

Untangling Trust, Transparency, and Uncertainty in AI Literature and Policy

Kai Nylund, Shani Spivak, Dakuo Wang, Lace Padilla

¹Khoury College of Computer Sciences, Northeastern University
Boston, MA, USA
nylund.k@northeastern.edu

Abstract

Trust, transparency, and uncertainty-awareness are widely cited as prerequisites for responsible and accountable machine learning (ML), yet their meanings and relationships remain inconsistent across research and policy. We present the first systematic review to examine how these concepts are defined and operationalized together across ML research, human–computer interaction, applied domains, AI policy, and technical standards. Analyzing 118 research papers, eight surveys of AI policy documents, and multiple national and international standards, we synthesize discussions on each concept and their relationships. We find that trust, transparency, and uncertainty are rarely defined explicitly in literature, limiting the transfer of meaning and findings between studies. In policy and standards, definitions are frequently circular or conflicting, creating ambiguous compliance expectations. Untangling the relationships between each concept, we argue that trust, transparency, and uncertainty are *co-produced* in ML, and show how this framing can improve conceptual clarity. Based on our findings, we provide recommendations to improve how we conceptualize, study, and implement ML systems in regard to each concept.

Code and extended version —

<https://github.com/KaiNylund/untangling-TTU>

Introduction

As black-box machine learning (ML) models are increasingly used as high-stakes decision-making aids, *trust* and *transparency* have become central requirements for deployment in both AI research and policy (Huang et al. 2023; Jobin et al. 2019). Implicit in discussions of these two concepts is a third: *uncertainty*. Trust, for example, is often defined as belief in another party under uncertainty (Mayer et al. 1995; Lee and See 2004); and a common goal for increasing transparency in ML models is to reduce uncertainty about how they arrive at outputs. Transparency and uncertainty communication are sometimes framed as two sides of the same coin, showing what models “know” and “don’t know” (Tomsett et al. 2020; Salvi et al. 2025), with the aim of helping users trust or distrust models in accordance with their capabilities. The importance, interrelatedness, and slipperiness of the three

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

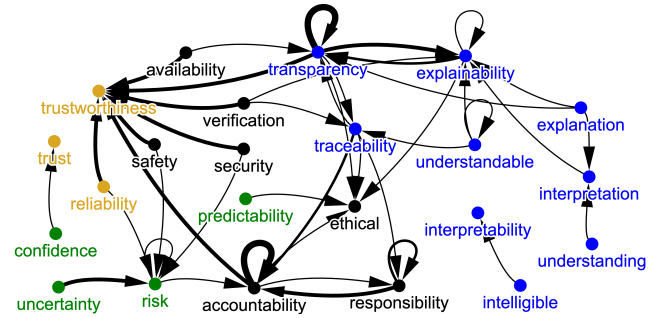


Figure 1: AI policy contains a complicated web of circular definitions for trust, transparency, and uncertainty. An arrow from *A* to *B* means *A* is used in *B*’s definition. The width of lines corresponds to count across policy documents (ranging from one to four). Notably, “transparency” and “explainability” are both used to define themselves, each other, and terms such as “trustworthiness” and “ethical”. Text color indicates terms related to **trust**, **uncertainty**, and **transparency**; black text indicates related terms.

concepts have created a complicated web of definitions that is hard to understand and put into practice. Figure 1 shows a dependency graph of definitions from a few of the most referenced AI technical standards, where we find that circular loops and broad language make following guidelines difficult and shirking regulation easy.

Policy and governance are shaped by ML research, where trust, transparency, and uncertainty are also inconsistently defined. In some works “transparency”, “interpretability”, and “explainability” are used interchangeably, and in others they have separate, even conflicting, meanings (Lipton 2018; Broniatowski 2021). In some studies, uncertainty quantification (UQ) is framed as separate from methods to interpret models, whereas in others it is proposed as the primary interpretability technique. Research papers on trust in ML models employ a wide array of measurement strategies that are not always comparable, and still other works define trust between humans and ML models as impossible. We document these recurring definitional problems and provide recommendations for improving conceptual clarity.

In addition to individual definitions, we find conflicting results about the *relationships* between concepts. In some

experiments, transparency is shown to improve trust, and in others it has the opposite effect. Uncertainty causes too much trust in models in some works, and aversion in others. We explore how these results interact with common assumptions about trust, transparency, and uncertainty, and describe other relationships that are currently understudied.

Although prior work has noted definitional ambiguity for individual concepts or examined pairwise connections (e.g., uncertainty-aware interpretability or trust calibration) (Lipton 2018; Jacovi et al. 2021; Brodlie et al. 2012; Boukhelifa et al. 2017), no review has systematically examined the full web of connections between trust, transparency, and uncertainty across ML research and policy. In this survey, we address this gap by: (1) examining how trust, transparency, and uncertainty are defined and used, with particular attention to works that discuss all three concepts together; (2) identifying recurring problems and inconsistencies in these definitions; and (3) providing recommendations for improving definitional clarity and coherence across research and policy contexts. To achieve these goals, we conduct a cross-cutting systematic review of 118 research papers, 8 surveys of AI policy documents, and 8 international AI standards documents that address trust, transparency, and uncertainty in ML. Using our findings, we simplify the complicated web of definitions from Figure 1 into a clear diagram mapping the interdependence of each concept (Figure 4). Ultimately, we argue that we can gain conceptual clarity by treating trust, transparency, and uncertainty as *co-produced*, rather than merely related.

Survey Methods

In this section, we describe our process for selecting and reviewing ML / AI papers and policy documents.

ML Paper Selection

To understand how trust, transparency, and uncertainty are defined and operationalized together across ML literature, we review papers in influential computer and natural sciences journals and conferences following the PRISMA guidelines (Page et al. 2021). We chose to include venues based on a combination of popularity (h5-index) and relevance to each of the three concepts in ML. In total, we search for papers in 18 journals and conferences across five domains: **Natural Sciences** (Nature, Cell Press), **ML** (ICLR, ICML, CVPR, NeurIPS, FaccT), **NLP** (ACL, EMNLP, NAACL), **HCI** (CHI, UIST, IUI, CSCW), and **Visualization** (VIS, EuroVis, PacificVis, InfoVis). To ensure that our review reflects current usage and definitions, we limit our search to papers in the last 10 years (1/1/2015 up to the date of our last search in June 2025). For each included source, we search for papers that mention trust, uncertainty, transparency, and ML each at least once. Our exact keywords and other details for this initial filtering step are documented in the Appendix.

Given the breadth of ML, trust, transparency, and uncertainty across fields, manually reviewing all relevant papers is infeasible. Rather than aiming for complete coverage, we curate a high-quality, representative sample of the literature. During our review, we take note of works such as Lee and See 2004 and Lipton 2018 that provide influential definitions or conceptual clarity and are highly cited, but do not

meet our criteria for inclusion. We discuss these works when applicable but do not include them in our coding.

To select relevant papers from this initial sampling, we manually review each abstract and sentence mentioning trust, transparency, and uncertainty or synonyms. We exclude works that do not discuss ML and all three concepts in at least one argument, contribution, and/or hypothesis. For example, we include papers that test different explanation methods under uncertainty to observe the effect on trust because the paper’s main contributions, hypothesis, and arguments would change significantly without each of the three concepts. We exclude works that propose uncertainty-aware interpretability methods without considering the potential impact on trust, and studies on trust under uncertainty in ML that do not discuss transparency methods. We also filter out papers that only reference a concept in passing for motivation such as “our method improves transparency, which is crucial for building trust”, or “uncertainty quantification is necessary for trust calibration”. Upon reviewing the full text of selected papers, we similarly exclude any that lack this criterion but were not caught in the first screening step. To address potential gaps caused by our narrow selection of journals and conferences, we also scan the citations of papers selected for full-text review and include relevant works. Based on these criteria, we expect that trust, transparency, and uncertainty are all defined in the papers selected for review.

ML Paper Coding

We seek to answer the following questions about trust, transparency, and uncertainty for each paper in our survey:

Q1 How is each concept defined and measured?

Q2 What types of each concept are discussed?

Q3 What are the goals of handling each concept?

To answer each question, for every paper and concept, we record corresponding measurement strategies, types, and goals, binned into the most commonly observed categories. We also divide definitions of each concept into one of three types: *explicit* (includes an explicit definition such as “We define trust as...”), *implicit* (measures the concept using, e.g., Likert-style questions or lists characteristics such as “marked by vulnerability”, but does not directly define), and *N/A* (does not measure, characterize, or define the concept).

We note that our assessment of these labels, particularly those that are not stated directly, is subjective. To measure the robustness of our results, the lead author read and coded each of the papers, and the second author coded half of the papers, randomly selected. The two coders discussed each instance where they differed, either coming to agreement and updating their labels to match or disagreeing. Our total agreement between the first and second author (using exact match only Cohen’s Kappa) is 0.9096. We discuss agreement in individual coding categories and implications for the robustness of our results in the Appendix. Throughout the paper we use the updated label counts from the first author.

ML Policy Document Selection and Review

To understand trust, transparency, and uncertainty definitions in technical standards, we systematically review all

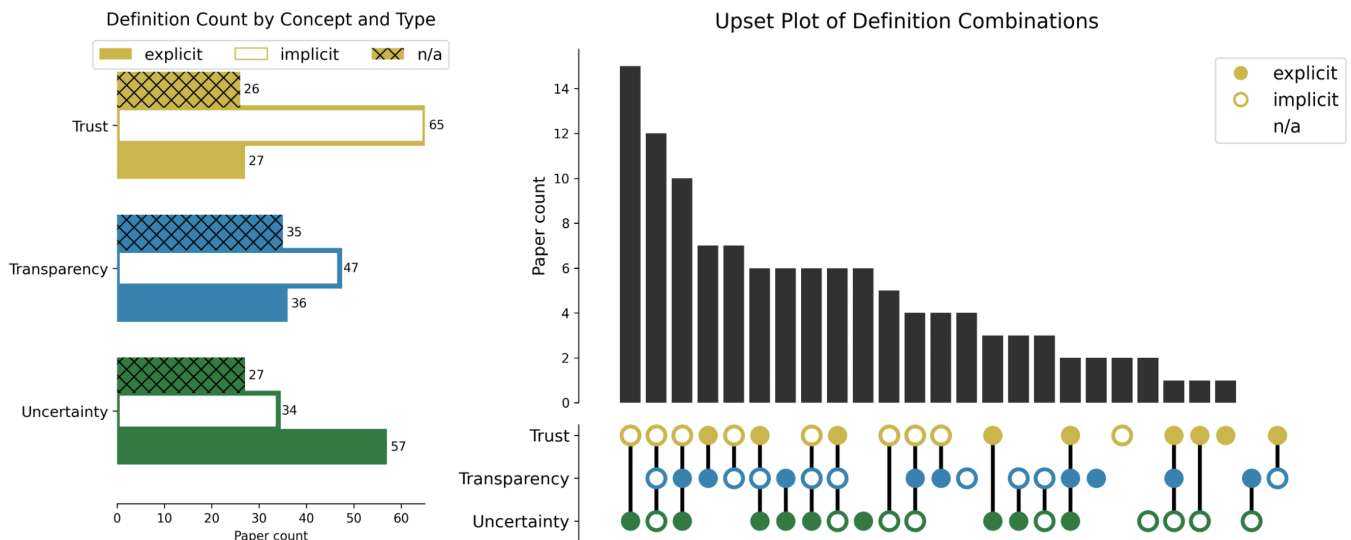


Figure 2: **Trust**, **transparency**, and **uncertainty** are mostly defined implicitly or not at all. Left: a bar chart showing the number of papers included in our review that provide explicit, implicit, and no definition for each concept. Right: an UpSet plot depicting the number of papers that define concepts using each permutation of explicit, implicit, and no definitions. We observe different definitional tendencies for each concept; uncertainty, for example, is defined explicitly the most and transparency the least. Only two papers in our survey provide explicit definitions for all three concepts.

relevant documents from the ISO (International Organization for Standardization)/IEC (International Electrotechnical Commission), which is the most widely recognized global standards-setting body. We also search for surveys of AI ethics principles and guidelines across policy and ML venues, and select works with both relatively high Google Scholar citation counts and applicability at state, national, and global levels. We excluded AI principles surveys that either did not include a list, count, or ranking of the principles themselves or were specific to a certain context, like the media (de Lima-Santos et al. 2025) or Generative AI (Laine et al. 2025), since these did not provide a broad sense of how trust, uncertainty, and transparency are used, prioritized, and defined. This list is not exhaustive and surveys of AI policies are often broad, but we believe their aggregation complements our review of ML literature without greatly increasing our scope. We also searched for guidance referencing trust, transparency, or uncertainty from authoritative sources such as the National Institute of Standards and Technology (NIST) and the European Commission’s high-level expert group on AI (AI HLEG). The second author reviewed the selected surveys, policy documents, guidelines, and standards for definitions related to trust, transparency, and uncertainty. See the Appendix for more on the motivation behind our focus on these global and regional principles, standards, policy, and law.

