https://pubsonline.informs.org/journal/isre

INFORMATION SYSTEMS RESEARCH
Vol. 34, No. 4, December 2023, pp. 1582–1602
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing

Kevin Bauer,[a,*] Moritz von Zahn,[b] Oliver Hinz[b]

[a] Information Systems Department, University of Mannheim, 68161 Mannheim, Germany; [b] Information Systems Department, Goethe University, 60323 Frankfurt am Main, Germany

*Corresponding author

Contact: kevin.bauer@uni-mannheim.de, https://orcid.org/0000-0001-8172-1261 (KB); vzahn@wiwi.uni-frankfurt.de, https://orcid.org/0000-0003-1160-1007 (MvZ); ohinz@wiwi.uni-frankfurt.de, https://orcid.org/0000-0003-4757-0599 (OH)

Received: June 11, 2021
Revised: June 2, 2022; October 28, 2022
Accepted: December 17, 2022
Published Online in Articles in Advance: March 3, 2023

https://doi.org/10.1287/isre.2023.1199

Abstract. Because of a growing number of initiatives and regulations, predictions of modern artificial intelligence (AI) systems increasingly come with explanations about why they behave the way they do. In this paper, we explore the impact of feature-based explanations on users' information processing. We designed two complementary empirical studies where participants either made incentivized decisions on their own, with the aid of opaque predictions, or with explained predictions. In Study 1, laypeople engaged in the deliberately abstract investment game task. In Study 2, experts from the real estate industry estimated listing prices for real German apartments. Our results indicate that the provision of feature-based explanations paves the way for AI systems to reshape users' sense making of information and understanding of the world around them. Specifically, explanations change users' situational weighting of available information and evoke mental model adjustments. Crucially, mental model adjustments are subject to the confirmation bias so that misconceptions can persist and even accumulate, possibly leading to suboptimal or biased decisions. Additionally, mental model adjustments create spillover effects that alter user behavior in related yet disparate domains. Overall, this paper provides important insights into potential downstream consequences of the broad employment of modern explainable AI methods. In particular, side effects of mental model adjustments present a potential risk of manipulating user behavior, promoting discriminatory inclinations, and increasing noise in decision making. Our findings may inform the refinement of current efforts of companies building AI systems and regulators that aim to mitigate problems associated with the black-box nature of many modern AI systems.

History: Alessandro Acquisti, senior editor; Jason Chan, associate editor.

Funding: This work was supported by the Deutsche Forschungsgemeinschaft (DFG) (Projek 449023539), Volkswagen Foundation (ML2MT), and LeibnizInstitute for Financial Research SAFE.

Supplemental Material: The online appendix is available at https://doi.org/10.1287/isre.2023.1199.

Keywords: explainable artificial intelligence • user behavior • information processing • mental models

1. Introduction

Contemporary artificial intelligence (AI) systems' high predictive performance frequently comes at the expense of users' understanding of why systems produce a certain output (Gunning et al. 2019, Meske et al. 2022). For AI systems that provide predictions to augment highly consequential processes such as hiring decisions (Hoffman et al. 2018), investment decisions (Ban et al. 2018), or medical diagnosing (Jussupow et al. 2021), this "black box" nature can create considerable downsides. These issues include impaired user trust, reduced error safeguarding, restricted contestability, and limited accountability (see Rosenfeld and Richardson 2019 for a review). Having recognized these problems, organizations developing AI and governments increasingly adopt principles and regulations (EU 2016, 2021; Google AI 2019; Meta AI 2021) effectively stipulating that AI systems need to provide meaningful explanations about why they make certain predictions (Goodman and Flaxman 2017, Cabral 2021). In light of these developments, the implementation and use of explainable AI (XAI) methods are becoming more widespread and mandated by law.

The purpose of XAI methods is to make AI systems' hidden logic intelligible to humans by answering the question: Why does an AI system make the predictions it does? Thereby, XAI methods aim to achieve high predictive performance and interpretability at the same time. Many state-of-the-art XAI techniques convey insights into AI systems' logic after training and explain behaviors by depicting the contribution of individual input features to the outputted prediction (Doshi-Velez and Kim 2017). Although there is reason to believe that XAI can mitigate black-box problems (Bauer et al. 2021), the pivotal question is how users respond to modern explanations, given that the human factor frequently creates unanticipated, unintended consequences even in well-designed information systems (Willison and Warkentin 2013, Chatterjee et al. 2015).

Nascent research on human-XAI interaction examines how explainability affects humans' perceptions, attitudes, and use of the system, for example, trust (Erlei et al. 2020), detection of malfunctioning (Poursabzi-Sangdeh et al. 2021), (over)reliance (Bussone et al. 2015), and task performance (Senoner et al. 2021). Prior research, however, does not consider the potential consequences of providing explanations for users' situational information processing (the use of currently available information in the given situation) and mental models (cognitive representations that encode beliefs, facts, and knowledge). By depicting the contribution of individual features to specific predictions, feature-based XAI enables users to recognize previously unknown relationships between features and ground truth labels that the AI system autonomously learned from complex data structures. In that sense, XAI may constitute the channel through which AI systems impact humans' conceptualization and understanding of their environment. This effect could reinforce the already considerable influence contemporary AI systems have on human societies (Rahwan et al. 2019) by, for better or worse, allowing human users to adopt systems' inner logic and problem-solving strategies. Despite the increasing (legally required) implementation of XAI methods, a systematic study of these effects is yet missing. The paper at hand aims to fill this important gap.

We ask three research questions. Does the additional provision of feature-based explanations affect AI system users' situational processing of observed information? Does it affect users' underlying mental models? What are important moderating factors? Consider, for instance, a loan officer who works with an AI system to predict an applicant's risk parameters and determine the credit approval. Because of legal requirements (e.g., Artificial Intelligence Act; EU 2021), the AI system recently started to provide feature-based explanations, showing that it strongly relies on people's smartphone charging behavior to predict creditworthiness.[1] Although previous research examines how this explanation may affect the loan officer's perceptions of the system, we conjecture that the explanation also, and maybe more importantly, affects his processing of currently available information and his underlying mental models of the determinants of creditworthiness. By changing mental models, explanations may even reshape the loan officer's behaviors in related domains beyond the loan approval decision, for example, assessing the faithfulness of his daughter's new boyfriend based on the smartphone charging behavior.[2]

Considerable challenges arise when trying to answer our research questions. First, measuring how XAI methods affect users' situational processing of information and mental models is extremely difficult because these cognitive processes are typically unobserved. Second, we need to control for possible external cues, unintended stimuli, additionally attainable information, and preferences that may affect these cognitive processes in any given situation. Third, whether people interact with an (X)AI system, let alone rely on it, is highly endogenous and depends on factors such as culture, technological literacy, and the socio-technological environment. Thus, isolating effects associated with the provision of explanations in addition to predictions is particularly demanding, if not outright impracticable, in a natural (organizational) setting. To address these challenges, we rely on two complementary, incentivized experimental studies.

In Study 1 (n = 607), laypeople played a series of investment games (Berg et al. 1995), making sequential economic transaction decisions in an intentionally abstract setting. In Study 2 (n = 153), experts from the real-estate industry predicted listing prices for real apartments located in Germany. Study 2 extends Study 1 by testing the generalizability of our findings and elaborating on mechanisms driving the results. In both studies, conditional on the treatment, participants either received no decision support, support from an AI system in the form of opaque predictions or an XAI system with predictions plus feature-based explanations. We answer our research questions by eliciting and comparing changes in both participants' decision-making patterns and their beliefs about feature-label relationships.

The two studies strongly complement each other for three reasons. First, laypeople (Study 1) and experts (Study 2) are the two diametrical archetypes of AI system users affected by growing explainability requirements. Studying both types' responses to XAI methods enables us to identify possibly differential effects and make inferences about the generalizability of our findings. Second, we consider two fundamental types of prediction problems where AI systems are frequently in use: transaction outcome predictions (Study 1) and price predictions (Study 2) (Ban et al. 2018, Rico-Juan and de La Paz 2021). Examining the two settings allows us to understand better whether the interplay between XAI and cognitive processes is task specific. Third, using local interpretable model-agnostic explanations (LIME) (Study 1)

and SHapley Additive exPlanations (SHAP) explanations (Study 2), the two most popular feature-based XAI methods (Gramegna and Giudici 2021), allows us to draw more general conclusions about the interplay between feature-based explainability and cognitive processes.

