard, our strategy is to bring these areas closer together and contribute to the design of a practical approach for assessing risk in the AI domain. Accordingly, we primarily focus on defining an approach for developing a risk analysis tool that can be used for AI risk assessment. We validate this approach with an implementation, which we test against a set of requirements inspired by relevant studies addressing AI-related risks [27,13]. The main objective of this contribution remains to advance the creation of tools, not necessarily accessible only to expert risk assessors, that foster greater risk literacy in the AI domain, thereby supporting increased awareness among both developers and users of these emerging technologies.

Our contributions, in summary:
- We present a reference conceptualisation of AI risk inspired by existing work[a] (section 2).
- We elicit requirements for assessing AI risk (section 4).
- We present an approach to create AI risk probabilistic models by: (i) extracting information from AI incident news; (ii) operationalising the predefined reference conceptualisation (section 5, section 6).
- We demonstrate the viability of the approach by a prototypical implementation, generating a probabilistic model in ProbLog [8] (subsection 6.2).

[a] [27], `https://artificialintelligenceact.eu/`, `https://www.ipcc.ch/`

The remainder of this manuscript is structured as follows: section 2 introduces a unified conceptualisation about AI risk assessment to be used as reference throughout the paper; section 3 and section 4 provide a brief description of how probabilistic models could be employed in AI risk assessment, and introduces a list of requirements looking at related literature; section 5 and section 6 illustrate the approach we propose; precisely, in subsection 6.2, we discuss a proof-of-concept validation of the proposed approach; section 7 and section 8 are dedicated to related work and final considerations, respectively.

[5] `https://airisk.mit.edu/`

2 Modeling AI Risk

In this section, we present a conceptual background for a principled approach to AI risk assessment, engaging directly with relevant policies and theoretical frameworks. Amid growing debate on AI risk, the scientific community seeks shared reference models and practical assessment tools for both developers and users. As in other risk domains, challenges stem from the polysemous nature of "risk" (as well as "AI"), whose meanings vary across contexts and lack a fully shared consensus. Consequently, efforts focus on widely adoptable specifications and documentation. A prominent example is the *AI Act*, which aims to create a common regulatory framework for AI within the *European Union*.

The AI Act distinguishes four risk levels: *(i) unacceptable*, *(ii) high*, *(iii) limited*, and *(iv) minimal*. As noted in prior work [14], classification criteria are primarily based on the *application domain*, the system's *purpose*, and the *subjects involved*, *exposed to*, or *targeted by* the technology. *Chapter II*[6] defines prohibited AI systems associated with unacceptable risk, while *Chapter III*[7] specifies criteria for high-risk systems. Limited-risk systems (e.g., chatbots) are those that may involve manipulation or deception and are subject to transparency requirements. The minimal-risk category covers all remaining systems.

That said, while the AI Act is central for AI risk modelling, its applicability remains widely debated. As argued in [27] and [28], its predefined criteria do not account for contextual factors such as harm probability, consequence severity, or the values of affected communities. This may result in norms that are overly strict or overly permissive, potentially undermining the EU's AI strategy. Additionally, the framework struggles to address general-purpose technologies, whose diverse and often unforeseeable uses complicate risk classification. [11,12]

Based on this critical analysis, recent efforts aim to support an implementation of the AI Act that enables contextual assessment of risk magnitude by accounting for multiple interacting factors and varied risk scenarios. In particular, [27] proposes integrating the *Intergovernmental Panel on Climate Change (IPCC)* risk framework [4,16], widely used in climate analysis, into the AI Act's terminology. Within this model, AI risk is understood through the interaction of three core *determinants*: *hazard (H)*, *vulnerability (V)*, and *exposure (E)*. As clarified in [1], a *hazard* is an event that may cause loss; *vulnerability* denotes the susceptibility of exposed elements to harm; and *exposure* refers to the condition in which an at-risk object may be adversely affected by an incident. When used to model a risk scenario, *Hs*, *Vs*, and *Es* may be characterized by correlations (or interactions [1,4]). For example, a hazard such as *a poor training of the model* may entail another hazard, namely the *impossibility of controlling or understanding the model's outputs*. Similarly, an exposure such as a *user's privacy* may be correlated with *age* or *low level of technological literacy*, which may be considered vulnerabilities.

[6] `https://artificialintelligenceact.eu/chapter/2/`
[7] `https://artificialintelligenceact.eu/chapter/3/`

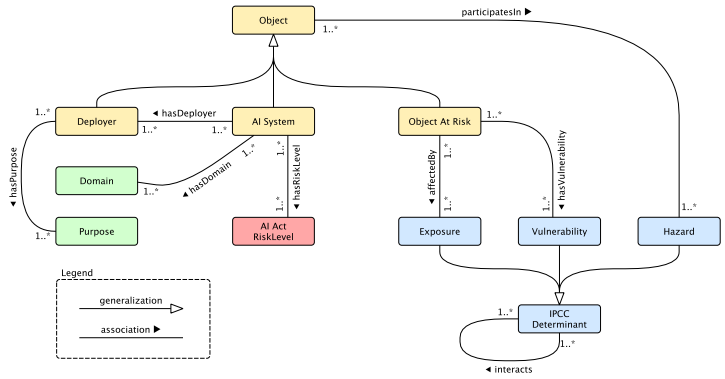

Fig. 1. Overview of the key concepts taken into consideration to model AI Risk.