Selected Papers and Guideline Documents

Appendix Figure A.1 is a PRISMA flow chart (Page et al. 2021) showing the number of papers we screened and included at each step in our review. Our initial search of 18 journals and conferences returned 586 papers, most of which only mention trust and uncertainty in passing, often as motivation for a proposed method or system. Screening abstracts

and keyword sentences, we flag 105 papers for full text review and include 95. After coding each paper and screening relevant citations, we include an additional 28 papers. Finally, during discussion between the first and second author, five papers were excluded due to a lack of focus on ML or uncertainty. We provide the full list of 118 papers included in our review and their corresponding coding labels in Appendix Table A.5. We additionally select 8 surveys of global AI principles (Jobin et al. 2019; Hagendorff 2020; Huang et al. 2023; Whittlestone et al. 2019; Schiff et al. 2020; Corrêa et al. 2023; de Lima-Santos et al. 2025; Tidjon and Khomh 2023) and 8 ISO/IEC standards documents for review (full list in Appendix Table A.3).

Definitions of Trust, Transparency, and Uncertainty

Trust, transparency, and uncertainty are all notoriously difficult to define and measure (Brodie et al. 2012; Boukhelifa et al. 2017; Lipton 2018; Afroogh et al. 2024). However, as many of the papers, guidelines, and standards we survey claim, each concept is also a crucial consideration for the design and implementation of ML systems. This combination of ambiguity and importance has led to a variety of definitions, measurement techniques, types, and goals associated with each concept. Figure 3 quantifies these uses across the 118 papers we survey, showing variations in concept types and quantitative measurement, and similar goals and methods for eliciting user judgments. In policy and standards documents, concept definitions are not set or used consistently within or across government entities, and we also see the impact of circular definitions. Terms are used to define themselves or depend on related terms, which might themselves depend on

the original term, as illustrated in Figure 1. In this section, we examine the definitions and uses of each concept individually, collecting and clarifying their range of meanings.

Trust, Transparency, and Uncertainty Are Mostly Defined Implicitly or Not at All

During our review, a lack of definitions consistently obfuscated uses of trust, transparency, and uncertainty. Figure 2 shows that most papers we survey do not explicitly state what each concept means. Definition type varied across the concepts, with uncertainty having the highest count of explicit definitions, followed by transparency and trust. Of the three, trust was defined implicitly the most, and transparency was least likely to be defined. One explanation for this discrepancy is a bias toward easily quantifiable measures. Uncertainty was frequently defined mathematically in terms of model or data characteristics, while trust and transparency were more often defined through influential prior work in sociology and human-computer interaction.

Although ISO/IEC standards commonly provide explicit definitions, policy documents tend to use these terms ad hoc, implicitly, or not at all. This pattern is especially pronounced for trust and trustworthiness. For example, four core policy documents that discuss the importance of trustworthy ML: U.S. Executive Order 13960, China’s AI principles, the EU AI Act, and the EU AI HLEG Assessment List, define neither term (NIST 2024; Tabassi 2023; Atherton et al. 2023; Commission 2024; HLEG 2020). Similarly, surveys of global AI principles often only identify related values such as fairness and accountability rather than defining trust directly (Jobin et al. 2019; Corrêa et al. 2023; Khan et al. 2022; Huang et al. 2023; Tidjon and Khomh 2023; Ibáñez and Olmeda 2022). In contrast, transparency is commonly defined in policy, while uncertainty is rarely mentioned.

● Trust

Regardless of context, almost all definitions of trust in ML literature require (1) belief in a trustee to act in the trustor’s best interests, and (2) uncertainty (e.g., (Jacovi et al. 2021; Afroogh et al. 2024; Zerilli et al. 2022)). Beyond these shared themes, there is a wide variety of definitions for trust in our surveyed papers, stemming most often from ongoing debate about how past theories of human trust should apply to human-AI interaction. Some definitions are intentionally narrow and AI-specific, such as “the extent to which the human believes the AI’s outputs are correct/useful for achieving their current goals in the current situation” (Tomsett et al. 2020). Other works attempt to capture the subject’s complexity through broader, general definitions (Zerilli et al. 2022; Zhang et al. 2020; Yang et al. 2024); for instance, describing trust as “a psychological state comprising the intention to accept vulnerability based on positive expectations of the intentions or behavior of another” (Zerilli et al. 2022).

These differences in scope often correspond to whether the paper focuses on trust in humans or machines. Kostick-Quenet et al. 2022, for example, define human-human trust in sociological terms as “influenced by how other people behave in contexts of reciprocity and exchange...” but note

that human-machine trust is still not well understood. Several works adapt the definition from “Trust in Automation” (Lee and See 2004) to human-AI trust, which is “an attitude that an [AI] agent will achieve an individual’s goal in a situation characterized by uncertainty and vulnerability.”

Alongside “trust”, debates about the difference between human-human and human-machine trust affect the terms “reliance” and “trustworthiness”. For example, zuchowski et al. 2024 (citing Baier 1986 and McGeer et al. 2017) define reliance as “the belief in a disposition and the willingness to act on this belief” and trust as “the reliance on the trustor’s care for [an aspect of] the trustee’s wellbeing.” From these definitions, the authors claim that people can only rely on *and not trust* AI systems because machines are incapable of caring. Other works argue the opposite, that the study of human-AI trust is more suitable than human-AI reliance because of peoples’ tendency to anthropomorphize of models (Jacovi et al. 2021). Similarly, although some works view trustworthiness as an objective system quality (Zerilli et al. 2022) (capable of being changed, e.g., by increasing model performance or displaying uncertainty information), others point out that trustworthiness is human-centered and can be manipulated independently from performance (Ehsan and Riedl 2024; Banovic et al. 2023).

Policy definitions of trust are rarer, emphasizing transparency and stakeholder satisfaction rather than uncertainty. The only direct definition of trust in ISO/IEC AI standards comes from a Technical Report, ISO/IEC TR 24028:2020, which in turn relies on a definition of trust from a defunct 2011 standard (ISO/IEC 2011): “Trust: degree to which a user or other stakeholder has confidence that a product or system will behave as intended.” Trustworthiness is also defined in ISO/IEC TR 24028:2020 as “the ability to meet stakeholders’ expectations in a verifiable way.” These conflict with other vaguer definitions of trustworthiness in policy such as “reliance can be justifiably placed on the service” (The IFIP Working Group 10.4 via IEEE and ISO/IEC). Confusingly, the Department of Defense (DoD) also claims that “There is currently no method to measure user trust in AI or measure what factors influence the users’ trust decisions”, which is disproved by many of the papers we survey.

Despite these conflicting definitions, we find that papers in our survey are unified in their focus on measuring human-AI trust. In our coding scheme, we categorize papers by whether they discuss trust, reliance, and/or trustworthiness, and by subject (such as users → ML models). For simplicity, we distinguish between trust and reliance, but not other related dichotomies like cognitive and behavioral trust, as they are often measured in the same ways (e.g., agreement and self-reported trust beliefs respectively). We find that the majority of papers in our survey focus on the first category: users’ trust in ML models (87), followed by the trustworthiness / reliability of (40), and users’ reliance on models (33).

Goals for handling trust, reliance, and trustworthiness are similarly model-focused with the most common being improving trust or reliance calibration (38), improving decision-making and/or human-AI team performance (32), and increasing adoption of technologies (31). Papers frequently had multiple connected goals associated with trust; for example,

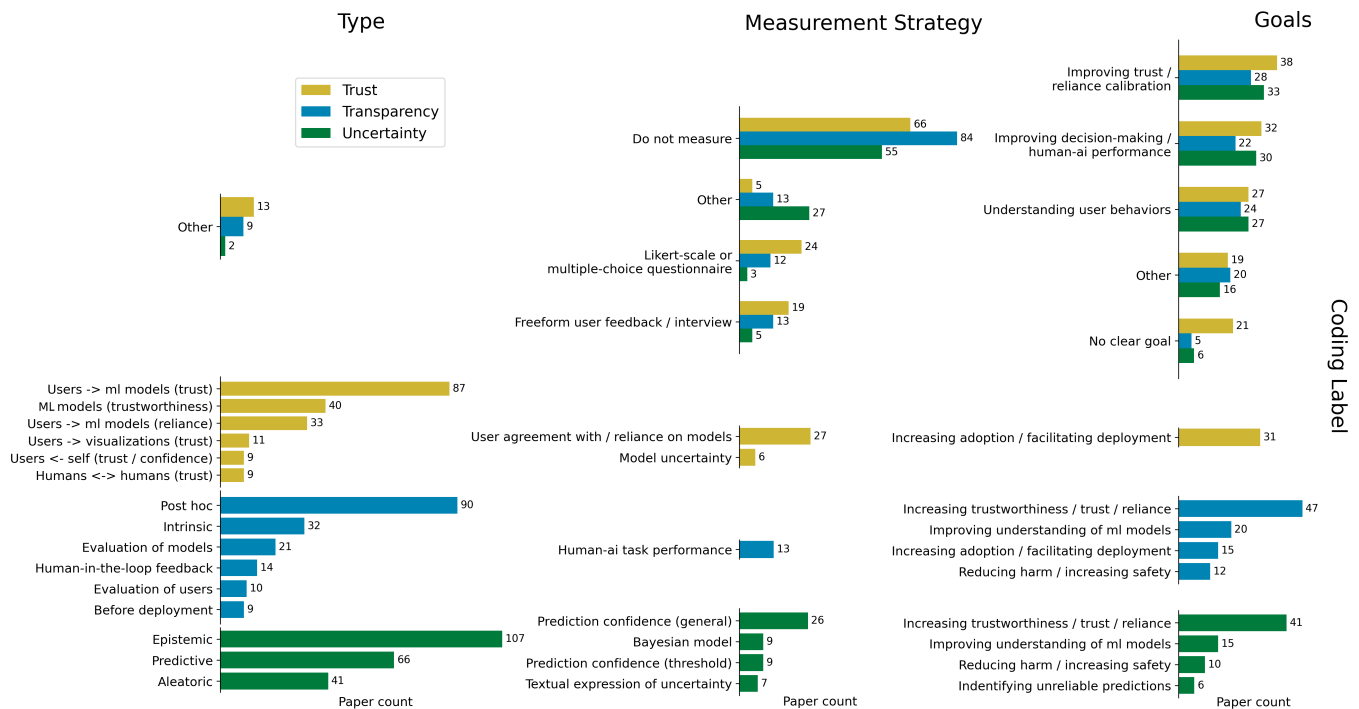


Figure 3: ML papers included in our survey study many different forms of trust, transparency, and uncertainty; use a wide range of measurement strategies; and have disparate goals for handling each concept. This figure shows the number of unique coding labels assigned to reviewed papers. Columns correspond to type, measurement strategy, and goals, and rows correspond to each concept. Bars for **trust**, **uncertainty**, and **transparency** are colored **yellow**, **green**, and **blue**, respectively.

