Explaining Bayesian Networks Reasoning to the General Public: Insights from the User Study

Nikolay Babakov[1,*], Ehud Reiter[2] and Alberto Bugarín[1]

[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Galicia, Spain
[2]University of Aberdeen, Aberdeen, UK

Abstract

Bayesian Networks (BNs) are widely used for modeling uncertainty and supporting decision-making in complex domains, but their reasoning processes are often challenging for non-experts to interpret. Providing clear, user-centered explanations for BN predictions is essential for building trust and enabling informed use of these models. We report the results of, to our knowledge, the largest user study to date evaluating the interpretability of BN reasoning among the general public. A total of 124 participants with varied backgrounds were introduced to basic BN concepts and asked to assess both non-explained and explained model predictions. Explanations were generated using a method that verbalizes the most meaningful separate paths of probability update. The majority of participants were able to understand fundamental BN ideas and provided insightful feedback on issues of model transparency and trust. Likert-scale results reveal that, while predictions without explanation were often viewed as justified, the addition of structured explanations significantly improved user understanding and trust. This study demonstrates that non-expert users can meaningfully engage with and evaluate BN explanations, providing valuable direction for the development of more accessible and user-centered explainable AI.

Keywords

Bayesian Networks, explanation, user study, explainable AI

1. Introduction

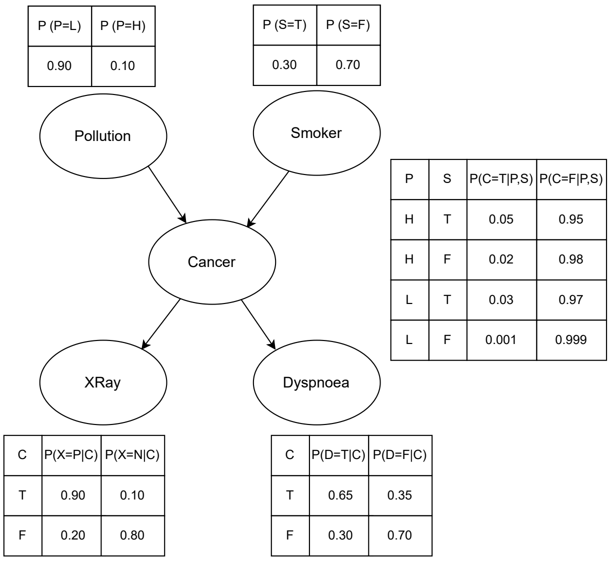
In many domains, users must make decisions with significant personal or societal consequences, often relying on Artificial Intelligence (AI)-generated predictions [1]. However, the value of these predictions fundamentally depends on how well the underlying reasoning can be justified and communicated to end users—especially when high personal responsibility is involved [2]. One essential feature for building trustworthy, actionable explanations is causality: understanding not just correlations but the underlying mechanisms that drive outcomes. Bayesian Networks (BNs), as probabilistic graphical models, provide a natural framework for encoding and communicating causal relationships between variables [3]. By making explicit the links among causes and effects, BNs offer a foundation for interpretable decision support. Yet, despite their theoretical suitability, generating explanations from BNs that are accessible and meaningful to non-expert users remains a significant challenge, as their reasoning can proceed in multiple directions and often involves complex, indirect relationships between variables.

Numerous explanatory methods have been developed to address these challenges [4, 5, 6], but the question of how intuitive these explanations are for the general public is still open. To date, most studies in this area propose original methods without systematically demonstrating them to potential end users [7, 8, 9, 10]. The lack of systematic evaluation is not unique to BNs; it reflects a broader challenge across the entire field of XAI [11]. To the best of our knowledge, there are only two published studies in which BN explanations were evaluated with human participants. In [4], the proposed method was introduced to 16 medical domain experts, while [6] compared their explanation approach with [4], engaging 25 participants with backgrounds in computer science and engineering research or development.

Multimodal, Affective and Interactive eXplainable AI

[*]Corresponding author.

nikolay.babakov@usc.es (N. Babakov)



P (P=L)	P (P=H)
0.90	0.10

P (S=T)	P (S=F)
0.30	0.70

P	S	P(C=T|P,S)	P(C=F|P,S)
H	T	0.05	0.95
H	F	0.02	0.98
L	T	0.03	0.97
L	F	0.001	0.999

C	P(X=P|C)	P(X=N|C)
T	0.90	0.10
F	0.20	0.80

C	P(D=T|C)	P(D=F|C)
T	0.65	0.35
F	0.30	0.70

Figure 1: A BN for the lung cancer problem described in [13].

In this paper, we present, to the best of our knowledge, the largest user study to date on the interpretability of BN explanations, engaging 124 participants from the general public without any filtering by prior knowledge of BNs or specific domain expertise. Our study design begins with a concise, accessible introduction to the essential concepts of BNs, ensuring that all participants acquire the minimal background required to follow the subsequent tasks. Participants first complete a five-question quiz to verify their understanding of these basics, and are then presented with a prediction scenario using a BN, shown both with and without an accompanying explanation. We collect both open-ended textual feedback and structured responses using Likert scales, enabling us to systematically capture the concerns, preferences, and intuitions of everyday users regarding BN explanations in practical, real-world contexts. The details of the study are available in our GitLab repository[1].

2. Preliminary concepts

A BN is a directed acyclic graph (DAG) model that captures the dependencies between variables, represented as nodes in the graph [12]. This structure provides a compact and intuitive way to model complex joint probability distributions. At its core, a BN is composed of two main components: a qualitative structure and quantitative parameters [3].

The qualitative aspect encompasses the collection of variables (nodes), which represent the factors examined within a BN and may be either discrete with multiple possible states or continuous. It also includes the directed arcs, which encode probabilistic (in)dependence relationships among these variables. While these arcs can often be interpreted as causal connections—especially in explicitly causal models [12]—in many data-driven BNs produced by structure learning, the arc directions are understood as indicators of conditional probabilistic dependence rather than definitive causal links.

The quantitative component pertains to the parameters of a BN, namely its Probability Distribution. This is typically expressed through Conditional Probability Tables (CPTs), which specify the conditional probabilities for every possible combination of discrete states between parent and child nodes. For continuous variables, the Probability Distribution can be represented differently, such as by using parametric forms like the mean and variance in a Bayesian conjugate distribution. Throughout this work, we use the term Probability Distribution to refer to any representation of conditional probabilities. The joint probability in a BN $P(X_1, X_2, \dots, X_n)$ can be factorized as $P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{Parents}(X_i))$, where $\text{Parents}(X_i)$ are the parent nodes of X_i in the BN.

[1] https://gitlab.nl4xai.eu/nikolay.babakov/bayesian-networks-reasoning-explanation-for-general-public

Efficient inference in a BN, such as computing marginal probabilities or propagating evidence, depends on message-passing algorithms [12]. For example, in the variable elimination approach [12], factors derived from the CPTs are successively summed or multiplied according to the query variables and the available evidence. For more intricate queries or scenarios requiring real-time inference, the Junction Tree Algorithm is commonly used; this involves transforming the BN into a clique tree and conducting belief propagation through message passing between cliques [14].

Figure 1 provides a well-known example of a simple BN [13], which describes a toy scenario involving possible causes (Pollution: Low or High, and Smoker: True or False) and consequences (XRay: Positive or Negative, and Dyspnoea: True or False) of Lung Cancer.

3. Related works

3.1. Bayesian Network Reasoning explanation methods

There are many explanation methods for BNs [15, 16] that can be delivered in different modalities.

