UNDERSTANDING UNCERTAINTIES IN EXPLAINABLE AI: A STRUCTURED LITERATURE REVIEW AND RESEARCH AGENDA

Completed Research Paper

Maximilian Förster, Karlsruhe Institute of Technology, Germany, maximilian.foerster@kit.edu
Michael Hagn, University of Regensburg, Germany, michael.hagn@ur.de
Nico Hambauer, University of Regensburg, Germany, nico.hambauer@ur.de
Paula Jaki, TU Darmstadt, Germany, jaki@ise.tu-darmstadt.de
Andreas Obermeier, University of Ulm, Germany, andreas.obermeier@uni-ulm.de
Andreas Schauer, University of Regensburg, Germany, andreas.schauer@ur.de
Alexander Schiller, University of Regensburg, Germany, alexander.schiller@ur.de
Julian Wohlschlegel, TU Darmstadt, Germany, julian.wohlschlegel@tu-darmstadt.de
Alexander Benlian, TU Darmstadt, Germany, benlian@ise.tu-darmstadt.de
Bernd Heinrich, University of Regensburg, Germany, bernd.heinrich@ur.de
Ekaterina Jussupow, TU Darmstadt, Germany, ekaterina.jussupow@tu-darmstadt.de
Mathias Klier, University of Ulm, Germany, mathias.klier@uni-ulm.de
Mathias Kraus, University of Regensburg, Germany, mathias.kraus@ur.de
Daniel Schnurr, University of Regensburg, Germany, daniel.schnurr@ur.de

Abstract

Artificial Intelligence (AI) is increasingly deployed in high-stakes domains such as healthcare, where decisions carry significant consequences for individuals and society. This amplifies the need for safe and trustworthy AI systems. While traditional explainable AI (XAI) methods aim to increase transparency, they often omit the uncertainties that are inherent in XAI-augmented decision-making, where AI supports human decision-makers. To better understand the sources and effects of uncertainty in XAI-augmented decision-making and structure the existing body of knowledge, we conduct a structured literature review focused on four main sources of uncertainty: data uncertainty, model uncertainty, XAI method uncertainty, and human uncertainty. Our review shows that these sources are largely examined in isolation, resulting in a fragmented understanding of how uncertainties interact and influence decision-making. Based on these findings, we outline how multiple sources of uncertainty can be incorporated into uncertainty-aware explanations and propose an integrated research agenda for developing uncertainty-aware XAI.

Keywords: Uncertainty, Explainable AI, Artificial Intelligence, XAI-augmented decision-making

1 Introduction

Artificial Intelligence (AI) is increasingly deployed in high-stakes domains such as healthcare, finance, and criminal justice, where decisions carry significant consequences for individuals and society (Meske

et al., 2022, Sunyaev et al., 2025). While AI already often matches or even surpasses the capabilities of human experts in these domains (McKinney et al., 2020), this growing reliance on AI also introduces new risks and challenges, thereby amplifying the importance of developing safe and trustworthy AI systems. This has prompted new AI regulations, most notably the European AI Act, which establishes explicit requirements for transparency, explainability, and human oversight of AI systems in high-risk applications (Schnurr, 2025). In doing so, the act reflects a broad societal consensus that AI in critical domains should be evaluated through a human-centric lens, beyond technical performance assessments alone, to ensure trustworthy outcomes. Central to these requirements is the use of explainable AI (XAI) methods that reveal the internal logic and inner workings of machine learning (ML) models to humans (Barredo Arrieta et al., 2020; Brasse et al., 2023). However, recent research reveals a persistent problem: despite various explanation techniques, humans still miscalibrate their trust in AI systems through both overreliance when systems are unreliable and underreliance when they are dependable (Ghassemi et al., 2021; Nourani et al., 2021). At the same time, traditional XAI methods often omit the uncertainties that are inherent in XAI-augmented decision-making, where (X)AI supports human decision-makers. Such uncertainties are defined as states "of not knowing whether a proposition is true or false" (Holton, 2004). Indeed, in XAI-augmented decision-making, AI predictions are never provably correct but hypothetical and potentially suffering from including incorrect model assumptions and noisy or imprecise data (Hüllermeier & Waegeman, 2021). Neglecting this uncertainty inherent to AI and its explanations is problematic and ignoring it may create an illusion of certainty, which can lead to misplaced confidence, poor decisions, or failures of accountability. In the worst case, explanations may foster a false sense of trustworthiness and undermine the potential benefits of AI in high-stakes decisions (Gawlikowski et al., 2023).

Uncertainty in XAI-augmented decision-making can arise from multiple sources, most notably the underlying data, the ML models themselves, the explanation methods employed, and the human decision-makers who interact with these systems. Accounting for these uncertainties has significant potential for empowering human users to better evaluate risks, calibrate appropriate trust, and mitigate negative consequences of AI. However, to date, a comprehensive overview of the impacts of these uncertainties, and integrated research efforts to address associated challenges, remain lacking. Indeed, while exhaustive prior surveys exist for individual sources of uncertainty, such as the data (e.g., Gong et al., 2023) or the ML model (e.g., Gawlikowski et al., 2023; Zhou et al., 2022), they predominantly examine these factors in isolation. Moreover, existing work often treats uncertainty as a purely technical property or focuses on abstract mathematical formalizations (e.g., Guo et al., 2024). They lack an integrated socio-technical perspective that considers how uncertainties propagate from data and models to explanations and ultimately to human decision-makers. This fragmented landscape hinders a comprehensive understanding of how uncertainties interact, potentially amplify each other and ultimately affect human decisions (Jussupow et al., 2021). To address this research gap and develop a more systematic understanding of how uncertainty arises within and across XAI methods and the human decisions made with their support, this paper investigates the following research questions:

RQ1: What are the main sources and effects of uncertainty in XAI-augmented decision-making?

RQ2: Which research gaps exist with respect to uncertainty-awareness in XAI-augmented decision-making, and how should they be addressed by future IS research?

To address RQ1, we conduct a structured literature review and find that most research to date has examined individual sources of uncertainty in AI separately rather than in combination. Data uncertainty has been investigated primarily in the context of data quality (Hagn et al., 2025; Heinrich et al., 2025), ML model uncertainty through quantification methods in ML (Gal & Ghahramani, 2016), XAI method uncertainty by XAI researchers concerned with stability and fidelity (e.g., Heinrich et al., 2024), and human uncertainty within behavioral research and cognitive science studies (He et al., 2023). Early work has begun to explore how to communicate singular sources of uncertainty to users (Antorán et al., 2020; Bobek & Nalepa, 2021; Wang et al., 2021). To advance toward a more integrated understanding of uncertainty in XAI-augmented decision-making, we examine the interdependencies of sources of uncertainty and propose the field to move towards uncertainty-aware explanations. Addressing RQ 2, we propose a research agenda on how to inform the design and evaluation of uncertainty-aware

explanations based on the research gaps identified in our literature review. This perspective should guide the development of methodologies and metrics that help users interpret and manage these uncertainties (Gong et al., 2023), ultimately advancing the development of uncertainty-aware XAI that supports calibrated trust and safer human-AI decision-making in high-stakes domains.

2 Conceptual Background

In XAI-augmented decision-making, the goal is typically to design and implement ML models that provide accurate predictions to assist human users in performing a particular task or making a particular decision. In typical use cases, human users are provided with the prediction of an ML model, possibly along with an explanation from an XAI method. However, this decision-making process is often shaped by uncertainty arising from multiple sources and across different stages of human-AI decision-making.