Our findings paint a consistent picture: Providing explanations is the critical factor that enables AI systems to influence the way people make sense of and leverage information, both situationally and more permanently. Crucially, we find an asymmetric enduring effect that can foster preconceptions and spill over to other decisions, thereby promoting certain (possibly biased) behaviors.

Our paper proceeds as follows. Section 2 presents theoretical foundations, whereas Section 3 explains our experimental studies and results. Section 4 concludes by discussing our results, the limitations of our work, and directions for future research.

2. Theory

In this section, we first discuss modern XAI methods (Section 2.1). Subsequently, we outline the relation between providing explanations and cognitive processes (Section 2.2) and discuss our work's contribution to the literature (Section 2.3).

2.1. Explainable AI

Following Doshi-Velez and Kim (2017), we conceptualize XAI as methods that possess the ability to present in understandable terms to a human why an AI system makes certain predictions. Over the last couple of years, researchers developed ample XAI methods that help elucidate the opaque logic of machine learning (ML)-based AI systems (Ribeiro et al. 2016, Lundberg and Lee 2017, Koh and Liang 2017, Lakkaraju et al. 2019). Very generally, XAI methods aim to alleviate problems associated with the black-box nature (e.g., distrust, lack of accountability, and error safeguarding) while maintaining a high level of prediction accuracy (Bauer et al. 2021).

Our study focuses on feature-based XAI methods, hereafter XAI methods, that can explain the behavior of any ML-based AI system by showing the contribution of individual features to the prediction. We do so for several reasons. First, these explanations are the most widespread in practice (Bhatt et al. 2020, Senoner et al. 2021, Gramegna and Giudici 2021). Second, they are highly intuitive and straightforward to interpret as they satisfy most requirements for human-friendly explanations (Molnar 2020). Third, they are typically applicable to systems using structured and unstructured data (Garreau and Luxburg 2020). Fourth, these methods can explain individual predictions, local explainability, which might be the only method legally compliant with (upcoming) regulations (Goodman and Flaxman 2017).

Many researchers recognize two related XAI methods as state-of-the-art: LIME and SHAP (Gramegna and Giudici 2021, Molnar 2020). LIME (Ribeiro et al. 2016) and SHAP (Lundberg and Lee 2017) provide explanations through additive feature attributions, that is, linear models that depict the numeric contribution of each feature value to the overall black box model prediction. Both approaches learn these interpretable "surrogate models" on input-prediction pairs of the black box model and are applicable to virtually all classes of ML models, that is, are model agnostic. On the individual level, SHAP and LIME provide contrastive explanations that inform users why predictions for a specific instance diverge from the prediction for an average instance (Molnar 2020). For example, if the SHAP value for the feature *Balcony* equals +500 (−200), it indicates that having a balcony marginally increases (decreases) the current apartment's listing price prediction by $500 ($200). The big difference between LIME and SHAP is the way of estimating the additive feature attributions. LIME creates synthetic, perturbed data points in the local neighborhood of the observation of interest and fits a weighted linear model to explain the relationship between the synthetic data and the relevant black box predictions. Importantly, LIME weights synthetic instances based on their proximity to the original data point. By contrast, SHAP is inspired by coalitional game theory and treats input features as a team of players that cooperate to generate a payoff (the prediction). The method essentially estimates the marginal contribution of each player to the overall payoff, Shapley values (Shapley 1953), using a linear model that weights instances based on characteristics of coalitions. Given these mathematical differences, the two methods can produce (slightly) different feature attributions for the same instance. However, from the perspective of a user who is not familiar with these details, the intuition and interpretation of the two methods' explanations are reasonably similar (Molnar 2020). Notably, LIME and SHAP closely relate to the seminal description of Gregor and Benbasat (1999) of "why and why not explanations" in the context of knowledge-based expert systems.

With the development of modern explainability methods, research on the impact of contemporary XAI on user behavior has become increasingly essential (Vilone and Longo 2021). Nascent research in this domain typically focuses on how explanations affect user attitudes and reliance on the AI system (Lu and Yin 2021). These studies produce mixed evidence on the consequences of XAI on decision performance, user trust, perception, and decision-making performance. Several studies depict that explanations can enhance trust in and positive perceptions of the system (Rader et al. 2018, Dodge et al. 2019, Yang et al. 2020), whereas others provide reversed evidence (Erlei et al. 2020, Poursabzi-Sangdeh et al. 2021). Although prior studies produce important insights regarding the interplay between XAI and user perceptions, none of them considers that the additional provision of explanations may also reshape users' information processing,

both situationally and more permanently. For instance, using SHAP to show the contribution of input features to a creditworthiness prediction may not only affect a loan officer's perception of the AI system in use. Instead, she may process currently available information about the applicant differently and develop a novel understanding of the determinants of creditworthiness, that is, adjust her mental model. With the increasing adoption of explainability principles by organizations (Google AI 2019, Meta AI 2021) and the growing number of regulatory transparency requirements (EU 2016, 2021), it is pivotal to understand how contemporary XAI methods influence cognitive processes that lie at the heart of people's knowledge, behavior, and problem-solving capabilities.

2.2. Cognitive Perspective on XAI Employment

Through feature-based explanations about an AI system's prediction, human users can observe possibly unknown feature-label relationships that the system learned from complex data structures by itself (Agarwal and Dhar 2014, Berente et al. 2021). Although providing explanations, in general, can have a variety of cognitive effects, researchers across disciplines generally agree that they primarily enhance people's understanding of someone or something, improve reasoning, and facilitate learning (Gregor 2006, Malle 2006). From a cognitive perspective, obtaining explanations can entail two effects: First, it may change people's situational processing of available information: their use of available information while observing explanations. Second, it can lead to an adjustment of their beliefs about feature-label relationships the AI system inherently models: their mental representation of real-world processes. In this paper, we follow previous work in information systems and rely on the "Mental Models Framework" to conceptualize relevant cognitive processes (Vandenbosch and Higgins 1996, Lim et al. 1997, Alavi et al. 2002).

Mental models are "all forms of mental representation, general or specific, from any domain, causal, intentional or spatial" (Brewer 1987, p. 193), encoding beliefs, facts, and knowledge (Jones et al. 2011). Through imaginary manipulations of model components, people can reason and make inferences about how to solve problems (Rouse and Morris 1986). Much of the people's decision making is based on these simulations that figuratively create informal algorithms for carrying out specific tasks (Johnson-Laird et al. 2017). For instance, real estate agents can mentally simulate how listing prices might change if an apartment for sale had a balcony.

When people perform tasks, they draw on relevant mental models that guide their processing of incoming information to form expectations and make (expectedly) optimal decisions. Working with an AI system that provides black box predictions, that is, information relevant to the task, allows people to reflect on their own expectations and compare it to the machine prediction (Schön 2017). This mental process might entice people to revise their expectations and thus make different decisions because the machine prediction effectively substitutes for people's own mental model driven formation of expectations (Agrawal et al. 2019). However, the black box nature does not allow users to directly compare their underlying beliefs and logic with that of the AI system. This comparison can only occur when they learn how the system combines available information to arrive at a prediction. In the previous example, the real estate agent may have access to an XAI system that provides a listing price prediction together with an explanation of how specific apartment attributes contribute to it. The agent can compare the explanation to her own initial perception of the individual attribute contributions to the listing price. As a result, the agent may detect inconsistencies that prompt her to revise her logic by putting more or less emphasis on specific information currently available to evaluate the apartment. This explanation-enabled situational process (Schön 2017) can reconcile the distinct logic that humans and machines apply to arrive at a certain assessment. From this perspective, providing explanations on top of predictions may constitute a pivotal factor in allowing users to reflect on how they leverage information to solve a problem and adapt it according to the AI system's logic for the given task.