Figure 1, building on the AI Act and the IPCC-based reworking introduced, aims to offer a partial, but unified and disambiguated conceptual *UML class-diagram*[8] representation of the key notions introduced above. Such notions can be directly mapped into the ones reported in [14], which is aimed at represent AI risk, according to the AI act; the ones discussed in [27], and the ones introduced in the work discussed in [32], which provides *COVER*, namely a *Core Ontology of Values and Risk* for risk representation and assessment, and was itself subject to validation and proper comparison to the literature of risk in risk analysis and management at large (e.g. [18,19]).

On the bottom right, the blueish boxes denote concepts inherited from the IPCC vocabulary. On the top left, the yellowish boxes encode some of the types of objects assumed by the AI Act classification; namely, the `deployer` of the target AI, the `target of risk` or `object at risk` (as defined in [32]), i.e., the object that is at risk with the target AI, and the `AI system` itself.[9] The AI system is associated with the `risk levels` established by the AI documentation, with a `domain of application` (or use-cases),[10] and indirectly with a `purpose` (through its relationships with the corresponding deployer). These concepts are, in turn, connected by some key relations, denoting interdependencies, such as the `participation` relation between objects and hazards [32], and the `impacts on` relation between AI systems and objects at risk.

Two observations follow. First, although this article does not aim to deliver a formal AI risk representation like [14], the outlined framework serves as an explicit reference for the concepts and relations used throughout. Second, the model is not exhaustive; it is designed to integrate established risk-analysis conceptualisations (e.g., [32]) and should be seen as evolvable. These extensions are left for future work. For present purposes, the framework offers a transparent minimal kernel to be operationalized in the following sections.

[8] `https://www.uml-diagrams.org/class-diagrams-overview.html`
[9] Here, more objects such as the provider or the user might be added.
[10] `https://artificialintelligenceact.eu/high-level-summary/`

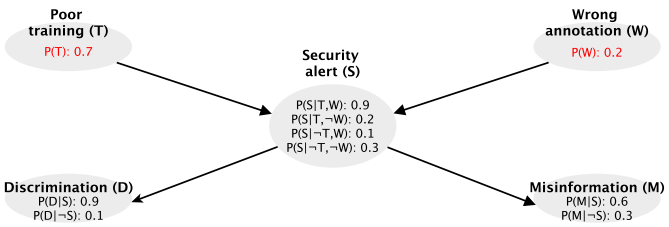

Fig. 2. An example of a Bayesian network-based risk model.[13]

3 Probabilistic Models for AI Risk Assessment

Under the AI Act and its IPCC-informed reinterpretation, AI risk is primarily expressed through categorical labels and concepts aimed at supporting understanding and classification. For example, Annex III[11] classifies as high-risk AI systems used in *recruitment* and *human resources* when deciding employee promotions. Likewise, the IPCC-based approach highlights how discrimination bias may result from interacting hazards, such as model opacity, limited dataset size, or poor-quality or misused training data (e.g., AIA, Art. 10).[12]

By contrast, safety engineering, security analysis, and formal methods typically adopt probabilistic models, defining risk as a function of event probability and impact [34,35]. These approaches use graph-based structures, where nodes denote events or states and edges encode causal or probabilistic dependencies, enabling risk propagation through simulation or analytical methods [10].

Well-established formalisms within this paradigm include Bayesian Networks [29,20], Markov Chains [6], and reliability- or security-oriented models such as fault trees (FTs) [31,25] and attack trees (ATs) [21,26]. These frameworks are widely used in safety-critical domains, including aviation, autonomous driving, and aerospace software, where their adoption is often mandated by standards and regulatory bodies such as the FAA, the NRC, and ISO 26262 [33,17].

Figure 3 illustrates a simple Bayesian Network that includes hypothetical events, states, and/or situations, and their correlations, within a scenario involving the use of AI systems.[14] The graph is directed and acyclic. Nodes without incoming edges, such as PoorTraining (T) and WrongAnnotation (W), are regarded as independent. Nodes that receive edges are considered dependent on the nodes from which the edges originate. For instance, securityAlert (S) depends on T and W, and the event Discrimination (D) depends on S. In such a model,

[11] `https://artificialintelligenceact.eu/annex/3/`

[12] `https://artificialintelligenceact.eu/article/10/`

[13] Complementary probabilities are compiled by default, e.g., $P(S \mid T, W) = 0.9$, then $P(\neg S \mid T, W) = 0.1$.

[14] Note that the semantics of Bayesian network nodes and relations are often ambiguous and may hide the presence of objects or other categories. For a detailed discussion, see the work presented in [10].

risk analysis consists of simulating specific events and updating the risk values associated with nodes that represent harmful outcomes.

The underlying rationale is straightforward. Each node is associated with a *probability value*, for example $P(\mathsf{T}) = 0.7$ and $P(\mathsf{W}) = 0.2$, which is usually extracted from historical data or is simply assigned to assess hypothetical scenarios. The risk associated with a node can then be derived as the product of its probability and a loss value [35], which is not explicitly represented in the graph, and it is usually assigned by an assessor or the person that is directly affected by the corresponding event. In Figure 3, if Misinformation entails a loss of 8 on a scale from 0 to 10, its risk would be $P(\mathsf{M}) \cdot 8$. Likewise, if Discrimination involves a loss of 10, its risk would be $P(\mathsf{W}) \cdot 10$.