First, *data uncertainty* emerges from issues such as incompleteness, noise, or imprecise measurements, which can undermine the reliability of both model training and subsequent predictions (Gudivada et al., 2017). Several surveys provide a broader overview of this challenge in the context of AI (Gong et al., 2023) and big data (Gudivada et al., 2017), although their focus remains limited to data-centric aspects rather than its integration with the broader decision-making context. Thus, its propagation to model predictions, explanations, and ultimately human users remains unclear.

Second, *ML model uncertainty* stems from the fact that ML models only approximate the real relationships between inputs and outputs, meaning they are constrained by limits in model capacity, optimization strategies, parameters, or lack of knowledge; while Bayesian approaches can improve calibration and predictive transparency, research has yet to explore how this uncertainty interacts with other sources or how it should be explained to human users (Tagasovska & Lopez-Paz, 2019; Zhou et al., 2022). Overviews such as Gawlikowski et al. (2023) and Zhou et al. (2022) summarize this stream of work, though their integration with XAI remains underexplored, which often results in uncertainty estimates being treated primarily as technical properties of models rather than as information that needs to be communicated to the human user.

Third, *XAI method uncertainty* arises when explanation techniques fail to adequately capture or convey the variability and limitations of model predictions, sometimes even masking critical errors of the underlying ML model by providing explanations that appear transparent yet are misleading or incorrect. Although several overview papers on XAI exist (Barredo Arrieta et al., 2020; Brasse et al., 2023; Meske et al., 2022), uncertainty-aware explanations remain in their infancy, with early work examining uncertain feature attributions (Wang et al., 2021) or robustness issues (Slack et al., 2021; Zhang et al., 2019). Without explicitly linking these concerns to uncertainties in the underlying data and ML models, XAI may create a false sense of reliability for human users.

Finally, *human uncertainty*, grounded in expertise, cognitive biases, or confidence levels, crucially shapes users' engagement in AI predictions. While structured overviews of human uncertainty in XAI are still lacking, prior work provides different perspectives on competencies and risk attitudes that impact how users manage uncertainty in XAI-augmented decision-making (Croskerry & Norman, 2008; He et al., 2023; Jussupow et al., 2020). As these sources of uncertainty emerge at different stages of the decision-making process and jointly influence the final outcome, it is important to understand how they interact and propagate through the overall XAI-augmented decision-making process, rather than considering them in isolation. Such a synthesis could inform the design of systems that communicate uncertainty more effectively.

3 Methodology

Our structured literature overview follows standard procedures and IS guidelines (Brocke et al., 2009; Webster & Watson, 2002). Table 1 summarizes our search strings. For each source of uncertainty (see Section 2) we first created a type-specific search term and concatenated it with a general XAI search string to ensure that results were grounded in the XAI context. The four sources of uncertainty as a basis for our literature search were defined a priori based on the existing literature on uncertainty in XAI. For

data uncertainty, known causes, such as inaccurate, outdated, and missing data are of particular interest. For ML model uncertainty, common ways of modeling uncertainty such as Bayes methods, boosting and bagging methods, or ensemble methods were included as search terms. For XAI method uncertainty, we included popular metrics to measure uncertainty in explanations in our search string. Finally, for human uncertainty, we focused specifically on constructs that capture how humans perceive, interpret and respond to explanations, such as confidence, overreliance or sensemaking, in our search string. To complement these specific search strings and to ensure that broader perspectives on uncertainties in XAI were not overlooked, we additionally constructed a general search string combining the term uncertainty with common XAI-related concepts. Using these terms, we conducted a title-abstract-keyword search. The focus of our structured literature search was on high quality, peer-reviewed outlets. Accordingly, we searched the AIS and ACM libraries and added twelve prestigious venues that regularly publish research on uncertainty in AI: IEEE TPAMI, ESWA, JMLR, Knowledge Based Systems, Pattern Recognition, Nature Machine Intelligence, ICML, NeurIPS, ICLR, UAI, AAAI, and IJCAI.

A total of 923 papers were identified using our search strings. Of these, 241 focused on data uncertainty, 196 on ML model uncertainty, 60 on XAI method uncertainty, 145 on human uncertainty, and 281 addressed uncertainty in general. Within a team of eight annotators, each of the 923 papers was individually labeled for the four sources of uncertainty, allowing multiple categories to be assigned per paper. Papers were excluded at the title and abstract screening stage if they did not address uncertainty related to the data, ML model, XAI method, or human decision-making. After this screening stage, 144 papers remained for further review. At the full-text screening stage, papers were further excluded if they did not substantively engage with at least one of the four sources of uncertainty. Based on this full-text screening, 52 papers were identified as relevant. A subsequent forward and backward search yielded an additional 40 papers, resulting in a final set of 92 papers included in this structured literature review. To verify the quality of our labels, we formed pairs of two annotators and measured inter-rater reliability using Cohen's Kappa, based on a 20% sample from each search string. We obtained an overall Cohen's Kappa of 0.814, indicating nearly perfect agreement (Landis & Koch, 1977).

Source of Uncertainty	**Search String**		
Data Uncertainty	"data quality" OR "data problems" OR "inaccurate data" OR "incorrect data" OR "incomplete data" OR "inconsistent data" OR "outdated data" OR "missing data" OR "defect data" OR "deficient data" OR (("completeness" OR "consistency" OR "currency") AND "data")	AND	"XAI" OR ((explain* OR "interpretable" OR "explanation") AND ("artificial intelligence" OR "machine learning" OR "AI" OR "ML"))
ML Model Uncertainty	Bayes* OR "boosting" OR "bagging" OR ensemble*		
XAI Method Uncertainty	fidel* OR faithful*		
Human Uncertainty	"mental model" OR "overreliance" OR "predictive scores" OR "confidence" OR "metacognition" OR "risk preferences" OR "sensemaking" OR "attitudes" OR "ambiguity preferences"		
Uncertainty in XAI	uncertain* OR certain*		

Table 1. Search strings used for our structured literature analysis.

4 Results

In this section, we present selected key findings from the literature search. The complete list of identified papers is provided in an online appendix[1]. Figure 1 illustrates the distribution of papers among the four sources of uncertainty identified in the systematic annotation process.

[1] https://osf.io/2wbua/

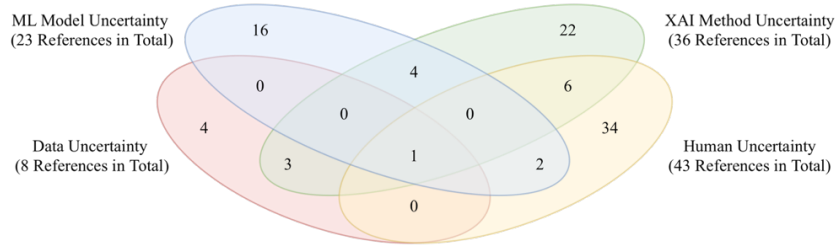

Figure 1. Distribution of the 92 reviewed papers across the four sources of uncertainty.

