Bus Inf Syst Eng
https://doi.org/10.1007/s12599-026-01015-y

RESEARCH NOTES

Beware of "Explanations" of AI

David Martens · Galit Shmueli · Theodoros Evgeniou · Kevin Bauer · Christian Janiesch · Stefan Feuerriegel · Sebastian Gabel · Sofie Goethals · Travis Greene · Nadja Klein · Mathias Kraus · Niklas Kühl · Claudia Perlich · Wouter Verbeke · Alona Zharova · Patrick Zschech · Foster Provost

Received: 15 November 2025 / Accepted: 22 June 2026
© The Author(s) 2026

Abstract Understanding the decisions made and actions taken by increasingly complex AI systems remains a key challenge. This has led to an expanding field of research in explainable artificial intelligence (XAI), highlighting the potential of explanations to enhance trust, support adoption, and meet regulatory standards. However, the question of what constitutes a "good" explanation is dependent on the goals, stakeholders, and context. At a high level, psychological insights such as the concept of mental model alignment can offer guidance, but success in practice is

Accepted after 2 revisions by Stefan Lessmann

D. Martens · S. Goethals
Department of Engineering Management, University of Antwerp, 2000 Antwerp, Belgium

G. Shmueli
Institute of Service Science, National Tsing Hua University, Hsinchu 30013, Taiwan

T. Evgeniou
Technology and Business, INSEAD, 77300 Fontainebleau, France

K. Bauer
Department of Information Systems, Goethe University Frankfurt, 60629 Frankfurt, Germany

C. Janiesch (✉)
Department of Computer Science, TU Dortmund University, 44227 Dortmund, Germany
e-mail: christian.janiesch@tu-dortmund.de

S. Feuerriegel
LMU Munich and Munich Center for Machine Learning, 80539 Munich, Germany

S. Gabel
Rotterdam School of Management, Erasmus University, 3062 Rotterdam, The Netherlands

T. Greene
Department of Digitalization, Copenhagen Business School, 2000 Copenhagen, Denmark

N. Klein
Scientific Computing Center, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany

M. Kraus
Faculty of Informatics and Data Science, University of Regensburg, 93053 Regensburg, Germany

N. Kühl
Faculty of Law, Business and Economics, University of Bayreuth, 95440 Bayreuth, Germany

N. Kühl
Fraunhofer Institute for Applied Information Technology FIT, 95440 Bayreuth, Germany

C. Perlich · F. Provost
Leonard N. Stern School of Business, New York University, New York, NY 10012, USA

W. Verbeke
Faculty of Economics and Business, KU Leuven, 3000 Leuven, Belgium

A. Zharova
School of Business and Economics, Humboldt-Universität zu Berlin, 10099 Berlin, Germany

P. Zschech
Faculty of Business and Economics, TUD Dresden University of Technology, 01069 Dresden, Germany

Published online: 25 August 2026

challenging due to social and technical factors. As a result of this ill-defined nature of the problem, explanations can be of poor quality (e.g., unfaithful, irrelevant, or incoherent), potentially leading to substantial risks. Instead of fostering trust and safety, poorly designed explanations can actually cause harm due to wrong decisions, privacy violations, manipulation, and reduced AI adoption. Therefore, we caution stakeholders to beware of explanations of AI: While they can be vital, they are not automatically a remedy for transparency or responsible AI adoption, and their misuse or limitations can exacerbate harm. Attention to these caveats can help guide future research to improve the quality and impact of AI explanations.

1 Introduction

As Artificial Intelligence (AI) systems[1] are increasingly being used in organizations and society, and have a growing impact on decision making and our daily lives, trusting their behavior is crucial. State-of-the-art AI models, such as XGBoost or Random Forests for structured data or deep neural networks for unstructured data like images and text, can achieve remarkable predictive performance. However, their complexity renders them *black boxes*, inherently uninterpretable by humans (Martens and Provost 2014; Janiesch et al. 2021; Rudin et al. 2019; Kruschel et al. 2026). This opacity introduces challenges in areas such as error understanding and prevention, fostering user trust, and meeting regulatory requirements not only in form but also in spirit (European Commission 2021). Unsurprisingly, the growing demand for safe and responsible AI adoption by organizations and individuals has heightened the need for clear, truthful and actionable explanations.

When AI outputs can be reliably traced back through the underlying processes, stakeholders – including regulators, managers, AI developers, and those affected by these outputs – can examine not only the outputs but also the inputs, data and methodologies driving them (Dwivedi et al. 2023). Safeguards are indispensable as AI systems increasingly impact critical aspects of society, from healthcare and criminal justice to hiring and financial services, where the stakes can be high (European Commission 2021). However, there is often a notable mismatch between how and when explainable AI (XAI) methods are deployed and the ways humans comprehend and make decisions (Poursabzi-Sangdeh et al. 2021; Alufaisan et al. 2021). This gap arises largely from socio-technical complexities of interpreting AI outputs (Herm et al. 2023; Babic et al. 2021). Current XAI research focuses predominantly on technical aspects, often overlooking the psychological, social, and contextual factors that shape human understanding and decision making. In fact, much XAI research is arguably driven by the interests of the researchers rather than actual AI application needs, a phenomenon aptly described as the "inmates running the asylum" (Miller et al. 2017).