to maximize human-AI team performance by increasing trust calibration, i.e., increasing users’ trust in the model when it is likely to be correct and decreasing trust when it is likely to err. Other works investigated the effect of factors such as accuracy and confidence on trust with the ultimate goal of facilitating adoption. Notably, only a single paper aimed to promote trust calibration and collaboration *between humans*. More than transparency or uncertainty, papers provide no clear goal for studying trust, reliance, or trustworthiness, reflecting that the concepts are often viewed as inherently beneficial, even though trust in models is not always desirable. Similarly, the majority (66/118) of papers that discuss trust/reliance/trustworthiness do not measure it. These results indicate an overwhelming emphasis on trust in ML models themselves rather than trust relationships between humans interacting with models and between users and model creators.

● Transparency

Most papers in our survey that provide an explicit definition of transparency broadly describe it as a way to help users understand models. Like trust and trustworthiness, however, there is disagreement on whether transparency is primarily subjective and context-dependent or a property of systems. Some works define transparency as “providing information to understand the inference logic of the advice-giving system” (Nunes and Jannach 2017), while others frame it as an *affordance* (Thornton et al. 2021; Liao and Sundar 2022) for user understanding: “not a property of the ML model but... a

relationship between algorithm and users” (Chen et al. 2022).

Another point of contention is the difference between “transparency,” “explainability,” and “interpretability”. For example, in one work explainability “refers to inspecting the AI model interior to confirm its output” (Föllmer et al. 2024), and another claims explainability methods “rely on a second, simpler model trained to approximate the predictions of the black-box AI” (Sagona et al. 2025). The second paper claims that “explainable AI, therefore, can offer only an approximation—not true transparency”, which contradicts other definitions of transparency that comprise “efforts to increase explainability, interpretability or other acts of communication and disclosure” (Jobin et al. 2019). Many well-known works such as Rudin 2019 and Babic et al. 2021 use interpretability to mean specifically intrinsic / inherent methods and explainability to mean post hoc methods for explaining model outputs. For brevity, we use “transparency” as a stand in for “transparency, interpretability, and explainability”, but acknowledge that these terms are not always interchangeable.

In policy, transparency, interpretability, and explainability are predominantly framed as system properties. Unlike trust and uncertainty, transparency is among the most frequently cited AI principles in global policies and guidelines (see Appendix for more detail). A sample of definitions for transparency include: “understanding the working logic of the model” (Atherton et al. 2023), and “making data, features, models, algorithms, training methods and quality assurance processes available for external inspection” (ISO/IEC 2020).

The meaning of transparency and related concepts are often interdependent, and cyclical definitions are common. Figure 1 depicts the complicated web of definitions extracted from 13 policy documents (full list in Appendix Table A.2), highlighting the difficulty of untangling the meaning of terms within and across guidelines and standards.

Many of these cyclical definitions are direct restatements such as “accountability: state of being *accountable*” (ISO/IEC 2022), and “Transparency: the data, system and AI business models should be *transparent*” (HLEG 2019). These are often benign in dictionaries, but can be undecipherable in technical standards when finding the source definition requires tracing back a convoluted chain of paywalled documents. Indirect circular definitions can be even harder to parse. For example, policymakers may consult the definition of explainability in the NATO Principles of Responsible Use of AI in Defence: “AI applications will be appropriately understandable and transparent...” Because the document does not define what “transparent” means, policymakers may next turn to the EU Guidelines on Ethics in AI, only to find that transparency is defined in terms of explainability, which is itself framed as a prerequisite for trustworthiness in other EU guidance (HLEG 2020). Consulting multiple policy documents is also common for companies servicing several jurisdictions. For example, Google must navigate differing guidelines from the US and EU for its cloud and AI services.

Inconsistent definitions of transparency bleed into how it is operationalized, with frequently assumed goals but few concrete measurement strategies. Of the three concepts, we find that transparency is measured the least, reflecting the common tendency in ML papers to axiomatically claim that proposed methods are transparent or increase transparency without further argument (Lipton 2018). In conflict with system-focused definitions for transparency, human-facing Likert scale questionnaires (12) and freeform feedback (13) were the most common measurement strategies. Authors typically solicited participant input on proposed methods through questions such as “to what extent do you understand the reasons behind the model’s predictions?” (Doula et al. 2025). Measuring transparency by human-AI task performance was also relatively common (13), usually under the assumption that higher team performance indicates better understanding. Compared to trust, transparency was rarely examined *without* a clearly articulated goal. In many works, transparency is motivated as a means of increasing trust or reliance on a system without justifying whether such increases are desirable. The most common goals for studying transparency were to modify trust and reliance, with papers split over whether they aim to increase *appropriate* trust / reliance (28), or simply increase trust / reliance without regard to calibration (47).

In our review, the most common classification of transparency methods is based on *when* they take place in the ML model pipeline, ranging from documentation before training, to using models that are inherently easier to understand, to post hoc methods that are applied after a model has made a prediction. Post hoc methods are by far the most prevalent in our survey (90/118), with uncertainty quantification and feature-based explanations appearing most often, followed by intrinsic methods (32).

● Uncertainty

The most common definition of uncertainty identified in our review divides it into *aleatoric* (uncertainty due to inherent randomness) and *epistemic* (uncertainty due to lack of knowledge) categories. Within ML, this dichotomy is often used as a proxy for uncertainty arising from data or ML models, respectively. However, as Marjanovic et al. 2024 observe, studies often blur the two by treating output probabilities as direct indicators of “predictive uncertainty.” Such usage implicitly conflates epistemic and aleatoric sources of uncertainty, making it difficult to disentangle errors due to either model limitations or stochasticity in the data.

Other terms such as risk, confidence, unpredictability, and uncertainty quantification were used conflictingly as either synonyms for uncertainty or distinct concepts. For example, Kostick-Quenet et al. 2022 define risk as having agreed-upon outcome probabilities while uncertainty does not, which clashes with other definitions of uncertainty as “probability that a given model output is accurate” (Loftus et al. 2022). Zhang et al. 2020 use a similar definition to describe confidence instead, writing “the model’s predicted probability for the most likely outcome.” Still other work equates *unpredictability* and uncertainty. For example, Salimzadeh et al. 2024 propose that “task uncertainty refers to the extent of unpredictability inherent in a task” (referring only to aleatoric uncertainty). Some authors describe uncertainty quantification (UQ) as a means of expressing confidence in model outputs (Toh et al. 2025), while others frame UQ more broadly and literally as “quantify[ing] sources of uncertainty” (Camps-Valls et al. 2025). This variation adds complexity to an already slippery concept, and obfuscates whether papers have compatible notions of uncertainty.

Notably, only two studies in our review mentioned *ontological* uncertainty, which is often included in the taxonomy alongside *aleatoric* and *epistemic* uncertainty (Spiegelhalter 2017), and refers to “unknown unknowns” or uncertainty about whether a system *can be* a reflection of reality.

For example, quantifying models’ lack of knowledge in terms of loss would be a measurement of epistemic uncertainty, whereas subjectively recognizing that all models are wrong would be an expression of ontological uncertainty. This omission is surprising given the pervasiveness of discussions of aleatoric and epistemic uncertainty in the papers we survey. More broadly, the lack of attention to ontological uncertainty highlights an open area for future research, particularly in understanding how ML applications contend with deep model limitations and the possibility that their underlying assumptions may fail to capture aspects of reality.

Unlike ML literature, policy documents rarely mention uncertainty, and are more likely discuss risk (Jobin et al. 2019; Corrêa et al. 2023; Khan et al. 2022; Huang et al. 2023; Tidjon and Khomh 2023; Ibáñez and Olmeda 2022). Across multilateral and national AI principles documents, we could not find a single definition of uncertainty. In their glossary of terms for trustworthy AI (Atherton et al. 2023), NIST refers to a definition of uncertainty from an IEEE systems and software engineering vocabulary: “Uncertainty: result of not having accurate or sufficient knowledge of a situation; state, even partial, of deficiency of information related to

understanding or knowledge of an event, its consequences, or likelihood.” Similarly, ISO/IEC 42001:2023 defines risk as an “effect of uncertainty” and also references the IEEE definition. This “deficiency” framing carries a negative connotation, capturing only epistemic sources and implying uncertainty is a problem to be solved.

In contrast, papers in our review frequently acknowledge the benefits of uncertainty as a source of information, where the most common associated goals are increasing trust, reliance, or trustworthiness (41), followed closely by improving trust and/or reliance calibration (33), and understanding the effect of uncertainty on user behavior (27). Unlike trust and transparency, however, we identified no papers that explicitly handled uncertainty to increase adoption. Of the three concepts, uncertainty was also measured the most using the widest variety of methods. Roughly in order of added effort, the most common strategy was directly assigning an uncertainty value from model outputs (26) (e.g., by taking the softmax of output logits), or creating discrete categories (e.g., by classifying all outputs with an output probability of < 0.7 as “low-confidence” or “uncertain”). These correspond to our “general” and “threshold” prediction confidence labels respectively. The next most common (more involved) measurement methods were Bayesian models (9), and counting textual expressions of uncertainty (7).

Relationships Between Trust, Transparency, and Uncertainty

In this section, we build on our individual review of the three concepts by discussing how they are jointly defined and used (summarized in Figure 4). Ultimately, we extend prior work by arguing that trust, transparency, and uncertainty are *co-produced* rather than individual design levers.

Trust and Transparency

Transparency is often assumed to increase trust and/or trust calibration, but the relationship between the two concepts is complex and context-dependent. The top two goals for transparency in papers we review are “increasing trustworthiness / trust / reliance” and “improving trust / reliance calibration”. In practice, however, the effect of different transparency methods on trust is complex and varies based on task, application, and user. This complexity has led to a range of sometimes conflicting results across studies. Some methods are shown to appropriately calibrate trust in systems (Wang and Yin 2021; Lavin et al. 2022; Vasconcelos et al. 2023), some cause overtrust in unreliable models (Smith et al. 2018; Bussone et al. 2015), some have no effect on trust (Zhang et al. 2020), and some even result in algorithmic aversion (Thomas et al. 2024; Gomez et al. 2024).

