Teaching Bayesian Reasoning as a Pathway Toward Active Thinking and Explainable AI

Dimitrios Lappas[1], Panagiotis Karampelas[2] and Giorgos Fesakis[1]
[1]University of the Aegean, Rhodes, Greece
[2]Hellenic Air Force Academy, Dekeleia, Greece

dlappas.aegean@gmail.com
panagiotis.karampelas@hafa.haf.gr
gfesakis@rhodes.aegean.gr

Abstract: Decision-making under uncertainty requires not only computational tools but also critical thinking skills that allow individuals to evaluate assumptions, weigh evidence, and mitigate automation bias. While many contemporary AI systems operate as opaque black-box models, Bayesian Networks (BNs) provide a transparent and explainable alternative, making them well-suited for both decision support and AI education. This paper introduces an educational framework where learners construct, parameterize, and interpret Bayesian models to address authentic problems, such as classifying suspicious emails in cybersecurity. By explicitly modelling variables, dependencies, and prior assumptions, BNs engage students in probabilistic reasoning while promoting metacognitive reflection and critical evaluation of their decision-making process. The contribution of this work is threefold: (1) it positions Bayesian Networks as both a mathematical reasoning tool and an accessible entry point into explainable AI; (2) it integrates probability theory, critical thinking, and transparency into a unified framework for Responsible AI education; and (3) it demonstrates how transparent reasoning can support human-in-the-loop decision-making and reduce automation bias. While the framework does not claim to solve the general challenges of explainability in complex AI models, it offers a concrete and transferable pathway for cultivating active thinkers capable of designing, interpreting, and questioning AI-assisted decisions.

Keywords: Bayesian networks, Explainable AI, Probabilistic reasoning

1. Introduction

Critical thinking in decision-making is defined as the ability to analyse, evaluate, and synthesize information in order to choose the best possible course of action among multiple alternatives (Facione, 2015). It involves identifying biases, uncovering underlying assumptions, and weighing alternative options (Paul & Elder, 2006). Within decision-making contexts, critical thinking is not merely a theoretical competence but a dynamic process that guides individuals in making informed judgments, especially under conditions of limited or uncertain information.

Decision-making under uncertainty represents a particularly demanding cognitive task, since decision-makers must often estimate potential outcomes based on incomplete or ambiguous data (Kahneman & Tversky, 1979). Uncertainty may arise from either missing information or the inherent stochasticity of the environment. In such situations, critical thinking functions as a filter for risk assessment, helping individuals to mitigate systematic judgment errors and cognitive biases.

Decision support tools are commonly employed to enhance critical thinking in uncertain contexts. These tools provide structured methods for analysing data, calculating probabilities, and evaluating scenarios, allowing decision-makers to combine quantitative and qualitative elements before reaching conclusions (Power, 2008). However, the increasing integration of artificial intelligence (AI) into decision-support systems has introduced both opportunities and challenges. A major concern is automation bias, the tendency of users to over-rely on AI recommendations even when they are incorrect (Parasuraman & Manzey, 2010; Goddard et al., 2012). This risk is exacerbated in "black-box" AI models, such as deep neural networks, which produce outputs without offering sufficient interpretability (Cabitza et al., 2017; Ribeiro et al., 2016).

In this regard, Bayesian Networks provide a transparent and explainable alternative, grounded in probabilistic reasoning and causal modelling. By engaging users in the explicit definition of variables, dependencies, and prior assumptions, Bayesian reasoning not only supports more reliable decision-making but also cultivates critical and metacognitive skills (Tonekaboni et al., 2019). Through this process, learners are not passive recipients of AI outcomes but active participants who construct, test, and reflect on the models themselves (Bansal et al., 2019).