Apart from situationally changing cognitive processes that shape the current decision, the interaction between mental models and explanations may also yield lasting effects because mental models possess the dynamic capacity to change (Jones et al. 2011). Repeatedly observing explanations about how feature X contributes to prediction $\hat{Y}$ and engaging in reflection processes may evoke adjustments of the underlying mental model in use. Following Vandenbosch and Higgins (1996), exposure to external stimuli, here explanations, can lead to two mental model adjustment processes: maintenance and building. Under mental model maintenance, people feel encouraged to maintain or reinforce current beliefs and decision-making rules. This process occurs when they perceive or select new information to fit into their current beliefs and routines. Under mental model building, individuals profoundly restructure or build new mental models in response to handling novel, disconfirming information. As a result of these processes, individuals may adopt different beliefs about how X contributes to the real label Y, enticing them to process information differently even when explanations are no longer present. Put differently: users may not merely combine situationally observed explanations with their own logic to solve a given task. Instead, observing the system's logic may more fundamentally reshape users' way of solving problems in general, that is, evoke learning. Therefore, users may exhibit different problem-solving strategies whenever they draw on the explanation-adjusted mental model, even in situations where they do not observe explanations anymore.

In sum, cognitive theories give reason to believe that providing explanations in addition to predictions can influence users' processing of information about feature X, both situationally and more fundamentally. Because of the latter effect, modern XAI methods may constitute a cornerstone of effective knowledge transfers from ML-based AI systems to human users, helping them to learn from the AI how X relates to Y. Hence, explanations could facilitate learning *machine knowledge*: new knowledge AI systems autonomously learned from Big Data and previously missed by domain experts (Teodorescu et al. 2021, van den Broek et al. 2021).

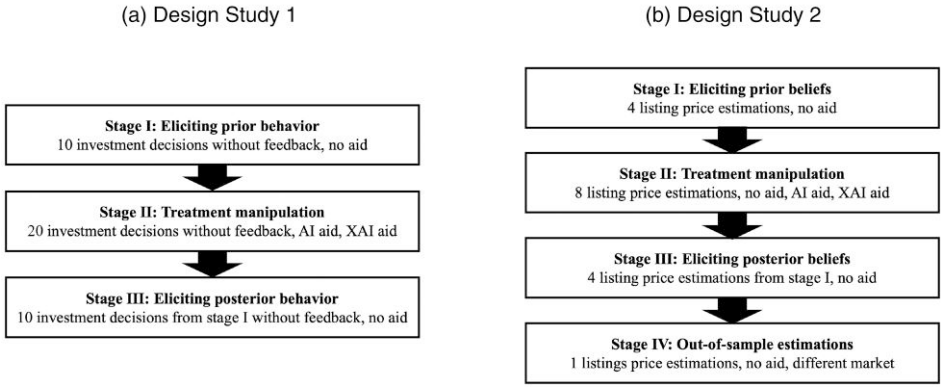
2.3. Contribution to the Literature

Our study complements three different streams of literature. The first and most closely related line of work studies the interplay between XAI techniques and user behavior (see Rosenfeld and Richardson (2019) and Vilone and Longo (2021) for an overview). About two decades ago, several studies found that suitably designed explanations about the functioning and purpose of legacy knowledge-based expert systems can increase users' trust in the systems, improve users' perceptions of the system, and enhance decision-making performance (Dhaliwal and Benbasat 1996, Gregor and Benbasat 1999, Ji-Ye Mao 2000, Wang and Benbasat 2007). However, these expert systems codify knowledge from human experts as explicit procedures, instructions, rules, and constraints in a digital format. They do not represent machine knowledge that modern ML-based AI systems learn independently of domain experts by training on large data sets (van den Broek et al. 2021). Given the inherent distinctions between expert systems and ML-based AI systems in terms of encoded knowledge, contemporary explainability methods present an entirely different form of reasoning to users, namely that of machines (Vilone and Longo 2021, Meske et al. 2022). More recent research on the impact of explainability on user behavior mainly focuses on how contemporary XAI methods impact users' perceptions of the AI system. This nascent literature shows that explainability often improves reliance on and trust in the system (Bussone et al. 2015), fairness perceptions (Dodge et al. 2019), human-AI collaboration (Yang et al. 2020), task efficiency (Senoner et al. 2021), and users' understanding of the system's malfunctions (Rader et al. 2018). However, there is also evidence of disadvantages relating to informational overload (Poursabzi-Sangdeh et al. 2021), reduced user trust (Erlei et al. 2020), and overreliance (Bussone et al. 2015). Moreover, explanations that are unstable and sensitive even to small perturbations to inputs have the potential to mislead human users into trusting a problematic black box, for example, by selectively providing explanations that conceal biased behaviors and malfunctions (Kaur et al. 2020, Lakkaraju and Bastani 2020). Hence, explanations may be a security concern if adversaries use perturbations of inputs and model attributes to produce intentionally misleading explanations that manipulate users' trust and behaviors (Ghorbani et al. 2019). We complement this pivotal and insightful work by examining the impact of contemporary XAI on users' situational information processing and mental models. Understanding how the provision of explanations about the workings of ML-based AI systems may reshape these cognitive processes is pivotal for anticipating the downstream consequences of this technology on human societies and designing effective transparency and explainability regulations.

The second set of literature we complement explores the mechanisms of learning in socio-technological environments. A common theoretical foundation builds on Bayes rule as a rational benchmark of how humans accommodate new information (Holt and Smith 2009). However, research has shown systematic deviations from Bayes' rule. Reasons include over- or underweighting of new information (Rabin and Schrag 1999) and a general tendency to asymmetrically discount information conflicting with prior beliefs while readily internalizing confirming information (Yin et al. 2016). We complement this research stream by showing how human users deviate from Bayes rule in the context of learning from modern AI systems. Notably, there exists a limited number of prior research examining how black box predictions change users' decision-making habits (Abdel-Karim et al. 2020, 2022; Fügener et al. 2021a, b; Jussupow et al. 2021). Relatedly, in a formal model, Agrawal et al. (2019) show that the predictions of black box AI systems can alter users' abilities by providing them with incentives to learn to assess the (negative) consequences of their actions for the task supported by the AI.[3] None of these studies, however, examines the role of feature-based explanations in learning, which could pave the way for more fundamental changes in the way users understand real-world processes. Our paper intends to fill this gap. We study how the provision of explanations about how an AI system solves prediction tasks allows users to integrate the presented machine knowledge into their mental models, that is, learn from XAI. A better understanding of how explainability may contribute to *machine teaching*, the notion that AI systems first learn novel knowledge that experts neither conceive nor anticipate from data and then transfer this knowledge to human users (Abdel-Karim et al. 2020), is particularly significant given the growing requirements to implement explainability methods when using AI systems.

The third stream of literature we add to studies how humans collaborate with computerized systems to solve

Figure 1. Structure of Empirical Studies



Notes. We provide an overview of the main sequence of our two empirical studies.

problems. Previous research in this area dates back decades. Several studies document that humans resist using computerized decision aids, despite possible performance benefits (Kleinmuntz 1990), whereas others find that humans possess a strong preference for using them (Dijkstra 1999). With the growing employment of modern AI systems in a broad range of domains, the examination of human-machine collaboration has seen a considerable resurgence, for example, in the domain of finance (Ge et al. 2021), medicine (Jussupow et al. 2021), customer service (Schanke et al. 2021), and on-demand tasks (Fügener et al. 2021a). Research on "centaur" systems (Goldstein et al. 2017, Case 2018) documents how hybrid human-AI systems (i.e., centaur systems) achieve superior results in comparison with the entities operating independently (Dellermann et al. 2019, Tschandl et al. 2020), promising considerable benefits from successful human-AI collaboration. Several factors moderate the interaction of humans and AI systems including the perceived subjectivity of the task (Castelo et al. 2019, Logg et al. 2019), seeing the system err (Dietvorst et al. 2015), being able to modify predictions (Dietvorst et al. 2018), the divergence between actual and expected predictive performance (Jussupow et al. 2020), and, most importantly for our research, understanding the system's internal logic (Gregor and Benbasat 1999, Hemmer et al. 2021). Following our conjecture that explanations pave the way for AI systems to affect people's cognitive processes, contemporary XAI methods introduce another layer of complexity in human-AI interaction and its success: an interaction between machine and human problem-solving strategies. Our work provides novel insights into whether and under what circumstances people prefer to rely on their own way of leveraging information or willingly adjust it according to machine explanations. In this sense, our work contributes to the literature on (hybrid) human-AI collaboration by analyzing the underlying cognitive processes that may facilitate or hinder the realization of the promise of this technology.