Several works point to context and human factors to explain why some transparency techniques are effective for trust and others are not. As Thornton et al. 2021 argue, transparency only has a positive effect on trust “when it lessens rather than increases the cognitive burden on the given user to interpret and understand information”. Papers in our review explore different aspects of how cognitive burden affects transparency and trust, including difficulty to verify

whether an explanation is correct (Smith et al. 2018), task complexity (Salimzadeh et al. 2024), and prior beliefs and experience (Cui et al. 2023; Carmichael et al. 2024; Gomez et al. 2024; Niraula et al. 2025). While some works claim that transparency can help build mental models of ML systems (Tomsett et al. 2020), others note that it only builds trust when it aligns with users’ preexisting mental models (Toh et al. 2025), and can lead to incorrect assumptions about how models work (Cau et al. 2023).

Although we do not advocate for using a single monolithic definition of trust or transparency, the complex, context-dependent, and human-affected relationship between them provides strong evidence that both concepts should not be treated as merely computational, and that no single transparency method will likely have the desired effect on trust for all situations and users.

The opposite directional relationship—how trust affects transparency—is an interesting but underexplored research direction. As a prior belief and part of users’ mental models, trust in ML systems is an important factor to consider for transparency methods, but only a few works in our review measure its effect. For example, clinicians’ preexisting trust towards ML systems affects how they chose to interact with model explanations (Sivaraman et al. 2023; Chanda et al. 2024). One paper similarly found that trust in the dataset and data source of a ML model affects how well people use transparency techniques to update their beliefs (Cui et al. 2023). A potentially interesting future direction is to measure how users’ preexisting trust in ML systems affects their tendency to view them as transparent.

Trust and Uncertainty

Trust and uncertainty are linked by definition. Calling a system trustworthy or placing trust in a model presupposes uncertainty about their ability and underlying intent. To understand user trust in a system, improve decision-making, or trust calibration, it is therefore important to jointly consider the sources of uncertainty that necessitate trust.

In the papers and policy documents we survey, uncertainty is often handled with the same trust goals as transparency, but with a less positive connotation. Works similarly assume that displaying uncertainty will improve trust calibration and/or trustworthiness, and the goals for handling transparency and uncertainty are mostly the same. One important difference is that no papers use uncertainty to directly improve adoption of ML systems. This is representative of a general trend where although both concepts are assumed to increase trustworthiness, transparency is assumed to increase trust calibration by *increasing* trust (reducing undertrust) (Prabhudesai et al. 2023; Wintersberger et al. 2019; Sermesant et al. 2021; Tomsett et al. 2020), while uncertainty is assumed to increase calibration by *decreasing* trust (reducing overtrust) (Tomsett et al. 2020; Sacha et al. 2016). This negative connotation is reflected in the framing that transparency describes what the model “knows” and uncertainty shows what the model “doesn’t know”, and the conflation that more certain models are more trustworthy.

The effects of uncertainty on trust are also varied and context-dependent. Like transparency, in some works

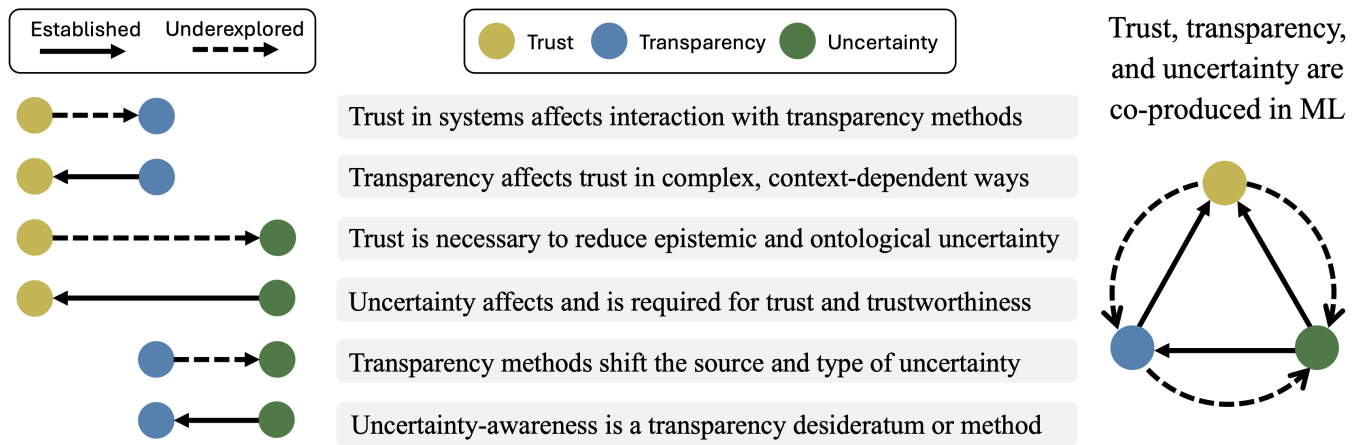


Figure 4: **Trust, transparency, and uncertainty are co-produced concepts in ML.** Left: a summary of the established and underexplored relationships between each pair of concepts. Right: a diagram depicting the connections between all three.

uncertainty is shown to either increase (Zhang et al. 2020; Barth et al. 2024), decrease (Schneider et al. 2019; Kim et al. 2024; Toh et al. 2025), or have no effect (Chanda et al. 2024; Carmichael et al. 2024; Bussone et al. 2015) on trust and trust calibration. Studies highlight a number of confounding factors, including whether a model is uncertain about “easy” or “difficult” problems (Koch et al. 2025), the number and ordering of high vs. low-confidence predictions (Thomas et al. 2024), how well users understand uncertainty (Cui et al. 2023; Prabhudesai et al. 2023; McGrath et al. 2023), differences in risk tolerance (Gomez et al. 2024), and users’ confidence in themselves (Smith et al. 2018; Bansal et al. 2021). The study of trust in these works necessitates accounting for the messiness of human relationships to uncertainty. This suggests that users’ trust calibration requires an understanding of both the model’s uncertainty and their own. An interesting direction for future work is to create systems that help users quantify these two sources together.

●---->● We observe that trust can also significantly affect uncertainty. Reducing aleatoric uncertainty with better measurement strategies or higher quality data requires trust in new tools and data collection practices. Ontological uncertainty concerning the limits of modeling depends on beliefs in the capability of current systems and the intent of the model makers. Epistemic uncertainty, which arises from a lack of information, is resolved only if we trust that a new source of information reflects reality. To increase transparency, methods such as textual explanations, counterfactuals, and saliency maps, which aim to resolve uncertainty about how the model works, only do so if the user observing these methods trusts that they reflect the model’s inner decision-making (which is often not the case (Adebayo et al. 2018; Ehsan and Riedl 2024)). To truly understand a source of uncertainty we must rely on, or in the case of other people, trust the source communicating it, lest it become another source of uncertainty.

●<-->● Transparency and Uncertainty

●<-->● Uncertainty affects both how people engage with transparency and what transparency methods are needed. Peo-

ple are more likely to rely on models with explanations when they are uncertain about the answer, uncertain about their own abilities, or in a situation marked by high uncertainty such as clinical decision-making (Kopka et al. 2025; Salimzadeh et al. 2024). The type and source of uncertainty also affects which types of transparency are needed. With simpler white-box models, for example, focus is often on resolving uncertainty about how to interpret the most important features, while deep learning approaches also introduce the need to extract relevant features, or even to determine unknown properties about the model.

●---->● The reverse relationship – that increasing transparency reduces uncertainty – is assumed in many of the papers we survey (Sacha et al. 2016; Ma et al. 2024; Huang et al. 2025). Although this is likely true when, for example, black-box model weights are released, we believe that the effects of transparency on uncertainty are underexplored and often complex. We observe that many transparency methods *shift* the source and type of uncertainty rather than straightforwardly reduce it. This is most obvious in post hoc methods such as LIME (Ribeiro et al. 2016) that train a smaller classifier to explain the prediction of a larger model. In this case, uncertainty about how a model works is transformed into uncertainty about whether the smaller model accurately reflects the larger. Similarly, black-box LLM explanations for their own behavior shift uncertainty about the initial generation to whether the second explanatory generation is correct. In this way, transparency methods that add complexity to understand model behavior also introduce new sources of uncertainty.

Replacing large, uninterpretable models with inherently transparent ones also transforms epistemic uncertainty. As Lipton argues, simpler models like decision trees can still be difficult to understand when they have thousands of nodes (not reducing epistemic uncertainty), and linear models can be brittle to feature selection and preprocessing (adding epistemic uncertainty) (Lipton 2018). Similarly, scaling models down may introduce ontological uncertainty about the class of problem that can be represented. The purpose of this argument is not to discourage the use of transparency

methods—shifting the source or type of uncertainty is often very useful—but rather to challenge the idea that they unproblematically reduce uncertainty.



Trust, Transparency, and Uncertainty are Co-produced

The conflicting results and confounding factors highlighted in our review show that the effects of trust, transparency, and uncertainty on each other are highly context-dependent, influenced by both human and computational factors. The papers we survey overwhelmingly focus on one-way relationships, with transparency and uncertainty modifying trust, and uncertainty-awareness as an important property of transparency; but we find that *all* pairwise relationships are relevant. Transparency methods shift the source and type of uncertainty and affect trust based on the method, task, complexity, uncertainty, and user. Trust influences how people interact with transparency methods and is needed to reduce sources of uncertainty. Uncertainty influences which kinds of transparency are needed, and is required for trust. Taken together, these findings suggest that future research can address persistent inconsistencies in definitions and outcomes by considering trust, uncertainty, and transparency not in isolation but as mutually shaping processes. Our solution is therefore to treat them in conjunction, recognizing that each concept acquires meaning through its interaction with the others. In the next section, we illustrate how this perspective provides greater conceptual stability by mapping the co-produced relationships among trust, uncertainty, and transparency, and by outlining how these connections operate in practice.

Discussion

In this paper, we systematically examine work at the intersection of trust, uncertainty, and transparency in ML. Although scholars and policymakers commonly acknowledge that each concept is foundational to the responsible deployment of ML our review shows their definitions remain fragmented and frequently conflict, ranging from narrow performance-based accounts to broader social and relational interpretations. Instead, their meanings and roles implicitly shift with context, goals, and measurement choices. We also find that the relationships between concepts are inconsistently theorized, leading to contradicting results that are difficult to resolve. We distill these findings into four definitional problems across ML policy and research:

P1 Lack of consistency leads to conflicting definitions:

Trust, uncertainty, transparency, explainability, and interpretability are sometimes defined in conflicting ways. For example, some definitions permit trust between humans and machines, while others rule it out entirely. Uncertainty is alternately framed as risk or confidence; and explainability, interpretability, and transparency have each been defined as specifically referring to post hoc or inherent methods. These inconsistencies are especially problematic in policy, where adhering to one guideline may implicitly violate another, creating ambiguous and potentially incompatible compliance obligations.

P2 Lack of explicit definitions: The lack of standard definitions for trust, uncertainty, and transparency in ML is made worse by the propensity to leave them undefined. Because each concept is too broad for a single monolithic definition, it is important to clarify precisely what we mean by trust, transparency, and uncertainty. Yet, most of papers in our survey do not provide explicit definitions for the terms, particularly of trust and transparency. As a result, studies operationalize each concept using ad hoc measures without specifying *what* they measure.

P3 Definitions are too broad to be implemented: Although explicit definitions are standard in policy documents, high-level language, particularly in AI principles, makes it difficult for stakeholders to understand or verify when they are in compliance. Technical standards define transparency at different levels but provide only broad definitions for trust and uncertainty.