While Bayesian Networks are inherently transparent due to their graphical structure and probabilistic reasoning, their educational and practical value can be further enhanced through integration with contemporary Explainable AI (XAI) frameworks. Modern libraries such as SHAP (Lundberg & Lee, 2017), and

LIME (Ribeiro et al., 2016) focus on generating local and global explanations for complex black-box models, but they often struggle to provide users with causal insights. Bayesian Networks can act as complementary tools within such frameworks by offering causal and probabilistic explanations that are both mathematically grounded and pedagogically intuitive. For instance, visualization dashboards that combine SHAP-like feature importance with Bayesian causal graphs can help learners and practitioners not only understand which variables influenced a decision, but also why these relationships exist within a broader causal structure. This integration bridges the gap between interpretable statistical modelling and state-of-the-art AI explainability, positioning Bayesian Networks as both a didactic instrument and a practical component of responsible AI pipelines.

This paper introduces an educational framework for teaching Bayesian reasoning as a pathway to transparent and responsible AI. By progressively integrating concepts such as probability theory, conditional dependence, and Bayes' theorem into real-world applications of Bayesian Networks, the approach aims to empower learners to both understand and critically evaluate AI-based decision-making. The ultimate goal is to encourage a shift from passive users of opaque AI systems to active thinkers capable of designing, interpreting, and questioning decision-support models in complex and uncertain environments.

Importantly, this work does not claim to fully resolve the broader challenge of AI explainability, nor to guarantee the development of fully "active thinkers." Instead, it positions Bayesian Networks as an illustrative case through which learners can engage with uncertainty, bias, and causal reasoning in an interpretable way. While the framework is grounded in Bayesian modelling, its principles—transparent reasoning, user involvement, and metacognitive reflection—can inform broader educational strategies for interacting with complex AI models, including deep learning systems when combined with XAI libraries. In this sense, the contribution should be viewed as a pathway: a modest but concrete step toward cultivating reflective and critically engaged users of AI.

The main contribution of this paper lies in the design of an educational framework that integrates Bayesian reasoning into responsible AI education. By leveraging Bayesian Networks as both probabilistic models and interpretable learning tools, the framework offers a structured pathway for cultivating critical thinking, metacognitive reflection, and transparent decision-making under uncertainty. This contribution is situated at the intersection of mathematics education, explainable AI, and human-centred decision support.

The innovation of the work is twofold. First, it introduces Bayesian Networks not only as mathematical constructs but also as explainable AI instruments that can be embedded in learning environments to promote active user engagement. Second, it demonstrates how these principles can complement broader XAI methodologies, positioning Bayesian reasoning as a pedagogically grounded yet technologically relevant bridge between statistical learning and modern AI systems. In doing so, the paper provides a novel perspective: fostering a shift from passive reliance on opaque AI outputs toward active engagement with interpretable, causally informed models.

The remainder of the paper is structured as follows: Section 2 reviews related work on critical thinking, probabilistic reasoning, and research in Explainable AI (XAI). Section 3 presents the theoretical foundations of Bayesian reasoning. Section 4 introduces the proposed educational framework, followed by Section 5 which illustrates its application through a cybersecurity decision-making scenario. Section 6 outlines the methodological approach, Section 7 discusses implications for responsible AI and future directions, and Section 8 concludes the paper.

2. Related Work

The integration of Artificial Intelligence (AI) into decision-making processes has been widely studied, particularly in domains such as healthcare, finance, and cybersecurity (Topol, 2019; Samek et al., 2017). While AI systems can improve accuracy and efficiency, their reliance on opaque "black-box" models such as deep neural networks has raised concerns regarding transparency, interpretability, and trustworthiness (Ribeiro et al., 2016; Cabitza et al., 2017). These challenges have been linked to automation bias, the human tendency to over-rely on automated recommendations even when they are incorrect (Parasuraman & Manzey, 2010; Goddard et al., 2012).

To address these issues, research in Explainable AI (XAI) has sought to provide frameworks that enable users to understand and critically evaluate model outputs. Approaches such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) aim to approximate explanations for complex models, but their post-hoc nature has been criticized for offering partial or misleading insights (Lipton, 2018). In contrast, probabilistic graphical

models, and particularly Bayesian Networks (BNs), have been highlighted as inherently interpretable alternatives, as they explicitly represent causal relationships and conditional dependencies among variables (Pearl, 2018; Jensen & Nielsen, 2007).