Based on our long and varied experience both as XAI scholars and as XAI practitioners, we advocate for a shift in both research and practice toward addressing critical questions necessary for enabling explanations to support the safe and responsible adoption of AI. We highlight how current explanations of AI decisions may create an illusion of accountability rather than fostering improved understanding, therefore also leaving the true reasons for the decisions inadequately scrutinized. Accordingly, we caution "Beware of Explanations of AI": They may not (always) provide the remedy sought, and may even cause downstream harms to individuals and groups, for example, violation of civil liberties or monetary loss (Wendt 2025). To address these limitations, we explore root causes of potential harm by identifying critical areas where increased awareness and scrutiny are needed to ensure that explanations contribute meaningfully to transparency, trust, accountability, and eventually the safe adoption of AI.

This research note contributes to the discourse on trustworthy and responsible AI by reframing AI explainability as a socio-technical design challenge rather than a purely technical one. We highlight that explanation quality depends fundamentally on stakeholder goals, context, and the alignment of explanations with human mental models. Our synthesized socio-technical root causes and associated caveats provide a structured lens for evaluating when and how explanations create value in real-world AI systems. By shifting attention from how to generate explanations to *when explanations are appropriate and beneficial*, this work offers actionable insights for researchers and practitioners designing, implementing, and using explainable AI in organizational contexts.

2 A Socio-Technical Understanding of AI Explanations

2.1 What Research in AI Explanations Provide Us so Far

Explainable AI can be defined as the research field that develops and applies algorithms to explain AI system outputs (Martens 2022). Although the term *XAI* is

[1] AI systems incorporate AI models, which are usually the product of machine learning. AI systems go beyond the mere prediction model, as they may use multiple models predicting different things, can have complex decision logic connecting the predictions to actions, and usually include further business rules, including guard rails, black lists and white lists, etc.

relatively recent (generally attributed to a DARPA project (Gunning et al. 2021)), the underlying need for explainability for AI system outputs has been recognized and addressed since AI systems started to make decisions that could affect people and the world. The 1970s early medical AI system MYCIN included a significant explanation component (Buchanan and Shortliffe 1984). In the 1980s, Michalski (1983) proposed the "comprehensibility postulate," emphasizing that the results of machine learning should be interpretable in natural language. In the 1990s, groundbreaking research was underway in understanding black-box machine learning models like neural networks (Craven and Shavlik 1995). Explainability is not merely a modern concern but a foundational aspect of building effective AI systems.

AI explanations of system behaviors – meaning the AI system's predictions, decisions, or actions – aim to render the behaviors of complex model-driven AI systems comprehensible to humans. Put differently: AI explanations show why or how a system produced a specific output from a specific input. The design of AI explanations typically aims to address challenges associated with the black-box nature of AI, such as distrust, lack of accountability, insufficient error safeguarding, and limited knowledge discovery, without sacrificing the quality of the AI predictions, decisions, or actions (Bauer et al. 2021; Martens 2022).

The irony is that contemporary XAI methods aim to address the issue of complex black-box models by adding another technical layer of complex XAI methods. Many XAI methods rely on a secondary algorithm designed to reveal (post hoc) important aspects of the behavior of an opaque primary AI system or the model driving it, such as a deep neural network. Thus, explainability is explicitly decoupled from the model itself, resulting in two distinct systems: the AI system (incorporating a model such as a neural network) and the algorithm responsible for generating explanations about the model or system's behavior (e.g., feature importance explanations like those produced by SHAP (Lundberg and Lee 2017)) or counterfactual explanation methods (Martens and Provost 2014; Wachter et al. 2018). These post-hoc explanation methods often are model-agnostic, meaning they can be applied to any predictive model regardless of its underlying architecture. Current explainability techniques are broadly categorized into *global* methods, which aim to explain the overall behavior of the model, and *local* methods, which focus on explaining specific outputs (Martens and Provost 2014). Popular approaches, such as counterfactual explanations (Martens and Provost 2014), saliency maps (Martens 2022), and feature importance explanations (Lundberg and Lee 2017; Ribeiro et al. 2016), establish relationships between the input features X and the model or system outputs $\hat{Y}$ to provide insights into why a particular output was – or was not – produced for a given input. By doing so, these methods enable individuals to better understand the model's input–output transformation.

However, despite decades of progress, much of current XAI research remains focused on developing technical solutions, often at the expense of answering, or even considering, the fundamental question: What constitutes a "good" explanation? For example, much of current XAI research remains centered on defining importance of a feature in some manner (e.g., highest gradient in target prediction (Selvaraju et al. 2017)), largest approximate Shapley value (Lundberg and Lee 2017), largest coefficient in a linear surrogate model (Ribeiro et al. 2016) and then developing algorithms to generate those. As we will illustrate in the following sections, this technical focus frequently neglects actual stakeholder goals, fails to adequately mitigate the challenges posed by black-box models, and leaves critical aspects of explanation quality, such as clarity, context, and relevance to stakeholders, insufficiently addressed or explored (Herm et al. 2022). This in turn limits the practical utility of XAI to address the important challenges we summarized at the outset. But why is it so difficult to simply define properly what a good AI explanation is, and develop an algorithm that universally solves this?

2.2 What are (Good) Explanations?

Definitions of what constitutes an explanation vary significantly across fields (Garfinkel 1982). Some scholars regard explanations as answers to "why" or "how" questions (Wellman 2011). Others conceptualize explanations as hypotheses linking causes to effects (Einhorn and Hogarth 1986; Thagard 2000). Still others emphasize that explanations should convey a sense of causality that helps to make sense of a given phenomenon.