Several methods of this kind have recently been proposed and adapted specifically to AI systems [37,15]. Nevertheless, efforts to explicitly quantify AI risk through probabilistic models continue to face significant unresolved challenges. For instance, Piorkowski et al. [30] review open problems in quantitative AI risk analysis, highlighting limitations in current metrics and emphasizing the need for more structured, computable representations of risk. Furthermore, most existing approaches remain largely domain-agnostic with respect to regulatory frameworks such as the AI Act and its grounding in the IPCC risk framework. As a result, they do not by themselves offer a principled conceptualisation of AI-specific risk levels, nor do they systematically incorporate key determinants of risk, including hazards, exposures, and vulnerabilities.

4 A (Non-exhaustive) List of Requirements

How can probabilistic risk assessment methods support AI risk analysis under frameworks such as the AI Act and its IPCC-based reinterpretation? If simulation techniques can model risk scenarios, how should AI risk levels, hazards, and related determinants be identified and represented? How can their interactions be formalized, and how can associated probabilities be estimated?

The current work aims to contribute to addressing these issues. The core challenge is to translate existing conceptual frameworks into a coherent probabilistic model. This requires collecting quantitative data to support statistical components and defining the operations expected from an AI risk analysis tool, while remaining aligned with established reference frameworks.

In what follows, we show how this objective can be advanced by outlining an approach for designing models that support operations grounded in key assumptions of the AI Act and the IPCC framework. As a first step, we define a set of requirements that the approach should meet.

R1. Users should be able to associate a given AI technology with a specific risk level, in line with the AI Act.

R2. It should be possible to identify the *hazard (H)*, *vulnerability (V)*, and *exposure (E)* determinants of a given risk level, following [27]. This requires a structured set of IPCC-aligned concepts and their correlations, including

hazard types, their possible *drivers* (intended as the other possible hazards that trigger/involve them), associated risk levels, and links to specific vulnerabilities and exposure conditions.

R3. A third requirement concerns the possibility of reporting detailed information about risk scenarios, including the AI system's *purpose*, *application domain*, *exposed subjects*, responsible actors (e.g., *deployers*), and other contextual elements necessary for classification under the AI Act.

R4. The relationships among objects, domains, purposes, and IPCC determinants must be explicit. This means that, for example, it should be possible to trace the vulnerabilities of exposed entities, as well as to identify the hazards in which specific AI targets participate, and to clarify how these elements are interconnected.

R5. It should be possible to navigate the network of interactions among the determinants identified under **R2**. Although the IPCC model does not provide a standardised strategy for identifying such correlations [24,1,10], it remains pivotal to trace, for instance, the drivers of a given hazard, as this may be key to identifying effective mitigation strategies.

R6. An additional requirement concerns the possibility of enabling the calculation of risk magnitude based on customizable thresholds, for example, by assessing how frequently a technology appears within a given risk level. This is especially relevant for broad categories such as "high-risk" where recurrence patterns may indicate differing degrees of concern.

R7. Finally, an additional requirement, orthogonal to the previous ones, is the ability to support the retrieval of statistical information associated with the defined concepts. For example, it should be possible to determine the probability that one hazard is correlated with another; to estimate how frequently a given AI system occurs in connection with a specific hazard; or to establish how often a particular AI system is associated with a given risk level.

Taken together, these requirements, identified through the ongoing debate on AI risk analysis, provide a reference point for designing a probabilistic model for AI risk assessment, based on AI act and IPCC specifications. In the next section, we present our approach for constructing such a model.

5 Characteristics of the Target Model

We preface the analysis of our approach that extracts the target model with a description of the main characteristics that our proposed model exhibits, also in light of the requirements outlined above and the conceptual framework introduced via Figure 1. These characteristics represent a fundamental parameter for the extraction process that will be described in the following section. Nevertheless, since this remains a parameter, we do not exclude the possibility that the model may be modified or further developed in the future. On the contrary, our decision to also elaborate on the extraction process aims precisely to emphasize the evolving nature of the proposed model, which can undoubtedly benefit from

collaboration with domain experts (e.g., experts on the AI Act, experts from the IPCC, and/or the authors of the related work taken into consideration).

```
% Deterministic facts
ai_system(scoring_model).
deployer(hr_department).
object_at_risk(job_applicant).
ai_domain(recruitment).
purpose(candidate_screening).
exposure(legal_principles).
vulnerability(ethnicity).
hazard(candidate_rejection).
hazard_driver(poor_training).
root_driver(annotation_error).
risk_level(high_risk).
% Concept relations
P::interacts(hazard(candidate_rejection), hazard_driver(poor_training)).
P::interacts(hazard_driver(poor_training),root_driver(annotation_error)).
P::interacts(root_driver(annotation_error),vulnerability(ethnicity)).
P::interacts(vulnerability(ethnicity),exposure(legal_principles)).
P::participates_in(ai_system(personality_scoring_model),hazard(discrimination)).
P::participates_in(object_at_risk(job_applicant),hazard(discrimination)).
P::affected_by(object_at_risk(job_applicant),exposure(legal_principles)).
P::has_vulnerability(object_at_risk(job_applicant),vulnerability(ethnicity)).
P::has_deployer(ai_system(scoring_model),deployer(hr_department)).
P::has_purpose(ai_system(scoring_model),purpose(candidate_screening)).
P::has_domain(ai_system(scoring_model),ai_domain(recruitment)).
P::ai_risk(ai_system(scoring_model), risk_level(high_risk)).
% Rules
purpose_ai_domain(X,Y,Z) :- has_purpose(X,Y),has_domain(X,Z).
hazard_ai_risk(X,Y,Z) :- ai_risk(X,Y),participatesIn(X,Z).
chain(X,Y,[X,Y]) :- interacts(X,Y).
chain(X,Y,[X|Rest]) :- interacts(X,Z),chain(Z,Y,Rest).
monitorWarning(AI) :- subquery(ai_risk(AI,high_risk),P),P > 0.3,P =< 0.5.
```

Listing 1: Examples of deterministic facts, concept relations, and rules.