Within the educational domain, Bayesian reasoning has been proposed as a valuable tool for teaching probabilistic thinking and decision-making under uncertainty (Konold, 1989; Gigerenzer & Hoffrage, 1995). Prior studies emphasize that engaging learners in the construction and evaluation of Bayesian models not only improves their statistical literacy but also fosters critical thinking by requiring them to assess assumptions, update beliefs with new evidence, and reflect on their reasoning processes (Conati et al., 2002; Bansal et al., 2019). Moreover, integrating Bayesian reasoning into AI education has been suggested as a pathway to developing more responsible and transparent uses of intelligent systems (Tonekaboni et al., 2019; Lyell et al., 2021).

The related literature was systematically reviewed using Google Scholar and Scopus databases combining search terms such as *"Bayesian reasoning in education," "Explainable AI," "critical thinking and AI,"* and *"automation bias."* Studies were included if they discussed probabilistic reasoning, AI explainability, or pedagogical frameworks involving uncertainty. This transparent search and selection process ensured that both foundational theories and recent XAI developments were represented.

Despite the rich literature on Explainable AI and Bayesian reasoning, few studies explicitly connect these domains within an educational context. Existing works often address either probabilistic reasoning or AI explainability separately, without proposing a unified pedagogical model that links the two. This gap highlights the need for an integrative framework that teaches explainable reasoning through hands-on Bayesian modelling—precisely the contribution of this study.

Taken together, these strands of research highlight the need for pedagogical approaches that move learners from passive consumers of AI outputs to active participants in the construction of interpretable, probabilistic models. Our work builds on this line of inquiry by proposing a didactic framework for teaching Bayesian Networks as both a mathematical tool and a means of cultivating critical thinking for transparent AI.

3. Theoretical Framework

3.1 Reasoning under Uncertainty: Human and AI Perspectives

Decision-making under uncertainty has long been analyzed through two complementary perspectives. The first is the intuitive and heuristic approach (Agor, 1986), which emphasizes how humans rely on mental shortcuts, intuition, and biases when probabilities are unclear. Classical studies by Tversky and Kahneman (1974) demonstrated that under uncertain conditions, individuals often deviate from rational models and adopt heuristics such as availability, representativeness, or anchoring. While these strategies enable fast judgments, they also introduce systematic errors. Interestingly, similar biases can be observed in the use of automated AI systems, where users may over-trust model outputs without scrutinizing underlying assumptions (Parasuraman & Manzey, 2010).

The second perspective is the probabilistic approach (Selvin, 1975), which treats uncertainty through formal probability models. Within this approach two dominant schools emerge: the frequentist view, where probability reflects long-run frequencies, and the Bayesian view, where probability encodes the degree of belief about a hypothesis given available evidence. The Bayesian framework is particularly relevant to AI, since it provides a formal mechanism for updating prior beliefs with new data, enabling dynamic inference rather than static prediction (Pearl, 1988; Fenton & Neil, 2013).

In the context of artificial intelligence, this theoretical distinction mirrors the contrast between black-box machine learning models, which often replicate heuristic reasoning without transparency, and Bayesian reasoning, which offers an explainable and statistically grounded methodology for handling uncertainty. Thus, understanding Bayesian inference is not only a matter of mathematical education but also a cornerstone for developing AI systems that promote transparency, accountability, and critical engagement.

3.2 Cultivating Probabilistic Reasoning in Education and AI

Developing probabilistic reasoning is widely recognized as a critical educational objective, particularly in data-driven societies where uncertainty permeates decision-making (Fenton & Neil, 2011; Garfield et al., 2008; Batanero et al., 2016). Traditional instruction often emphasizes formulaic computation of probabilities, but

research suggests that fostering deeper understanding requires engaging learners in the cognitive process of reasoning under uncertainty rather than in mechanical calculation alone (Zieffler et al., 2008).

From an AI perspective, this pedagogical insight parallels current debates in explainable AI (XAI). Just as students must move beyond rote learning of probability rules, AI users must move beyond passive acceptance of black-box predictions. Tools such as Bayesian Networks can serve both as computational models for inference and as didactic instruments that make reasoning under uncertainty visible and interpretable (Salter-Townshend et al., 2012; Khuda et al., 2024).