This diversity in definitions underscores the complexity of explanations. The role and purpose of explanations can become more apparent when viewed at a higher level through the lens of cognitive psychology. Central to the conceptualization of explanations is the concept of a *mental model*, rooted in cognitive psychology. Mental models are "all forms of mental representation, general or specific, from any domain, causal, intentional, or spatial" (Brewer 1987), encoding beliefs, facts, and knowledge. Explanations can thus be effectively understood as

(mental) models that replicate the structure and dynamics of real-world phenomena, enabling both *understanding* and *prediction* of the phenomena in question (Craik 1967). Consistent with this view, Johnson-Laird emphasized that "understanding consists in having a working model of the phenomenon in your mind" (Johnson-Laird 1983).

Now, what are good *AI explanations*? Wachter et al. (2018) define an AI explanation as "an attempt to convey the internal state or logic of an algorithm that leads to a decision". This definition of an explanation as an 'attempt' already indicates the complex nature of the problem. If we build on the prior, more general definitions, we can define a good AI explanation as one where there is a good fit between the mental model of the explanation recipient and the provided explanation (Kayande et al. 2009; Martens and Provost 2014). This fit naturally depends on the particular recipient, the goal of the recipient in using the (X)AI system, and the particular circumstances (Wanner et al. 2022), which motivates why coming up with a "good" explanation is currently an ill-defined problem. This challenge is compounded by the technical fact that even given a particular notion of what makes for a satisfactory explanation, there can be multiple satisfactory explanations for the same input–output pair (Martens and Provost 2014), and in addition, the set of explanations produced by some methods may not be robust to minor input changes. Let us examine how a good explanation depends on the goal, stakeholder, and context.

2.2.1 Societal Dimensions

There are at least five key purposes for providing AI explanations (Martens and Provost 2014): (1) Increasing acceptance and likelihood of adoption of an AI system (Gregor and Benbasat 1999), where explanations increase trust by validating the discovered patterns, and relatedly, improving the relationship between managers and system developers; (2) justifying individual system decisions (at inference/deployment time), e.g., to ensure or increase customer satisfaction and/or to satisfy ethical principles, such as providing user privacy and control (Chen et al. 2017); (3) revealing erroneous reasoning (at inference time), and thereby allowing it to be corrected; (4) improving AI models and improving the machine learning that produces them (e.g., by adding guardrails to a model or by improving training data); (5) improving long-term learning (Gregor and Benbasat 1999) by all stakeholders (learning about the world, the business domain, the company, the models' reasoning processes, machine learning).

As explanations are never given or received in a social vacuum, it is pivotal to also consider the diversity in *stakeholders* throughout the entire machine-learning life cycle, encompassing an interconnected XAI ecosystem with partially competing interests (see Fig. 1). During the development phase, *AI researchers*, *data scientists*, and *domain experts* rely on explanations to refine models, typically prioritize technical performance rather than utility for users, and drive innovation. For these stakeholders, explanations serve as tools for debugging, feature selection, training data improvement, and understanding system/model behavior to improve decision making, overall reliability and model accuracy.

As AI systems move from development to deployment, the range of stakeholders expands notably. Practitioners such as (non-machine learning) *engineers* and *managers* use explanations to ensure operational reliability, troubleshoot unexpected system behaviors, and justify AI system outputs to *customers, clients, users, decision-makers,* and *other stakeholders*. In this phase, explanations are often required to bridge the gap between the technical intricacies of the AI system and the practical demands of organizational and business environments. Those affected by the decisions (which may include some of the aforementioned) represent another critical group to consider during deployment, with needs that vary based, for example, on their level of expertise and the context in which they interact with the system. For some, simplified narratives (Martens et al. 2025) may be sufficient to build acceptance of the system's actions, while others may require more detailed, technically rigorous justifications.

Beyond internal stakeholders and those directly affected by the system, there are other external stakeholders, including *data subjects*, *regulators*, and *auditors*. They introduce another layer of complexity. These stakeholders may demand explanations that demonstrate compliance with ethical, legal, and business standards, often requiring rigorous and transparent documentation. The needs of this group can diverge significantly from those of other stakeholders, emphasizing a level of formalization and accountability that may conflict with the simplicity sought by end-users, the trust-building needed by managers, or the efficiency and efficacy desired by developers.

And thirdly, the *context* is a key driver of what explanation is considered suitable or good from a societal perspective. For example, in high-risk applications, such as credit scoring or recruitment, much higher demands will be set on explanations than in low-risk applications such as targeted advertising. This contextual issue is even the basis of the risk-based approach taken in the European AI Act (European Commission 2021). These regulatory and societal requirements are yet another contextual dimension that will determine the needs for an AI explanation.