3. Empirical Studies

We now present the design and results of Studies 1 and 2. In both studies, participants made decisions under uncertainty (providing loans and predicting apartment listing prices) either with the aid of an opaque AI, an explainable AI, or without any support. We paid participants according to their decision-making performance to reveal actual preferences and beliefs.[4] We implemented both studies using oTree, Python, and HTML and ran them online. In Study 1, we recruited 607 participants on Prolific and let them engage in deliberately abstract investment games (Berg et al. 1995). Results allow us to observe how the provision of explanations on top of predictions shapes information processing and mental models for laypeople in a very general sequential transaction domain. Study 2 extends the first study by testing the generalizability of mental model adjustments regarding the task domain (listing price predictions), decision-maker expertise, and the explanation presentation, and elaborates on important asymmetric effects. With the help of our industry partner, the *Real Estate Association Germany (IVD)*, we recruited 153 experts from the real estate industry to participate in Study 2. We report the designs and results of the two studies consecutively. Figure 1 portrays an overview of the experimental designs.

3.1. Study 1

3.1.1. Design. In Study 1, participants repeatedly engaged in one-shot investment games (Berg et al. 1995) that possess the following structure. An investor receives 10 monetary units (MU). The investor initially observes 10 deliberately abstract borrower characteristics and decides whether to invest her 10 MU with the borrower. If she does not invest, the game ends without the borrower making a decision, and both the investor and borrower earn a payoff of 10 MU. If she invests, the borrower possesses 30 MU and can keep the whole amount without repercussions. Crucially, the borrower can repay the

investor 10 MU, thereby reciprocating the investor's initial trust. In case of repayment, the investor receives 20 MU (we double the amount); otherwise, the investor earns 0 MU while the borrower gets 30 MU. The borrower, in the absence of sufficiently strong social motives, for example, altruism, egalitarian concerns, or moral preferences (Miettinen et al. 2020), will not make a repayment and maximize his personal income. As a result, the payoff structure of the investment game is of an adversarial nature from the investor's perspective because her material well-being is at the mercy of the borrower if she invests. The investor loses her initial investment of 10 MU whenever the borrower pursues pure income-maximizing or adversarial motives like wanting to minimize the investors' payoffs. Given this payoff structure, an income-maximizing investor in the experiment will only invest if (i) her belief that the borrower's motive leads him to repay her is sufficiently strong, and (ii) she ultimately judges that the prospect of doubling her income is worth risking the loss of her investment.[5] Study 1 participants always played as investors. Borrowers are subjects from a previous incentivized field study who had to decide on repayment assuming an initial investment; that is, they have already committed to a repayment decision and cannot strategically change this choice ex post. We did not provide intermediary feedback to prevent the development of idiosyncratic expertise, experience, or investment strategies that may confound our results. We randomly matched investor and borrower decisions to determine game outcomes at the end of the study and pay both according to the earned MU.

Study 1 comprised a baseline (AI) and a treatment (XAI) condition, each with three stages.[6] In Stage I, each participant made 10 investment decisions for distinct, randomly drawn borrowers without intermediary feedback. They always observed the 10 characteristics of a borrower and did not obtain any aid. The idea is that the 10 borrower characteristics allow investors to get an idea of the likelihood that an individual borrower will make a repayment, for whatever motives, and to assess whether it is worth taking the risk of losing their investment. We deliberately chose 10 unintuitive traits correlated with a person's repayment inclination so that participants did not possess strong prior beliefs about the informativeness of characteristics for someone's repayment behavior (see Table 4 in the online appendix). The main reason for choosing just these characteristics is that previous empirical tests have shown that they are appropriate features for developing an AI system that accurately predicts repayment with which participants interact in Stage II. Importantly, participants learned that the AI system makes predictions based on the same 10 borrower characteristics they also observe, mitigating concerns that they believed the AI system to have access to more information.

Stage II introduced our treatment variation. Participants made 20 decisions for new random borrowers observing all 10 borrower traits. Additionally, baseline participants saw an AI system's prediction about whether borrowers will repay an initial investment. Again, we did not provide intermediary feedback. We trained the AI system on 1,054 distinct data points collected in a previous field study, the same data set that the borrowers that participants encounter in the experiment stem from (see the online appendix for details).[7] The system did not continue to learn during the experiment. Treatment participants, on top of predictions, observed LIME explanations (Ribeiro et al. 2016) for each borrower characteristic, informing them of its contribution to the repayment prediction. Revealing LIME values on top of identical predictions constituted the treatment variation. As is often the case, we depicted LIME values graphically using colored bars of different lengths. Participants received detailed information about the model, input features, performance on a representative test set, and how to interpret LIME explanations.

Stage III perfectly mirrored Stage I. Importantly, participants engaged with the same borrowers from Stage I in random order. We did not draw participants' attention to this fact to alleviate concerns about the experimenter's demand effect. The study concluded with a brief questionnaire on socio-economic control variables.

3.1.2. Results. Throughout our analyses of Study 1, we mainly rely on the following regression model:

$$Y_{ijs} = \beta_1 \cdot X_j + \beta_2 \cdot (X_j \times I_s) + \beta_3 \cdot (X_j \times Expl_i) + \beta_4 \cdot (X_j \times Expl_i \times I_s) + \gamma_{ijs} + \epsilon. \quad (1)$$

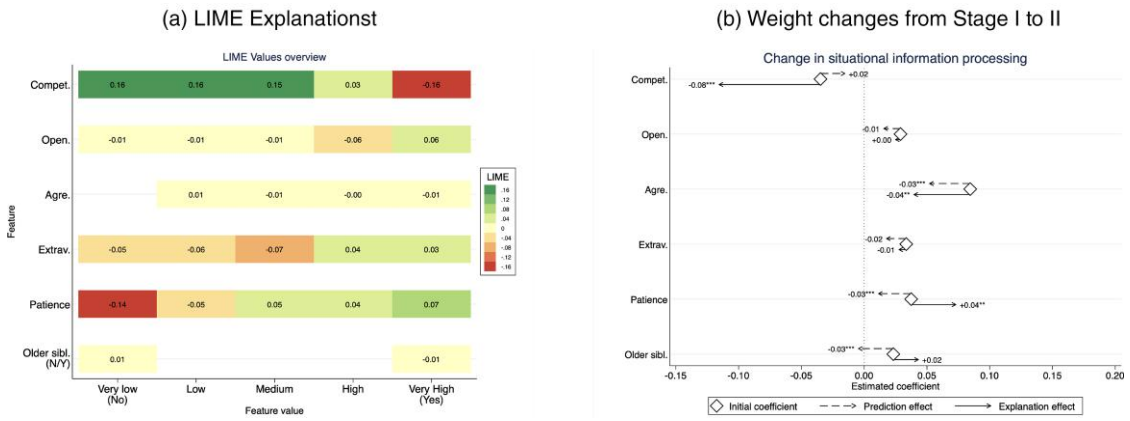
Y_{ijs} is a dummy indicating whether participant i invested with borrower j in Stage s. Hence, β coefficients measure variation in the probability to invest with a borrower, and X_j is a vector reflecting the 10 observed borrower traits, the overall prediction, and LIME values.[8] Most relevant to our analyses, I_s and $Expl_i$ are dummy variables, respectively, indicating whether a decision takes place in Stage s compared with Stage I (i.e., Stage I serves as the reference category) and whether participant i is in the XAI treatment (observes explanations on top of predictions in Stage II), and γ_{is} represents individual-state fixed effects. We report standardized regression coefficients with robust standard errors. Our main interest lies in the interaction terms β_3 and β_4, respectively, capturing the isolated effects of observing the prediction and additionally observing LIME explanations. As β_4 constitutes a difference-in-difference (DiD) estimator, it is pivotal to check that before the intervention, there are no treatment differences (parallel trends assumption). Regression analyses reveal that baseline and treatment participants in Stage I did not place significantly different weight on any trait; hence, the use of a

DiD identification strategy appears generally valid. Nevertheless, because participants placed significant weight on *Gender*, *Conscientiousness*, *Neuroticism*, and *Younger Siblings* in only one of the two conditions, there is still some concern about the appropriate interpretation of DiD estimates for these traits.[9] To avoid drawing incorrect conclusions, we conservatively refrain from interpreting these traits' estimates while still including them as controls in the model.