Inquiry-based and discovery-oriented approaches (Bruner, 2004) align well with this goal. By constructing Bayesian models, learners not only simulate uncertainty but also observe how new evidence reshapes prior beliefs in real time. This mirrors the way transparent AI systems should allow end-users to interrogate, adjust, and reflect on model assumptions. In addition, realistic case studies — such as cybersecurity or medical diagnosis — situate probabilistic reasoning in authentic AI-driven decision contexts, reinforcing both statistical logic and metacognitive awareness (Biehler et al., 2015).

Ultimately, Bayesian reasoning bridges the gap between mathematics education and responsible AI design. It transforms uncertainty from an obstacle into a structured problem-solving process, equipping both learners and AI users with the capacity to critically evaluate assumptions, revise judgments, and resist automation bias. In this sense, the theoretical framework of Bayesian reasoning serves a dual role: as a cognitive scaffold for human learners and as a paradigm for building AI systems that are inherently explainable, transparent, and trustworthy**.**

3.3 Theoretical Foundation of Bayesian Networks

At the core of Bayesian Networks lies Bayes' theorem, which formalizes how beliefs about the probability of a hypothesis are updated in light of new evidence. Formally:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

where P(H) is the prior probability of a hypothesis, P(E|H) is the likelihood of observing evidence E given H, and P(H|E) is the posterior probability once new information is incorporated. This updating mechanism positions Bayesian reasoning as a dynamic learning process, rather than a static estimation of probabilities (Pearl, 1988; Fenton & Neil, 2013).

Bayesian Networks extend this principle by structuring conditional dependencies among multiple variables through a directed acyclic graph (DAG). Each node represents a random variable, while edges denote causal or probabilistic relationships. Crucially, each node is equipped with a conditional probability table (CPT) that quantifies how its value depends on its parents in the graph. This compact representation allows for efficient reasoning in complex domains where uncertainty is pervasive.

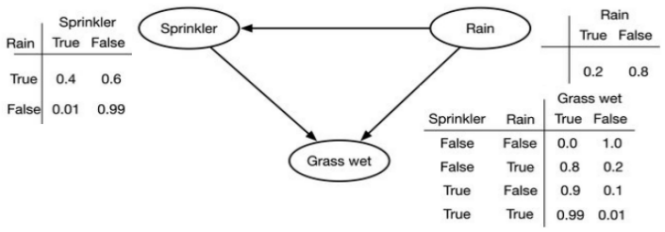

Figure 1: Example of a Bayesian network with conditional probabilities (Zhang et. al., 2021)

From an artificial intelligence perspective, Bayesian Networks are particularly valuable because they bridge two traditionally separate areas:

- Statistical inference, by enabling robust probabilistic reasoning under uncertainty.
- Causal modeling, by making explicit assumptions about how variables influence one another (Pearl, 2009).

This dual role makes Bayesian Networks central to the current movement toward explainable and responsible AI. Unlike neural networks or other black-box models, Bayesian Networks allow users to trace the reasoning process: one can examine which variables drive the outcome, how evidence updates prior assumptions, and

where uncertainty remains. This transparency fosters user trust and mitigates automation bias by encouraging critical engagement with model outputs (Tonekaboni et al., 2019).

Furthermore, Bayesian Networks have been successfully applied in domains where both accuracy and interpretability are essential, including medical diagnosis, cybersecurity, environmental management, and decision-support systems (Fenton & Neil, 2013; Bansal et al., 2019). Their theoretical foundation thus provides not only a mathematical framework but also a paradigm for AI that is both transparent and pedagogically powerful, making them ideal for contexts where human decision-makers must remain active evaluators rather than passive recipients of algorithmic outputs.