4.1 Data Uncertainty

Data uncertainty refers to indeterminacy of data used for the training, testing, and application of ML. In the context of XAI, data uncertainty stems from data quality defects in the data used for the ML model and the XAI method, with data quality defects indicating a disagreement between the real world and the representing data (Heinrich et al., 2025). These data quality defects can take on different forms, such as inaccurate, outdated, or missing data (Hagn et al., 2025). For instance, sensor errors in clinical monitoring or biased training data from underrepresented patient populations are typical sources of data uncertainty in healthcare. It is well known that data quality defects can degrade the performance of ML (Mohammed et al., 2025), resulting in suboptimal models, incorrect predictions and invalid explanations (Morrison et al., 2024). While data quality is considered by some approaches in (X)AI-augmented decision-making, data uncertainty may persist or even increase when such approaches are employed, e.g., when incomplete samples are deleted (Peng et al., 2023). Several papers have conducted overviews regarding data uncertainty in (X)AI-augmented decision-making. While some of these works (e.g., Paleyes et al. 2022) provide general discussions of data uncertainty as one of multiple relevant challenges in AI, most focus on data uncertainty in a certain stage of the (X)AI-augmented decision-making process. First, some works examine the data uncertainty itself, aiming to give an overview of data quality defects and their impact on ML (Gong et al., 2023). Next, there are reviews of the ML stage, highlighting ML model modifications to increase robustness against data uncertainty (Gawlikowski et al., 2023). In addition, Hagn et al. (2025) provide a taxonomy on ML on data with data quality defects, which enables both a classification of the defects and the approaches to address the defects. While these overviews offer valuable insights into data uncertainty in the context of AI, they do not address the relationship between data uncertainty and other sources of uncertainty. Moreover, they do examine the interaction of data uncertainty with explanations. The latter is tackled by an explorative paper by Bertossi & Geerts (2020), which, however, is not a comprehensive survey. Thus, the guidance required for developing uncertainty-aware explanations is not provided.

Our literature review on data uncertainty in XAI further revealed that the existing research on this topic often analyzes specific, relatively isolated problem settings regarding both general context as well as characteristics of data uncertainty. For example, one of the identified papers discussed the explainable detection of poisoning attacks in vehicular networks (Bataineh et al., 2024) while another paper suggested an XAI method for diabetes prediction under data uncertainty (B & N, 2024). This highlights in particular that data uncertainty affects XAI across a variety of different domains. Regarding characteristics of data uncertainty, defects of different data quality dimensions were responsible for the data uncertainty, with accuracy (e.g., Ghorbani et al., 2019) and completeness (e.g., Dey & Lee, 2022) being especially prevalent. Mostly, feature values were affected by data uncertainty, but some papers also discussed the uncertainty of e.g. labels (Bataineh et al., 2024) Similarly, data uncertainty affects both training data (e.g., Yik et al., 2022) and test data (e.g., Ghorbani et al., 2019), which also leads to different requirements for uncertainty-aware explanations. Despite the differences, a common theme amongst the identified papers was that data uncertainty was typically considered in isolation.

4.2 ML Model Uncertainty

ML model uncertainty arises from the model itself and how it is trained (Tagasovska & Lopez-Paz, 2019). Models remain inherently uncertain approximations of the true functional relationship between

input and output, even in case data uncertainty is remedied (Kläs & Vollmer, 2018). ML model uncertainty relates to a models' fitting ability, optimization strategy, parameterization, or lack of knowledge and has been viewed in different research settings (Zhou et al., 2022). However, little effort has been made to assess ML model uncertainty with respect to explaining it effectively to human users (Gawlikowski et al., 2023). An overview paper notes that this interaction between ML model uncertainty and the human factor is limited to belief theory, which is one way to capture ML model uncertainty, but this neglects the effect on XAI and human users (Guo et al., 2024).

ML model uncertainty is driven by various model assumptions with respect to model parameters, model capacity, and type of model parameterization, e.g., thresholds in decision trees versus weights in neural networks (Marzari et al., 2024). These sources differ not only in their origin but also in their consequences: while some can be mitigated through better model design, others persist regardless of data quality or training effort. A fundamental reason for ML model uncertainty is model misspecification, where the chosen architecture or its assumptions are inadequate to capture the ground truth function, which cannot be resolved with better data but requires changing the model class. In healthcare, for example, applying a model class ill-suited to clinical data structures introduces model uncertainty that persists regardless of data quality. Model bias occurs when the chosen model class does not suit the complexity of the problem or makes incorrect assumptions. In contrast, model opacity or complexity describes uncertainty resulting from the intrinsic complexity of models like deep convolutional neural networks (Roelofs et al., 2022), where high predictive performance and interpretability are often in tension. For example, models tend to interpolate well in densely populated regions of the feature space, whereas regions characterized by sparse or out-of-range feature values typically yield high uncertainty (Kim et al., 2023). Distribution shifts include discrepancies between training and test data distributions, covariate, temporal, or label shifts (Ovadia et al., 2019). For instance, the training procedure, configuration, or hyperparameters can be insufficiently chosen (Tizpaz-Niari et al., 2022). Computational constraints introduce uncertainty when limitations in computing power or time necessitate simpler methods compared to more complex approaches like Bayesian inference (Lakshminarayanan et al., 2017). Yet the interaction with other uncertainty sources is underexplored, e.g., if a model's prediction is unreliable in an extrapolated region, feature attributions in the XAI method will be unfaithful to the underlying phenomenon, even if they accurately reflect the model's internal logic (Adlam & Pennington, 2020).

Various methods exist to quantify ML model uncertainty, which the literature broadly categorizes as intrinsic methods that inherently provide uncertainty estimates, such as Bayesian inference or ensemble approaches, post-hoc methods applied after training, including calibration techniques, bootstrapping, and specialized methods targeting specific model types or uncertainty sources (Ovadia et al., 2019). Conformal prediction offers a frequentist approach with statistically valid prediction intervals and guaranteed coverage rates (Marx et al., 2023). Finally, regarding the choice of representation, uncertainty may be expressed using nominal categories (e.g., certain vs. uncertain), ordinal scales (e.g., low, medium, high), or interval-scaled measures (e.g., variance or confidence intervals), which allow for more nuanced and quantitative uncertainty communication in XAI contexts (Peixoto & Azim, 2024).

4.3 XAI Method Uncertainty

Uncertainty in explanations may arise not only from the underlying data or the ML model, but also from the XAI methods themselves. This uncertainty stems from imperfections, abstractions, and simplifications inherent in XAI methods (Marx et al., 2023; Zhang et al., 2019). The literature distinguishes between different sources of uncertainty introduced by XAI methods. A first stream of work highlights uncertainty in post-hoc methods such as LIME or SHAP (Slack et al., 2021; Zhang et al., 2019). These approaches approximate black-box models through perturbations or surrogate models, but their explanations depend strongly on sampling strategies, parameter settings, and locality assumptions. Thus, multiple instantiations of the same method may yield different explanations for the same prediction (Zhang et al., 2019). In a clinical decision support system, for instance, this stochasticity may cause LIME or SHAP to assign inconsistent importance scores to the same patient features across runs. A second line of research examines intrinsically interpretable models such as linear models,

shallow decision trees, or rule-based systems. Although transparent by design, such models can still produce divergent but equally plausible explanations depending on training data, initialization, or hyperparameter choices (Bykov et al., 2020). Finally, uncertainty can also emerge from aggregation and presentation (Łajewska et al., 2024). For example, ensembles like random forests or model simplifications for human consumption merge multiple reasoning paths into one explanation, creating interpretative ambiguity even when the underlying model is itself interpretable. To communicate uncertainties arising from these sources, the literature distinguishes between different scopes of XAI method uncertainty: Some studies focus on global explanations, such as feature importance rankings (Smith et al., 2022), while others address local explanations for individual predictions (Heinrich et al., 2024; Kim et al., 2023). Local uncertainty communication is especially relevant when users need to assess the reliability of explanations in high-stakes decision contexts. However, most widely used methods only provide point estimates, failing to represent variability or instability, thereby risking unwarranted user confidence (Nauta et al., 2023; Zhang et al., 2019).