Probabilistic logic language. The language selected to represent the model is *ProbLog*, [8],[15] a well-established *probabilistic logic programming language* in which facts and rules may be annotated with probabilities. ProbLog enables information retrieval from the model in a style typical of classical knowledge bases, while also supporting queries like probabilistic models such as Bayesian networks. Furthermore, it provides mathematical operators that allow the construction of rules, the definition of thresholds, and the execution of computations based on the probabilistic values embedded in the theory.

Deterministic facts. Concerning the identification of key concepts to be included, we manually analyzed and derived a set of terms, defining the boundaries of what can be extracted. Each term can represent an instance of the concepts defined in Figure 1. In particular, from the AI Act Chapter II (Article 5) and Chapter III (see, for example, Articles 6, 8–17, and Annexes I and III), also following the work in [14], we derived a list of: *AI systems*, *intended scopes* (or purposes), *application domains*, potential *object at risk*, and possible *deployers*. We also derived a set of potential *hazards*, *hazard drivers*, *root hazard drivers*, *vulnerabilities*, and *exposures*, by also taking into consideration the AI risk taxonomy provided by MIT as well, and comparing it with the data contained in the

[15] `https://ProbLog.readthedocs.io/en/latest/`.

AI Act documentation. This process enabled us to acquire a reasonably broad reference vocabulary of concepts for an initial case study, currently available in a corresponding file (`'conceptualisation.py'`), which can be exploited by the prototype implementation. Each derived concept can then be represented in the model via a deterministic fact, as illustrated in Listing 1 below (records 2-13).

Concept relations. The model allows for conceptualising the types of constructs necessary to represent the relationships among the derived terms. Regarding the interactions among the risk determinants, namely hazards (Hs), exposures (Es), and vulnerabilities (Vs), for the version proposed in this paper, we represented them with a binary relation *"interacts"* [27,4]. Accordingly, Hs, Es, and Vs derived from the input data can be chained as in Listing 1, (records 14–17). Each relation is also associated with a probability value ("`P`"), derived from the number of times the encoded interaction occurs. More precisely, given a "target" and a "source" of the interaction, $P(Source|Target) = \frac{Count(Source,Target)}{Count(Source)}$. The criteria for identifying such interactions, the associated probabilities, and the corresponding extraction methods may vary significantly. A detailed analysis of the available options and optimal strategies is beyond the scope of this work, but it represents a task for future research.

Regarding the relationships among objects, purposes, domains, determinants, and risk levels, the model has the relations presented in Listing 1 (records 18-25) (as from Figure 1). Also in this case, each relation is associated with a conditional probability value derived from the reference data. Considering the association between an AI system and the risk level defined in the AI Act, in order to extract the relevant information and to remain consistent with the classification criteria provided, we assumed as a constraint that an AI system must be linked to the concepts characterizing its risk level. Such a risk level depends on what is related to that AI system. For example, if the purpose of the AI system is "assessing personality traits", the application domain is "law enforcement", and the target subjects are "persons" or "groups of persons," the system is classified as high-risk, as described in Annex III, point 6(d) of the AI Act. Accordingly, the theory is informed by a series of conditions (e.g., *"iff domain is X and purpose is Y then risk level is Z"*), derived from the AI Act. Those conditions can be adapted and extended by updating the reference conceptualisation to be used for creating the target model. Finally, the relationship between the AI system and its risk level is also associated with a probability value, based on the number of times that a specific AI system is linked to that risk level (over the total number of times that the same AI system is associated with any risk level).

Rules. As a fourth aspect, the model supports the formulation of rules to query the defined facts. For example, it can retrieve all purposes and application domains linked to a specific AI system, together with their associated probability values (Listing 1, record 27). It can also identify all hazards involving a high-risk AI system (Listing 1, record 28) and trace the chain of interacting determinants (Listing 1, records 29–30).[16] As a final example, by setting a probability thresh-

[16] Note that the `chain` predicate can be associated with a probability value that is derived as the joint probability of the probabilities of the given interactions.

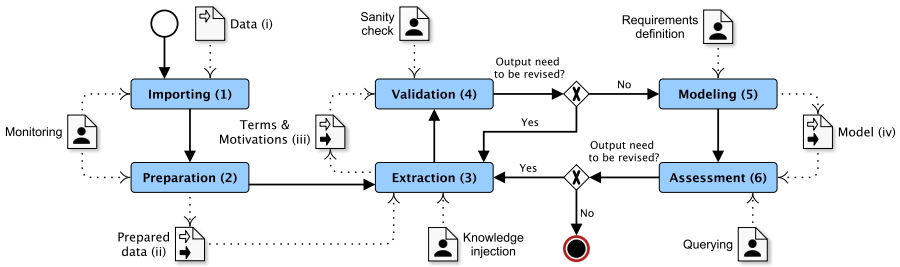

Fig. 3. The model generation workflow.

old for the relationship between an AI system and the AI Act risk level, the model can determine the system's recidivism level (Listing 1, record 31).