3.4 Problem Statement

Despite the rapid progress of Artificial Intelligence in decision-support systems, a persistent challenge remains: the opacity of black-box models. Neural networks and other complex architectures often deliver accurate outputs but provide limited insight into their reasoning processes. This lack of transparency fosters automation bias, where users over-rely on algorithmic recommendations without questioning their validity (Parasuraman & Manzey, 2010; Goddard et al., 2012). Such dependence weakens critical thinking and undermines responsible AI adoption, particularly in high-stakes domains such as cybersecurity, healthcare, and environmental management.

While Explainable AI (XAI) research has advanced significantly, most approaches focus on post-hoc explanations, attempts to justify or approximate the reasoning of opaque models (Ribeiro et al., 2016; Cabitza et al., 2017). These explanations, however, often remain abstract and insufficient for cultivating users' ability to critically evaluate AI outputs. What is missing is a systematic pedagogical framework that trains decision-makers not only to interpret explanations but also to construct and reason with transparent models themselves.

Bayesian Networks (BNs) offer a promising foundation for bridging this gap. By explicitly encoding variables, dependencies, and conditional probabilities, BNs provide a transparent mechanism for probabilistic reasoning (Pearl, 1988; Fenton & Neil, 2013). Unlike black-box models, they allow users to inspect assumptions, update beliefs with new evidence, and trace causal pathways. Yet, their potential as an educational tool for teaching explainability and fostering critical thinking in AI contexts remains underexplored.

This paper addresses this gap by proposing an innovative instructional intervention that employs Bayesian Networks not only as a statistical tool but also as a didactic framework for Responsible AI education. By engaging learners in building and testing their own BN models, the approach shifts them from passive users of opaque systems to active thinkers capable of transparent reasoning and informed decision-making.

Figure 2 illustrates the overall research design and process, outlining the methodological steps from theoretical foundation to data interpretation.

Figure 2: Research Design and Process Flow

4. Educational Framework: Bayesian Reasoning as a Pathway to XAI

The proposed educational framework positions Bayesian Networks not only as a statistical tool but as a structured pathway toward cultivating explainable reasoning in Artificial Intelligence (AI). While traditional probability instruction often focuses on abstract formulas and mechanical calculations, this framework emphasizes the progressive development of critical and metacognitive skills through authentic problem-solving. The core innovation lies in explicitly linking Bayesian reasoning with the principles of transparency and interpretability that underpin Explainable AI (XAI).

The framework is organized into seven interconnected stages (table 1), each designed to build upon the previous one and guide learners from foundational probability concepts toward the construction and reflection on Bayesian models in real-world contexts. This staged progression ensures that learners move step by step from understanding uncertainty mathematically to practicing transparent reasoning in applied scenarios.

Table 1: Pedagogical structure of the intervention

Stage	Description	Student Activity	Learning Objective
Introduction to Probability	Basic concepts, definitions, and computations	Solving simple probability problems	Establishing a common mathematical foundation
Independent & Dependent Events	Relationships between events	Identifying event relations in real-world scenarios	Connecting abstract concepts with empirical observations
Conditional Probability	Updating probabilities with new information	Analysing examples	Understanding the influence of observations
Bayes' Theorem	Application in belief revision	Calculating simple Bayesian problems	Introducing the logic of Bayesian updating
Bayesian Networks – Theory	Nodes, causal links, conditional probability tables	Designing structures in software	Linking theory with practical model building
Application to Authentic Problems	Case studies (e.g., cybersecurity, environment, diagnosis)	Constructing and analysing Bayesian models	Developing modelling and decision-making skills
Presentation & Reflection	Interpretation and discussion of results	Presenting models to peers	Fostering metacognitive awareness and communication skills

Figure 3 presents the internal flow of the proposed educational framework, describing the sequential learning stages through which participants engage with Bayesian reasoning and explainable AI concepts.