3.1.2.1. Situational Information Processing. We start analyzing how participants' weighting of borrower characteristics changed from Stage I to II, that is, changes in participants' situational information processing. Figure 2 illustrates our results. Figure 2(a) depicts the average LIME values (color saturation) participants observed for different feature values (y and x axis). Higher positive (negative) LIME values depict a higher positive (negative) contribution of a given feature value to the predicted probability that a borrower makes a repayment. Figure 2(b) portrays how the provision of predictions and explanations affected the weighting of a given borrower trait. The diamond marker represents the original weighting in Stage I (β_1). The dashed and solid arrows, respectively, illustrate the isolated effects of observing predictions (β_3) and additional explanations (β_4). Depicted results stem from regressions reported in Table 9 in the online appendix.

There are two main insights. First, prediction effects in Figure 2(b) suggest that the provision of opaque predictions generally decreased the weight participants placed on observed borrower traits. On average, the absolute magnitude of coefficients changed by 63.6%. Although only the estimates for *Agreeableness*, *Patience*, and *Older Siblings* are significant, predictions reduced the absolute magnitude of all variables. Second, the provision of explanations on top of predictions entailed significant weight changes that mirror the relationship between borrower traits and repayment behavior as depicted by the LIME values. Here, the average magnitude of absolute weight changes equals 73.9%. Figure 2(a) shows that the predicted repayment probability markedly decreases (increases) with a borrower's level of *Competitiveness* (*Patience*). Figure 2(b) reveals that these are the two traits whose weighting the provision of explanations significantly fostered: observing explanations rendered the relationship between a borrower's *Competitiveness* (*Patience*) and a participant's investment likelihood significantly more negative (positive). Notably, explanations as such increased the absolute magnitude of the coefficient for *Competitiveness* (*Patience*) by 240.0% (94.6%). LIME values reveal that *Agreeableness*, the trait participants initially weighted the most, has almost no impact on the repayment prediction. Accordingly, we find that the provision of explanations led to a significant decrease in the magnitude of the weight participants placed on this trait (−44.7%). Additional analyses confirm that LIME values for these three characteristics had a significantly positive influence on participants' investment decisions, corroborating the notion that participants paid attention to and adjusted their weighting of traits according to observed explanations (see Table 11 in the online appendix). Taken together, participants significantly adjusted their weighting of information in the direction of observed explanations for (i) the trait they initially perceived as most important and (ii) the traits LIME highlighted as most important.[10] Finally, although not shown in the Figure 2 for ease of interpretation, regression analyses further reveal that explanations significantly reduced the weight participants placed on the prediction as such (magnitude of coefficient decreased

Figure 2. (Color online) Prediction and Explanation Effects on Situational Information Processing



Notes. We illustrate how the provision of opaque predictions and LIME explanations on top of predictions affect participants situational information processing. (a) LIME values (z axis) for different feature values (x axis) participants observed in the study. For the binary feature Older siblings, we show the LIME values for No and Yes at the outer limits of the continuous feature scale. (b) Estimated prediction and explanation effects, respectively, of β_3 and β_4 in Model (1) with $s = 2$. Initial values represent β_1. We denote significance levels by $^*p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

by 26.8%); that is, they were less likely to follow a prediction that a borrower makes a repayment.[11]

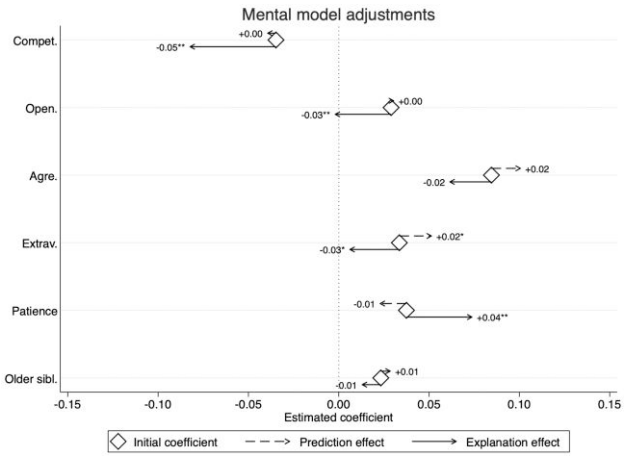
Result 1.1. *Observing explanations changed participants' situational processing of the overall prediction and borrower traits that explanations or they themselves consider most important. The direction of adjustments mirrors explanations.*

Result 1.1 agrees with our theoretical elaborations: People adjust their situational information processing in response and according to explanations they currently observe. Notably, elicited expectations about the prediction accuracy did not differ significantly for predictions with or without explanations (71.8% and 70.6%, respectively; $p = 0.751$, Wilcoxon rank-sum test). Therefore, changes in the weighting of predictions do not seem to result from lower performance expectations. Next, we test the conjecture that explanations affect beliefs about the relationship between borrower characteristics and repayment behavior, that is, mental models.

3.1.2.2. Mental Model Adjustments. We compare participants' information weighting across Stages I and III to test the conjecture that explanations affect mental models about the relationship between borrower traits and repayment behavior. We rely on the regression model (1), setting $s = 3$ and excluding controls for the prediction and LIME values. Figure 3 illustrates regression results that we report in Table 12 in the online appendix.

Figure 3 portrays how the provision of predictions and explanations lastingly changed the weighting of a given borrower trait across Stages I and III, where participants had no (X)AI aid. The diamond marker depicts the original weighting in Stage I (β_1). The dashed and solid arrows, respectively, show how having observed predictions (β_3) and explanations on top of predictions (β_4) did fundamentally alter participants' information processing, that is, mental models.

Figure 3. Mental Model Adjustments



Notes. We depict participants' mental model adjustments as measured by their change in the weighting of borrower traits across Stages I and III. The estimated prediction and explanation effects respectively represent β_3 and β_4 in Model (1) with $s = 3$. Initial values represent β_1. We denote significance levels by $^*p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Observing opaque predictions did not result in a significant change in participants' weighting of borrower traits. By contrast, depicted results suggest that providing explanations did entail an adjustment of mental models with the absolute magnitude of coefficients changing by 61.8% on average. Importantly, this adjustment was asymmetric. Observing explanations led participants to place significantly more weight on borrowers' *Competitiveness* (+148.6%) and *Patience* (+59.4%) in Stage III than in Stage I. The weight changes again mirror the observed LIME explanations. After observing explanations that the AI system places the most weight on borrowers' *Competitiveness* and *Patience*, participants increased their weighting of these attributes even for investment decisions where they no longer observed explanations. Intriguingly, we do not find that explanations about the low relevance of *Agreeableness* led participants to adjust their marked weighting of this trait significantly. Although participants weighted *Agreeableness* significantly less while observing explanations, they returned to their original weighting of it once they lost access to the XAI system. Naturally, one may wonder about this asymmetry's origins. One plausible interpretation is that explanations are less likely to evoke pronounced mental model adjustments when they conflict with strong preconceptions. Put differently, people are more inclined to engage in mental model maintenance rather than building because it is less cognitively demanding and creates less psychological distress (Vandenbosch and Higgins 1996). In Stage I, participants put by far the most emphasis on a borrower's *Agreeableness* to decide on investing. LIME values, however, suggested that this conception is incorrect because it is among the least relevant predictors for borrowers' repayment inclination. Although one would expect that participants engaged in mental model building to reshape their beliefs about the relationship between *Agreeableness* and repayment behavior, we do not find significant adjustments. For *Competitiveness* (*Patience*), explanations depicted an important negative (positive) influence, which, given their initial weighting of it, confirmed participants' prior beliefs. Following the Mental Models framework, confirming explanations should evoke the maintenance or reinforcement of prior beliefs. Given the significant explanation effects, it seems that participants willingly engaged in this process. This inclination to engage in mental model maintenance rather than building more generally concurs with the frequently documented confirmation bias (Yin et al. 2016), that is, the tendency to selectively process information in a way that allows for the continuation or strengthening of beliefs. We elaborate on this issue in Study 2 and the discussion.[12]

Result 1.2. *Machine explanations entailed asymmetric mental model adjustments. Participants reinforced priors that explanations confirmed but did not abandon priors that explanations markedly contradicted.*

3.1.2.3. Investment Performance. Thus far, it remains open how providing explanations on top of predictions affected participants' decision-making performance in our setting. Table 1 summarizes participants' performance measured by the accuracy (share of payoff maximizing decisions) and recall (share of investments with repaying borrowers). We also report p values of F tests to illustrate significant treatment differences.[13]

Although there are no differences in Stage I, treatment participants performed significantly worse than baseline ones in Stage II (−8.9% and −11.0% for accuracy and recall, respectively).[14] Treatment participants' relatively lower performance in Stage II stems from not investing with the most competitive borrowers (with most negative LIME values), whereas the overall prediction implies doing so, that is, from overruling positive predictions.[15]

They overruled positive predictions and refrained from investing in 46.5% of these cases, resulting in a decision accuracy of merely 53.5%. Baseline participants, for most competitive borrowers, overruled positive predictions only in 21.2% of the cases and achieved a decision accuracy of 78.9%; that is, they are 47.5% more likely to make an income maximizing decision than treatment participants. For all other borrowers, treatment (baseline) participants overruled positive predictions and made optimal decisions in 23% (19.4%) and 69.6% (71.1%) of the cases, respectively. Hence, treatment participants seem to have placed too much weight on very high competitiveness, leading them to overrule the overall prediction inefficiently often.