In the XAI literature, several concepts have emerged that approximate XAI method uncertainty (see Mohseni et al., 2021; Nauta et al., 2023 for overviews). The correctness of XAI methods has been explored (Nauta et al., 2023), where faithfulness describes how well explanations reflect the model's internal logic (Qiu et al., 2022), and fidelity encompasses the alignment between explanations and model outputs (Mohseni et al., 2021). High fidelity does not necessarily imply faithfulness; simple surrogate models may reproduce ML model outputs without capturing the model's true reasoning (Qiu et al., 2022). Reliability captures how stable XAI methods are to small perturbations and consistency across identical inputs (Löfström et al., 2024). Unstable explanations undermine trust, especially when they remain persuasive despite being unreliable (Mehdiyev et al., 2023; Nauta et al., 2023). Finally, completeness captures the extent to which an explanation conveys the model's underlying reasoning. While incomplete explanations simplify interpretation by highlighting only top-ranked features, they risk omitting relevant details and amplifying uncertainty (Nauta et al., 2023). Emerging research seeks to address the challenges of uncertainty in XAI methods. First, scholars develop approaches to reduce uncertainty in specific classes of XAI methods (Kim et al., 2023; Marx et al., 2023; Schwab & Karlen, 2019). Modifications and extensions of traditional XAI methods have also been proposed, such as incorporating Bayesian contexts to LIME and SHAP to account for uncertainty in model predictions (Hung & Lee, 2024; Slack et al., 2021). Second, researchers work on measuring and disclosing uncertainty through metrics such as fidelity scores or deletion tests (Bhatt et al., 2021; Nauta et al., 2023). Yet these approaches remain partial, often ignoring interactions with other uncertainty sources such as data or ML model uncertainty (cf. Figure 1).

4.4 Human Uncertainty

Human uncertainty can be understood along two complementary dimensions: first, humans can be inherently uncertain in their decision-making stemming from cognitive limitations, judgment processes, or incomplete domain knowledge; second, collaborating with AI or XAI on a specific task can externally introduce human uncertainty. This relates to uncertainty about the task, the internal mechanisms of the system, or one's own knowledge and ability to make judgments (Fügener et al., 2021; Pinski et al., 2023; Tversky & Kahneman, 1974). Both dimensions of human uncertainty can affect humans' reliance on provided AI support and, in turn, how well the AI system is used for a specific decision. However, most existing work implicitly assumes that humans act as oracles, overlooking uncertainty in their feedback and corrections (Collins et al., 2023).

Regarding the first dimension, humans face different types of inherent uncertainty when making decisions, shaped by individual heterogeneities in information processing, risk preferences, and prior beliefs. While prior knowledge and expertise generally reduce uncertainty, limited understanding or high accountability can increase it. In healthcare, for example, a clinician's level of prior training directly shapes how much uncertainty they experience when interpreting XAI outputs for unfamiliar diagnoses. Such uncertainty may lead to inappropriate reliance on AI, as users might either accept AI suggestions too readily or reject them without sufficient justification (Cau et al., 2023; He et al., 2023). It can also foster overconfidence in one's own decisions (Taudien et al., 2022b) and degrade decision quality,

particularly when probability information is inconsistently presented (Rastogi et al., 2022). Although typical model explanations alone may not substantially mitigate AI overreliance (Vasconcelos et al., 2023), communicating AI uncertainty can help users calibrate their confidence and form a more accurate understanding of system capabilities (He et al., 2023; Taudien et al., 2022a). Finally, individual competencies and risk attitudes shape how users handle AI-related uncertainty, influencing algorithmic appreciation or aversion (Schecter et al., 2023; Sindermann et al., 2020). Overall, moderate uncertainty could encourage reflection and calibrated reliance, whereas excessive or misplaced uncertainty can lead to dismissing accurate recommendations or following incorrect ones. Conversely, underestimated uncertainty may result in unwarranted trust precisely where human oversight is most essential in XAI contexts. Human uncertainty can manifest in various forms, which differ in their origin and reducibility. Epistemic uncertainty results from a lack of knowledge or insufficient information and is therefore reducible (Hüllermeier & Waegeman, 2021). Calibration gaps occur when subjective confidence does not align with actual accuracy (Tomsett et al., 2020). Perceived decision risk, trust-related uncertainty (Ma et al., 2024), and normative uncertainty reflect anticipated consequences or concerns regarding outcomes, system reliability, and ethical standards. Interpretative and causal attribution uncertainty concern difficulties in understanding the meaning of AI outputs and their underlying causes (Chen et al., 2025). Behaviorally, it manifests as information-seeking, the delay of decisions, fluctuating confidence, or repeated explanations requests. Cognitive overload can occur when users feel overwhelmed by complex information (Rezaeian et al., 2025), while misplaced familiarity describes treating uncertain problems as if they were well understood.

Regarding the second dimension, interaction with AI systems can constitute an external source of human uncertainty, depending on contextual, task-related, and individual factors. At the individual level, traits such as risk tolerance, trust propensity, metacognitive ability, and the need for cognitive closure shape how people respond to uncertain situations. Moreover, human uncertainty can evolve over time as people gain experience with AI systems. Consistent patterns such as algorithmic appreciation or aversion may develop (Jussupow et al., 2024; Rosenbacke, 2024), while trust drift and repair describe how trust may decline after negative experiences and recover following positive ones (Kahr et al., 2024). Understanding the triggers, forms, and influencing factors of human uncertainty in XAI-augmented decision-making is central to designing explanations that foster calibrated trust, appropriate reliance, and effective human-XAI collaboration.

5 Discussion

Across the reviewed studies, uncertainty emerges as a phenomenon with both technical and socio-cognitive dimensions, influencing not only system outputs but also how humans interpret and act upon them. Despite increasing attention to this topic, the current research landscape remains fragmented with respect to the different sources of uncertainty and their interdependencies. Our structured review reveals that the four sources of uncertainty have largely been analyzed in isolation, with very little cross-referencing (cf. Figure 1), limiting the understanding of how uncertainties interact and potentially strengthen across the decision-making process. While advances in XAI provide transparency, most techniques fail to communicate the uncertainty inherent in AI systems, thereby risking overconfidence or distrust among human users. Moreover, there is little empirical evidence on how uncertainty, whether communicated or not, affects user trust, reliance, and decision behavior. To help address these gaps, we propose an integrated conceptual framework for uncertainty-aware explanations (see Figure 2). This framework integrates the four sources of uncertainty into a unified view of the XAI-augmented decision-making process, rather than treating them as independent sources. The framework positions uncertainty-aware explanations as the central construct that synthesizes and communicates all four sources of uncertainty to the human decision-maker. It further embeds the XAI-augmented decision-making process within two overarching contextual layers: *organizations*, which shape the goals, constraints, and accountability structures under which (X)AI systems operate, and *society*, which influences this process through regulation and governance (e.g., such as the EU AI Act), and is, in turn, affected by decisions made with (X)AI support. Building on this framework, we propose a research agenda that consolidates research questions across uncertainty sources (see Figure 3). The research agenda is mapped onto the

framework, thereby advancing an uncertainty-aware paradigm in XAI-augmented decision-making (see Figure 2), and encompasses the entire lifecycle of modeling, designing, and investigating uncertainty (see Figure 3 and the corresponding research directions referenced by the numbers in Figure 2).