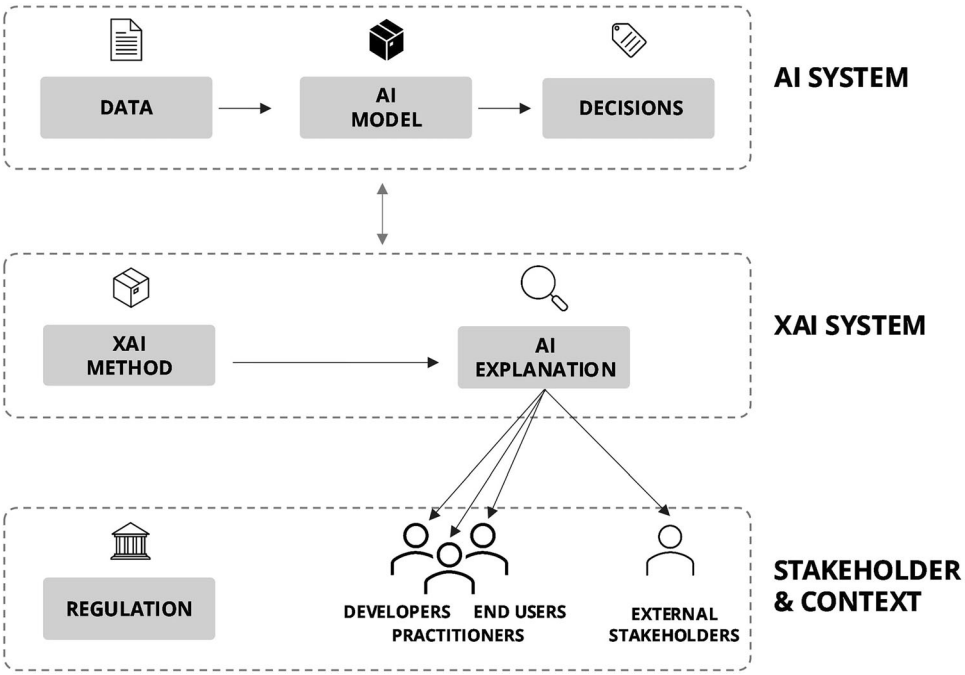

Fig. 1 The XAI ecosystem includes the AI and XAI systems with their data and algorithmic components, as well as different human stakeholders and context

2.2.2 *Technical Dimensions*

A fundamental disconnect exists between the probabilistic nature of AI systems and the expectation that explanations provide crisp causal clarity. Further, even data scientists experienced in causal inference have trouble with the key explanatory distinction between causality at the AI system/model level and causality "in the real world." The question, "why did the system deny me credit," is different from the question "why am I not credit-worthy." AI systems do not necessarily faithfully represent underlying causal processes, and often do not need to (e.g., relying instead on accurate proxies). Thus, while an AI model might produce accurate predictions using probabilistic methods, explanation recipients might be expecting a revelation of the latent causal mechanisms driving the actual phenomenon being modeled.

This limitation is worsened by the selective nature of explainability methods. A single model (global explanation) or an individual output (local explanation) can yield multiple, distinct explanations (Martens and Provost 2014; Barocas et al. 2020; Brughmans et al. 2024). While multiple explanations may be reasonable, such variability can make them seem inconsistent and unreliable, raising concerns about objectivity and interpretability. These methods, though capable of generating plausible narratives, may disappoint when users expect definitive or universal insights – an expectation often unrealistic (or even not desired (Thomas and Uminsky 2022)) due to context-dependent explanation goals. Providing a single explanation is further complicated by competing priorities: Granular explanations useful for developers may overwhelm end-users, while simplified versions may not meet regulatory needs. Achieving balance requires flexible, adaptable systems, presenting significant design and computational challenges.

3 The XAI Causes and Caveats: Stakeholders Beware

Given that the literature does not provide a clear understanding of what constitutes a universally "good" explanation, and given the wealth of different explanation goals and stakeholders, there are likely many context-dependent answers to that question. Thus, AI explanations in practice may not be tailored to the context. This leads to a set of root causes of "bad" explanations, some related to the quality of explanations in general (as noted in the literature beyond AI), and some due to particular characteristics of XAI such as the ill-defined nature of the problem, the statistical nature of AI, or the key role of individual data in AI systems' decision making. In particular, the following characteristics of AI explanations are root *causes* of possible downstream harms. We illustrate them each with an example from credit scoring.

AI explanations may be:

a. *Unfaithful*: The explanation is inconsistent with the AI model and data, or inconsistent more generally.

> Example: The explanation cites "employment stability" while this is not used by the model.

b. *Irrelevant*: The explanation does not match the context or user needs, or provides excessive or insufficient detail.

> Example: The explanation cites all available features, while most do not have an impact on the decision made for the user, or even for the prediction score.

c. *Unstable*: Explanations produced by XAI may differ even when the same data and AI model are used.

> Example: The approximate SHAP explanation, generated twice with different random seeds, provides two different sets of 'important' features.

d. *Incoherent*: The explanation is too complex or unclear for the user, or is in conflict with the user's mental model without being clear why, or contains inconsistent elements that undermine understanding.

> Example: A SHAP explanation provided to a non-technical lay user, as the reason for rejected credit.

e. *Revealing*: The explanation exposes personal, sensitive, or proprietary information.

> Example: A counterfactual explanation: "If your income were higher than $3,000, your loan would have been approved" reveals the bank's binning threshold for income, which in this case determines the approval outcome.

f. *Unnecessary*: The explanation is unnecessary or redundant for the user's needs.

> Example: A SHAP explanation that lists the top 100 important features to a rejected applicant, while a counterfactual explanation would reveal that increasing income by $200 alone would lead to the AI system accepting the applicant.

g. *Cherry-picked*: The explanation selectively highlights evidence, ignoring contradictory, sensitive, or proprietary information not intended for sharing with the public.

> Example: A counterfactual explanation: "If your income were higher than $3,000, your job tenure 10 years longer, and your amount of loan $100,000 lower, you would have been approved for the loan" while the explanation "If your gender was male instead of female you would be approved" is intentionally not shown.