In [17] authors proposed two main approaches for explaining BNs: tracing evidence propagation through the network and constructing narrative scenario-based explanations. [18, 19] expanded on these ideas by generating argument graphs, which represent BN reasoning as subgraphs, to improve interpretability.

INSITE [20] generates explanations for BN inference by highlighting the most influential evidence and their direct impact on the hypothesis, focusing on clarity and relevance. BANTER [21] expands this idea into a medical tutoring context, but its explanations remain closely tied to complex inference chains, limiting user accessibility. B2 [22] further improves on this by incorporating natural language, discourse filtering, and a graphical-text interface to create more intuitive and context-sensitive explanations.

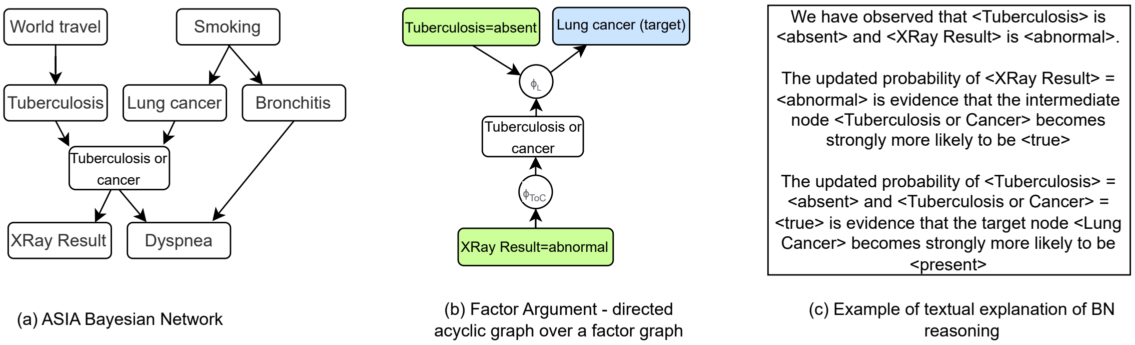
Elvira [23] is an interactive software environment for constructing and explaining BNs and Influence Diagrams, emphasizing intuitive graphical and interactive explanations. It supports both structural and reasoning explanations, featuring visual evidence propagation, scenario comparison, and extensions for decision-support, such as Influence Diagram explanations and what-if analyses [24]. Compared to earlier systems like INSITE and BANTER, Elvira offers enhanced visualizations, interactive debugging, and sensitivity analysis, making complex probabilistic reasoning more accessible.

In [25], the BN is restructured so the target node's Markov blanket forms its parent set, with CPTs condensed into decision trees to generate dynamic, context-specific explanations. Related methods [26, 7] extract a support graph—a directed subgraph representing inference chains—to elucidate the reasoning path to the target node.

In [7], a two-phase argument extraction approach converts the support graph into structured arguments, connecting probabilistic inference with legal reasoning. Building on this, [26] applies Natural Language Generation and qualitative probability annotations to automatically produce clear, context-aware textual explanations, enhancing comprehensibility for non-experts in legal and forensic domains.

The works in [27, 28] enhance the interpretability and trust of BN-based AI, particularly for clinical applications. [27] propose a multi-level explanation framework that progressively details key evidence, information flow, and impact. This is refined in [4] with automated evidence selection and faster, adaptable explanations. Additionally, [27] introduce a comprehensive evaluation framework using quantitative and qualitative criteria, such as fidelity, actionability, and user trust. Together, these studies advance transparent and user-centered BN explanations for complex decision-making.

The recent approach in [6] generates natural language explanations for BN reasoning using factor arguments—structured graphs that trace how evidence from observed nodes influences a target variable. By introducing factor argument independence, the method decides whether to combine or separate explanatory chains and ranks them by strength. User studies show these explanations are more helpful than previous methods.



We have observed that <Tuberculosis> is <absent> and <XRay Result> is <abnormal>.

The updated probability of <XRay Result> = <abnormal> is evidence that the intermediate node <Tuberculosis or Cancer> becomes strongly more likely to be <true>

The updated probability of <Tuberculosis> = <absent> and <Tuberculosis or Cancer> = <true> is evidence that the target node <Lung Cancer> becomes strongly more likely to be <present>

(a) ASIA Bayesian Network

(b) Factor Argument - directed acyclic graph over a factor graph

(c) Example of textual explanation of BN reasoning

Figure 2: Vizualization of the explanation method used in the study [6]. a) widely known ASIA BN [41]

b) Example of Factor Argument concept - the initial BN is represented as a factor graph, and the subset of this graph from evidence nodes to target node in the form of directed subgraph, i.e. Factor Argument, shows the path of significant probability updates. c) The example of verbalization of Factor Argument.

3.2. Explainable AI evaluation

Although there are well-established metrics for evaluating the predictive accuracy of models, there is still no unified framework for assessing the quality of explanations in XAI. This lack of consensus highlights the inherent difficulty in objectively measuring how effective AI-generated explanations are [29]. [30] identified several key dimensions for evaluating XAI, including the quality of explanations, user satisfaction, trust, the user's mental model, the influence of curiosity in seeking explanations, and the effectiveness of user-XAI collaboration. [11] identified twelve conceptual properties relevant to XAI evaluation, conducted a structured survey, and analyzed which of these properties are most frequently assessed in studies involving XAI evaluation. [31] suggested adapting user-centric evaluation frameworks from recommender systems to promote human-centered standardization in XAI evaluation. [32] introduced a taxonomy to guide researchers and practitioners in the design and implementation of XAI evaluation studies. Buçinca et al. [33] demonstrated that XAI evaluation should be grounded in authentic decision tasks rather than relying on artificial proxy tasks. Rosenfeld and Richardson [34] argued that, in addition to expert feedback, various objective, user-agnostic methods are available for evaluating XAI techniques. [35] proposed a human-centered demand framework that categorizes XAI users into five primary roles, each with distinct needs, based on a comprehensive review of the literature. They also identified six widely used human-centered XAI evaluation measures that are instrumental in assessing the impact of XAI. [36] offered a set of recommendations for designing user studies in XAI and conducted an extensive user evaluation examining the effects of rule-based and example-based contrastive explanations. There are also numerous surveys aimed at collecting existing practices in XAI evaluation from different points of view [37, 38, 39, 40].

4. Methodology

4.1. Demonstrated explanation method

In our study, we chose to demonstrate a single explanation method, Factor Argument Explanation (FAE), as introduced in [6]. The primary motivation for this choice is that, to the best of our knowledge, it is the most recently proposed explanation approach for BN reasoning. This method has already been evaluated against alternative explanation strategies in a prior user study [6], which showed positive results regarding its understandability and usefulness. As a result, we consider it reasonable to demonstrate only this method in our evaluation, in order to minimize the cognitive load on participants and focus on the most promising explanatory approach.

The FAE method provides natural language explanations for BN reasoning by explicitly tracing the paths through which evidence affects a target variable. The core idea is to represent these explanatory chains as “factor arguments” (FAs): directed acyclic subgraphs over the BN’s factor graph, which show how information flows from observed evidence nodes to the node of interest. Each FA captures a specific chain of reasoning, detailing the intermediate variables and their updates along the way. See Figure 2 for the example explanation generation with FAE.