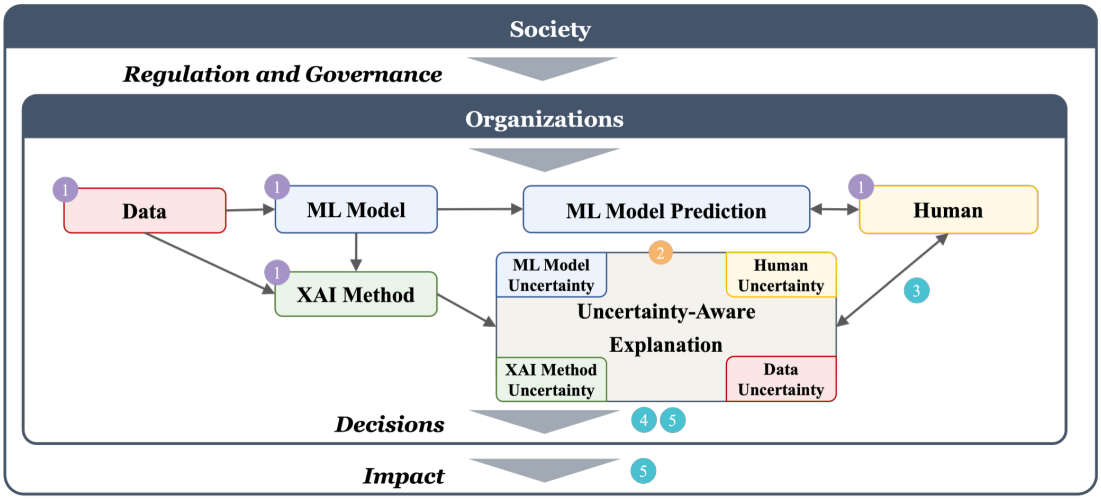

Figure 2. Envisioned state: An integrated view of uncertainty-aware explanations.

Foundational work is needed to model the interdependencies between data, ML model, XAI method, and human uncertainty (Research Direction (RD) 1) before integrated, uncertainty-aware explanations can be effectively designed (RD 2). Further efforts are required to investigate how such explanations shape user cognition and behavior (RD 3) and to evaluate their value across different stakeholders and contexts, particularly in relation to decision-making (RD 4). Finally, the iterative integration of modeling, design, and behavioral evaluation across distinct application settings is needed to develop generalizable principles for uncertainty-aware XAI and its societal impact (RD 5). Within the presented lifecycle (Figure 3), we conceptualize these RDs as sequentially layered, while acknowledging their mutual interdependence (e.g., RD2 "Designing" before RD3 "Investigating"). The overall research agenda is illustrated in Figure 3, and Table 2 provides a summary of example research.

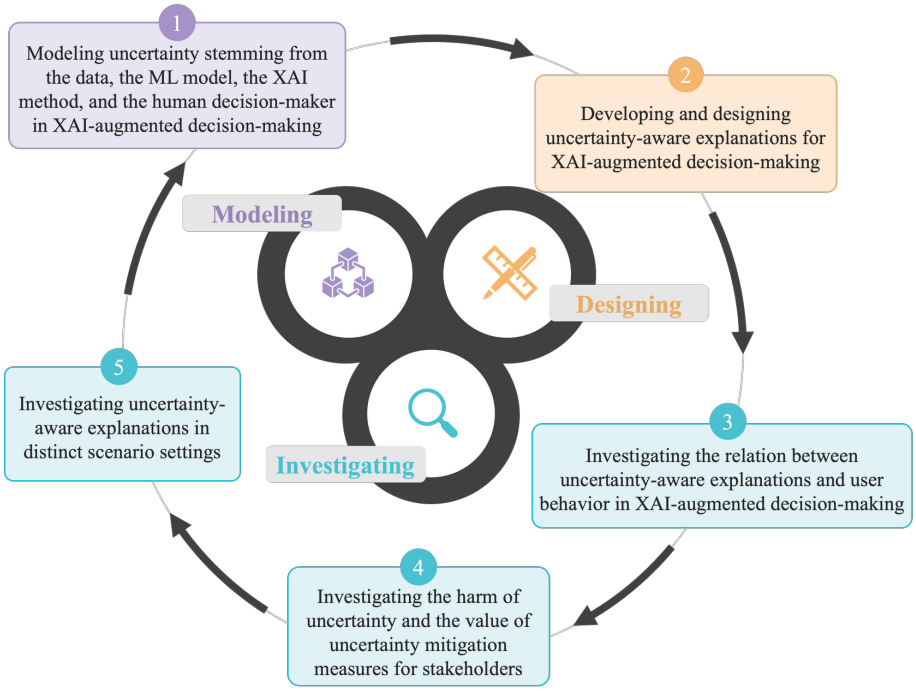

Figure 3. A research agenda for modeling, designing, and investigating uncertainty-aware explanations in AI-augmented decision-making.

RD 1 - Modeling uncertainty stemming from the data, the ML model, the XAI method, and the human decision-maker in XAI-augmented decision-making: Assessing, quantifying, and handling the sources of uncertainty while accounting for their interdependencies is essential for developing meaningful uncertainty-aware explanations. Future research in this direction should build on established guidelines and methods for assessing uncertainty (Batini & Scannapieco, 2016) and for managing it during model training and propagation (Tsang et al., 2011), while also exploring innovative approaches, such as incorporating ground truth instances to illustrate how uncertainty affects predictions and explanations. On this basis, approaches for quantifying XAI method uncertainty (Nauta et al., 2023) and human-related uncertainty can be integrated. This may enable the development of sensitivity measures (Heinrich et al., 2024) to identify robust predictions and decisions, providing a statistically grounded basis for decision support. Moreover, analyzing how feature changes affect both stable and unstable predictions could support the creation of uncertainty-aware feature attributions.

RD2 - Developing and designing uncertainty-aware explanations for XAI-augmented decision-making: This RD aims to develop approaches that visualize and integrate the uncertainties modeled in RD1 into meaningful, user-centric explanations. To this end, existing XAI methods can be adapted, and new ones developed, to represent multiple uncertainty sources and their interdependencies (see, e.g., Schwalbe & Finzel, 2023). A key challenge concerns how to design explanations that effectively foster user understanding, given that explanations do not always achieve their intended cognitive or behavioral outcomes (Hamm et al., 2023). This requires considering differences in prior knowledge, cognitive abilities, and preferences to address distinct user types and their uncertainty-related traits (Schneider & Handali, 2019). Furthermore, contextual factors such as task complexity and decision risk should be incorporated (see Section 4.4). Open questions remain regarding which explanation scope (local or global), type (e.g., feature importance, counterfactual), and presentation format (e.g., textual, visual) are most suitable for communicating uncertainty to different users (see Section 4.3). Furthermore, initial efforts toward personalization (Schröppel & Förster, 2024) and interactive explanations that allow users to explore and request uncertainty information (Bertrand et al., 2023) require further evaluation and refinement.