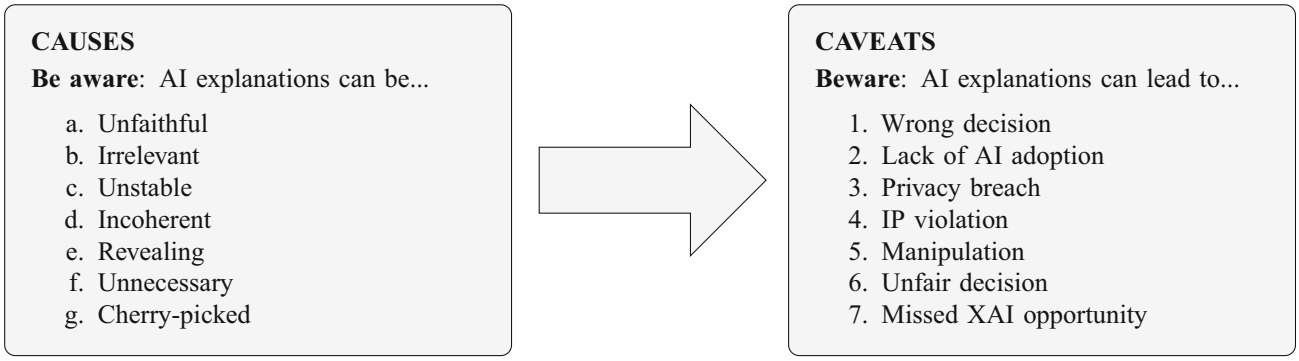

Fig. 2 Due to the ill-defined nature of AI explanations, their quality can be poor for various reasons (causes), resulting in multiple caveats to beware of

These limitations and inferior AI explanations can lead to risks – which we collectively call *caveats*.[2] Each root cause highlights an underlying characteristic of an explanation, such as unfaithfulness or irrelevancy, that (possibly in combination) gives rise to the caveats (see Fig. 2). The relationships between causes and caveats are complex, involving multiple causes for each caveat and interactions between them. For instance, a stakeholder might make a wrong decision (Caveat 1) because the explanation is unfaithful (Cause a) to the underlying model, or conversely, because the explanation is technically accurate but incoherent (Cause d), causing the stakeholder to misunderstand the output. Thus, we argue that prudent users of XAI should beware of the set of caveats and their possible root causes. We now elaborate why, illustrating with examples from multiple domains.

1. *Explanations can lead to wrong decisions:* Automation bias, combined with convincing but poorly chosen explanations, can prompt stakeholders to trust and follow flawed AI outputs, even when those may be harmful to them. This overtrust can lead to wrong decisions with serious consequences, such as a doctor relying on flawed AI advice for treatment, simply because the explanation is not in disagreement with people's own mental model (while the AI model is). The explanation can thus unintentionally create an illusion of understanding, where users believe they grasp the underlying logic of a predictive model despite its complexity remaining beyond their comprehension (Rozenblit and Keil 2002; Cabitza et al. 2024). This illusion can be further compounded by hindsight bias, where users, after receiving an AI explanation, falsely assume they could have predicted the output (Keil 2006). Finally, explanations that fail to offer actionable insights may also result in confusion or fail to guide decision making (Liao et al. 2022), once more potentially leading to poor decisions.

> Example: A judge using an AI system to predict probability of recidivism throws out a plea agreement between the prosecutor and the defendant, purely based on the AI explanation indicating that the most important features in the prediction are the defendant's unstable employment, lack of community support, and his criminal record. All this while the AI model predicts a low risk of recidivism.

2. *Explanations can lead to lack of adoption of AI and missed opportunities:* Unfaithful or overly complex explanations can cause misplaced distrust, leading users to question quality AI outputs that may in fact be helpful. While explanations can build trust and positively impact AI adoption, poorly chosen explanations can have counterproductive effects. When AI explanations fail to align with stakeholders' mental models, they may hinder rather than help adoption, even if the AI system has the potential to improve outcomes. Explanations that are irrelevant or incomprehensible can exacerbate this distrust (Papenmeier et al. 2019). They may even provide users with an excuse to discard relevant or helpful AI outputs that they do not agree with for strategic or subjective reasons.

[2] Similar to the term *caveat emptor* meaning 'buyer beware,' the implication is that the burden rests on the buyer to assess the quality of goods.

Example: A credit officer receives SHAP explanations of the 20 most important features, but doesn't understand the meaning of the SHAP values, and is overwhelmed with the complexity of the set of features. In the officer's mind, income of the applicant is the only reason for rejection. Therefore, the credit officer does not trust the AI model to work well and ignores its predictions, though it would substantially improve its lending practices.

3. *Explanations can lead to privacy violations:* AI explanations, especially when generated interactively or repeatedly for multiple input data points, could expose personal, sensitive information that was used to train the AI model. For instance, when explanations are generated using instance-based strategies, they may reveal sensitive details about individuals in the training data. A notable risk is the explanation linkage attack, where adversaries link counterfactual explanations to background information, potentially identifying and extracting private attributes of individuals (Goethals et al. 2023b). Another type of attack is a membership inference attack, where it is determined whether a specific individual is part of the training data (Shokri et al. 2020; Naretto et al. 2022).

Example: The hospital technician repeatedly requests explanations from the AI diagnostic system for the same patient. Each time, different features such as age, family history, or genetic markers are highlighted. By combining the explanations with publicly known patient attributes, the technician can re-identify the person as a certain colleague and can infer that the colleague carries a high-risk genetic condition.