To quantify and prioritise explanations, the method defines the strength of each FA based on its impact on the target variable’s probability. The algorithm automatically identifies all maximal, proper, and independent FAs connecting evidence to the target, ensuring that overlapping or redundant chains are combined only when their effects are not independent. Once the key FAs are selected, the method generates natural language explanations by narrating each step: describing the observations, the inferred updates at each intermediate node, and the cumulative effect on the target. The explanation is delivered both in visual and textual form.

We used the implementation of the method provided in the GitLab repository in the original paper [6].

4.2. Study structure

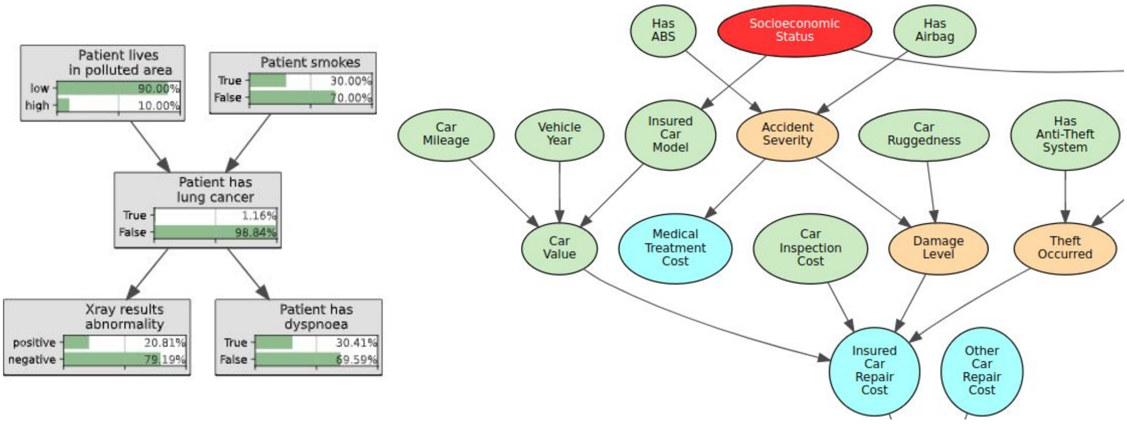
The user study was developed using the Qualtrics [2] platform and consists of several sections: a general introduction to the study, informed consent, a questionnaire on participant background, an introduction to the fundamentals of BNs, a pre-survey quiz, and the main section. In the core part of the study, participants are presented with a scenario in which certain variables are known and the task is to determine how these affect a selected target variable. Participants first see the BN prediction without any explanation, and then with an explanation, after which their feedback is collected. The screenshots and the final raw results of the study are available in our repository[3].

General introduction The general introduction welcomes participants, explains that the study aims to evaluate a new method for making BN predictions more understandable, and briefly describes BNs as models of how causes affect probabilities. Participants are informed they will help assess a verbal explanation approach, are given an overview of the study process, and are assured their responses will remain confidential and used only for research.

Background questionnaire In the background collection section, participants are asked to provide information about their age range, level of education, and professional field. The survey also gathers data on participants’ computer skills and any previous knowledge or experience they may have with BNs. This information helps contextualize the results and understand the diversity of the study sample.

Basic introduction to BNs The BN introduction section begins by explaining the basic intuition of the BNs, why they are useful, and why explanation is necessary for effectively using them in real practice. Next, participants are introduced to the sample BN, specifically the ASIA network (as illustrated in Figure 2). Using this example, key concepts such as nodes, edges, and CPTs are explained in an accessible manner that does not require prior expertise in probability theory. The introduction also covers two fundamental types of reasoning: predictive reasoning (from parent to child nodes, which tends to be more intuitive) and diagnostic reasoning (where evidence about a child node updates the probability of its parent). We exclude intercausal reasoning, because it may be quite difficult ti understand for the general public, but, arguably, less important than explicit explanation of the diagnostic one. All explanations are accompanied by graphical tips for better understanding. At the end of this section, a video demonstration is provided using the Bayes Server interface [4], showing how probabilities update as more evidence is added to the BN.

[2] www.qualtrics.com
[3] Will be available after acceptance
[4] www.bayesserver.com



a) “Cancer” BN used for pre-study quiz

b) Subset of “insurance” BN used for main study

Figure 3: Examples of BNs used in the study. a) BN used as a reference for a pre-study quiz. b) A subset of BN used in the main study for BN reasoning demonstration. Nodes are coloured for initial demonstration according to the semantic group of factors.

Pre-survey quiz The pre-survey quiz is designed as a short set of five multiple-choice questions, with only one option selectable for each. These questions serve several purposes: they verify that participants are attentive and genuinely engaged, and confirm that they have understood the essential basics of BNs and their reasoning. The quiz is intentionally straightforward, not intended to exclude those without advanced knowledge, but rather to ensure a minimal, working understanding necessary for meaningful participation in the main part of the study. The questions presented to participants are as follows:

1. What is this study aimed for? *(options: You will design Bayesian Networks from scratch; You will evaluate the usefulness of explanations of Bayesian Network reasoning; You will try to predict a diagnosis of an imaginary patient; You will pass the probability theory test)*
2. What does a Bayesian Network help to model? *(options: The relationships between different variables and their probabilities; Random guesses about uncertain events; It works like a dialogue agent; It predicts future events with 100% certainty)*
3. Which of the following is an example of a factor (node) in a Bayesian Network? *(options: An arrow between ”Smoking” and ”Lung cancer”, ”Patient lives in a polluted area”, A doctor’s diagnosis)*
4. According to the Bayesian Network shown above, which of the following statements is true? *(options: ”Patient smokes” directly causes ”Patient has lung cancer”; ”Patient has lung cancer” directly causes ”Patient smokes”; ”Patient has dyspnoea” directly causes ”Patient has lung cancer”)*
5. According to the Bayesian Network above, if we learn that the patien’s XRay results are abnormal could it possibly affect other factors (nodes)? *(options: This was not explained in the introduction; No, because there are no outgoing arrows from ”Xray results abnormality”; Yes, because probabilities could be updated on both sides after we learn some facts, even against the direction of the arrows in the network)*

The first two questions (1 and 2) are general in nature and primarily serve as attention checks, being straightforward for anyone who has read the material carefully and requiring no specialized knowledge.

The next two questions (3 and 4) reference a simple illustrative BN with five nodes (the “cancer” BN[5] shown in Figure 3a) and are slightly more technical. Question 4 engages the participant to check the edges of the BN and to infer which causal statement is correct. Although these questions are a bit more challenging than the initial ones, they are still easily answerable for participants who have engaged with the introductory explanations and diagrams.

[5] www.bnlearn.com/bnrepository/discrete-small.html#cancer

The final question (5) is likely the most challenging, as it addresses a concept that is directly emphasized in the introduction section. Although this may initially appear to be an advanced topic, it is essential for a correct understanding of BN reasoning: probabilities can be updated in both directions, not just along the direction of the arrows. Grasping this bidirectional flow—encompassing both predictive and diagnostic reasoning—is crucial for meaningful participation in the study, even for those without technical expertise.

Core Study Task We selected the well-known "insurance" BN [42][6] as the basis for our main study task and made slight modifications to simplify it for participant comprehension. The subset of the BN used for reasoning demonstrations is illustrated in Figure 3b. This network models car accident scenarios and the associated potential costs for an insurance company.

To keep the scenarios both engaging and non-trivial, we chose two cases that incorporate predictive, diagnostic, and intercausal reasoning paths. Additionally, we designed the cases so that the available evidence could be partially contradictory, making the need for a clear explanation especially important. One participant was demonstrated with only one random scenario to ensure maximal engagement and prevent excessive cognitive load, which may worsen the quality of the replies.