Also in this case, multiple modeling choices are possible. At this stage, our primary objective is to emphasize the model's level of expressiveness and flexibility, together with its fidelity to the reference terminology.

6 Obtaining the Target Model

6.1 A Principled Process

In Figure 3 we provide a complete representation of the process we propose. This is divided into six tasks, each involving both automated procedures and user interventions, where a prototypical user can be assumed to be both the individual responsible for creating and the person employing the final model for AI risk assessment.

(1) Importing. The first task concerns the acquisition of the information needed to construct the model (see *Data (i)* in Figure 3). The data sources may consist of textual information downloaded from platforms that collect reports of incidents related to the use of AI or other sources, such as expert-generated reports or technical documentation describing incidents observed in specific applications. At the current stage, we do not require the source textual information to follow a standardized format or structure, although such standardization could be introduced to optimize subsequent steps of the process. The importing task may thus be understood either as a combination of automated web scraping and parsing activities, supervised by a user, or as a purely manual collection of relevant textual data.

(2) Preparation. This task aims to clean the imported data and format it into a structure suitable for the subsequent extraction phase. In addition to defining the necessary fields a key part of this step concerns the removal or consolidation of duplicate information referring to the same event. This is essential to prevent the final model's statistics from being compromised. To mitigate this issue, in addition to user supervision, we include automated support

for deduplication, assuming that its output (see *Prepared data (ii)* in Figure 3) is always validated by a user. The final output of this task is a set of records with clean and deduplicated information.

(3) Extraction. This task comprises two main aspects. Firstly, it requires defining the conceptual schema that guides the information extraction process and will be populated with the extracted data. Secondly, it relies on *automated support to extract the concepts* needed to build the risk assessment model. Regarding the conceptual schema, although here this is primarily grounded in the AI Act and its IPCC reworking (section 5), it may incorporate any relevant background knowledge, whether provided by domain experts or derived from existing models and taxonomies in the literature. Such knowledge should specify the key concepts and relationships (e.g., hazard types and their links to objects at risk) required to construct the final assessment model.

For the second aspect, the proposed approach leverages the knowledge extraction capabilities of existing *Large Language Models (LLMs)* [23]. The data prepared in the previous task are supplied to an LLM together with the background knowledge defining the conceptual schema. The output consists of two structured objects: one containing all extracted information, along with brief justifications for each extraction decision; and another including all the terms needed to construct the model, ready for direct use (see *Concepts & motivations (iii)* in Figure 3).

(4) Validation. Considering the potential limitations in reliability and accuracy of the LLM used for extraction,[17] as well as the variability of the conceptual schemas that may be adopted, this task involves a detailed validation of the output produced in step (3). If the extracted set of terms is not deemed acceptable, the process may return to the previous task.[18] This validation step may include selecting a different LLM, updating the contextual knowledge, or substituting it entirely to better guide the extraction process. We assume that a user always reviews the results to ensure quality, although validation methods based on gold standards (when available) may also be employed, especially to assess the accuracy of the extracted concepts.

(5) Modeling. If the output generated by the extraction task (3) is successfully validated in (4), it is used as input for the construction of the target probabilistic model. The goal of this task is to take the list of terms associated with all processed news items and derive the set of facts and rules that can support the queries required to analyze the target AI risk scenario.

(6) Assessment. The final task of the process concerns the use of the generated probabilistic model. At this stage, user intervention is again possible if the functionalities of the resulting model do not meet expectations. In such cases, the process is resumed starting from the extraction task, ensuring that any necessary adjustments can be made, assuming that the input data validation remains satis-

[17] Note that several studies on the reliability of LLMs in data and event extraction has demonstrated that, although those might be sometimes subject of "hallucinations", the level of accuracy of the extraction task is very high [5].

[18] For instance, the LLM may associate the wrong type of incident with a given news.

factory. The possibility of returning to task (3) from this point is essential, since certain limitations, such as issues arising from the conceptualisation adopted during extraction, may become apparent only once the final model is used. The same consideration applies to the definition of the requirements formulated in task (5). If, instead, the model is deemed adequate for its intended purpose, the user may proceed to employ it for the analysis.

6.2 Proof of Concept Implementation

To validate the feasibility of the proposed approach, we developed a prototype application in Python, which was executed on a *MacBook Pro* (Retina, 13-inch, Early 2022) with CPU 8-core Apple M2 chip, and is available here: `https://github.com/Matt-81/RAIN`. As the primary information source, we relied on the archive of news on AI-related incidents provided by the *OECD.AI* portal.[19]

Regarding the extraction task and its subsequent validation, we adopted the following strategy. Using a combination of Ollama and LangChain,[20] we locally installed three small models (*gemma3*, *llama3.2*, and *mistral*)[21] and integrated them into a simple pipeline to extract structured information from the list of input news items.[22] The most important part of this phase involved designing a prompt to guide the extraction by defining the reference conceptualisation (i.e., the terms derived from the reference literature that can be used for creating the target model).

To determine which large language model was the most suitable for the target task, we validated the installed models with the defined prompt, using a set of 100 news items manually annotated with the concepts to be extracted, and evaluated their extraction precision. We also run an additional cross-check manual validation.[23] The results of the validation are available in the repository with the data to replicate the experimentnt.[24] Taking into account both the precision values and the quality of the generated explanations, we selected *gemma3* as the model for extracting the information subsequently used in the construction of the risk assessment model.