4. *Explanations can lead to IP leakage:* Not only sensitive information about persons might be revealed, but also proprietary information can be revealed through AI explanations, providing competitors with sensitive details that undermine the explainer's strategic position. For example, trade secrets embedded in the AI model or unique patterns in the data may be exposed via explanations, giving rivals an unintended advantage. Also, given enough explanations, adversaries may be able to reverse engineer the AI model used (Aïvodji et al. 2020; Yan et al. 2023). This risk is rooted in AI explanations being overly transparent, where they reveal more than intended about the underlying system.

Example: A credit card company uses AI models to detect fraud based on transaction amount, frequency, location, and account history. Flagged transactions trigger detailed explanations, such as "round amounts" (e.g., exactly $500.00) or "abnormal overseas transfers" (e.g., a transaction in China within 10 hours of an in-store purchase in the U.S.). These explanations help fraudsters adapt, by avoiding round amounts, splitting transfers, spacing out transactions, or using VPNs, allowing them to evade future detection while still committing fraud.

5. *Explanations can be used to manipulate stakeholders:* Explanations, particularly when misused by malicious actors or those with conflicting incentives, can become tools for manipulation. Those actors might intentionally mislead stakeholders, possibly at scale. Given the multiplicity of possible explanations, providers could exploit this flexibility to manipulate while still claiming correctness in a technical XAI sense (Goethals et al. 2023a). The possibility of cherry-picked or unfaithful AI explanations exacerbates this risk, as they allow malicious actors to control narratives for personal gain.

Example: After an accident involving its autonomous vehicle, the manufacturer releases explanations pointing to poor road conditions or human errors as the primary causes. The explanation omits any mention of the vehicle's delayed braking system, and as such shifts blame away from the AI system. The explanations can still be correct, but are not the only possible explanations or contributing factors for the accident.

6. *Explanations can lead to unfair decision making:* Explanations may support intentional or unintentional unethical behavior, such as discrimination (Deck et al. 2024). For example, AI explanations might cherry-pick details to intentionally mislead, such as omitting gender or a clearly correlated feature when the underlying model makes decisions that are unfavorable, for example, to a protected class. Similarly, they might fabricate unfaithful

explanations via a surrogate model that are not aligned with the actual AI model's reasoning, such as falsely attributing a decision to insufficient income when income has no impact on the AI system's output. This kind of misuse can also include efforts to avoid liability (fairwashing (Aïvodji et al. 2019)), or provide minimal information to deter further inquiries ("keeping users at bay"). These tactics exploit the multiplicity of possible explanations, enabling providers to choose (or manipulate) narratives while maintaining an appearance of technical correctness and compliance with AI regulations.

> Example: Automatic discarding of explanations that reveal the use of sensitive information that is illegal to use in credit scoring. For example, in many countries, it is illegal to use information on gender in credit risk modeling. To conceal the use of this information, banks could cherry-pick explanations that do not include any sensitive attributes. These alternative explanations allow the bank to maintain a misguided perception of fairness and transparency.

7. *Explanations can lead to missed opportunity for XAI value creation:* Overemphasis on explainability can lead to significant resource allocation for AI explanations that may not be needed in a given context. For example, providing detailed counterfactual explanations for low-stakes decisions, such as product recommendations, or to users who do not desire them (e.g., those who were provided credit), can detract from other priorities. This risk is linked to caveats such as irrelevant or not-needed explanations, where the provided explanations are misaligned with user needs or the decision's importance. Developing processes to assess when and how explanations add value is critical to avoid wasting resources.

> Example: A credit card company uses AI models to detect fraud based on transaction patterns. To enhance transparency, it provides SHAP-based explanations not only for flagged fraudulent transactions but also for all non-fraudulent transactions, informing customers why their transactions were not flagged. Given that each explanation costs $0.01 in computing power and the company processes 100 million transactions daily, this results in excessive costs for explanations that customers do not want.

4 Discussion and Conclusion

Developers and users of AI and XAI should beware of the caveats of AI explanations and their root causes, keeping in mind the various ways explanations can cause harms. While XAI holds substantial promise for enhancing trust and transparency in an age of complex AI systems – and can do so if employed sensibly – it also introduces new or scales up existing socio-technical challenges. Research and development of XAI is still not focused on what constitutes a "good" explanation for the intended context, which risks oversimplification or misuse. We acknowledge that XAI is vital, but we caution against indiscriminate adoption. Instead, we urge all stakeholders, be they consumers, data scientists, researchers, managers, or policy makers, to recognize the complexity and caveats. Approach XAI with caution, as AI explanations are not a silver bullet for transparent AI-driven decision making.

Moving forward, instead of merely focusing on technical XAI methods and tools, we should aim for explanations that genuinely serve their intended purpose as the harm from a poor explanation could exceed that of a black-box prediction. The caveats underscore the importance of interdisciplinary collaboration, combining insights from machine learning, psychology, philosophy, information systems, and human-computer interaction. Specifically, the claim that the "inmates running the asylum" highlights the need for more investigation into the actual, varied needs for AI explanations in practice, and how well they map to XAI research. Ideally such an investigation would involve substantial fieldwork.

As a result, we specifically foresee important areas for future research: First, we should empirically study the frequency and severity of each cause and caveat, as well as the mapping between them. Second, we should advance technical work (inspired by the interdisciplinary insights) that specifically detects, measures and resolves a certain cause or caveat. Existing approaches such as differential privacy (Patel et al. 2022) or *k*-anonymity (Goethals et al. 2023b) for explanations are good pointers in that direction. Third, we should again zoom out and better define the problem given the purpose, stakeholder, and context, motivating the development of both, context-aware explanation systems, which adaptively select or generate

explanations accordingly, and objective metrics to quantify the mental-fit concept.