RD3 - Investigating the relation between uncertainty-aware explanations and user behavior in XAI-augmented decision-making: Future studies in this RD should focus on how the effects of uncertainty-aware explanations depend on user characteristics and perceptions, including trust in technology (Mcknight et al., 2011), AI literacy (Pinski & Benlian, 2024), and state, effect, and response uncertainty (Ashill & Jobber, 2010). Building on early work such as Walter (2024), which investigates how different uncertainty sources and XAI techniques affect trust and reliance, research should further explore how uncertainty-aware explanations shape user experience, trust development, behavior, and task performance across diverse decision-making contexts. A second key focus lies in understanding how users interact with and adapt to uncertainty-aware explanations during decision-making. Future studies should investigate how explanations can be tailored to varying levels of human uncertainty, building on initial studies such as Cau et al. (2023). Experimental approaches could, for instance, compare the effects of uncertainty-aware explanations on decision accuracy, trust, and reliance across expertise levels. Finally, as most prior work has examined only short-term effects (e.g., Wang et al., 2021), future research should also explore the long-term influence of uncertainty-aware explanations on user adaptation and collaboration with AI systems.

RD4 - Investigating the harm of uncertainty and the value of uncertainty mitigation measures for stakeholders: This RD focuses on understanding the consequences of non-transparent and uncommunicated uncertainty in XAI-augmented decision-making and on exploring how uncertainty-aware explanations can mitigate these risks. Unaddressed uncertainty may lead to overreliance on unreliable AI predictions and thus to suboptimal or harmful decisions, particularly in high-stakes domains such as healthcare. Building on prior research, future work should quantify these impacts and develop strategies to reduce them through uncertainty-aware explanations. Moreover, the value and implications of such explanations should be examined for different stakeholders, including direct users and affected third parties, such as physicians (Neri et al., 2020) and patients or insurance companies (Park et al., 2021) in medical applications, as emphasized in prior discussions on stakeholder

perspectives (Deshpande & Sharp, 2022). Finally, with the implementation of regulations such as the EU AI Act (European Commission, 2024), which mandate risk assessment and transparency, future research should investigate how legal and ethical frameworks can promote uncertainty-awareness and foster safer XAI-augmented decision-making practices.

RD5 - Investigating uncertainty-aware explanations in distinct scenario settings: This RD addresses the generalizability of findings on uncertainty-aware explanations, recognizing that their effectiveness is highly context-dependent. While prior studies have identified promising strategies in specific settings (Bertrand et al., 2023), these approaches may not transfer seamlessly across different domains or risk levels. For example, in high-risk contexts such as medical diagnosis, detailed uncertainty-aware explanations can support appropriate trust calibration and verification of AI outputs (Taudien et al., 2022a), whereas in lower-risk settings, simpler explanations may be sufficient. Future research should examine how contextual factors such as task risk, complexity, and decision type influence the usefulness, design, and reception of uncertainty-aware explanations. Decision tasks differ in nature and cognitive demand (Salimzadeh et al., 2024), ranging from preference-based to inference-based and from routine to non-routine, where complex or non-routine decisions may require more detail to support human understanding and accountability (Taudien et al., 2022b). Key considerations include how risk perception and task complexity influence explanation requirements, how stakeholder needs differ across domains, and how findings can be scaled and generalized beyond specific contexts to serve society as a whole. By addressing these aspects, future work can refine the contextual design of uncertainty-aware explanations, ensuring their effectiveness across diverse XAI-augmented decision-making environments.

Research Direction	**Example Research Questions**
RD 1: Modeling uncertainty stemming from the data, the ML model, the XAI method, and the human decision-maker in XAI-augmented decision-making	• How can uncertainties from data, models, explanations, and human decision-makers be modeled and integrated to capture their interdependencies and propagation throughout the XAI-augmented decision process? • How can ground truth instances be considered for quantifying and explaining uncertainty? • How can the impact of (combinations of) uncertainties on (X)AI models and their performance be quantified and analyzed (e.g., through upper bounds)?
RD 2: Developing and designing uncertainty-aware explanations for XAI-augmented decision-making	• How can uncertainty-aware explanations be generated that visualize and integrate different sources of uncertainty and their interdependencies? • How can uncertainty-aware explanations be designed and personalized in a user-centric way that accounts for different user types and contextual factors? • How can interactivity of explanations contribute to the user's understanding of uncertainty in XAI-augmented decision-making?
RD 3: Investigating the relation between uncertainty-aware explanations and user behavior in XAI-augmented decision-making	• How do uncertainty-aware explanations influence user experience, user behavior, trust, and task performance in XAI-augmented decision-making? • How do uncertainty-aware explanations relate to user states (e.g., positive and negative affect), traits (e.g., AI literacy, need for control), and how users manage their own uncertainties? • How does the display of ML model uncertainty affect user engagement with explanations, especially in complex situations where situational uncertainty is high, and does it lead to a U-shaped relationship between uncertainty and adaptive reliance on the AI system?
RD 4: Investigating the harm of uncertainty and the value of uncertainty mitigation measures for stakeholders	• How does uncommunicated uncertainty influence user behavior and decision outcomes in XAI-augmented decision-making, and how can uncertainty-aware explanations mitigate resulting risks such as under- or overreliance? • What are potential risks of incorrect but socially persuasive explanations in terms of the outcome of XAI-augmented decision-making?

	• What is the value of uncertainty-aware explanations for third parties (e.g., credit scoring)? • What regulatory obligations and governance mechanisms are suitable to improve uncertainty-awareness in critical decision-making domains where XAI-augmentation is implemented?
RD 5: Investigating uncertainty-aware explanations in distinct scenario settings	• How does uncertainty stemming from the human decision-maker and the impact of uncertainty-aware explanations on human-AI-interaction vary depending on the risk associated with a decision? • How does the impact of uncertainty-aware explanations on human-AI-interaction vary depending on the type of decision task? (e.g., preference vs. inference task, routine vs. non-routine task)

Table 2. Example research questions across the proposed research directions.

6 Conclusion

Increasing uncertainty-awareness in XAI-augmented decision-making holds substantial potential for advancing both research and practical applications. A systematic and integrated examination of uncertainty across data, ML models, XAI methods, and the human decision-maker, as well as their interactions, can help identify and mitigate unsafe or unreliable system behavior more effectively than approaches that treat each source in isolation. Uncertainty originating in the data may propagate to model outputs, affect the faithfulness of XAI explanations, and ultimately influence the epistemic state of the human decision-maker. However, these propagation pathways remain theoretically underspecified, and their practical consequences for real-world applications are rarely considered, as the existing literature largely addresses the four sources of uncertainty in isolation. Building on this observation, our integrated conceptual framework addresses this research gap by treating uncertainty as an interdependent phenomenon, rather than as a property of any single component.

For practitioners, this implies the need to implement concrete mitigation measures grounded in an integrated understanding of how different sources of uncertainties interact throughout the XAI-augmented decision-making process. This includes communicating ML model uncertainty, making data uncertainty and its propagation into explanations explicit, and designing uncertainty-aware explanation interfaces that support users in interpreting XAI method uncertainty and calibrating their own human uncertainty during decision-making. Such explicit consideration of different sources of uncertainty aligns with the growing demand for trustworthy and safe AI, as reflected in regulations for high-stakes domains such as the EU AI Act.

Beyond these practical implications, our proposed research agenda is directly derived from the framework and redirects scholarly attention to interdependencies that are currently overlooked in the literature. In doing so, we offer a theoretical reconceptualization of uncertainty as an interdependent, cross-layer phenomenon – one that can benefit organizations and society as a whole. The research agenda further serves as a call to action for the IS community, providing a first step towards advancing uncertainty-aware explanations to promote trustworthy and safe AI. The topic of uncertainty-aware XAI is particularly relevant to the IS community, which is uniquely positioned to explore the complex socio-technical intersections between AI systems and human users, and therefore the implications arising from different sources of uncertainty at this interface. By leveraging the field's diverse set of methodologies, including theory building, design science, ML modeling, as well as quantitative and qualitative empirical approaches, IS research can contribute to a holistic understanding of uncertainty-aware explanations and their implementation in practice.