The file of concepts extracted using *gemma3*, and subsequently validated, is then used as input for the generation of the final model together, which was generated through a script that produces a theory in ProbLog similar to the one presented in section 5. Below, we present some features of the generated theory,

[19] `https://oecd.ai/en/incidents`

[20] `https://ollama.com/`, `https://docs.langchain.com/`

[21] `https://ollama.com/library`

[22] The choice of models was guided primarily by their relatively small size compared to other available models, and because they are recommended for structured information extraction from text.

[23] The reliability of LLMs in knowledge extraction has already been proved [5]. In a more structured experimental setup, with more computational capabilities, we can use LLMs that are more powerful and accurate compared to the ones we selected.

[24] To derive the precision value, we checked whether the manually extracted concept corresponded to the automatically extracted one.

which is available in the repository with the implementation files, grouping some of the admissible queries and reporting some example results, considering the requirements introduced in section 4.

R1, R2, R3, and R7. The generated model enables the retrieval of risk levels for a selected AI system (**R1**). In addition, it allows the identification of the *hazards*, *vulnerabilities*, and *exposures* that determine the associated risk level (**R2**). The model also supports the reporting of detailed information about risk scenarios (**R3**), including, for example, the AI system's purpose and application domain. Accordingly, the theory supports queries like the following:

$$\begin{aligned} (i)\ & \exists a \exists r\ ai\_risk(a, r). \\ (ii)\ & \exists a \exists h\ ai\_hazard(a, h). \\ (iii)\ & \exists a \exists h \exists r\ ai\_hazard\_risk(a, h, r). \\ (iv)\ & \exists o \exists v \exists e\ object\_vuln\_exp(o, v, e). \\ (v)\ & \exists a \exists h \exists p \exists d\ ai\_purpose\_domain(a, h, p, d). \end{aligned} \tag{1}$$

An example of a rule supporting, for instance, query (3-i) is the following:

$$\forall a \forall r \left(ai(a) \wedge risk\_level(r) \rightarrow\ ai\_risk(a, r) \right). \tag{2}$$

Examples of the results that can be returned by (3-i) selecting "facial recognition system" as an AI system are the following:

```
▶ 0.333::ai_risk(ai(facial_recognition_system), risk_level(unacceptable)).
▶ 0.1667::ai_risk(ai(facial_recognition_system), risk_level(high)).
▶ 0.1667::ai_risk(ai(facial_recognition_system), risk_level(limited)).
▶ 0.333::ai_risk(ai(facial_recognition_system), risk_level(minimal)).
```

where the probability values on the left (**R7**) represent the distribution of the corresponding risk levels across the input news dataset (e.g., if the dataset contains 100 news articles, the "facial recognition system" is classified with an unacceptable level of risk in ∼33 of them).

R4, R5, R6, and R7. The model allows querying the relationships of the reference conceptualisation presented in Figure 1 (**R4**), thereby enabling navigation of the network of interactions among risk determinants (**R5**). Moreover, it allows the calculation of risk magnitude based on customizable thresholds (**R6**).

Accordingly, the theory supports queries like the following:

$$\begin{aligned} (i)\ & \exists h \exists d \exists r\ hazard\_to\_root(h, d, r) \\ (ii)\ & \exists a \exists h \exists d \exists r \exists p \exists x\ full\_pathway(a, h, d, r, p, x) \\ (iii)\ & \exists a\ monitor\_warning(a) \end{aligned} \tag{3}$$

An example of a rule supporting, for instance, query (3-iii) is the following:

$$\forall a \forall r \left(P(ai\_risk(ai(a),\ risk(unacceptable))) \geq 0.3 \rightarrow\ monitor\_warning(a) \right)^{25} \tag{4}$$

By binding the ''`risk`'' predicate to ''`unacceptable`'' and keeping the example threshold, query (3-iii) may return, for instance, ''`monitor_warning(facial_-recognition_system)`'' as a system to be monitored, since its classification as "unacceptable" in the given dataset occurs with probability 0.333, namely in more

[25] P represents the probability value associated with the predicate "has_risk".

than 30% of cases (**R7**). Similarly, queries analogous to (3-i) and (3-ii) enable the retrieval of information such as *"the hazards in which AIs associated with a high risk level are involved"*, *"the purposes of AIs involved in a given hazard"*, and *"the vulnerabilities, the AIs, the exposure, and the objects involved in a given hazard"*. All queries supported by the proof-of-concept model we generated are available in the corresponding ProbLog file (`'ai_risk_concepts_problog.py'`), which can be executed for testing and adaptation.

7 Related Work

Our approach spans multiple dimensions, making it challenging to identify a single, well-defined body of related work. Nonetheless, the research that is most closely aligned with ours in spirit consists of efforts to develop models for assessing AI risk. Within this landscape, we can distinguish three main categories of relevant work.

Data-driven approaches. A representative example is [2], which examines how AI risks are framed through news media analysis. Similarly, Felländer et al. [9] propose *DRESS-eAI*, a data-driven risk-assessment method based on organisational challenges in ethical AI implementation. These works aim to bridge high-level guidelines and actionable risk management, emphasizing practical workflows aligned with our goal of operationalizable assessments. However, they do not offer probabilistic models for reasoning over the analysed data. Our approach complements this line of work by converting news-reported incidents into operational probabilistic structures that enable systematic analysis and querying.