The first case (demonstrated on Figures 4 and 5) presents the following situation: the car did not have airbags, it was equipped with an ABS anti-lock system, and the driver's medical costs were around ten thousand dollars. Participants are asked how these factors influence the car's damage level. In this case, although the presence of ABS would generally suggest a lower risk of severe damage, the high medical costs increase the probability of a severe outcome, creating a non-obvious inference.

The second case presents the following scenario: the medical treatment cost is considerable (tens of thousands of dollars), the repair cost for the insured car is small (only a few thousand dollars), and the car is described as highly rugged or "tank"-like. Participants are asked to determine how these factors influence the severity of the accident. In this situation, while the high medical costs suggest a severe accident, the combination of a durable car and low repair costs provides contradictory evidence that argues against high accident severity.

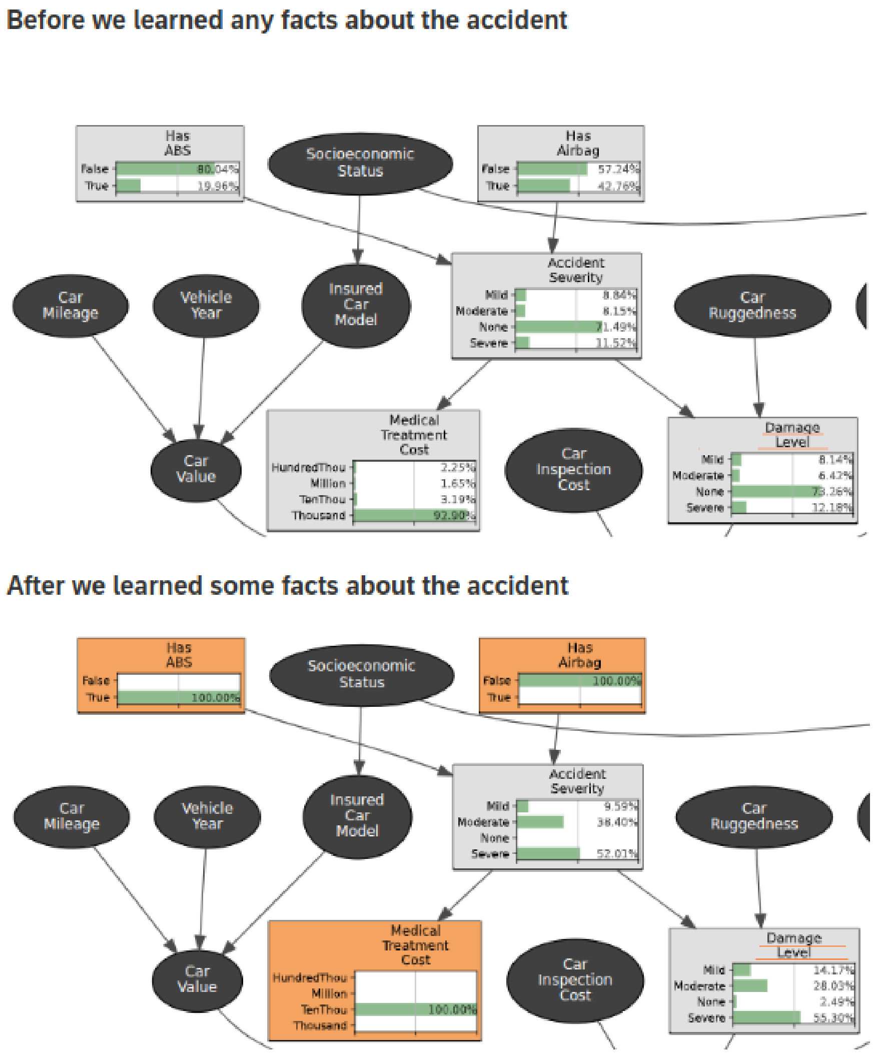
Prediction without explanation In the first part of the core study task, participants were introduced to the case scenario, including the relevant evidence and the target variable. They were then shown the subset of nodes involved, both before and after the evidence was applied, and received a verbalized prediction generated by the BN—however, no explanation for the prediction was provided at this stage.

Figure 4 shows an example of the prediction without explanation. The demonstrated cases was shown together with the following verbalization of prediction: `Initially, before any facts were known, the most likely state of the Damage Level was "None", meaning that in most cases, no damage would occur. This had a probability of 73.26%. However, after incorporating the known facts about the accident, the probability shifted significantly: The likelihood of Severe Damage increased from 12.8% to 55.3%. This means that, based on the new known fact, the model now believes it is more probable that the car suffered severe damage.`

Following this, participants were asked to respond to three Likert-scale statements (with five options ranging from strongly disagree to strongly agree) regarding their understanding and perception of the prediction. They were also invited to provide optional open-ended feedback about their experience. The Likert-scale statements were as follows:

- The prediction is clear.
- The prediction is well justified based on the given information.
- An explanation is necessary to better understand the prediction.

[6] www.bnlearn.com/bnrepository/discrete-small.html#insurance


Before we learned any facts about the accident

After we learned some facts about the accident

Figure 4: Example of interface without explanation

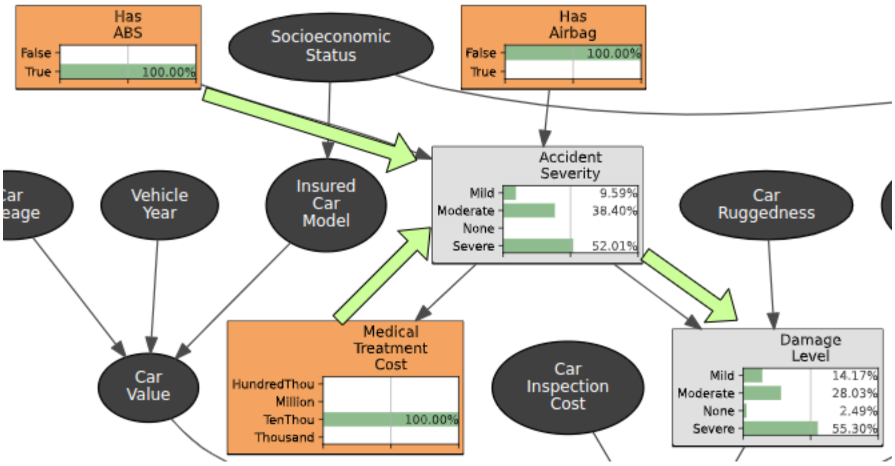


Figure 5: Example of interface with explanation (textual part of explanation is shown in Sec 4.2)

Prediction with explanation In the second and final part of the core study, participants were presented with an explanation of the prediction based on the FAE method, in addition to the initial non-explained prediction. For each case, two explanatory paths were provided, both visualized within the network diagram and described in natural language below the schematic. After reviewing the explanation, participants were asked to rate statements about the explanation using a Likert scale and to provide required written feedback reflecting their perception and understanding of the explanation. The statements were as follows:

- The explanation of the Bayesian Network reasoning is clear.
- The explanation helps me understand how probability updates influence the prediction.
- The explanation increases my trust in the model's prediction.