References

Adlam, B., & Pennington, J. (2020). Understanding double descent requires a fine-grained bias-variance decomposition. *Advances in Neural Information Processing Systems*, *33*, 11022–32.

Antorán, J., Bhatt, U., Adel, T., Weller, A., & Hernández-Lobato, J. M. (2020). *Getting a CLUE: A Method for Explaining Uncertainty Estimates*.

Ashill, N. J., & Jobber, D. (2010). Measuring State, Effect, and Response Uncertainty: Theoretical Construct Development and Empirical Validation. *Journal of Management*, *36*(5), 1278–1308.

B, A., & N, K. (2024). Leveraging Machine Learning Models for Trustworthy Prediction of Diabetes. *Companion Proceedings of the ACM Web Conference 2024*, 1872–1875.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Bataineh, A. S., Zulkernine, M., Abusitta, A., & Halabi, T. (2024). Detecting Poisoning Attacks in Collaborative IDSs of Vehicular Networks Using XAI and Shapley Value. *Journal on Autonomous Transportation Systems*, *2*(3), 1–21.

Batini, C., & Scannapieco, M. (2016). *Data and Information Quality*. Springer International Publishing.

Bertossi, L., & Geerts, F. (2020). Data Quality and Explainable AI. *Journal of Data and Information Quality*, *12*(2), 1–9.

Bertrand, A., Viard, T., Belloum, R., Eagan, J. R., & Maxwell, W. (2023). On selective, mutable and dialogic xai: A review of what users say about different types of interactive explanations. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., & others. (2021). Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 401–413.

Bobek, S., & Nalepa, G. J. (2021). Introducing Uncertainty into Explainable AI Methods. *International Conference on Computational Science*, 444–457.

Brasse, J., Broder, H. R., Förster, M., Klier, M., & Sigler, I. (2023). Explainable artificial intelligence in information systems: A review of the status quo and future research directions. *Electronic Markets*, *33*(26).

Brocke, J. vom, Simons, A., Niehaves, B., Niehaves, B., Riemer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the giant: On the importance of rigour in documenting the literature search process. *Proceedings of the 17th European Conference on Information Systems.*

Bykov, K., Höhne, M. M.-C., Müller, K.-R., Nakajima, S., & Kloft, M. (2020). *How Much Can I Trust You?–Quantifying Uncertainties in Explaining Neural Networks*. Working Paper.

Cau, F. M., Hauptmann, H., Spano, L. D., & Tintarev, N. (2023). Effects of AI and Logic-Style Explanations on Users' Decisions Under Different Levels of Uncertainty. *ACM Transactions on Interactive Intelligent Systems*, *13*(4), 22:1-22:42.

Chen, R., Wang, R., Fang, F., & Sadeh, N. (2025). Missing Pieces: How Do Designs that Expose Uncertainty Longitudinally Impact Trust in AI Decision Aids? An In Situ Study of Gig Drivers. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 790–816.

Collins, K. M., Barker, M., Espinosa Zarlenga, M., Raman, N., Bhatt, U., Jamnik, M., Sucholutsky, I., Weller, A., & Dvijotham, K. (2023). Human uncertainty in concept-based ai systems. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 869–889.

Croskerry, P., & Norman, G. (2008). Overconfidence in clinical decision making. *The American Journal of Medicine*, *121*(5), S24–S29.

Deshpande, A., & Sharp, H. (2022). Responsible ai systems: Who are the stakeholders? *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 227–236.

Dey, S., & Lee, S.-W. (2022). Are we training with the right data? Evaluating collective confidence in training data using Dempster Shafer theory. *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results*, 11–15.

European Commission. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU)*

2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research*, *33*(2), 678–696.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd International Conference on Machine Learning* (Vol. 48, pp. 1050–1059). PMLR.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., & Zhu, X. X. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, *56*(S1), 1513–1589.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745–e750.

Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 3681–3688.

Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, *162*, 107268.

Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, *10*(1), 1–20.

Guo, Z., Wan, Z., Zhang, Q., Zhao, X., Zhang, Q., Kaplan, L. M., Jøsang, A., Jeong, D. H., Chen, F., & Cho, J.-H. (2024). A survey on uncertainty reasoning and quantification in belief theory and its application to deep learning. *Information Fusion*, 101987.

Hagn, M., Heinrich, B., Krapf, T., & Schiller, A. (2025). Handling imperfection: A taxonomy for machine learning on data with data quality defects. *Decision Support Systems*, 114493.

Hamm, P., Klesel, M., Coberger, P., & Wittmann, H. F. (2023). Explanation matters: An experimental study on explainable AI. *Electronic Markets*, *33*(1).

He, G., Kuiper, L., & Gadiraju, U. (2023). Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18.

Heinrich, B., Klier, M., Obermeier, A., & Schiller, A. (2025). Different but the Same? An Event-driven Approach to determine Probabilities of Data Duplication. *MIS Quarterly*, *49*(4), 1539–1566. https://doi.org/10.25300/MISQ/2025/18178.

Heinrich, B., Krapf, T., & Miethaner, P. (2024). EXPLORE: A Novel Method for Local Explanations. *Proceedings of the International Conference on Information Systems*.

Holton, G. A. (2004). Defining risk. *Financial Analysts Journal*, *60*(6), 19-25.

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457–506.

Hung, Y.-H., & Lee, C.-Y. (2024). BMB-LIME: LIME with modeling local nonlinearity and uncertainty in explainability. *Knowledge-Based Systems*, *294*, 111732.

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. *Proceedings of the 28th European Conference on Information Systems.*

Jussupow, E., Benbasat, I., & Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. *MIS Quarterly*, *48*(4), 1575–1590.

Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence. *Information Systems Research*, *32*(3), 713–735.

Kahr, P. K., Rooks, G., Snijders, C., & Willemsen, M. C. (2024). The Trust Recovery Journey. The Effect of Timing of Errors on the Willingness to Follow AI Advice. *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 609–622.

Kim, B., Srinivasan, K., Kong, S. H., Kim, J. H., Shin, C. S., & Ram, S. (2023). ROLEX: A Novel Method for Interpretable Machine Learning Using Robust Local Explanations. *MIS Quarterly*, *47*(3), 1303–1332.

Kläs, M., & Vollmer, A. M. (2018). Uncertainty in Machine Learning Applications: A Practice-Driven Classification of Uncertainty. *Proceedings of Computer Safety, Reliability, and Security.* 431–438.

Łajewska, W., Spina, D., Trippas, J., & Balog, K. (2024). Explainability for transparent conversational information-seeking. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1040–1050.

Lakkaraju, H., & Bastani, O. (2020). "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 79–85.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6405–6416.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.

Löfström, H., Löfström, T., Johansson, U., & Sönströd, C. (2024). Calibrated explanations: With uncertainty information and counterfactuals. *Expert Systems with Applications*, *246*, 123154.

Ma, Z., Mei, Y., Gajos, K. Z., & Arawjo, I. (2024). Schrödinger's Update: User Perceptions of Uncertainties in Proprietary Large Language Model Updates. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–9.

Marx, C., Park, Y., Hasson, H., Wang, Y., Ermon, S., & Huan, L. (2023). *But Are You Sure? An Uncertainty-Aware Perspective on Explainable AI. 206*, 7375–7391.