Examining Stage III, we find that this overweighting of the highest competitiveness level persisted even when participants did not observe explanations anymore (see Table 13 in the online appendix). In Stage III, treatment (baseline) participants invested with most competitive borrowers in 44.7% (54.7%, $p < 0.01$, F test) of the cases and with other borrowers in 68.2% (67.6%, $p < 0.7$, F test) of the cases. As a result, treatment and baseline participants respectively achieved a decision accuracy of 51.7% and 57.2% (−9.6%, $p < 0.01$, F test) for most competitive borrowers and 59.5% and 62.8% (−5.3%, $p < 0.05$, F test) for other borrowers. Notably, participants already associated very high competitiveness with a low repayment likelihood in Stage I: Most competitive borrowers received an investment in 56.3% of the cases, whereas all others did so in 69.5% of the cases (there do not exist treatment differences). Against this background, explanations seem to have exacerbated this inaccurate pattern[16] to an extent that treatment participants made significantly worse decisions than before. Put differently, confirming explanations inappropriately reinforced preconceptions about most competitive borrowers not repaying an investment in our setting.

Result 1.3. *Participants excessively increased the isolated weighting of a trait they already believe to be evidence against repayment. This reaction inefficiently decreased participants' likelihood to invest with repaying borrowers that were highly competitive.*

In sum, the results for Study 1 are highly consistent with the notion that the provision of explanations creates a novel channel through which AI systems may reshape users' way of processing information, both situationally and more permanently. For the latter effect, we observe an asymmetry that is reminiscent of a confirmation bias and, in our setting, decreased participants' decision-making performance by excessively reinforcing inaccurate preconceptions.

3.2. Study 2

The goal of Study 2 is twofold. First, we extend Study 1 results by testing the generalizability of mental model adjustment findings regarding the task domain, user expertise, and explanation presentation and examining whether the asymmetry we found for explanation-driven mental model adjustments in Study 1 is indeed a manifestation of the confirmation bias. Second, we explore if mental model adjustments spill over to related but disparate domains.

Table 1. Investment Performance Across Stages

	Stage I (no aid)		Stage II (with aid)		Stage III (no aid)	
	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall
Baseline (AI) (%)	60.3	64.9	63.1	64.6	62.7	65.1
Treatment (XAI) (%)	60.7	67.4	57.5	57.5	56.5	60.2
F test: Baseline versus treatment	$p = 0.79$	$p = 0.31$	$p < 0.01$***	$p < 0.01$***	$p < 0.02$**	$p < 0.04$**

Notes. We depict participants' investment performance as measured by their accuracy (share of payoff maximizing decisions) and recall (share of investments with repaying borrowers) in Stages I, II, and III. We report results separately for baseline (AI) and treatment (XAI) participants. F tests reveal the significance of treatment differences per measure and stage.

*$p < 0.1$; **$p < 0.05$; and ***$p < 0.01$.

3.2.1. Design. Study 2 comprises four consecutive stages, where recruited real estate experts estimated the listing price per square meter in Euros of apartments that we previously collected from a large online platform.[17] Participants saw 10 apartment characteristics to make an informed guess and did not receive intermediate feedback. To reduce the task complexity and avoid informational overload, we fixed seven apartment characteristics across all stages, that is, apartments only differed regarding the same three characteristics: Location (Frankfurt/Cologne), Balcony (Yes/No), and Green voter share in the district (Below city average/City average/Above city average).[18] We provide screenshots of the interfaces from each stage in the online appendix.

In Stage I, we elicited participants' initial beliefs about the relationship between the three variable apartment characteristics and listing prices. Participants estimated the listing price of four random apartments with different combinations of the variable attributes by entering their marginal contributions to the price using a slider. Sliders ranged from minus to plus 2.500€ in steps of 50€. We initially set the marginal contributions and overall price estimation to 0€ and the average listing price (9,600€), respectively. Participants additionally stated their confidence in the entered marginal contributions and the resulting price estimation on a five-point scale.

Stage II introduced our treatment variations. In all variations, participants estimated listing prices for eight random apartments with different combinations of variable attributes they did not encounter in Stage I. In contrast to Stage I, participants directly entered the estimated listing price. As a reference point, they again observed the average listing price for an apartment. Participants stated their confidence on a five-point scale. In our baseline condition (NoAid), participants estimated the price without any aid. Participants in the AI condition observed opaque listing price predictions of a steady, that is, nonlearning, AI system trained on 4,975 collected observations.[19] In our XAI condition, in addition to observing these predictions, participants also saw numerically presented SHAP values for the three variable apartment characteristics, that is, marginal contributions to the prediction in Euros. After they entered all eight listing price estimates, participants in treatments with decision support filled out a survey containing items on their trust, degree of reliance, and perceived transparency of the AI system (and explanations).

Stage III replicated Stage I to measure posterior beliefs. Independent of the condition, participants again made decisions without any aid for the same apartments.

Finally, in Stage IV, participants estimated the listing price for one last apartment without any decision aid. Across participants, we varied the balcony and green voter attribute of the apartment, whereas the seven fixed attributes were identical to the previous listings. Most importantly, the apartment was in a midsize city in eastern Germany (Chemnitz). For historical, demographic, and socioeconomic reasons, Chemnitz is very different from "A-cities" such as Frankfurt and Cologne, so the housing market is also very different. Germans in general and real estate agents in particular are usually aware of this East-West disparity.[20] The study concluded with a questionnaire on participants' sociodemographics.

3.2.2. Results. We report our results in three steps. First, we outline the experts' belief adjustments from Stage I to Stage III. Second, we examine the occurrence of confirmation bias in these adjustment processes. Finally, we analyze experts' listing price estimates in Stage IV.

3.2.2.1. Mental Model Adjustments. Figure 4 shows the distribution of absolute differences between experts' beliefs about the marginal contribution of the three variable attributes before and after the treatment intervention. We show results for the NoAid, AI, and XAI conditions. The distributions for the NoAid and AI conditions are remarkably similar and skewed toward zero, indicating that experts frequently did not adjust beliefs. The distribution for XAI participants is considerably less right-skewed; that is, they adjusted their beliefs across Stages I and III more. On average, NoAid, AI, and XAI participants' absolute belief adjustments equaled 166.4€, 165.4€, and 299.1€, respectively. Only the differences between NoAid versus XAI (+79.7%, $p < 0.01$, F test), and AI versus XAI (+80.8%, $p < 0.01$, F test) conditions are statistically significant (see Table 24 in the online appendix), that is, observing explanations led to remarkably stronger adjustments of beliefs. Our notion is that real estate experts updated initially held mental models about the relationship between apartment attributes and listing prices as they encountered SHAP explanations. Contrasting our first study, we directly measure participants' prior and posterior beliefs about the contribution of distinct apartment characteristics to listing prices in Study 2. This design facet enables us to estimate mental model adjustments directly, leveraging the accepted framework by DeGroot (1974). Specifically, we assume that agent i's posterior belief about the relationship of characteristic j and the listing price $Post_{i,j} = a_{i,j} \cdot Prior_{i,j} + (1 - a_{i,j}) \cdot Expl_{i,j}$ is a weighted combination of the corresponding prior belief $Prior_{i,j}$ and the personally observed explanation $Expl_{i,j}$; $1 - a_{i,j}$ represents the extent of belief adaptation in the direction of the explanation, whereas $a_{i,j}$ describes the anchoring of the previous belief. For instance, in the extreme case of $1 - a_{i,j} = 1$, individual i completely abandons her prior mental model and adopts the observed explanation as her new one. We estimate the weights $(1 - a_i)$ and a_i for our three study conditions using a

regression model comprising treatment interactions that has the following form:

$$Pos_{ijk} = \beta_1 \cdot Pri_{ijk} + \beta_2 \cdot (AI_i \times Pri_{ijk}) + \beta_3 \cdot (Expl_i \times Pri_{ijk}) + \beta_4 \cdot SV_{ij} + \beta_5 \cdot (AI_i \times SV_{ij}) + \beta_6 \cdot (Expl_i \times SV_{ij}) + \gamma_i + \delta_k + \epsilon. \quad (2)$$

The variables Pos_{ijk} and Pri_{ijk}, respectively, represent expert i's posterior and prior beliefs about attribute j's contribution to apartment k's listing price in Euros. Most importantly, AI_i is a dummy variable indicating that expert i observed a prediction, whereas the dummy $Expl_i$ equals one if a participant additionally observed explanations; SV_{ij} represents the average SHAP value for apartment attribute j of the apartments participant i encountered in Stage II; and γ_i and δ_k are expert and apartment controls, respectively.