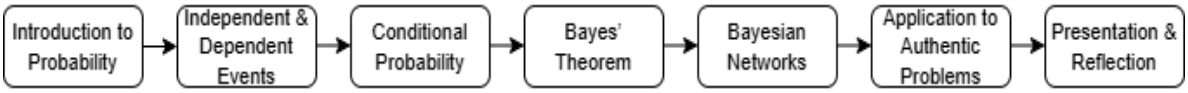

Figure 3: Educational Flow within the Framework

At each stage, learners are encouraged to not only compute probabilities but also to make their reasoning explicit. This emphasis on transparency mirrors the goals of XAI: making decision processes visible, interpretable, and open to critique. For example, when students update beliefs using Bayes' theorem, they practice a transparent form of reasoning where assumptions (priors), evidence (likelihood), and updated conclusions (posterior) are all explicitly represented. Similarly, the graphical structure of Bayesian Networks allows learners to see how evidence propagates through causal dependencies, offering a direct analogy to the explanatory mechanisms that AI researchers aim to achieve in complex machine learning models.

Thus, the educational framework goes beyond teaching probability theory. It instills an *explainable mindset*, a way of thinking where learners not only generate outputs but also understand and articulate the reasoning that leads to them. This dual orientation, combining mathematical rigor with explainability, prepares learners to engage critically with AI systems and provides a didactic microcosm of XAI principles. In this way, Bayesian reasoning becomes both a pedagogical bridge and a conceptual foundation for building responsible, transparent AI practices.

5. Application: Bayesian Networks in Cybersecurity Decision-MakingProblem Setting

To demonstrate the practical and pedagogical value of the proposed framework, this section presents a hands-on example situated in the domain of cybersecurity. Decision-making under uncertainty is a daily challenge for security analysts, who must evaluate large volumes of incoming data—such as emails, logs, and network activity—in order to determine whether a potential incident is benign or malicious. This scenario provides a fertile context for engaging learners with Bayesian reasoning, since it requires combining incomplete and often ambiguous evidence while avoiding systematic cognitive biases.

5.1 Problem Setup

Using GeNIe Academic software, learners are asked to build a Bayesian network for email triage with one parent node and three evidence nodes:

- Spam (parent; states: *Yes*, *No*)
- Suspicious Sender (child of *Spam*; *Yes/No*)

- Unsafe Link (child of *Spam*; *Yes/No*)
- Unusual Language (child of *Spam*; *Yes/No*)

The model assumes conditional independence of the three evidences given *Spam*.

Learners are also given the following probabilities (table 2)

Table 2: Probabilities extracted from historical data

Evidence node	Condition	P(Yes)	P(No)
Spam email		0.16	0.84
Suspicious Sender	given **Spam=Yes**	0.70	0.30
	given **Spam=No**	0.05	0.95
Unsafe Link	given **Spam=Yes**	0.65	0.35
	given **Spam=No**	0.10	0.90
Unusual Language	given **Spam=Yes**	0.80	0.20
	given **Spam=No**	0.15	0.85

Learners must then compute the probability that a message is spam if it shows a suspicious sender and unusual language, but does not contain a link. Using the Bayesian network they developed, it turns out that the probability of an email with the above characteristics being spam is 0.85 (figure 4).

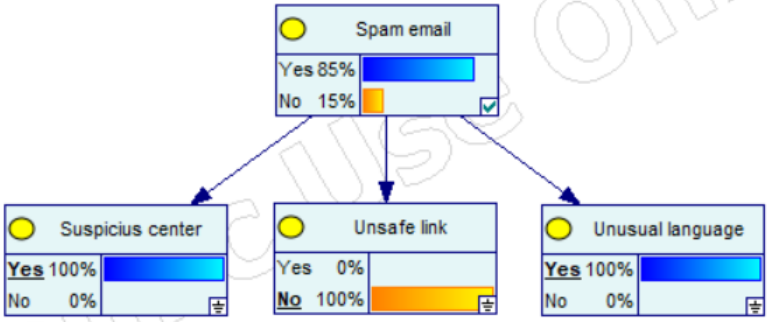

Figure 4: The Bayesian network that the trainees are asked to build

Through the constructed network, students apply Bayes' theorem to update prior beliefs in light of new observations. For example, the absence of an unsafe link reduces the posterior probability of spam, but the co-occurrence of a suspicious sender and unusual language significantly increases it. This dynamic revision illustrates how Bayesian reasoning handles uncertainty in a structured, interpretable way.