P4 Cyclical definitions: Cyclical definitions are also common in both policy and technical standards. This creates another barrier to implementation as trust, transparency, and uncertainty are defined self-referentially, in terms of each other, or through related high-level terms with their own tangled web of definitions (see Figure 1)

This misalignment has consequences that extend beyond conceptual clarity. When trust, transparency, uncertainty, and the relationships between them are assumed, they shape how systems are built, assessed, and governed in ways that are difficult to interpret or audit. In research settings, this limits cumulative knowledge-building and obscures the effects of design interventions on user behavior. In policy and standards, it weakens accountability by allowing shared principles to be interpreted in incompatible ways. Ultimately, we argue that framing trust, transparency, and uncertainty as *co-produced*, rather than merely connected can help untangle their meanings, providing conceptual clarity and explaining conflicting results. This perspective also highlights gaps in how the three concepts are studied together, including how transparency methods shift the source and type of uncertainty, and the affects of trust on transparency and uncertainty. In the following sections, we provide examples of how this framing can improve conceptual clarity, and offer recommendations to improve how trust, transparency, and uncertainty are conceptualized, studied, and implemented in ML systems.

Utility of Conceptualizing Trust, Transparency, and Uncertainty as Co-produced

As we have shown in the previous sections, trust, transparency, uncertainty, and the relationships between them are incredibly complex. We argue that framing these concepts as *co-produced*, rather than merely connected, can help provide both conceptual clarity and a novel lens for future research. We offer two examples to illustrate this point.

Example 1: Gomez et al. 2024 and Sivaraman et al. 2023 present two seemingly conflicting results on the effect of SHAP (Lundberg and Lee 2017) feature contribution scores on clinician trust. In Gomez et al., clinicians diagnose strep throat from patient images using one of four explanation methods including: the Centor score, a well-established method for calculating the likelihood of strep, and the SHAP

feature importance of multiple factors. The authors find that clinicians had significantly lower trust in the system when given the SHAP contributions compared to the Centor score, both in terms of self-reported trust on a Likert questionnaire, and in the fraction of negative cases where participants suggested a confirmatory test (Gomez et al. 2024). In Sivaraman et al., the authors similarly measure how SHAP explanations affect clinician’s trust when predicting sepsis treatment plans from patient data, and find that it moderately increases participants’ confidence when using the ML model prediction, and has no significant effect on overall prediction confidence (Sivaraman et al. 2023).

Although there are meaningful differences between the two experiments, viewing transparency, uncertainty, and trust as co-produced can also help explain this conflicting result. In Gomez et al., the Centor score reduces clinicians’ uncertainty because it is a trusted and simple metric. Although SHAP aims to explain opaque models, it adds another source of epistemic uncertainty because the SHAP model *itself* is not transparent to participants, causing them to request more confirmatory tests. In Sivaraman et al., on the other hand, the shift in uncertainty source from the ML model to SHAP explanation has a negligible effect compared to not using the model at all, which is confirmed by the paper’s finding that most clinicians ignore or selectively rely on model predictions. Seeing the SHAP explanation after only the ML model prediction (which is not an option in the strep prediction task), however, manifests as an increase in information or knowledge, reducing epistemic uncertainty. Taken together, these studies suggest that the effect of SHAP on clinicians’ trust is not fixed. Instead, it depends on whether the explanation reduces the uncertainty for users or introduces new sources of uncertainty, illustrating how transparency, uncertainty, and trust are co-produced in this context.

Example 2: Many works claim that transparency is necessary for trustworthiness (Hazeleger et al. 2024; Lo Piano 2020; Wu et al. 2024; Jobin et al. 2019; Antoran et al. 2021), and others claim trustworthiness is necessary for transparency (Föllmer et al. 2024; Bykov et al. 2020). Both claims are reasonable, but lead to a definitional loop. Trustworthiness is treated as requiring transparency, while transparency is treated as requiring trustworthy explanations, confidence information, or data, thereby generating a potentially unbounded cycle in which each concept is invoked to justify the other. Framing trust and transparency as co-produced rather than merely connected can help escape this cycle. Taken together, transparency produces trust by providing necessary information, and trust produces transparency by allowing that information to improve understanding (rather than becoming another source of uncertainty). Both transparency *and* trust (and by extension uncertainty) are therefore necessary for trustworthy models, one does not lead to the other.

Recommendations

Several of the challenges we identify stem from treating trust, transparency, and uncertainty as self-evident or universally agreed-upon concepts (P1, P2), or from relying on definitions that are too broad to meaningfully guide practice (P3). Instead of seeking a single definition of trust, transparency,

or uncertainty, researchers should specify the purpose for invoking each construct (e.g., calibration, safety, adoption) and define it accordingly. This reframes definitional diversity as acceptable variability rather than inconsistency, provided the purpose is explicit. Further, rather than treating trust, transparency, and uncertainty as separable variables, studies should acknowledge they are co-produced and test their joint effects. We collect evidence that trust, transparency, and uncertainty should *all* be studied as subjective and context-dependent rather than solely computational. We encourage future work to investigate underexplored relationships between the three concepts, such as the effects of trust on transparency and how transparency can transform uncertainty.

Another definitional problem we identified involves circular or internally referential reasoning, in which trust, transparency, and uncertainty are defined primarily in terms of one another (P4). In system design, this circularity can be broken by grounding these concepts in concrete system properties and limitations. For example, we recommend prioritizing designs that communicate when and why a model may fail, rather than relying solely on abstract assurances of trustworthiness or transparency. Alternatively, framing the three concepts as co-produced but distinct allows separating their meanings without having to rely on narrow, monolithic definitions. An open research gap that we identified was the near absence of ontological uncertainty, which suggests an opportunity to explicitly communicate when models may fail in principle rather than merely in prediction.

Conclusion

Although trust, transparency, and uncertainty are widely regarded as foundational to responsible ML, their definitions remain fragmented and their relationships inconsistently theorized across research and policy. We synthesize discussions on each concept and their relationships, challenging assumptions about how they interact, and identifying gaps in how they are studied together. Ultimately, we argue that trust, transparency, and uncertainty are co-produced in ML, and demonstrate how this framing can improve conceptual clarity. We urge researchers, policymakers, and system designers to make definitional assumptions explicit, align concepts with their intended purposes, and recognize trust, transparency, and uncertainty as interdependent design goals.

Limitations This work has several important limitations. Our analysis of policy and standards documents is not exhaustive and focuses on highly cited surveys, major U.S. and international guidelines, and widely adopted technical standards. Similarly, our survey of the research literature prioritizes representativeness over completeness by focusing on influential venues and papers that explicitly engage with trust, transparency, and uncertainty together. Our focus on papers that consider all three concepts together was a necessary step to narrow the scope of our review, but it may skew our sampling towards a specific subliterate. For example, our survey likely contains more papers with human-subjects studies compared to ML literature in general.

Acknowledgments

We would like to thank Lav Varsheny for their input and feedback, and Northeastern University's Khoury College of Computer Sciences for funding this work. This work was supported by the U.S. National Science Foundation under Grant Nos. 2428149 and 2313998 and by the National Institutes of Health under Grant No. 1R01AI188576-01. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of funding agencies.

References

- Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity Checks for Saliency Maps. *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, 9525–9536.
- Afroogh, S.; Akbari, A.; Malone, E.; Kargar, M.; and Alambeigi, H. 2024. Trust in AI: progress, challenges, and future directions. *Humanities and Social Sciences Communications*, 11(1): 1–30. Publisher: Palgrave.
- Aichberger, L.; Schweighofer, K.; Ielanskyi, M.; and Hochreiter, S. 2025. Improving Uncertainty Estimation through Semantically Diverse Language Generation.
- Alvarado Garcia, A.; Candello, H.; Badillo-Urquiola, K.; and Wong-Villacres, M. 2025. Emerging Data Practices: Data Work in the Era of Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, 1–21. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-1394-1.
- Antoran, J.; Bhatt, U.; Adel, T.; Weller, A.; and Hernández-Lobato, J. M. 2021. Getting a {CLUE}: A Method for Explaining Uncertainty Estimates. In *International Conference on Learning Representations*.
- Atherton, D.; Schwartz, R.; Fontana, P.; and Hall, P. 2023. The Language of Trustworthy AI: An In-Depth Glossary of Terms.
- Babic, B.; Gerke, S.; Evgeniou, T.; and Cohen, I. G. 2021. Beware explanations from AI in health care. *Science*, 373(6552): 284–286.
- Banovic, N.; Yang, Z.; Ramesh, A.; and Liu, A. 2023. Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1).
- Bansal, G.; Nushi, B.; Kamar, E.; Lasecki, W. S.; Weld, D. S.; and 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: 2–11.
- Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. Yokohama Japan: ACM. ISBN 978-1-4503-8096-6.
- Barth, C.-M.; Bernard, J.; and Huang, E. M. 2024. “It’s like a glimpse into the future”: Exploring the Role of Blood Glucose Prediction Technologies for Type 1 Diabetes Self-Management. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, 1–21. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0330-0. Event-place: Honolulu, HI, USA.
- Bhatt, U.; Antorán, J.; Zhang, Y.; Liao, Q. V.; Sattigeri, P.; Fogliato, R.; Melançon, G.; Krishnan, R.; Stanley, J.; Tickoo, O.; Nachman, L.; Chunara, R.; Srikumar, M.; Weller, A.; and Xiang, A. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 401–413. Virtual Event USA: ACM. ISBN 978-1-4503-8473-5.
- Boukhelifa, N.; Perrin, M.-E.; Huron, S.; and Eagan, J. 2017. How Data Workers Cope with Uncertainty: A Task Characterisation Study. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3645–3656. Denver Colorado USA: ACM. ISBN 978-1-4503-4655-9.
- Brodie, K.; Allendes Osorio, R.; and Lopes, A. 2012. *A review of uncertainty in data visualization*. London: Springer.
- Broniatowski, D. A. 2021. Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. NISTIR 8367.
- Bussone, A.; Stumpf, S.; and O’Sullivan, D. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*, 160–169. Dallas, TX, USA: IEEE. ISBN 978-1-4673-9548-9.
- Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–21.
- Bykov, K.; Höhne, M. M.-C.; Müller, K.-R.; Nakajima, S.; and Kloft, M. 2020. How Much Can I Trust You? – Quantifying Uncertainties in Explaining Neural Networks. ArXiv:2006.09000 [cs].
- Camps-Valls, G.; Fernández-Torres, M.-A.; Cohrs, K.-H.; Höhl, A.; Castelletti, A.; Pacal, A.; Robin, C.; Martinuzzi, F.; Papoutsis, I.; Prapas, I.; Pérez-Aracil, J.; Weigel, K.; Gonzalez-Calabuig, M.; Reichstein, M.; Rabel, M.; Giuliani, M.; Mahecha, M. D.; Popescu, O.-I.; Pellicer-Valero, O. J.; Ouala, S.; Salcedo-Sanz, S.; Sippel, S.; Kondylatos, S.; Happé, T.; and Williams, T. 2025. Artificial intelligence for modeling and understanding extreme weather and climate events. *Nature Communications*, 16(1): 1919. Publisher: Nature Publishing Group.
- Carmichael, J.; Costanza, E.; Blandford, A.; Struyven, R.; Keane, P. A.; and Balaskas, K. 2024. Diagnostic decisions of specialist optometrists exposed to ambiguous deep-learning outputs. *Scientific Reports*, 14(1): 6775. Publisher: Nature Publishing Group.
- Cau, F. M.; Hauptmann, H.; Spano, L. D.; and Tintarev, N. 2023. Supporting High-Uncertainty Decisions through AI and Logic-Style Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI

- Chakraborti, T.; Banerji, C. R. S.; Marandon, A.; Hellon, V.; Mitra, R.; Lehmann, B.; Bräuninger, L.; McGough, S.; Turkay, C.; Frangi, A. F.; Bianconi, G.; Li, W.; Rackham, O.; Parashar, D.; Harbron, C.; and MacArthur, B. 2025. Personalized uncertainty quantification in artificial intelligence. *Nature Machine Intelligence*, 7(4): 522–530. Publisher: Nature Publishing Group.
- Chanda, T.; Hauser, K.; Hobelsberger, S.; Bucher, T.-C.; Garcia, C. N.; Wies, C.; Kittler, H.; Tschandl, P.; Navarrete-Dechent, C.; Podlipnik, S.; Chousakos, E.; Crnaric, I.; Majstorovic, J.; Alhajwan, L.; Foreman, T.; Peternel, S.; Sarap, S.; Özdemir, i.; Barnhill, R. L.; Llamas-Velasco, M.; Poch, G.; Korsing, S.; Sondermann, W.; Gellrich, F. F.; Heppt, M. V.; Erdmann, M.; Haferkamp, S.; Drexler, K.; Goebeler, M.; Schilling, B.; Utikal, J. S.; Ghoreschi, K.; Fröhling, S.; Krieghoff-Henning, E.; and Brinker, T. J. 2024. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *Nature Communications*, 15(1): 524. Publisher: Nature Publishing Group.
- Chandu, K.; Li, L.; Awadalla, A.; Lu, X.; Park, J. S.; Hessel, J.; Wang, L.; and Choi, Y. 2025. CertainlyUncertain: A Benchmark and Metric for Multimodal Epistemic and Aleatoric Awareness.
- Chen, H.; Gomez, C.; Huang, C.-M.; and Unberath, M. 2022. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *npj Digital Medicine*, 5(1): 1–15. Publisher: Nature Publishing Group.
- Chen, Y.; Verbert, K.; Gerling, K.; Vanden Abeele, M.-E.; and Vanden Abeele, V. 2025. Will Health Experts Adopt a Clinical Decision Support System for Game-Based Digital Biomarkers? Investigating the Impact of Different Explanations on Perceived Ease-of-Use, Perceived Usefulness, and Trust. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 1581–1597. Cagliari Italy: ACM. ISBN 979-8-4007-1306-4.
- Commission, T. E. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council.
- Corrêa, N. K.; Galvão, C.; Santos, J. W.; Del Pino, C.; Pinto, E. P.; Barbosa, C.; Massmann, D.; Mambrini, R.; Galvão, L.; Terem, E.; and de Oliveira, N. 2023. Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance.
- Cui, Y.; Liu, Z.; Liu, X.; Liu, X.; Wang, C.; Kuo, T.-W.; Xue, C. J.; and Chan, A. B. 2023. Bayes-MIL: A New Probabilistic Perspective on Attention-based Multiple Instance Learning for Whole Slide Images.
- de Lima-Santos, M.-F.; Yeung, W. N.; and Dodds, T. 2025. Guiding the way: a comprehensive examination of AI guidelines in global media. *AI & SOCIETY*, 40: 2585–2603.
- Doshi-Velez, F.; and Kim, B. 2017. Towards A Rigorous Science of Interpretable Machine Learning. ArXiv:1702.08608 [stat].
- Doula, A.; Mühlhäuser, M.; and Sanchez Guinea, A. 2025. AEXL: Enhancing Path Prediction with Active Explainable Learning via Large Language Models. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713958.
- Ehsan, U.; and Riedl, M. O. 2024. Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns*, 5(6): 1–8. Publisher: Elsevier.
- Esteva, A.; Chou, K.; Yeung, S.; Naik, N.; Madani, A.; Motlaghi, A.; Liu, Y.; Topol, E.; Dean, J.; and Socher, R. 2021. Deep learning-enabled medical computer vision. *npj Digital Medicine*, 4(1): 1–9. Publisher: Nature Publishing Group.
- Fischer, M. T.; Hirsbrunner, S. D.; Jentner, W.; Miller, M.; Keim, D. A.; and Helm, P. 2022. Promoting Ethical Awareness in Communication Analysis: Investigating Potentials and Limits of Visual Analytics for Intelligence Applications. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 877–889. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9352-2. Event-place: Seoul, Republic of Korea.
- Föllmer, B.; Williams, M. C.; Dey, D.; Arbab-Zadeh, A.; Maurovich-Horvat, P.; Volleberg, R. H. J. A.; Rueckert, D.; Schnabel, J. A.; Newby, D. E.; Dweck, M. R.; Guagliumi, G.; Falk, V.; Vázquez Mézquita, A. J.; Biavati, F.; Išgum, I.; and Dewey, M. 2024. Roadmap on the use of artificial intelligence for imaging of vulnerable atherosclerotic plaque in coronary arteries. *Nature Reviews Cardiology*, 21(1): 51–64. Publisher: Nature Publishing Group.
- Gardiner, L.-J.; Carrieri, A. P.; Wilshaw, J.; Checkley, S.; Pyzer-Knapp, E. O.; and Krishna, R. 2020. Using human in vitro transcriptome analysis to build trustworthy machine learning models for prediction of animal drug toxicity. *Scientific Reports*, 10(1): 9522. Publisher: Nature Publishing Group.
- Gomez, C.; Smith, B.-L.; Zayas, A.; Unberath, M.; and Canares, T. 2024. Explainable AI decision support improves accuracy during telehealth strep throat screening. *Communications Medicine*, 4(1): 1–11. Publisher: Nature Publishing Group.
- Hagele, D.; Krake, T.; and Weiskopf, D. 2022. Uncertainty-Aware Multidimensional Scaling. *IEEE Transactions on Visualization and Computer Graphics*, 1–10.
- Hagendorff, T. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1): 99–120.
- Hasan, M. M.; Abdar, M.; Khosravi, A.; Aickelin, U.; Lio, P.; Hossain, I.; Rahman, A.; and Nahavandi, S. 2025. Survey on Leveraging Uncertainty Estimation Towards Trustworthy Deep Neural Networks: The Case of Reject Option and Post-training Processing. *ACM Computing Surveys*, 57(9): 1–35.
- Hazeleger, W.; Aerts, J. P. M.; Bauer, P.; Bierkens, M. F. P.; Camps-Valls, G.; Dekker, M. M.; Doblas-Reyes, F. J.; Eyring, V.; Finkenauer, C.; Grundner, A.; Hachinger, S.; Hall, D. M.; Hartmann, T.; Iglesias-Suarez, F.; Janssens, M.; Jones, E. R.; Kölling, T.; Lees, M.; Lhermitte, S.; van Nieuwpoort, R. V.; Pahker, A.-K.; Pellicer-Valero, O. J.; Pijpers, F. P.; Siibak, A.; Spitzer, J.; Stevens, B.; Vasconcelos, V. V.; and Vossepoel, F. C. 2024. Digital twins of the Earth with and for humans. *Communications Earth & Environment*, 5(1): 1–11. Publisher: Nature Publishing Group.

- HLEG, T. E. A. 2019. Ethics guidelines for trustworthy AI.
- HLEG, T. E. A. 2020. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment.
- Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2019. Metrics for Explainable AI: Challenges and Prospects.
- Hou, B.; Liu, Y.; Qian, K.; Andreas, J.; Chang, S.; and Zhang, Y. 2024. Decomposing Uncertainty for Large Language Models through Input Clarification Ensembling.
- Huang, C.; Zhang, Z.; Mao, B.; and Yao, X. 2023. An Overview of Artificial Intelligence Ethics. *IEEE Transactions on Artificial Intelligence*, 4(4): 799–819.
- Huang, Z.; Yang, E.; Shen, J.; Gratzinger, D.; Eyerer, F.; Liang, B.; Nirschl, J.; Bingham, D.; Dussaq, A. M.; Kunder, C.; Rojansky, R.; Gilbert, A.; Chang-Graham, A. L.; Howitt, B. E.; Liu, Y.; Ryan, E. E.; Tenney, T. B.; Zhang, X.; Folkins, A.; Fox, E. J.; Montine, K. S.; Montine, T. J.; and Zou, J. 2025. A pathologist–AI collaboration framework for enhancing diagnostic accuracies and efficiencies. *Nature Biomedical Engineering*, 9(4): 455–470. Publisher: Nature Publishing Group.
- Ibáñez, J. C.; and Olmeda, M. V. 2022. Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study. *AI & SOCIETY*, 37(4): 1663–1687.
- ISO/IEC. 2011. 25010:2011 Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models. <https://www.iso.org/standard/35733.html>. Global standards setting body.
- ISO/IEC. 2020. TR 24028:2020 Overview of trustworthiness in artificial intelligence. <https://www.iso.org/standard/77608.html>. Global standards setting body.
- ISO/IEC. 2022. 22989:2022 Artificial intelligence concepts and terminology. <https://www.iso.org/standard/74296.html>. Global standards setting body.
- Jacovi, A.; Marasović, A.; Miller, T.; and Goldberg, Y. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635. Virtual Event Canada: ACM. ISBN 978-1-4503-8309-7.
- Jobin, A.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1: 389–399.
- Kapoor, S.; Gruver, N.; Roberts, M.; Collins, K.; Pal, A.; Bhatt, U.; Weller, A.; Dooley, S.; Goldblum, M.; and Wilson, A. G. 2024. Large language models must be taught to know what they don’t know. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS ’24. Red Hook, NY, USA: Curran Associates Inc. ISBN 9798331314385.
- Khan, A. A.; Akbar, M. A.; Fahmideh, M.; Liang, P.; Waseem, M.; Ahmad, A.; Niazi, M.; and Abrahamsson, P. 2023. AI Ethics: An Empirical Study on the Views of Practitioners and Lawmakers. *IEEE Transactions on Computational Social Systems*, 10(6): 2971–2984.
- Khan, A. A.; Badshah, S.; Liang, P.; Waseem, M.; Khan, B.; Ahmad, A.; Fahmideh, M.; Niazi, M.; and Akbar, M. A. 2022. Ethics of AI: A Systematic Literature Review of Principles and Challenges. In *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering*, EASE ’22, 383–392. New York, NY, USA: Association for Computing Machinery. ISBN 9781450396134.
- Kim, S. S. Y.; Liao, Q. V.; Vorvoreanu, M.; Ballard, S.; and Vaughan, J. W. 2024. “I’m Not Sure, But...”: Examining the Impact of Large Language Models’ Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’24, 822–835. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0450-5. Event-place: Rio de Janeiro, Brazil.
- Kim, Y.-S.; Kayongo, P.; Grunde-McLaughlin, M.; and Hullman, J. 2021. Bayesian-Assisted Inference from Visualized Data. *IEEE Transactions on Visualization and Computer Graphics*, 27(2): 989–999.
- Koch, M.; Cao, N.; Weiskopf, D.; and Kurzhals, K. 2025. Active Gaze Labeling: Visualization for Trust Building. *IEEE Transactions on Visualization and Computer Graphics*, 31(7): 3942–3956.
- Kopka, M.; Wang, S. M.; Kunz, S.; Schmid, C.; and Feufel, M. A. 2025. Technology-supported self-triage decision making. *npj Health Systems*, 2(1): 1–11. Publisher: Nature Publishing Group.
- Kostick-Quenet, K. M.; and Gerke, S. 2022. AI in the hands of imperfect users. *npj Digital Medicine*, 5(1): 1–6. Publisher: Nature Publishing Group.
- Lai, V.; Chen, C.; Smith-Renner, A.; Liao, Q. V.; and Tan, C. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *2023 ACM Conference on Fairness Accountability and Transparency*, 1369–1385. Chicago IL USA: ACM. ISBN 979-8-4007-0192-4.
- Laine, J.; Minkinen, M.; and Mäntymäki, M. 2025. Understanding the Ethics of Generative AI: Established and New Ethical Principles. *Communications of the Association for Information Systems*, 56: 1–25. Copyright - Copyright Association for Information Systems 2025; Last updated - 2025-12-24.
- Laux, J.; Wachter, S.; and Mittelstadt, B. 2024. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1): 3–32.
- Lavin, A.; Gilligan-Lee, C. M.; Visnjic, A.; Ganju, S.; Newman, D.; Ganguly, S.; Lange, D.; Baydin, A. G.; Sharma, A.; Gibson, A.; Zheng, S.; Xing, E. P.; Mattmann, C.; Parr, J.; and Gal, Y. 2022. Technology readiness levels for machine learning systems. *Nature Communications*, 13(1): 6039. Publisher: Nature Publishing Group.
- Lee, H.; Choi, Y. R.; Lee, H. K.; Jeong, J.; Hong, J.; Shin, H.-W.; and Kim, H.-S. 2025. Explainable vision transformer for automatic visual sleep staging on multimodal PSG sig-