Conceptual approaches to AI risks and governance. Golpayegani et al. [14] introduce the AIRO ontology to represent AI-system risks under the EU AI Act, partly grounded in incidents from the AIAAIC repository. Their incident-based conceptual formalisation aligns with our aim of grounding structured representations in real-world failures, but focuses on regulatory compliance rather than probabilistic modelling or data-driven querying. Similarly, the JRC report [36] develops a taxonomy of AI types through mixed methods. While not risk-focused, it provides useful classificatory scaffolding, yet lacks probabilistic reasoning and direct linkage to hazard data. Novelli et al. [27] explore applying the IPCC framework to assess AI risk magnitude, offering a conceptual and normative perspective without practical assessment tools. In contrast, our work advances these approaches in a data-driven direction by structuring and quantifying real-world AI incidents within a probabilistic model.

Quantitative approaches to risk assessment. Piorkowski et al. [30] review the technical challenges of quantitative AI risk analysis, surveying existing metrics and noting methodological limitations. They stress the need for structured, metric-ready representations but do not offer a concrete implementation—an opening our framework addresses. Likewise, Nagbøl et al. [22] present the AIRA tool to support AI system design and assessment. Complementing this work, we propose a scalable method for building structured probabilistic risk models directly from news-reported incidents.

8 Conclusion

In this article, our primary goal is to propose an approach that supports the creation of usable, data-driven probabilistic models for AI risk assessment, leveraging existing AI risk conceptualisations such as the one provided by the AI Act. As illustrated, the approach involves several tasks. The implementation we present is intended to demonstrate feasibility; however, it does not imply that the specific solutions adopted for each task are optimal.

Limitations and perspectives. In our view, the limitations arising from the design of the target model and its proof-of-concept implementation help highlight areas that deserve further attention in future work. A first threat to validity concerns the reliability of the news sources used as input in the reference example. Because the data used to construct the model shape the statistical patterns embedded within it, an insufficient or biased corpus may compromise the robustness of the results. Future work should therefore substantially expand the dataset and diversify the sources considered. In addition to incorporating further authoritative news portals, expert-generated material could be introduced to enrich the dataset. Experts may even be guided in drafting the textual input to enhance the quality and consistency of the information extraction process. Another aspect concerns the use of the language model. To limit the number of variables, we assessed and deployed a few small models; however, it would be valuable to validate the extraction potential of much larger models. It would also be useful to better structure the validation task, for example by creating a gold standard on a substantially larger dataset. This could support the development of a benchmark for extracting risk graphs from texts describing AI-related incidents. There is also considerable room for improvement in the conceptualisation and the associated vocabulary of terms (e.g., hazard types, drivers, vulnerabilities, and exposure) that guide the extraction process and the population of the probabilistic model. At this stage, we provide an initial reference vocabulary that could be refined and extended, potentially with the support of domain experts. The rules used to classify AI risk levels were derived from a preliminary analysis and may therefore require further refinement. It would also be worthwhile to explore whether a more precise conceptual theory, such as an ontology defining concepts and properties, could support more accurate extraction and enable greater expressiveness in the resulting assessment model. Furthermore, alternative retrieval-augmented generation (RAG) techniques could be investigated, since in the current implementation knowledge is incorporated through a conventional prompt. Importantly, the proposed approach has been designed so that the reference input conceptualisation and the classification rules can be adapted, improved, and extended through more systematic requirements elicitation, such as expert interviews or a broader literature review, without affecting the core process we introduce.

Nevertheless, given the current lack of practical applications for AI risk assessment, the results reported in this paper represent a step toward bridging this gap and aim to stimulate collaboration with domain experts who can further develop these contributions by addressing the identified open issues.

Acknowledgments

Mattia Fumagalli and Diego Calvanese have been partially supported by the HEU project *CyclOps (GA n. 101135513)*, by the Province of Bolzano and FWF through project *OnTeGra (DOI 10.55776/PIN8884924)*. By the Province of Bolzano and EU through projects *ERDF-FESR 1078 CRIMA* and *ERDF-FESR 1047 AI-Lab*, and by MUR through the PRIN project *2022XERWK9 S-PIC4CHU*. Stefano Nicoletti was partially funded by the European Union's Horizon 2020 research and innovation programme under the *Marie Skłodowska-Curie grant agreement (101008233)*, and the ERC Proof of Concept grant *101187945* (*RUBICON*). Giancarlo Guizzardi was supported by the research project *"DECIDE: Democratizing AI – Empowering Citizens through Transparent Decision-making"* (`https://www.nwo.nl/en/projects/iweog51094`).

References

1. Adamo, G., Sperotto, A., Fumagalli, M., Mosca, A., Sales, T.P., Guizzardi, G.: Unpacking the semantics of risk in climate change discourses. In: 14th International Conference on Formal Ontology in Information System, FOIS 2024. pp. 163–177. IOS (2024)
2. Allaham, M., Kieslich, K., Diakopoulos, N.: Informing ai risk assessment with news media: Analyzing national and political variation in the coverage of ai risks. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. vol. 8, pp. 90–103 (2025)
3. Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y.N., Zhang, Y.Q., Xue, L., Shalev-Shwartz, S., et al.: Managing extreme ai risks amid rapid progress. Science **384**(6698), 842–845 (2024)
4. Change, I.P.O.C.: Ipcc. Climate change (2014)
5. Chen, R., Qin, C., Jiang, W., Choi, D.: Is a large language model a good annotator for event extraction? In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 17772–17780 (2024)
6. Ching, W.K., Ng, M.K.: Markov chains. Models, algorithms and applications (2006)
7. Christian, B.: The alignment problem: Machine learning and human values. WW Norton & Company (2020)
8. De Raedt, L., Kimmig, A., et al.: ProbLog: A probabilistic Prolog and its application in link discovery. In: Veloso, M.M. (ed.) Proceedings of IJCAI 2007. pp. 2462–2467 (2007)
9. Felländer, A., Rebane, J., Larsson, S., Wiggberg, M., Heintz, F.: Achieving a data-driven risk assessment methodology for ethical ai. Digital Society **1**(2), 13 (2022)
10. Fumagalli, M., Engelberg, G., Sales, T.P., Oliveira, Í., Klein, D., Soffer, P., Baratella, R., Guizzardi, G.: On the semantics of risk propagation. In: International Conference on Research Challenges in Information Science. pp. 69–86. Springer (2023)
11. Fumagalli, M., Ferrario, R.: Leveraging teleological explanation to support general-purpose ai assessment. AI & SOCIETY **41**(2), 825–842 (2026)
12. Fumagalli, M., Ferrario, R., Guizzardi, G.: A teleological approach to information systems design: M. fumagalli et al. Minds and Machines **34**(3), 23 (2024)