The formulation of the final obligatory free-form question was - `As the final step of the study, please provide detailed feedback on the quality of the explanation. What aspects were clear or unclear? Did it help you understand the reasoning behind the prediction? Any suggestions for improvement are highly valuable. Your response is essential for properly finalizing the study.`

Figure 5 shows an example of the prediction with explanation with one of the two explanatory paths generated with the FAE method. This path was shown together with the following text: `We observe that the car is equipped with an ABS anti-lock system (Has ABS is True) and that the Medical Treatment Cost is around $10,000 (Medical Treatment Cost is TenThou). Having ABS increases the likelihood that the accident was mild (Accident Severity is Mild). A high medical treatment cost suggests that the accident was more likely severe (Accident Severity is Severe). Overall, these updates slightly shift the probability of Accident Severity toward a severe accident. As a result, the increased probability of Accident Severity=Severe weakly raises the likelihood that the car suffered severe damage (Damage Level is Severe).`

4.3. Participant Recruitment and Compensation

Participants were recruited through the Prolific platform[7], which provides access to a broad and diverse population for research studies. The only preliminary screening applied was based on the highest acquired educational level, with eligibility restricted to individuals who had completed at least a technical or community college; this filtration was implemented using Prolific's internal filters. The median participation time for the study was 35 minutes, and participants were compensated at an average hourly rate of £12 per hour (which was fair payment according to Prolific in-platform tips). Submissions were reviewed manually prior to approval, and payments were issued promptly after verification of completion. Throughout the study, participants' anonymity and informed consent were ensured in accordance with ethical research guidelines.

5. Results

5.1. Demographic Profile of Participants

We engaged 124 participants in our study. It is important to understand the general background of the participants in order to clarify what constitutes the "general public" in the obtained results. As shown in Figure 6, most participants were between 25 and 44 years old, with the largest group in the 25–34 range. The majority reported medium to high computer skills, while prior experience with BNs was mostly low, highlighting that our sample largely consists of individuals without specialized knowledge of BNs.

To gain insight into the professional backgrounds of our participants, we collected their fields of work using a free-form text response. As a result, we report only a basic analysis due to the variety and inconsistency in the responses. The most common fields represented include finance, IT, software engineering, and other computer-related areas, followed by participants from engineering, healthcare,

[7] www.prolific.com/

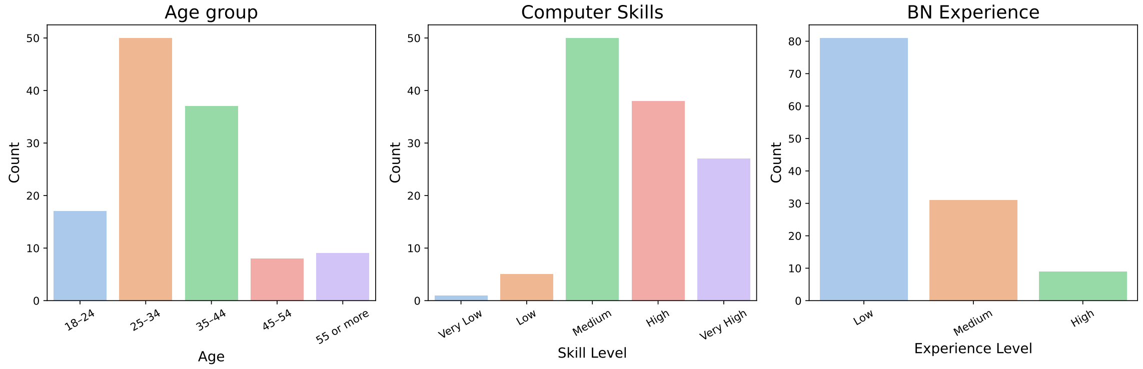


Figure 6: Demographic information of the study participants

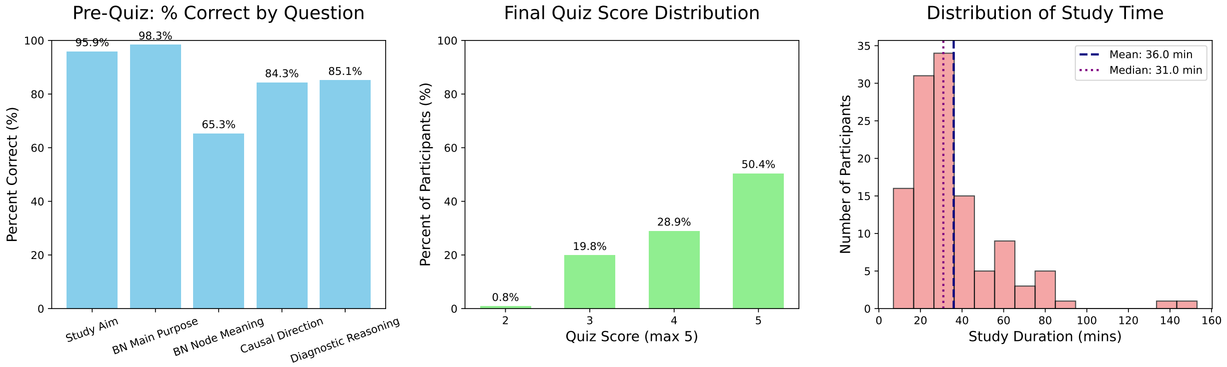


Figure 7: Detalization of pre-study quiz responses.

education, business, and economics. There were also smaller numbers from fields such as communication, public administration, biology, humanities, law, and environmental sciences. Overall, the participant pool reflects a diverse mix of technical, scientific, business, and public service backgrounds, consistent with a broad general public sample.

5.2. Pre-study quiz results

Figure 7 presents the results of the pre-study quiz. Attention-check questions, such as those about the study aim and the main purpose of BNs, were answered correctly by the vast majority of participants, with accuracy rates above 95%. The more technical questions yielded slightly lower scores but were still answered correctly by most respondents. Specifically, 65% of participants correctly identified the meaning of a BN node, while over 80% selected the correct causal statement in the BN structure diagram and demonstrated the correct intuition in the diagnostic reasoning question.

Most mistakes in the more technical pre-quiz questions were due to misunderstandings about BN semantics and reasoning. For the BN node meaning question, while 79 participants correctly identified ”Patient lives in a polluted area” as the meaning of the node, 22 instead selected ”The probability that a patient has lung cancer,” 16 chose ”An arrow between ’Smoking’ and ’Lung cancer’,” and 4 picked ”A doctor’s diagnosis.” For the causal direction question, 102 participants correctly identified ”Patient smokes directly causes Patient has lung cancer,” but 10 selected ”Patient has dyspnoea directly causes Patient has lung cancer,” 7 mistakenly reversed the relationship (”Patient has lung cancer directly causes Patient smokes”). For the diagnostic reasoning question, 103 answered correctly that probabilities can be updated in both directions in the network, while 15 believed that inference is blocked due to the absence of outcoming arrows, and 3 indicated that the concept was not explained in the introduction.