Marzari, L., Leofante, F., Cicalese, F., & Farinelli, A. (2024, July). *Rigorous Probabilistic Guarantees for Robust Counterfactual Explanations*. Working Paper.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., & others. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89–94.

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, *2*(2), 1–25.

Mehdiyev, N., Majlatow, M., & Fettke, P. (2023). Explainable Artificial Intelligence Meets Uncertainty Quantification for Predictive Process Monitoring. *International Joint Conference on Artificial Intelligence*, 29–32.

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, *39*(1), 53–63.

Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2025). The effects of data quality on machine learning performance on tabular data. *Information Systems*, *132*, 102549.

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *11*(3–4), 1–45.

Morrison, K., Spitzer, P., Turri, V., Feng, M., Kühl, N., & Perer, A. (2024). The impact of imperfect XAI on human-AI decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–39.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, *55*(13s), 1–42.

Neri, E., Coppola, F., Miele, V., Bibbolino, C., & Grassi, R. (2020). Artificial intelligence: Who is responsible for the diagnosis? *La Radiologia Medica*, *125*(6), 517–521.

Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E., & Gogate, V. (2021). Anchoring bias affects mental model formation and user reliance in explainable AI systems. *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 340–350.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, *32*.

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: A survey of case studies. *ACM Computing Surveys*, *55*(6), 1–29.

Park, S. H., Choi, J., & Byeon, J. S. (2021). Key Principles of Clinical Validation, Device Approval, and Insurance Coverage Decisions of Artificial Intelligence. *Korean Journal of Radiology*, *22*(3), 442–453.

Peixoto, M. J., & Azim, A. (2024). Explainable Artificial Intelligence (XAI) Approach for Reinforcement Learning Systems. *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, 971–978.

Peng, J., Hahn, J., & Huang, K.-W. (2023). Handling missing values in information systems research: A review of methods and assumptions. *Information Systems Research*, *34*(1), 5–26.

Pinski, M., Adam, M., & Benlian, A. (2023). AI Knowledge: Improving AI Delegation through Human Enablement. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Pinski, M., & Benlian, A. (2024). AI literacy for users – A comprehensive review and future research directions of learning methods, components, and effects. *Computers in Human Behavior: Artificial Humans*, *2*(1), 100062.

Qiu, L., Yang, Y., Cao, C. C., Zheng, Y., Ngai, H., Hsiao, J., & Chen, L. (2022). Generating perturbation-based explanations with robustness to out-of-distribution data. *Proceedings of the ACM Web Conference 2022*, 3594–3605.

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., & Tomsett, R. (2022). Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–22.

Rezaeian, O., Bayrak, A. E., & Asan, O. (2025). Explainability and AI confidence in clinical decision support systems: Effects on trust, diagnostic performance, and cognitive load in breast cancer care. *International Journal of Human–Computer Interaction*, 1–21.

Roelofs, L., Weissman, C., Raicu, D., Furst, J., & Tchoua, R. (2022). Similarity-based uncertainty scores for interpretable deep learning in scientific applications. *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, 1152–1155.

Rosenbacke, R. (2024). Heuristics and errors in XAI-Augmented clinical decision-making: Moving beyond algorithmic appreciation and aversion. *Proceedings of the 2024 European Conference on Information Systems*.

Salimzadeh, S., He, G., & Gadiraju, U. (2024). Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision Making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Schecter, A., Bogert, E., & Lauharatanahirun, N. (2023). Algorithmic appreciation or aversion? The moderating effects of uncertainty on algorithmic decision making. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems.*

Schneider, J., & Handali, J. (2019). Personalized explanation in machine learning: A conceptualization. *Proceedings of the 2019 European Conference on Information Systems*.

Schnurr, D. (2025). European Data Regulation: Between Data Protection and Free Flow of Data in a Global Digital Economy. In M. Denga & L. Hornuf (Eds.), *Regulatory Competition in the Digital Economy: Artificial Intelligence, Data, and Platforms* (pp. 101–117). Springer, Cham.

Schröppel, P., & Förster, M. (2024). Exploring XAI users' needs: A novel approach to personalize explanations using contextual bandits. *Proceedings of the 2024 European Conference on Information Systems*.

Schwab, P., & Karlen, W. (2019). CXPlain: Causal Explanations for Model Interpretation under Uncertainty. *Advances in Neural Information Processing Systems*, *32*.

Schwalbe, G., & Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, *38*(5), 3043–3101.

Sindermann, C., Sha, P., Zhou, M., Wernicke, J., Schmitt, H. S., Li, M., Sariyska, R., Stavrou, M., Becker, B., & Montag, C. (2020). Assessing the Attitude Towards Artificial Intelligence: Introduction of a Short Measure in German, Chinese, and English Language. *KI - Künstliche Intelligenz*, *35*(1), 109–118.

Slack, D., Hilgard, A., Singh, S., & Lakkaraju, H. (2021). Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems*, *34*, 9391–9404.

Smith, M., Acquesta, E., Smutz, C., Moss, B., & Rushdi, A. (2022). Assessing the Fidelity of Explanations with Global Sensitivity Analysis. *Proceedings of the 2023 Hawaii International Conference on System Sciences*.

Sunyaev, A., Benlian, A., Pfeiffer, J., Jussupow, E., Thiebes, S., Maedche, A., & Gawlitza, J. (2025). High-Risk Artificial Intelligence. *Business & Information Systems Engineering*, 1–14.

Tagasovska, N., & Lopez-Paz, D. (2019). Single-model uncertainties for deep learning. *Advances in Neural Information Processing Systems*, *32*.

Taudien, A., Fuegener, A., Gupta, A., & Ketter, W. (2022a). Calibrating Users' Mental Models for Delegation to AI. *Proceedings of the 2022 International Conference on Information Systems*.

Taudien, A., Fuegener, A., Gupta, A., & Ketter, W. (2022b). The Effect of AI Advice on Human Confidence in Decision-Making. In *Proceedings of the 55th Annual Hawaii International Conference on System Sciences* (pp. 245–253).

Tizpaz-Niari, S., Kumar, A., Tan, G., & Trivedi, A. (2022). Fairness-aware configuration of machine learning libraries. *Proceedings of the 44th International Conference on Software Engineering*, 909–920.

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., & Kaplan, L. (2020). Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. *Patterns*, *1*(4), 100049.

Tsang, S., Kao, B., Yip, K. Y., Ho, W.-S., & Lee, S. D. (2011). Decision Trees for Uncertain Data. *IEEE Transactions on Knowledge and Data Engineering*, *23*(1), 64–78.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–38.

Walter, M. C. (2024). Follow Me, Everything Is Alright (or Not): The Impact of Explanations on Appropriate Reliance on Artificial Intelligence. In *Proceedings of the 57th Hawaii International Conference on System Sciences* (pp. 1287–1296).

Wang, D., Zhang, W., & Lim, B. Y. (2021). Show or suppress? Managing input uncertainty in machine learning model explanations. *Artificial Intelligence*, *294*, 103456.

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2), xiii–xxiii. JSTOR.

Yik, W., Serafini, L., Lindsey, T., & Montañez, G. D. (2022). Identifying bias in data using two-distribution hypothesis tests. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 831–844.

Zhang, Y., Song, K., Sun, Y., Tan, S., & Udell, M. (2019, June). "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. *International Conference on Machine Learning, AI for Social Good Workshop*.

Zhou, X., Liu, H., Pourpanah, F., Zeng, T., & Wang, X. (2022). A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications. *Neurocomputing*, *489*, 449–465.