13. Giudici, P., Centurelli, M., Turchetta, S.: Artificial intelligence risk measurement. Expert Systems with Applications **235**, 121220 (2024)
14. Golpayegani, D., Pandit, H.J., Lewis, D.: Airo: An ontology for representing ai risks based on the proposed eu ai act and iso risk management standards. In: Towards a knowledge-aware AI, pp. 51–65. IOS Press (2022)
15. He, L., Jia, Q.S., Li, A., Sang, H., Wang, L., Lu, J., Zhang, T., Zhou, J., Zhang, Y., Wang, Y., et al.: Towards provable probabilistic safety for scalable embodied ai systems. arXiv preprint arXiv:2506.05171 (2025)
16. Hulme, M., Mahony, M.: Climate change: What do we know about the ipcc? Progress in Physical Geography **34**(5), 705–718 (2010)
17. International Standardization Organization: ISO/DIS 26262: Road vehicles, functional safety. `https://www.iso.org/standard/68383.html` (2018)
18. ISO: Risk Management - Vocabulary, ISO Guide 73:2009 (2009)
19. ISO: ISO 31000:2018 - Risk management – Guidelines (2018)
20. Kabir, S., Papadopoulos, Y.: Applications of bayesian networks and petri nets in safety, reliability, and risk assessments: A review. Safety science **115**, 154–175 (2019)
21. Mauw, S., Oostdijk, M.: Foundations of attack trees. In: International Conference on Information Security and Cryptology. pp. 186–198. Springer (2005)
22. Nagbøl, P.R., Müller, O., Krancher, O.: Designing a risk assessment tool for artificial intelligence systems. In: International Conference on Design Science Research in Information Systems and Technology. pp. 328–339. Springer (2021)
23. Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A.: A comprehensive overview of large language models. ACM Transactions on Intelligent Systems and Technology **16**(5), 1–72 (2025)
24. Nicoletti, S.M., Hahn, E.M., Fumagalli, M., Guizzardi, G., Stoelinga, M.: WATCHDOG: an ontology-aWare risk AssessmenT approaCH via object-oriented DisruptiOn Graphs. In: International Conference on Advanced Information Systems Engineering. pp. 314–331. Springer (2025)
25. Nicoletti, S.M., Hahn, E.M., et al.: BFL: a logic to reason about fault trees. In: 2022 52nd Annual IEEE/IFIP DSN. pp. 441–452. IEEE (2022)
26. Nicoletti, S.M., Lopuhaä-Zwakenberg, M., Hahn, E.M., Stoelinga, M.: Atm: a logic for quantitative security properties on attack trees. Software and Systems Modeling pp. 1–21 (2025)
27. Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., Floridi, L.: Taking AI risks seriously: a new assessment model for the AI act. AI & Society pp. 1–5 (2023)
28. Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., Floridi, L.: Ai risk assessment: a scenario-based, proportional methodology for the ai act. Digital Society **3**(1), 13 (2024)
29. Pearl, J., Russell, S.: Bayesian networks (2000)
30. Piorkowski, D., Hind, M., Richards, J.: Quantitative ai risk assessments: Opportunities and challenges. Seton Hall J. Legis. & Pub. Pol'y **49**, 644 (2025)
31. Ruijters, E., Stoelinga, M.: Fault tree analysis: A survey of the state-of-the-art in modeling, analysis and tools. Computer science review **15**, 29–62 (2015)
32. Sales, T.P., et al.: The common ontology of value and risk. In: Conceptual Modeling. ER 2018. pp. 121–135. Springer (2018)
33. Stamatelatos, M., Vesely, W., Dugan, J., Fragola, J., Minarick, J., Railsback, J.: Fault tree handbook with aerospace applications. Prepared for NASA Office of Safety and Mission Assurance (2002)
34. Stearns, S.C.: Daniel bernoulli (1738): Evolution and economics under risk. Journal of biosciences **25**(3), 221–228 (2000)

35. Von Neumann, J., Morgenstern, O.: Theory of games and economic behavior. In: Theory of games and economic behavior. Princeton university press (2007)
36. Watch, A.: Defining artificial intelligence. Towards an operational definition and taxonomy of artificial intelligence. Unter Mitarbeit von Europäische Kommission (2020)
37. Wisakanto, A.K., Rogero, J., Casheekar, A.M., Mallah, R.: Adapting probabilistic risk assessment for ai. arXiv preprint arXiv:2504.18536 (2025)