Overall, half of the participants achieved the top quiz score, and an additional 29% made only a single mistake. This indicates that most participants started the main study with a good understanding of the

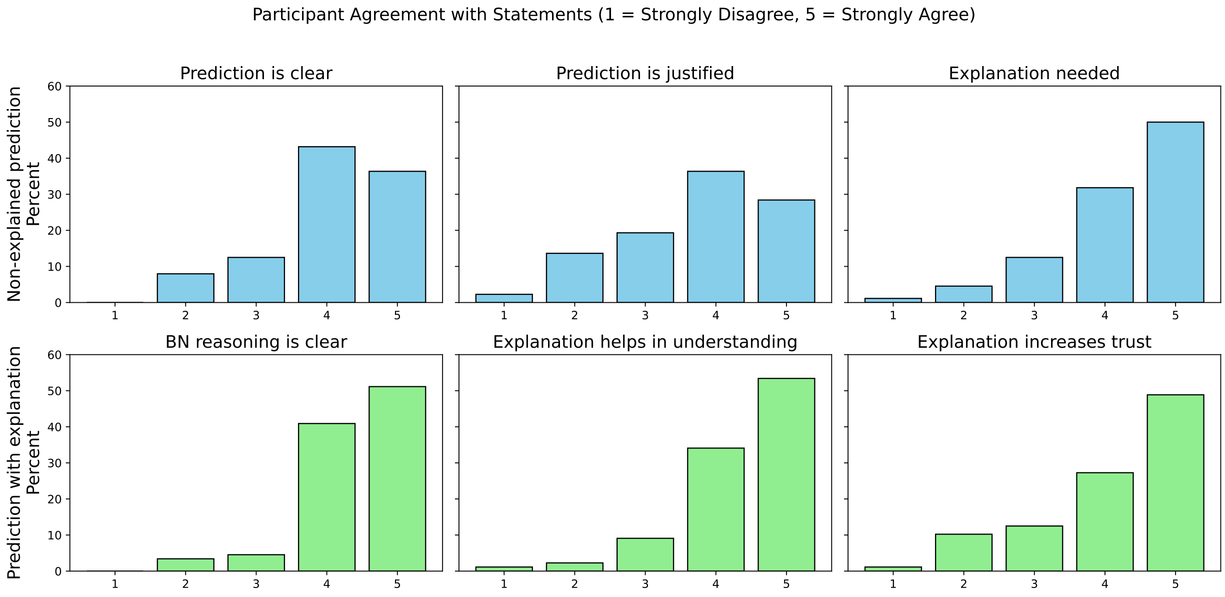


Figure 8: Answers to Likert-scale questions related to the BN predictions without explanations (first row) and with explanations (second row).

basic BN concepts and the quiz content.

Figure 7 also shows the statistics of time spent in the study. The median completion time was 31 minutes, while the mean was slightly higher at 36 minutes, reflecting a few participants who took substantially longer. Most participants completed the study in under an hour, though a small number spent significantly more time, as indicated by the long right tail in the distribution.

5.3. Filtering the Results

Prior to reporting the final statistics, we applied several filtering steps to ensure data quality and reliability. First, three participants were manually excluded via Prolific: two because they submitted low-quality, nearly identical and meaningless responses within an unrealistically short time (less than five minutes), and one whose answers were clearly generated by artificial intelligence.

Next, we applied automatic filters based on study duration and quiz performance. Specifically, we removed eight participants who completed the study in less than 15 minutes, as well as 25 participants who scored less than 4 out of 5 on the pre-study quiz. After these filtering steps, the final dataset included 88 high-quality participant entries used for subsequent analysis.

5.4. Likert scale responses

Figure 8 summarizes participants' responses to the Likert scale questions. The plots in the first row of the Figure indicate that most participants found the predictions to be clear (with the majority selecting 4 or 5 on the scale) and generally considered them justified, though a noticeable portion gave more moderate ratings. Notably, when asked whether an explanation was needed, the responses strongly shifted toward agreement, with the vast majority selecting the highest options (4 or 5). This indicates that, despite a general perception of clarity and justification, participants still felt a significant need for further explanation to fully understand or trust the prediction.

The second row of Figure 8 displays participant responses after they were provided with explanations for the BN predictions. The results show a pronounced shift toward highly positive ratings across all three questions. The majority of participants rated the BN reasoning as clear, and even more strongly agreed that the explanation helped them understand the prediction. Additionally, most participants felt that the explanation increased their trust in the result, with responses heavily concentrated at the highest end of the scale. This suggests that providing structured explanations not only clarifies the reasoning process but also enhances user confidence and trust in the BN's output.

5.5. Free-form comments about predictions without explanation

In this subsection, we analyze the free-form voluntary comments provided by participants in response to the predictions without explanation. Participants were prompted with the question: "Please share any comments or thoughts you have about the prediction, including any uncertainties or aspects that stood out to you."

Need for Causal Attribution and Factor Weighting Participants repeatedly expressed a strong interest in understanding how much each specific piece of evidence contributed to the model's prediction. Several comments pointed out the dramatic increase in probability and questioned how the model determined the impact of factors such as serious injuries, lack of airbags, and high medical costs. While many found the prediction itself plausible, they wanted clarity about the range of the probability update and sought detailed insight into how each individual fact was weighted in the model.

Need for Explanation and Transparency A key recurring theme was the need for transparent explanations. Many participants indicated that, while the overall prediction felt reasonable, it was difficult to fully trust or understand the outcome without knowing how much each input (for example, no airbags, presence of ABS, or high medical costs) influenced the result. These comments underscore that, even when the final probability shift makes intuitive sense, a step-by-step breakdown is essential for users to feel confident in the model's reasoning and conclusions.

Contradictory Evidence and the Need for Explanation Numerous comments highlighted confusion when the model was presented with conflicting facts—for example, high medical costs (suggesting severity) versus low car repair costs and ruggedness (suggesting less severity). Participants found it difficult to reconcile these opposing signals and often stated that a detailed explanation was essential to understand how the model weighed such contradictory evidence.

Requests for Interactivity or Additional Information A number of participants suggested that interactive features—such as the ability to modify evidence or visualize how changing inputs affects predictions—would enhance understanding. Others asked for access to more detailed background information or alternative scenarios to better grasp the model's workings and the reasoning behind its outputs.

5.6. Free-form comments about predictions with explanation

In this subsection, we examine the detailed feedback provided by participants regarding the quality of the explained prediction. Participants were required to reflect on which aspects of the explanation were clear or unclear, whether it aided their understanding of the model's reasoning, and to suggest any improvements. For clarity, our analysis is divided into comments that are likely general to any explanation method and those that are specific to the presented Factor Argument Explanation approach.

5.6.1. Method-agnostic comments

Questions on Baseline Probabilities and Data Sources A common area of natural doubt for BN users relates to the origin and interpretation of initial probabilities and the content embedded in the network. Some participants questioned the "default" probability states—such as why the BN initially predicts "no damage" as most likely, even in situations where minor damage seems almost inevitable. Others wanted greater transparency regarding the source and rationale behind the numerical values and cost figures used in the model, as well as clarification on how these input values specifically affect predictions. These comments highlight the importance of clearly communicating not only the process of probabilistic updating, but also the underlying assumptions, baseline probabilities, and data sources used to construct the BN, especially when users are looking deeply into how outcomes are derived.

Importance of Textual Explanation Participants emphasized the crucial role of detailed textual explanations in making BN predictions understandable. While visual aids and color-coding were seen as helpful, some noted that the main driver of comprehension was the written explanation, which enabled logical thinking and made the reasoning process clearer. Several participants suggested further enhancements, such as adding a concise summary at the end of each explanation to reinforce the main takeaway and explicitly tie together conflicting pieces of evidence. Additionally, some recommended including a brief overview of the initial probabilities and their significance, helping users form a more complete and coherent understanding of how the BN arrives at its conclusions.