The exercise also highlights how small changes in observed features can drastically shift decision outcomes, reinforcing the importance of data quality and feature evaluation in AI-driven decision support.

5.2 Transparent Reasoning vs. Black-Box AI

The strength of this activity lies not only in its mathematical rigor but also in its alignment with the principles of Explainable AI. Unlike opaque neural networks, Bayesian Networks make the reasoning process explicit. Students can clearly observe how each piece of evidence (e.g., unusual language) modifies the posterior probability of spam. The transparency of priors, likelihoods, and causal dependencies fosters metacognitive reflection, allowing learners to question their assumptions, identify potential biases, and understand why certain outcomes emerge.

This transparency directly addresses the problem of automation bias: rather than blindly trusting algorithmic outputs, students learn to interrogate the decision-making process itself. In this way, the exercise not only enhances statistical literacy but also cultivates an explainable mindset, equipping learners with the ability to critically evaluate AI-driven decision-support systems.

5.3 Pedagogical and AI Implications

From an educational perspective, this example reinforces the framework presented in Section 4 by bridging abstract probability theory with authentic, high-stakes decision-making scenarios. From an AI perspective, it highlights how Bayesian reasoning serves as a microcosm of XAI, demonstrating the value of transparent, causal, and interpretable modelling. By engaging in this process, students are not passive users of opaque algorithms but active thinkers who construct, test, and explain their models—mirroring the broader goals of responsible AI.

6. Discussion

This study highlights the pedagogical potential of Bayesian reasoning as both a method for structured decision-making under uncertainty and as a model of explainable AI. By engaging learners in the explicit construction of Bayesian Networks, the framework reduces automation bias and emphasizes human-in-the-loop decision-making. Rather than relying uncritically on opaque algorithmic outputs, students learn to interpret causal structures, evaluate assumptions, and iteratively update beliefs in response to new data.

The contribution of this approach is therefore twofold. First, it empowers learners to become active thinkers, capable of questioning, testing, and refining models instead of acting as passive recipients of AI-generated outputs. Second, it provides a concrete example of explainable AI, where the reasoning process is transparent and interpretable. In contrast to black-box models, Bayesian Networks show how evidence combines to shape conclusions, fostering not only mathematical understanding but also a mindset of reflective and responsible decision-making.

Nevertheless, some limitations must be acknowledged. The proposed framework currently focuses on Bayesian reasoning and does not directly extend to more complex models such as deep neural networks, which dominate many AI applications. While Bayesian Networks are uniquely suited for illustrating explainability through causal modelling, the generalizability of this approach requires further investigation. Future work should explore how similar pedagogical principles can be applied to other AI paradigms, for instance by integrating Bayesian explanations with post-hoc XAI tools such as SHAP or LIME, or by developing hybrid teaching modules where Bayesian reasoning complements the interpretability of more complex architectures.

In this sense, teaching Bayesian reasoning acts as a bridge: it cultivates learners' ability to critically reason under uncertainty ("active thinkers") while simultaneously introducing them to transparent models that exemplify the principles of explainable AI. This dual orientation situates the intervention at the intersection of education, decision science, and responsible AI.

7. Conclusion

This paper introduced a pedagogical framework for teaching Bayesian reasoning as a pathway toward active thinking and explainable AI. By guiding learners through the progressive construction of Bayesian Networks, from probability fundamentals to real-world applications, the framework enables them to move beyond passive use of AI tools and toward active engagement with the reasoning process itself.

The educational value lies not only in enhancing probabilistic literacy but also in cultivating critical and metacognitive skills that reduce automation bias and strengthen responsible decision-making. At the same time, the framework positions Bayesian reasoning as a paradigmatic example of explainable AI, demonstrating how transparent and causal models can complement modern AI practices.

By positioning Bayesian reasoning as both a cognitive tool and an AI paradigm, the proposed framework illustrates how education can simultaneously nurture active thinkers and advance the goals of explainable and responsible AI.

Ethics Declaration: No ethical clearance was required for this study.

AI Declaration: AI tools were used solely for English language editing.