On an individual level, Model (2) estimates how observed SHAP values affected participants' adjustments of beliefs about the relationship between a given characteristic and the listing price. It enables us to quantify the "stickiness" of prior beliefs ($\beta_1 - \beta_3$) and "gravitational pull" of explanations ($\beta_4 - \beta_6$) and directly test the occurrence of confirmation bias. Importantly, this estimation is only possible for Study 2, where we elicited prior and posterior beliefs about distinct feature-label relationships. In Study 1, we measured the ultimate investment decisions only and observed belief changes indirectly through changes in those decisions. As a result, we cannot individually quantify the impact of observed explanations on specific beliefs nor can we analyze confirmation bias: a key contribution of our second study.

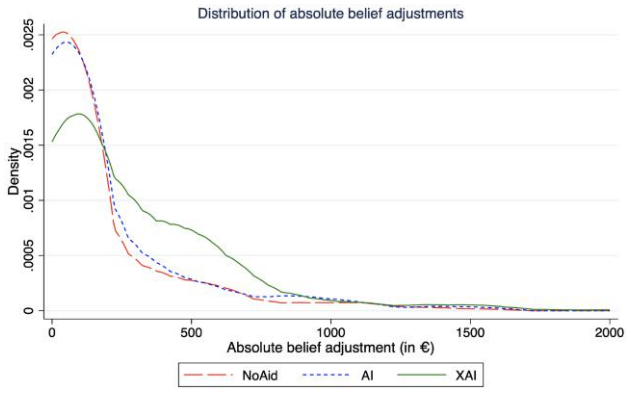
Table 2 depicts regression results for Model (2). Results show that in our NoAid and AI conditions where participants did not observe explanations, SHAP values (unsurprisingly) have no significant explanatory power regarding posterior beliefs (see β_4 and β_5).[21] When participants did not obtain machine aid or only observed predictions, their prior and posterior beliefs were more than 60% positively correlated (β_1 and β_2); that is, participants barely adjusted their beliefs. Only when participants observed explanations in addition to predictions did the displayed SHAP values have positive, statistically significant effects. β_6 reveals that XAI participants significantly adjusted their beliefs in the direction of observed explanations. According to the estimate, posterior beliefs resembled SHAP values more closely in the XAI treatment condition compared with the NoAid and AI conditions (approximately +25 percentage points). Observing explanations also caused XAI participants' posterior beliefs to resemble their prior significantly less (β_3), that is, prior beliefs became less "sticky" compared with the NoAid and AI conditions (approximately −25 percentage points). In sum, these results suggest that observing SHAP explanations led participants to adjust their beliefs in the direction of explanations and abandon their priors. This insight corroborates our Result 1.2 in Study 1 on an individual level, revealing that explanation-driven mental model adjustments also occur for experienced experts, who are arguably familiar with apartment traits and listing price predictions.[22]

Table 2. Posterior Belief Formation

Dependent variable: *Posterior belief*	(1)	(2)
Prior belief (β_1)	0.634***	0.782***
	(0.060)	(0.063)
Prior belief × *AI* (β_2)	0.070	−0.027
	(0.104)	(0.084)
Prior belief × *Expl.* (β_3)	−0.276***	−0.240***
	(0.084)	(0.075)
Avg. SHAP (β_4)	0.025	0.033
	(0.040)	(0.039)
Avg. SHAP × *AI* (β_5)	0.078	0.083
	(0.053)	(0.050)
Avg. SHAP × *Expl.* (β_6)	0.265***	0.249***
	(0.053)	(0.052)
Fixed effects	No	Yes
N	1,836	1,836
R^2	0.740	0.787

Notes. We depict results from OLS regression models with robust standard errors reported in parentheses. The dependent variable equals participants' posterior belief about the marginal contribution of apartment attributes to the listing price in euros. The main independent variables of interest are participants' prior beliefs, the average SHAP values for apartment attributes in Stage II, a dummy indicating that participants observed a prediction in Stage II (AI), a dummy indicating that participants observed explanations in Stage II (XAI), and interaction terms. We further control for the overall posterior listing price participants entered for the apartment and its interaction with treatment dummies, and the average prediction they observed in Stage II. In column (2), we additionally include individual and apartment fixed effects.
*$p < 0.1$; **$p < 0.05$; and ***$p < 0.01$.

Figure 4. (Color online) Distribution of Absolute Belief Changes



Notes. We depict the distribution of experts' absolute belief adjustments across Stages I and III. We aggregate the belief adjustments over all apartment attributes. Different distributions show results separately for NoAid, AI, and XAI participants.

3.2.2.2. Confirmation Bias. In Study 1, we observed asymmetric mental model adjustments that are reminiscent of the confirmation bias. The design of Study 2 allows us to test for confirmation bias in mental model adjustment processes more directly by examining whether XAI participants' adjustments depended on the alignment of explanations and prior beliefs.

We define that explanations confirmed an expert's preconception about the price contribution of a specific apartment attribute if the prior and the observed average SHAP value for the corresponding attribute have the same sign. With this definition, observed explanations confirm prior beliefs in 49.6% of the cases.[23] We analyze differences in belief adjustments with respect to confirming and conflicting explanations using a modified version of Model (2). Specifically, we are interested in whether the convergence of XAI participants' posterior beliefs toward observed SHAP values only occurred when explanations confirmed prior beliefs. Therefore, we focus on the subsample of XAI participants allowing us to omit treatment dummies and interaction terms which facilitates the interpretation of results. Along the lines of Model (2), we regress XAI participants' posterior beliefs about the relationship between apartment characteristics and the listing price on their prior beliefs and observed SHAP values. Most importantly, we now add a dummy variable (*Confirm*) indicating whether explanations confirmed prior beliefs and its interaction with average SHAP values and prior beliefs as independent variables. The interaction *Avg. SHAP* × *Confirm* will provide insights into whether the influence of observed SHAP values on belief adjustments depended on the alignment of explanations and prior beliefs, which are insights we cannot obtain from Study 1 using Model (1).

Corroborating our interpretation of Result 1.2 from Study 1, we find that explanation-driven belief adjustment processes depended on whether explanations confirmed or conflicted with prior beliefs. The estimate for the interaction term *Avg. SHAP* × *Confirm* is positive and statistically significant (see column (1) in Table 3). Following the estimate, posterior beliefs resembled observed SHAP values significantly more closely (about 50% more) if they confirmed their prior beliefs. Hence, consistent with confirmation bias, the belief adjustment was asymmetric regarding the confirmatory nature of explanations. If participants had updated beliefs rationally according to Bayes rule, the interaction term should be insignificant as Bayesian observers would not weight explanations conditional on their alignment with prior beliefs (Rabin and Schrag 1999).

To elaborate on the notion that these asymmetric belief adjustments are a manifestation of confirmation bias, we further consider the role of experts' confidence in their prior beliefs. Prior research shows that confirmation bias is strongest for entrenched beliefs (Pyszczynski and Greenberg 1987, Knobloch-Westerwick and Meng 2009). To test the existence of such heterogeneity, we consider experts' reported confidence in prior beliefs and define that an expert possessed low (high) confidence in a prior, if, on a five-point scale, they reported a confidence level of less than 4 (at least 4). In columns (2) and (3) of Table 3, we, respectively, repeat the regression analysis reported in column (1) for the subsamples of low- and high-confidence prior beliefs.