Feedback on the Visual Component of Explanations Participants generally found the visual aspects of the explanations helpful, but suggested several areas for improvement to enhance clarity and accessibility. Many valued visual aids like tables and flowcharts, noting that these formats supported their understanding, though some found flowcharts overwhelming without additional context. Suggestions included displaying more precise probability values for each parameter, rather than just categorical outcomes, and incorporating impact visualizations and counterfactual comparisons to make the reasoning process more transparent. Several participants highlighted the benefits of interactive features, such as the ability to explore the network or view video summaries, to make the system more engaging and intuitive. Others recommended clearer highlighting of the target node, ordering options in a logical sequence, and using less technical, more conversational language to broaden accessibility. Overall, while visual aids were considered essential, participants emphasized the need for improved readability, interactivity, and contextual cues to fully support non-expert users.

5.6.2. Method-specific comments

Feedback about paths of probability updates (Factor Argument) Many participants praised the structure of the explanation, particularly the way it broke the reasoning process into clear, step-by-step causal paths. This division helped users see how different factors—such as the presence or absence of airbags, medical costs, and ABS—each contributed to the final prediction. The clarity and realism of the approach were highlighted, especially when the explanation showed how various pieces of evidence could support conflicting outcomes. Participants found that this method made the BN's reasoning more understandable and gave them a better sense of how the model updates its predictions based on multiple, sometimes opposing, facts.

Despite the positive feedback, some users were uncertain about the logic behind the creation and sequencing of explanatory paths. Several questioned why specific factors, such as Medical Cost, were not included in certain paths, or how the model determined which nodes to group together in the explanation, given that some nodes (like ABS and Airbag) both influence the same outcome. There were requests for more explicit justification or explanation of the criteria for splitting or combining explanatory paths, as well as suggestions for visualizing the numerical impact of each step.

Inconsistency between probability updates on the graph and in explanations Several participants noted confusion arising from the difference between the visualized probability updates in the graphical support and the verbal descriptions provided in the explanation. While visual aids, such as highlighted arrows and updated node values, were appreciated for indicating which factors influenced the outcome, there was uncertainty about the magnitude and timing of probability changes at each explanatory step. Specifically, participants pointed out that the graphics often displayed only the final probabilities after all evidence was considered, rather than showing incremental changes as each piece of evidence was introduced. This sometimes made it difficult to correlate the narrative of how probabilities shift with the corresponding visual representation, and to judge the true impact of individual factors.

Participants suggested that more granular or stepwise visual feedback, as well as clearer quantification of probability updates at each stage, would make the explanation process more transparent and easier to follow.

6. Discussion

Our study demonstrates that, with carefully designed materials and introductory explanations, members of the general public can effectively engage with basic BN concepts and reasoning tasks. The pre-survey quiz results indicate that most participants, despite limited prior experience with BNs, were able to answer the majority of questions correctly. This suggests that well-structured, accessible questions—focused on fundamental concepts such as the role of nodes and edges or the possibility of diagnostic reasoning—can enable non-expert users to meaningfully participate in BN-related studies. This finding is encouraging for the broader goal of making probabilistic modeling and explainable AI tools accessible beyond specialist audiences.

Beyond successfully completing the quiz, most participants offered meaningful feedback that sheds light on key sources of potential mistrust when BN reasoning is used in more specialized domains. Some of them expressed doubts about the origins of initial probabilities and the construction of the BN, raising questions about baseline assumptions and data sources, which is completely natural in real-world applications of BNs. Their responses also highlighted the critical importance of combining clear textual explanations with effective graphical representations to enhance understanding. Furthermore, there was a strong call for increased interactivity.

What particularly supports the conclusion that the general public is capable of engaging with these tasks is that many participants were able to reasonably identify and articulate specific shortcomings in both the explanation method used and its presentation. For instance, several questioned how the specific explanatory paths were determined, noting that these were simply presented to them without context on how or why they were selected. This suggests a need to clarify that such paths are sorted by their calculated strength of effect on the final probability update. Additionally, participants pointed out inconsistencies between the verbal descriptions of probability changes (e.g., ”weakly” or ”significantly” increased) and the actual probability shifts shown in the graphical output. This discrepancy arises because the explanation describes updates along isolated reasoning paths, while the visualization displays only the final state of the BN after all evidence is incorporated, which can be misleading. This highlights an important direction for future work: the verbal explanation should be better synchronized with the probability values shown to users, possibly by reflecting incremental changes at each step rather than only the overall result.

Overall, the Likert scale responses indicate that while most participants found the basic, non-explained prediction to be clear and generally justified, the majority also agreed that an explanation would be helpful. When provided with the Factor Argument Explanation, participants overwhelmingly reported that the reasoning behind the BN prediction became clearer, more understandable, and easier to trust. This demonstrates the added value of structured explanations in making probabilistic models accessible and credible to the general public.

7. Conclusion

This study presents a large-scale evaluation of the comprehensibility and perceived usefulness of natural language explanations for Bayesian Network reasoning among the general public. By carefully introducing BN fundamentals and designing accessible scenarios, we show that non-expert participants can meaningfully engage with both the underlying concepts and the interpretive tasks associated with probabilistic models. Participants not only succeeded in answering basic BN questions, but also provided thoughtful feedback that revealed nuanced concerns about model transparency, the origins of probabilities, and the importance of combining textual and graphical explanations.

Our findings highlight that structured, stepwise explanations—such as those provided by the Factor Argument Explanation method—significantly improve users' clarity, understanding, and trust in BN-based predictions compared to non-explained outputs. At the same time, user feedback identified areas for further development, including the need for greater transparency about the selection and impact of explanatory paths, better alignment between verbal and graphical information, and more interactive, user-driven exploration of model reasoning.

Taken together, these results underscore the feasibility and value of involving the general public in the evaluation of explainable AI techniques for BNs. They also point to clear directions for advancing explanation methods and interfaces, with the ultimate goal of making probabilistic AI systems more interpretable, trustworthy, and accessible in real-world applications.

Limitations

This study has several limitations that should be considered when interpreting the results. First, the questions posed to participants were primarily perceptive in nature, focusing on their immediate understanding and trust in BN explanations. It is possible that participant responses could differ if they were required to rely on BN predictions in real-world, high-stakes scenarios where personal or professional responsibility is involved. Second, our study demonstrated only a single explanation method without presenting alternatives for direct comparison, which could introduce bias in participant feedback. However, we considered this approach reasonable, as the selected method had previously been benchmarked against existing alternatives. Future studies should explore more ecologically valid settings and include comparative evaluations of multiple explanation methods.

Acknowledgments

This paper is part of the R+D+i project TED2021-130295B-C33, funded by MCIN/AEI/10.13039/501100011033/ and by the "European Union NextGenerationEU/PRTR". This research also contributes to the projects PID2020-112623GB-I00 and PID2023-149549NB-I00 funded by MCIN/AEI/10.13039/501100011033/ and by ERDF A way of making Europe. The support of the Galician Ministry for Education, Universities and Professional Training and the "ERDF A way of making Europe" is also acknowledged through grants "Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04" and "Reference Competitive Group accreditation 2022-2025 ED431C 2022/19"

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5 to check grammar and spelling.

References

[1] V. Dignum, Responsibility and Artificial Intelligence, The Oxford handbook of ethics of AI 4698 (2020) 215.

[2] D. Doran, S. Schulz, T. R. Besold, What does explainable AI really mean? a new conceptualization of perspectives, arXiv preprint arXiv:1710.00794 (2017).

[3] D. Koller, N. Friedman, Probabilistic graphical models: principles and techniques, MIT press, 2009.