References

Agor, W. H. (1986). *The logic of intuitive decision-making.* Westport, CT: Quorum Books.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human–AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*(1), 2–11. https://doi.org/10.1609/hcomp.v7i1.5285

Batanero, C., Chernoff, E., Engel, J., Lee, H., & Sánchez, E. (2016). *Research on teaching and learning probability.* ICME-13 Topical Surveys. Springer. https://doi.org/10.1007/978-3-319-31625-3_1

Biehler, R., Frischemeier, D., & Gould, R., & Pfannkuch, M. (2024). Impacts of digitalization on content and goals of statistics education. In B. Pepin, G. Gueudet, & J. Choppin (Eds.), *Handbook of digital resources in mathematics education* (pp. 547–583). Springer.

Bruner, J. (2004). The narrative creation of self. In L. E. Angus & J. McLeod (Eds.), *The handbook of narrative and psychotherapy: Practice, theory, and research* (pp. 3–14). Sage Publications. https://doi.org/10.4135/9781412973496.d3

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *JAMA*, *318*(6), 517–518. https://doi.org/10.1001/jama.2017.7797

Conati, C., Gertner, A., & VanLehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction*, *12*, 371–417. https://doi.org/10.1023/A:1021258506583

Facione, P. (2015). *Critical thinking: What it is and why it counts.* Insight Assessment.

Fenton, N., & Neil, M. (2011). The use of Bayes and causal modelling in decision making, uncertainty and risk. *Journal of Risk Research*, *12*, 1–20.

Fenton, N., & Neil, M. (2018). *Risk assessment and decision analysis with Bayesian networks* (2nd ed.). CRC Press. https://doi.org/10.1201/b21982

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Jensen, F. V., & Nielsen, T. D. (2007). *Bayesian networks and decision graphs* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-68282-2

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision making under risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10.2307/1914185

Khuda, I. (2021). Innovative teaching pedagogy for teaching and learning of Bayes' theorem. *Cankaya University Journal of Science and Engineering*, *18*(1), 61–71.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction*, *6*(1), 59–98. https://doi.org/10.1207/s1532690xci0601_3

Lyell, D., Coiera, E., Chen, J., Shah, P., & Magrabi, F. (2021). How machine learning is embedded to support clinician decision-making: An analysis of FDA-approved medical devices. *BMJ Health & Care Informatics*, *28*(1), e100301. https://doi.org/10.1136/bmjhci-2020-100301

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4766–4777.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Paul, R., & Elder, L. (2006). *Critical thinking: Learn the tools the best thinkers use.* Pearson Prentice Hall.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* Morgan Kaufmann.

Pearl, J. (2018). *The book of why: The new science of cause and effect.* Basic Books.

Power, D. J. (2008). Decision support systems: A historical overview. In *Handbook on decision support systems* (pp. 121–140). Springer. https://doi.org/10.1007/978-3-540-48713-5_7

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Salter-Townshend, M., White, A., Gollini, I., & Murphy, T. B. (2012). Review of statistical network analysis: Models, algorithms, and software. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *5*(4), 243–264. https://doi.org/10.1002/sam.11146

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing, and interpreting deep learning models. *ITU Journal: ICT Discoveries*, *1*(1), 39–48.

Selvin, S. (1975). On the Monty Hall problem (Letter to the editor). *The American Statistician*, *29*(3), 134. https://doi.org/10.2307/2683689

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. *Proceedings of the 4th Machine Learning for Healthcare Conference*, 359–380. https://proceedings.mlr.press/v106/tonekaboni19a.html

Topol, E. J. (2019). *Deep medicine: How artificial intelligence can make healthcare human again.* Basic Books.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Zhang, W., Ramezani, R., & Naeim, A. (2021). An introduction to causal reasoning in health analytics. *arXiv preprint* arXiv:2105.04655. https://doi.org/10.48550/arXiv.2105.04655

Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., & Chang, B. (2008). What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature. *Journal of Statistics Education*, *16*(2), 1–22. https://doi.org/10.1080/10691898.2008.11889566