- nals. *npj Digital Medicine*, 8(1): 1–14. Publisher: Nature Publishing Group.
- Lee, J. D.; and See, K. A. 2004. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1): 50–80.
- Li, B.; Han, Z.; Li, H.; Fu, H.; and Zhang, C. 2022. Trustworthy Long-Tailed Classification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6960–6969. New Orleans, LA, USA: IEEE. ISSN: 2575-7075.
- Liang, W.; Tadesse, G. A.; Ho, D.; Fei-Fei, L.; Zaharia, M.; Zhang, C.; and Zou, J. 2022. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*, 4(8): 669–677. Publisher: Nature Publishing Group.
- Liao, Q.; and Sundar, S. S. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *2022 ACM Conference on Fairness Accountability and Transparency*, 1257–1268. Seoul Republic of Korea: ACM. ISBN 978-1-4503-9352-2.
- Lin, H.; Akbaba, D.; Meyer, M.; and Lex, A. 2023. Data Hunches: Incorporating Personal Knowledge into Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 29(1): 504–514.
- Lipton, Z. C. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10): 36–43.
- Liu, W.; Chen, Y.; and Yue, X. 2025. Enhancing Multi-View Classification Reliability with Adaptive Rejection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18): 18969–18977.
- Lo Piano, S. 2020. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanities and Social Sciences Communications*, 7(1): 1–7. Publisher: Palgrave.
- Loftus, T. J.; Shickel, B.; Ozrazgat-Baslanti, T.; Ren, Y.; Glicksberg, B. S.; Cao, J.; Singh, K.; Chan, L.; Nadkarni, G. N.; and Bihorac, A. 2022. Artificial intelligence-enabled decision support in nephrology. *Nature Reviews Nephrology*, 18(7): 452–465.
- Lu, W. 2024. Inevitable challenges of autonomy: ethical concerns in personalized algorithmic decision-making. *Humanities and Social Sciences Communications*, 11(1): 1–9. Publisher: Palgrave.
- Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Ma, H.; Han, Z.; Zhang, C.; Fu, H.; Zhou, J. T.; and Hu, Q. 2021. Trustworthy Multimodal Regression with Mixture of Normal-inverse Gamma Distributions. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.
- Ma, Z.; Mei, Y.; Gajos, K. Z.; and Arawjo, I. 2024. Schrödinger’s Update: User Perceptions of Uncertainties in Proprietary Large Language Model Updates. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–9. Honolulu HI USA: ACM. ISBN 979-8-4007-0331-7.
- MacDonald, S.; Foley, H.; Yap, M.; Johnston, R. L.; Steven, K.; Koufariotis, L. T.; Sharma, S.; Wood, S.; Addala, V.; Pearson, J. V.; Roosta, F.; Waddell, N.; Kondrashova, O.; and Trzaskowski, M. 2023. Generalising uncertainty improves accuracy and safety of deep learning analytics applied to oncology. *Scientific Reports*, 13(1): 7395. Publisher: Nature Publishing Group.
- Marjanovic, S.; Augenstein, I.; and Lioma, C. 2024. Investigating the Impact of Model Instability on Explanations and Uncertainty. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 11854–11879. Bangkok, Thailand: Association for Computational Linguistics.
- Marx, C.; Park, Y.; and Hasson, H. 2023. But Are You Sure? An Uncertainty-Aware Perspective on Explainable AI. *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- Maslej, N.; Fattorini, L.; Perrault, R.; Gil, Y.; Parli, V.; Kariuki, N.; Capstick, E.; Reuel, A.; Brynjolfsson, E.; Etchemendy, J.; Ligett, K.; Lyons, T.; Manyika, J.; Niebles, J. C.; Shoham, Y.; Wald, R.; Walsh, T.; Hamrah, A.; Santarlasci, L.; Lotufo, J. B.; Rome, A.; Shi, A.; and Oak, S. 2025. Artificial Intelligence Index Report 2025. arXiv:2504.07139.
- Mayer, R. C.; Davis, J. H.; and Schoorman, F. D. 1995. An integrative model of organizational trust. *Academy of management review*, 20(3): 709–734.
- McGrath, S.; Mehta, P.; Zytek, A.; Lage, I.; and Lakkaraju, H. 2023. When Does Uncertainty Matter?: Understanding the Impact of Predictive Uncertainty in ML Assisted Decision Making. ArXiv:2011.06167 [cs].
- Meyer, A. P.; Albarghouthi, A.; and D’Antoni, L. 2023. The Dataset Multiplicity Problem: How Unreliable Data Impacts Predictions. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’23*, 193–204. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0192-4. Event-place: Chicago, IL, USA.
- Mohseni, S.; Zarei, N.; and Ragan, E. D. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3-4): 1–45.
- Niraula, D.; Cuneo, K. C.; Dinov, I. D.; Gonzalez, B. D.; Jamaluddin, J. B.; Jin, J. J.; Luo, Y.; Matuszak, M. M.; Ten Haken, R. K.; Bryant, A. K.; Dilling, T. J.; Dykstra, M. P.; Frakes, J. M.; Liveringhouse, C. L.; Miller, S. R.; Mills, M. N.; Palm, R. F.; Regan, S. N.; Rishi, A.; Torres-Roca, J. F.; Yu, H.-H. M.; and El Naqa, I. 2025. Intricacies of human–AI interaction in dynamic decision-making for precision oncology. *Nature Communications*, 16(1): 1138. Publisher: Nature Publishing Group.
- NIST. 2024. About NIST. <https://www.nist.gov/about-nist>. U.S. federal agency conducting research and standards development.
- Nunes, I.; and Jannach, D. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27(3-5): 393–444.

- Page, M. J.; McKenzie, J. E.; Bossuyt, P. M.; Boutron, I.; Hoffmann, T. C.; Mulrow, C. D.; Shamseer, L.; Tetzlaff, J. M.; Akl, E. A.; Brennan, S. E.; Chou, R.; Glanville, J.; Grimshaw, J. M.; Hróbjartsson, A.; Lalu, M. M.; Li, T.; Loder, E. W.; Mayo-Wilson, E.; McDonald, S.; McGuinness, L. A.; Stewart, L. A.; Thomas, J.; Tricco, A. C.; Welch, V. A.; Whiting, P.; and Moher, D. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews*, 10(89).
- Pahud de Mortanges, A.; Luo, H.; Shu, S. Z.; Kamath, A.; Suter, Y.; Shelan, M.; Pöllinger, A.; and Reyes, M. 2024. Orchestrating explainable artificial intelligence for multi-modal and longitudinal data in medical imaging. *npj Digital Medicine*, 7(1): 1–10. Publisher: Nature Publishing Group.
- Panagiotidou, G.; Lamqaddam, H.; Poblome, J.; Brosens, K.; Verbert, K.; and Moere, A. V. 2022. Communicating Uncertainty in Digital Humanities Visualization Research. *IEEE Transactions on Visualization and Computer Graphics*, 29(1): 1–11.
- Prabhudesai, S.; Yang, L.; Asthana, S.; Huan, X.; Liao, Q. V.; and Banovic, N. 2023. Understanding Uncertainty: How Lay Decision-makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 379–396. Sydney NSW Australia: ACM. ISBN 979-8-4007-0106-1.
- Qiu, X.; and Miikkulainen, R. 2022. Detecting Misclassification Errors in Neural Networks with a Gaussian Process Model. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI-2022)*. (Also arXiv:2010.02065, which also includes the appendices).
- Qiu, X.; and Miikkulainen, R. 2024. Semantic Density: Uncertainty Quantification for Large Language Models through Confidence Measurement in Semantic Space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 134507 – 134533. Red Hook, NY, USA: Curran Associates Inc.
- Ransan-Cooper, H.; Sturmberg, B. C. P.; Shaw, M. E.; and Blackhall, L. 2021. Applying responsible algorithm design to neighbourhood-scale batteries in Australia. *Nature Energy*, 6(8): 815–823. Publisher: Nature Publishing Group.
- Rebitschek, F. G.; Gigerenzer, G.; and Wagner, G. G. 2021. People underestimate the errors made by algorithms for credit scoring and recidivism prediction but accept even fewer errors. *Scientific Reports*, 11(1): 20171. Publisher: Nature Publishing Group.
- Rechkemmer, A.; and Yin, M. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*, 1–14. New Orleans LA USA: ACM. ISBN 978-1-4503-9157-3.
- Reichstein, M.; Benson, V.; Blunk, J.; Camps-Valls, G.; Creutzig, F.; Fearnley, C. J.; Han, B.; Kornhuber, K.; Rahaman, N.; Schölkopf, B.; Tárraga, J. M.; Vinuesa, R.; Dall, K.; Denzler, J.; Frank, D.; Martini, G.; Nganga, N.; Maddix, D. C.; and Weldemariam, K. 2025. Early warning of complex climate risk with integrated artificial intelligence. *Nature Communications*, 16(1): 2564. Publisher: Nature Publishing Group.
- Ribeiro, M. T.; Singh, S.; and 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*, 1135–1144.
- Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215. Publisher: Nature Publishing Group.
- Sacha, D.; Senaratne, H.; Kwon, B. C.; Ellis, G.; and Keim, D. A. 2016. The Role of Uncertainty, Awareness, and Trust in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 22(1): 240–249.
- Sagona, M.; Dai, T.; Macis, M.; and Darden, M. 2025. Trust in AI-assisted health systems and AI's trust in humans. *npj Health Systems*, 2(1): 1–5. Publisher: Nature Publishing Group.
- Salimzadeh, S.; He, G.; and 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*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.
- Saltelli, A.; Bammer, G.; Bruno, I.; Charters, E.; Di Fiore, M.; Didier, E.; Nelson Espeland, W.; Kay, J.; Lo Piano, S.; Mayo, D.; Pielke Jr, R.; Portaluri, T.; Porter, T. M.; Puy, A.; Rafols, I.; Ravetz, J. R.; Reinert, E.; Sarewitz, D.; Stark, P. B.; Stirling, A.; Van Der Sluijs, J.; and Vineis, P. 2020. Five ways to ensure that models serve society: a manifesto. *Nature*, 582(7813): 482–484.
- Salvi, M.; Seoni, S.; Campagner, A.; Gertych, A.; Acharya, U. R.; Molinari, F.; and Cabitza, F. 2025. Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. *International Journal of Medical Informatics*, 197: 105846.
- Schiff, D.; Biddle, J.; Borenstein, J.; and Laas, K. 2020. What's Next for AI Ethics, Policy, and Governance? A Global Overview. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, 153–158. New York, NY, USA: Association for Computing Machinery. ISBN 9781450371100.
- Schlesinger, D. E.; Alam, R.; Ringel, R.; Pomerantsev, E.; Devireddy, S.; Shah, P.; Garasic, J.; and Stultz, C. M. 2025. Artificial intelligence for hemodynamic monitoring with a wearable electrocardiogram monitor. *Communications Medicine*, 5(1): 1–8. Publisher: Nature Publishing Group.
- Schmidt, S.; Rolff, T.; Voigt, H.; Offe, M.; and Steinicke, F. 2024. Natural Expression of a Machine Learning Model's Uncertainty Through Verbal and Non-Verbal Behavior of Intelligent Virtual Agents. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 1–15. Pittsburgh PA USA: ACM. ISBN 979-8-4007-0628-8.