[4] E. Kyrimi, S. Mossadegh, N. Tai, W. Marsh, An incremental explanation of inference in Bayesian Networks for increasing model trustworthiness and supporting clinical decision making, Artificial Intelligence in medicine 103 (2020) 101812.

[5] J. Keppens, Explainable Bayesian Nnetwork query results via natural language generation systems, in: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, 2019, pp. 42–51.

[6] J. Sevilla, N. Babakov, E. Reiter, A. Bugarín, Explaining Bayesian Networks in natural language using factor arguments. evaluation in the medical domain, arXiv preprint arXiv:2410.18060 (2024).

[7] S. T. Timmer, J.-J. C. Meyer, H. Prakken, S. Renooij, B. Verheij, A two-phase method for extracting explanatory arguments from Bayesian Networks, International Journal of Approximate Reasoning 80 (2017) 475–494. URL: https://www.sciencedirect.com/science/article/pii/S0888613X16301402. doi:https://doi.org/10.1016/j.ijar.2016.09.002.

[8] A. de Waal, J. W. Joubert, Explainable Bayesian Networks applied to transport vulnerability, Expert Systems with Applications 209 (2022) 118348. URL: https://www.sciencedirect.com/science/article/pii/S0957417422014671. doi:https://doi.org/10.1016/j.eswa.2022.118348.

[9] T. Koopman, S. Renooij, Persuasive contrastive explanations for Bayesian Networks, in: European Conference on Symbolic and Quantitative Approaches with Uncertainty, Springer, 2021, pp. 229–242.

[10] C. S. Vlek, H. Prakken, S. Renooij, B. Verheij, A method for explaining Bayesian Networks for legal evidence with scenarios, Artificial Intelligence and Law 24 (2016) 285–324.

[11] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI, ACM Computing Surveys 55 (2023) 1–42.

[12] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan kaufmann, 1988.

[13] K. B. Korb, A. E. Nicholson, Bayesian Artificial Intelligence, CRC press, 2010.

[14] S. L. Lauritzen, D. J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, Journal of the Royal Statistical Society: Series B (Methodological) 50 (1988) 157–194.

[15] C. Lacave, F. J. Díez, A review of explanation methods for Bayesian Networks, The Knowledge Engineering Review 17 (2002) 107–127.

[16] C. Hennessy, A. Bugarín, E. Reiter, Explaining Bayesian Networks in Natural Language: State of the Art and Challenges, in: 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence, Association for Computational Linguistics, Dublin, Ireland, 2020, pp. 28–33. URL: https://aclanthology.org/2020.nl4xai-1.7.

[17] M. Henrion, M. J. Druzdzel, Qualitative propagation and scenario-based approaches to explanation of probabilistic reasoning, in: Uncertainty in Artificial Intelligence, volume 6, 1990, pp. 17–32.

[18] I. Zukerman, K. Korb, R. McConachy, Perambulations on the way to an architecture for a nice argument generator, in: Notes of the ECAI-96 Workshop on Gaps and Bridges:"New Directions in Planning and Natural Language Generation, 1996, pp. 31–36. URL: https://dl.acm.org/doi/abs/10.5555/295240.295901.

[19] I. Zukerman, R. McConachy, K. B. Korb, Bayesian reasoning in an abductive mechanism for argument generation and analysis, in: AAAI/IAAI, 1998, pp. 833–838.

[20] H. J. Suermondt, Explanation in Bayesian Belief Networks, Stanford University, 1992.

[21] P. Haddawy, J. Jacobson, C. E. Kahn Jr, BANTER: a Bayesian network tutoring shell, Artificial Intelligence in Medicine 10 (1997) 177–200.

[22] S. McRoy, A. Liu-Perez, J. Helwig, S. Haller, B2: A tutoring shell for Bayesian Networks that supports natural language interaction, in: Working Notes, 1996 AAAI Spring Symposium on Artificial Intelligence and Medicine, Stanford, Calif: AAAI, Citeseer, 1996, pp. 114–8.

[23] C. Lacave, R. Atienza, F. J. Díez, Graphical explanation in Bayesian Networks, in: International Symposium on Medical Data Analysis, Springer, 2000, pp. 122–129.

[24] C. Lacave, M. Luque, F. J. Diez, Explanation of Bayesian Networks and influence diagrams in Elvira, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 37 (2007) 952–965.

[25] G.-E. Yap, A.-H. Tan, H.-H. Pang, Explaining inferences in Bayesian Networks, Applied Intelligence 29 (2008) 263–278.

[26] J. Keppens, Explaining Bayesian Belief Revision for Legal Applications., in: JURIX, 2016, pp. 63–72.

[27] E. Kyrimi, W. Marsh, A progressive explanation of inference in 'hybrid' Bayesian networks for supporting clinical decision making, in: A. Antonucci, G. Corani, Campos (Eds.), Proceedings of the Eighth International Conference on Probabilistic Graphical Models, 2016, pp. 275–286.

[28] E. Pisirir, J. M. Wohlgemut, E. Kyrimi, R. S. Stoner, Z. B. Perkins, N. R. M. Tai, D. W. R. Marsh, A process for evaluating explanations for transparent and trustworthy ai prediction models, in: 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI), 2023, pp. 388–397. doi:10.1109/ICHI57859.2023.00058.

[29] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).

[30] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable AI: Challenges and prospects, arXiv preprint arXiv:1812.04608 (2018).

[31] I. Donoso-Guzmán, J. Ooge, D. Parra, K. Verbert, Towards a comprehensive human-centred evaluation framework for explainable AI, in: World Conference on Explainable Artificial Intelligence, Springer, 2023, pp. 183–204.

[32] M. Chromik, M. Schuessler, A taxonomy for human subject evaluation of black-box explanations in XAI., Exss-atec@ iui 1 (2020).

[33] Z. Buçinca, P. Lin, K. Z. Gajos, E. L. Glassman, Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems, in: Proceedings of the 25th international conference on intelligent user interfaces, 2020, pp. 454–464.

[34] A. Rosenfeld, Better metrics for evaluating explainable Artificial Intelligence, in: Proceedings of the 20th international conference on autonomous agents and multiagent systems, 2021, pp. 45–50.

[35] X. Kong, S. Liu, L. Zhu, Toward human-centered XAI in practice: A survey, Machine Intelligence Research (2024) 1–31.

[36] J. van der Waa, E. Nieuwburg, A. Cremers, M. Neerincx, Evaluating XAI: A comparison of rule-based and example-based explanations, Artificial Intelligence 291 (2021) 103404. doi:https://doi.org/10.1016/j.artint.2020.103404.

[37] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, ACM Trans. Interact. Intell. Syst. 11 (2021). URL: https://doi.org/10.1145/3387166. doi:10.1145/3387166.

[38] R. Visser, T. M. Peters, I. Scharlau, B. Hammer, Trust, distrust, and appropriate reliance in (x) AI: a survey of empirical evaluation of user trust, arXiv preprint arXiv:2312.02034 (2023).

[39] J. Zhou, A. H. Gandomi, F. Chen, A. Holzinger, Evaluating the quality of machine learning explanations: A survey on methods and metrics, Electronics 10 (2021) 593.

[40] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (2019) 832.

[41] S. L. Lauritzen, D. J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, Journal of the Royal Statistical Society. Series B (Methodological) 50 (1988) 157–224. URL: http://www.jstor.org/stable/2345762.

[42] J. Binder, D. Koller, S. Russell, K. Kanazawa, Adaptive probabilistic networks with hidden variables, Machine Learning 29 (1997) 213–244.