Finally, on a social level, we should guide thoughtful regulation designed to ensure that what is desirable is indeed achieved with minimal negative side effects, and to heighten awareness to ensure explanations are applied meaningfully and responsibly. Currently, end-users must do their own due diligence to assess the quality of explanations given to them. However, our position is consistent with legal proposals (Yew and Hadfield-Menell 2022) that would shift the burden of proof on XAI developers to preemptively disclose the context-sensitive risks and harms of AI explanations to regulators. *Caveat procurator!*

Funding Open Access funding enabled and organized by Projekt DEAL. Galit Shmueli received funding from the National Science and Technology Council (114-2410-H-007-092-MY3). Stefan Feuerriegel received funding from the Deutsche Forschungsgemeinschaft (530066172). Sebastian Gabel received funding from the Dutch Research Council (VI.Veni.221E.073). Nadja Klein received funding from the Bundesministerium für Forschung, Technologie und Raumfahrt (16IS24082B). Mathias Kraus and Patrick Zschech received funding from the Bundesministerium für Forschung, Technologie und Raumfahrt (01IS22080). Wouter Verbeke received funding from Research Foundation Flanders (G0AU025N & G038723N). Foster Provost thanks Ira Rennert and the NYU/Stern Fubon Center for support.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Aïvodji U, Arai H, Fortineau O, Gambs S, Hara S, Tapp A (2019) Fairwashing: the risk of rationalization. In: International Conference on Machine Learning (PMLR). pp 161–170

Aïvodji U, Bolot A, Gambs S (2020) Model extraction from counterfactual explanations. arxiv:2009.01884 preprint

Alufaisan Y, Marusich LR, Bakdash JZ, Zhou Y, Kantarcioglu M (2021) Does explainable artificial intelligence improve human decision-making? In: Proc AAAI Conf Artif Intell 35:6618–6626. https://doi.org/10.1609/aaai.v35i8.16819

Babic B, Gerke S, Evgeniou T, Cohen IG (2021) Beware explanations from AI in health care. Science 373(6552):284–286. https://doi.org/10.1126/science.abg1834

Barocas S, Selbst AD, Raghavan M (2020) The hidden assumptions behind counterfactual explanations and principal reasons. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*), ACM, New York, pp 80–89, https://doi.org/10.1145/3351095.3372830

Bauer K, Hinz O, van der Aalst W, Weinhardt C (2021) Expl(AI)n it to me – Explainable AI and information systems research. Bus Inf Syst Eng 63(2):79–82. https://doi.org/10.1007/s12599-021-00683-2

Brewer WF (1987) Schemas versus mental models in human memory. In: Morris P (ed) Modelling cognition. Wiley, Chichester, pp 187–197

Brughmans D, Melis L, Martens D (2024) Disagreement amongst counterfactual explanations: how transparency can be misleading. TOP 32:1–33. https://doi.org/10.1007/s11750-024-00670-2

Buchanan BG, Shortliffe EH (1984) Rule based expert systems: the MYCIN experiments of the Stanford heuristic programming project (The Addison-Wesley Series in Artificial Intelligence). Addison-Wesley, London

Cabitza F, Fregosi C, Campagner A, Natali C (2024) Explanations considered harmful: the impact of misleading explanations on accuracy in hybrid human-AI decision making. In: World Conference on Explainable Artificial Intelligence. Springer, Heidelberg, pp 255–269. https://doi.org/10.1007/978-3-031-63803-9_14

Chen D, Fraiberger SP, Moakler R, Provost F (2017) Enhancing transparency and control when drawing data-driven inferences about individuals. Big Data 5(3):197–212. https://doi.org/10.1089/big.2017.0074

Craik K (1967) The nature of explanation, vol 445. CUP Archive

Craven M, Shavlik J (1995) Extracting tree-structured representations of trained networks. In: Proceedings of the 12th International Conference on Machine Learning. pp 24–32

Deck L, Schoeffer J, De-Arteaga M, Kühl N (2024) A critical survey on fairness benefits of explainable AI. In: 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT), Rio de Janeiro, pp 1579–1595, https://doi.org/10.1145/3630106.3658990

Dwivedi R, Dave D, Naik H, Singhal S, Rana OF, Patel P, Qian B, Wen Z, Shah T, Morgan G, Ranjan R (2023) Explainable AI (XAI): core ideas, techniques, and solutions. ACM Comput Surv 55(9):194:1–194:33, https://doi.org/10.1145/3561048

Einhorn HJ, Hogarth RM (1986) Judging probable cause. Psychol Bull 99(1):3–19. https://doi.org/10.1037/0033-2909.99.1.3

European Commission (2021) Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

Garfinkel A (1982) Forms of explanation: rethinking the questions in social theory. Yale University Press

Goethals S, Martens D, Evgeniou T (2023a) Manipulation risks in explainable AI: the implications of the disagreement problem. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Heidelberg, pp 185–200. https://doi.org/10.1007/978-3-031-74633-8_12

Goethals S, Sörensen K, Martens D (2023b) The privacy issue of counterfactual explanations: explanation linkage attacks. ACM Transact Intell Syst Technol 14(5):1–24. https://doi.org/10.1145/3608482

Gregor S, Benbasat I (1999) Explanations from intelligent systems: theoretical foundations and implications for practice. MIS Q 23(4):497–530. https://doi.org/10.2307/249487