- Schneider, H.; Wayrauther, J.; Hassib, M.; and Butz, A. 2019. Communicating Uncertainty in Fertility Prognosis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11. Glasgow Scotland Uk: ACM. ISBN 978-1-4503-5970-2.
- Schoonderwoerd, T. A.; Jorritsma, W.; Neerincx, M. A.; and Van Den Bosch, K. 2021. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies*, 154: 102684.
- Schwabe, D.; Becker, K.; Seyferth, M.; Klaß, A.; and Schaeffter, T. 2024. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. *npj Digital Medicine*, 7(1): 1–30. Publisher: Nature Publishing Group.
- Sel, K.; Hawkins-Daarud, A.; Chaudhuri, A.; Osman, D.; Bahai, A.; Paydarfar, D.; Willcox, K.; Chung, C.; and Jafari, R. 2025. Survey and perspective on verification, validation, and uncertainty quantification of digital twins for precision medicine. *npj Digital Medicine*, 8(1): 1–12. Publisher: Nature Publishing Group.
- Sermesant, M.; Delingette, H.; Cochet, H.; Jaïs, P.; and Ayache, N. 2021. Applications of artificial intelligence in cardiovascular imaging. *Nature Reviews Cardiology*, 18(8): 600–609. Publisher: Nature Publishing Group.
- Shad, R.; Cunningham, J. P.; Ashley, E. A.; Langlotz, C. P.; and Hiesinger, W. 2021. Designing clinically translatable artificial intelligence systems for high-dimensional medical imaging. *Nature Machine Intelligence*, 3(11): 929–935.
- Silva, A.; Schrum, M.; Hedlund-Botti, E.; Gopalan, N.; and Gombolay, M. 2023. Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. *International Journal of Human-Computer Interaction*, 39(7): 1390–1404.
- Simmonds, E. G.; Adjei, K. P.; Cretois, B.; Dickel, L.; González-Gil, R.; Laverick, J. H.; Mandeville, C. P.; Mandeville, E. G.; Ovaskainen, O.; Sicacha-Parada, J.; Skarstein, E. S.; and O'Hara, B. 2024. Recommendations for quantitative uncertainty consideration in ecology and evolution. *Trends in Ecology & Evolution*, 39(4): 328–337. Publisher: Elsevier.
- Sivaraman, V.; Bukowski, L. A.; Levin, J.; Kahn, J. M.; and Perer, A. 2023. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 1–18. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9421-5. Event-place: Hamburg, Germany.
- Slack, D.; Hilgard, S.; Singh, S.; and Lakkaraju, H. 2021. Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*.
- Smith, A.; Kumar, V.; Boyd-Graber, J.; Seppi, K.; and Findlater, L. 2018. Closing the Loop: User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, IUI '18, 293–304. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-4945-1. Event-place: Tokyo, Japan.
- Spiegelhalter, D. 2017. Risk and uncertainty communication. *Annual Review of Statistics and Its Application*, 4(1): 31–60.
- Sun, H.; Breugel, B. v.; Crabbé, J.; Seedat, N.; and Schaar, M. v. d. 2023. What is Flagged in Uncertainty Quantification? Latent Density Models for Uncertainty Categorization. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Suresh, H.; Lewis, K. M.; Gutttag, J.; and Satyanarayan, A. 2022. Intuitively Assessing ML Model Reliability through Example-Based Explanations and Editing Model Inputs. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, 767–781. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9144-3. Event-place: Helsinki, Finland.
- Tabassi, E. 2023. NIST Trustworthy and Responsible AI Risk Management Framework. NIST AI 100-1.
- Thomas, I.; Oh, S. Y.; and Szafr, D. A. 2024. Assessing User Trust in Active Learning Systems: Insights from Query Policy and Uncertainty Visualization. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, 772–786. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0508-3. Event-place: Greenville, SC, USA.
- Thornton, L.; Knowles, B.; and Blair, G. 2021. Fifty Shades of Grey: In Praise of a Nuanced Approach Towards Trustworthy Design. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 64–76. Virtual Event Canada: ACM. ISBN 978-1-4503-8309-7.
- Tidjon, L. N.; and Khomh, F. 2023. The Different Faces of AI Ethics Across the World: A Principle-to-Practice Gap Analysis. *IEEE Transactions on Artificial Intelligence*, 4(04): 820–839.
- Toh, S. Y.; Cai, C.; Wang, L. R.; Bai, X.; Ngeow, J.; and Fan, X. 2025. The Effect of Explainable AI and Uncertainty Quantification on Medical Students' Perspectives of Decision-Making AI: A Cancer Screening Case Study. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–13. Yokohama Japan: ACM. ISBN 979-8-4007-1395-8.
- Tomsett, R.; Preece, A.; Braines, D.; Cerutti, F.; Chakraborty, S.; Srivastava, M.; Pearson, G.; and Kaplan, L. 2020. Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. *Patterns*, 1(4). Publisher: Elsevier.
- Tonekaboni, S.; Joshi, S.; McCradden, M. D.; and Goldberg, A. 2019. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *Proceedings of Machine Learning Research*.
- Ulmer, D.; Gubri, M.; Lee, H.; Yun, S.; and Oh, S. 2024. Calibrating Large Language Models Using Their Generations Only. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15440–15459. Bangkok, Thailand: Association for Computational Linguistics.

- Vaccaro, K.; Agarwalla, T.; Shivakumar, S.; and Kumar, R. 2018. Designing the future of personal fashion. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–11. New York, NY, USA: Association for Computing Machinery.
- van Breugel, B.; Liu, T.; Oglic, D.; and van der Schaar, M. 2024. Synthetic data in biomedicine via generative artificial intelligence. *Nature Reviews Bioengineering*, 2(12): 991–1004. Publisher: Nature Publishing Group.
- Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M. S.; and 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.
- Vishwakarma, G.; Sonpal, A.; and Hachmann, J. 2021. Metrics for Benchmarking and Uncertainty Quantification: Quality, Applicability, and Best Practices for Machine Learning in Chemistry. *Trends in Chemistry*, 3(2): 146–156. Publisher: Elsevier.
- Wang, X.; and Yin, M. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, IUI '21, 318–328. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8017-1. Event-place: College Station, TX, USA.
- Wang, Z.; Duan, J.; Cheng, L.; Zhang, Y.; Wang, Q.; Shi, X.; Xu, K.; Shen, H. T.; and Zhu, X. 2024. ConU: Conformal Uncertainty in Large Language Models with Correctness Coverage Guarantees. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 6886–6898. Miami, Florida, USA: Association for Computational Linguistics.
- Whittlestone, J.; Nyrup, R.; Alexandrova, A.; and Cave, S. 2019. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, 195–200. New York, NY, USA: Association for Computing Machinery. ISBN 9781450363242.
- Wiedeman, C.; and Wang, G. 2024. Decorrelative network architecture for robust electrocardiogram classification. *Patterns*, 5(12). Publisher: Elsevier.
- Wintersberger, P.; Dmitrenko, D.; Schartmüller, C.; Frison, A.-K.; Maggioni, E.; Obrist, M.; and Riener, A. 2019. S(C)ENTINEL: monitoring automated vehicles with olfactory reliability displays. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, 538–546. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6272-6. Event-place: Marina del Ray, California.
- Wu, D.; Gu, J.-C.; Yin, F.; Peng, N.; and Chang, K.-W. 2024. Synchronous Faithfulness Monitoring for Trustworthy Retrieval-Augmented Generation. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9390–9406. Miami, Florida, USA: Association for Computational Linguistics.
- Xu, Z.; Yue, X.; Lv, Y.; Liu, W.; and Li, Z. 2023. Trusted Fine-Grained Image Classification through Hierarchical Evidence Fusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9): 10657–10665.
- Yang, F.; Mortenson, C. R.; Nisbet, E.; Diakopoulos, N.; and Kay, M. 2024. In Dice We Trust: Uncertainty Displays for Maintaining Trust in Election Forecasts Over Time. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, 1–24. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0330-0. Event-place: Honolulu, HI, USA.
- Yin, F.; Srinivasa, J.; and Chang, K.-W. 2024. Characterizing truthfulness in large language model generations with local intrinsic dimension. In *Proceedings of the 41st International Conference on Machine Learning*, ICML '24. Vienna, Austria: JMLR.org.
- Yin, M.; Wortman Vaughan, J.; and Wallach, H. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. Glasgow Scotland Uk: ACM. ISBN 978-1-4503-5970-2.
- Yona, G.; Aharoni, R.; and Geva, M. 2024. Can Large Language Models Faithfully Express Their Intrinsic Uncertainty in Words? In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7752–7764. Miami, Florida, USA: Association for Computational Linguistics.
- Zaidan, E.; and Ibrahim, I. A. 2024. AI Governance in a Complex and Rapidly Changing Regulatory Landscape: A Global Perspective. *Humanities and Social Sciences Communications*, 11(1): 1–18. Publisher: Palgrave.
- Zerilli, J.; Bhatt, U.; and Weller, A. 2022. How transparency modulates trust in artificial intelligence. *Patterns*, 3(4): 100455. Publisher: Elsevier.
- Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. E. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. Barcelona Spain: ACM. ISBN 978-1-4503-6936-7.
- Zhao, J.; Wang, Y.; Mancenido, M. V.; Chiou, E. K.; and Maciejewski, R. 2024. Evaluating the Impact of Uncertainty Visualization on Model Reliance. *IEEE Transactions on Visualization and Computer Graphics*, 30(7): 4093–4107.
- Zhou, K.; Hwang, J.; Ren, X.; and Sap, M. 2024. Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3623–3643. Bangkok, Thailand: Association for Computational Linguistics.
- Zuchowski, L. C.; Zuchowski, M. L.; and Nagel, E. 2024. A trust based framework for the envelopment of medical AI. *npj Digital Medicine*, 7(1): 1–11. Publisher: Nature Publishing Group.