Gunning D, Vorm E, Wang JY, Turek M (2021) DARPA's explainable AI (XAI) program: a retrospective. Appl AI Lett 2(4):e61. https://doi.org/10.1002/ail2.61

Herm L, Steinbach T, Wanner J, Janiesch C (2022) A nascent design theory for explainable intelligent systems. Electron Mark 32(4):2185–2205. https://doi.org/10.1007/S12525-022-00606-3

Herm L, Heinrich K, Wanner J, Janiesch C (2023) Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. Int J Inf Manag 69:102538. https://doi.org/10.1016/J.IJINFOMGT.2022.102538

Janiesch C, Zschech P, Heinrich K (2021) Machine learning and deep learning. Electron Mark 31(3):685–695. https://doi.org/10.1007/S12525-021-00475-2

Johnson-Laird P (1983) Mental models: Towards a cognitive science of language, inference, and consciousness. Harvard University Press

Kayande U, De Bruyn A, Lilien GL, Rangaswamy A, van Bruggen GH (2009) How incorporating feedback mechanisms in a DSS affects DSS evaluations. Inf Syst Res 20(4):527–546. https://doi.org/10.1287/isre.1080.0198

Keil FC (2006) Explanation and understanding. Annu Rev Psychol 57(1):227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100

Kruschel S, Hambauer N, Weinzierl S, Zilker S, Kraus M, Zschech P (2026) Challenging the performance-interpretability trade-off: an evaluation of interpretable machine learning models. Bus Inf Syst Eng 68(1):159–183. https://doi.org/10.1007/S12599-024-00922-2

Liao QV, Zhang Y, Luss R, Doshi-Velez F, Dhurandhar A (2022) Connecting algorithmic research and usage contexts: a perspective of contextualized evaluation for explainable AI. arxiv:220610847

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4765–4774

Martens D, Provost F (2014) Explaining data-driven document classifications. MIS Q 38(1):73–99. https://doi.org/10.25300/MISQ/2014/38.1.04

Martens D (2022) Data science ethics: concepts, techniques, and cautionary tales. Oxford University Press, Oxford

Martens D, Hinns J, Dams C, Vergouwen M, Evgeniou T (2025) Tell me a story! Narrative-driven XAI with large language models. Decis Support Syst 191:114402. https://doi.org/10.1016/j.dss.2025.114402

Miller T, Howe P, Sonenberg L (2017) Explainable AI: beware of inmates running the asylum or: how I learnt to stop worrying and love the social and behavioural sciences. arxiv:1712.00547 preprint

Michalski RS (1983) A theory and methodology of inductive learning. Artif Intell 20(2):111–161. https://doi.org/10.1016/0004-3702(83)90016-4

Naretto F, Monreale A, Giannotti F (2022) Evaluating the privacy exposure of interpretable global explainers. In: 2022 IEEE 4th International Conference on Cognitive Machine Intelligence (CogMI). IEEE, pp 13–19

Papenmeier A, Englebienne G, Seifert C (2019) How model accuracy and explanation fidelity influence user trust. arxiv:1907.12652 preprint

Patel N, Shokri R, Zick Y (2022) Model explanations with differential privacy. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT), ACM, New York, p 1895–1904, https://doi.org/10.1145/3531146.3533235

Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan JW, Wallach H (2021) Manipulating and measuring model interpretability. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp 1–52. https://doi.org/10.1145/3411764.3445315

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

Rozenblit L, Keil F (2002) The misunderstood limits of folk science: an illusion of explanatory depth. Cogn Sci 26(5):521–562. https://doi.org/10.1207/s15516709cog2605_1

Rudin C et al (2019) Stop explaining black box models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–215. https://doi.org/10.1038/s42256-019-0048-x

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 618–626, https://doi.org/10.1109/ICCV.2017.74

Shokri R, Strobel M, Zick Y (2020) Exploiting transparency measures for membership inference: a cautionary tale. In: The AAAI Workshop on Privacy-preserving Artificial Intelligence (PPAI). AAAI. p 13

Thagard P (2000) Explaining disease: Correlations, causes, and mechanisms. In: Keil FC, Wilson RA (eds) Explanation and cognition. MIT Press, Cambridge, pp 255–276

Thomas R, Uminsky D (2022) Reliance on metrics is a fundamental challenge for AI. Patterns 3(5):100476. https://doi.org/10.1016/j.patter.2022.100476

Wachter S, Mittelstadt B, Russell C (2018) Why a right to explanation matters. Harvard J Law Technol 31(2):202–238

Wanner J, Herm L, Heinrich K, Janiesch C (2022) A social evaluation of the perceived goodness of explainability in machine learning. J Bus Anal 5(1):29–50. https://doi.org/10.1080/2573234X.2021.1952913

Wellman HM (2011) Reinvigorating explanations for the study of early cognitive development. Child Dev Perspect 5(1):33–38. https://doi.org/10.1111/j.1750-8606.2010.00154.x

Wendt DW (2025) AI strategy and security: a roadmap for secure, responsible, and resilient AI adoption. Apress, Columbus

Yan A, Huang T, Ke L, Liu X, Chen Q, Dong C (2023) Explanation leaks: explanation-guided model extraction attacks. Inf Sci 632:269–284. https://doi.org/10.1016/j.ins.2023.03.020

Yew RJ, Hadfield-Menell D (2022) A penalty default approach to preemptive harm disclosure and mitigation for AI systems. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES). pp 823–830. https://doi.org/10.1145/3514094.3534130

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.