Table 3. Confirmation Bias and Posterior Belief Formation

Dependent variable: *Posterior belief*	(1) Overall	(2) Low confidencebeliefs	(3) High confidencebeliefs
Prior belief	0.492***	0.483***	0.496***
	(0.091)	(0.105)	(0.136)
Avg. SHAP	0.303***	0.344***	0.145**
	(0.043)	(0.055)	(0.067)
Confirm	12.039	−10.838	115.724
	(27.949)	(39.552)	(73.702)
Avg. SHAP	0.166***	0.107	0.301***
× *Confirm*	(0.059)	(0.077)	(0.094)
N	708	481	222
R^2	0.746	0.725	0.843

Notes. We depict results from OLS regression models with individual and apartment fixed effects. We report robust standard errors reported in parentheses. The dependent variable equals XAI participants' posterior belief about the marginal contribution of apartment attributes to the listing price in euros. The main independent variables of interest are participants' prior beliefs, the average SHAP values for apartment attributes in Stage II, a dummy indicating that observed SHAP values in Stage II confirmed participants' priors, measured by an equal sign of prior beliefs and average SHAP values for a given attribute, and interaction terms. We further control for the overall posterior listing price participants entered for the apartment and the average prediction they observed in Stage II. Column (1) presents results for all decisions. Columns (2) and (3) respectively depict results for the shares of decisions where XAI participants report low and high confidence in their prior.

*$p < 0.1$; **$p < 0.05$; and ***$p < 0.01$.

Reported estimates provide further evidence that explanation-enabled mental model adjustments were subject to confirmation bias. According to the estimated coefficient of *Avg. SHAP* × *Confirm*, for low-confidence priors, the influence of observed SHAP values on posterior beliefs did not depend on whether explanations confirmed prior beliefs (see column (2)). Considering the positive and significant estimate of *Avg. SHAP*, the belief updating was in line with Bayes rule. By contrast, for high-confidence priors, belief adjustments were highly sensitive to whether SHAP values confirmed priors (see column (3)). The estimate for *Avg. SHAP* × *Confirm* suggests that the magnitude of the adjustment of high-confidence priors was about two times larger when observed explanations were in line with them.

Result 2.1. *Study* 1 *findings extend to expert users, SHAP explanations, and the domain of apartment price predictions: SHAP explanations led real estate experts to adjust prior beliefs about the relation between apartment attributes*

and listing prices. Adjustment processes were subject to the confirmation bias.

3.2.2.3. Spillover Effects. Although we observe that real estate experts (asymmetrically) adjusted prior beliefs, all previously reported results pertain to the same market: Participants observed SHAP explanations for the same two A-cities in Western Germany, for which we elicited prior and posterior beliefs. What remains open is whether explanation-driven belief adjustments spilled over to the listing price estimation for apartments in different markets. We put this idea to the test by examining the distribution of participants' final price predictions for an apartment in a medium-sized eastern German city that is not an "A city": Chemnitz.[24]

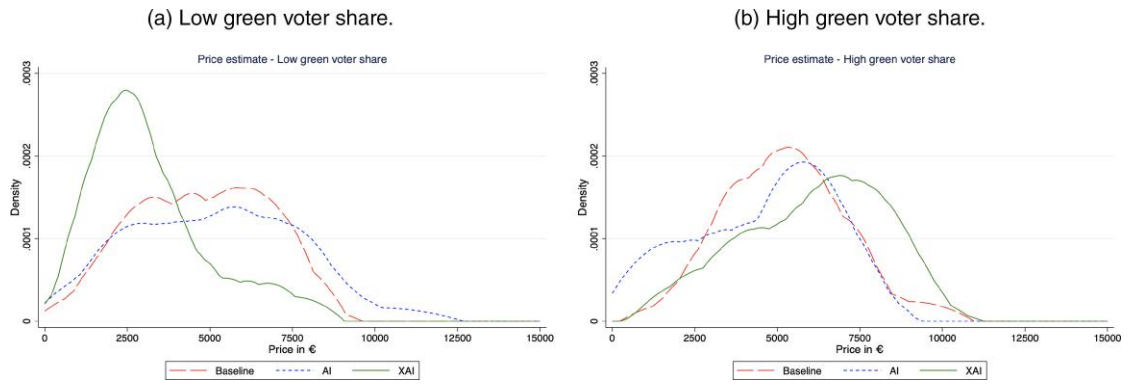
Figure 5 shows the distribution of listing price estimates conditional on the share of green voters in the district for NoAid, AI, and XAI participants. The results indicate that observing explanations impacted participants' price estimates for Chemnitz apartments in neighborhoods with high and low proportions of green voters. Figure 5(a) shows that the distribution of listing prices for an apartment in a district with a low green voter share is considerably more right-skewed for XAI than NoAid or AI participants; that is, they estimate relatively low prices more frequently. NoAid, AI, and XAI participants on average estimated a listing price of 4,752€, 5,141€, and 3,140€, respectively. Only the differences between NoAid versus XAI and AI versus XAI are statistically significant in regression analyses ($p < 0.05$, F test, for both). The distribution of price estimates in districts with high shares of green voters has a stronger left-skew for XAI participants than their NoAid and AI counterparts (Figure 5(b)). On average, NoAid, AI, and XAI participants estimated a listing price of 5,231€, 4,600€, and 6,092€, respectively, for an apartment in a district with a high percentage of green voters. Again, we only find significant explanation effects ($p < 0.1$, F test, for both). These results reveal the economic significance in the changes of price distributions. Specifically, compared with observing no predictions (opaque predictions), observing explained predictions decreased Chemnitz price estimates by 33.9% (38.9%) if the share of green voters was low and increased price estimates by 16.5% (32.4%) if the share of green voters was high. As one might expect, the direction of the difference in experts' evaluation of the green voter share attribute is in line with explanations observed in Stage II: SHAP values indicated that in Frankfurt and Cologne, a high (low) share of green voters marginally contributes to listing prices by about +652€ (−613€). We do not find any effect for experts who only observed opaque predictions in Stage II.

To elaborate on these findings, we also perform a median split and analyze the subsamples of experts whose average absolute belief adjustment for the attribute "Green voter" is below and above the median. Consistent with the idea that belief spillover effects drive differences in listing price estimates in Chemnitz, experts who strongly adjusted their beliefs about the relevance of "Green voters" from Stage I to III drive our aggregate-level results. We do not find significant treatment differences in the accuracy of participants' listing price estimates as measured by the absolute deviation from actual prices. Nevertheless, our results show that using XAI as a decision support tool in one market can affect aggregate listing prices in another market in an economically considerably way (average absolute change: approximately 20%), which is not the case for opaque systems. This result demonstrates that XAI methods can link disparate decision-making tasks.

Result 2.2. *Pronounced explanation-driven belief adjustments spill over to experts' listing price estimation in a fundamentally different market.*

In summary, our results from Study 2 (i) demonstrate the robustness of our results from Study 1 on mental model adjustments in terms of system user expertise, explanation representation, and decision domain; (ii)

Figure 5. (Color online) Price Distributions in Chemnitz



Notes. We depict the distribution of experts' listing price estimates in Chemnitz. (a) and (b) Price distribution for apartments in a district with a low and high share of green voters, respectively. Different distributions show results separately for NoAid, AI, and XAI participants.

provide strong evidence that explanation-driven mental model adjustments are subject to confirmation bias; and (iii) show that explanation-driven mental model adjustments generate significant spillover effects.

4. Discussion and Conclusion

We report results from two empirical studies that provide novel insights into the interplay between the use of feature-based XAI methods and users' cognitive processes. Our main contribution is the identification of considerable side effects of providing feature-based explanations, the most popular form of XAI methods, on users' situational information processing and mental models. We find that the latter effect (i) is subject to the confirmation bias so that misconceptions can persist and even accumulate, possibly leading to suboptimal decisions, and (ii) can create spillover effects into other decision domains. These overarching results suggest that the growing, partially legally required, use of feature-based XAI methods opens a new channel through which AI systems may fundamentally reshape the way humans understand real-world relationships between features X and target variables Y. In the following, we discuss our results, present implications for organizations and society, and, based on the limitations of our studies, provide directions for future research.

4.1. Discussion of Results

Study 1 demonstrates that the provision of explanations can situationally lead lay users to adjust their weighing of features accordingly, the average absolute change in estimates equals 73.9%, and to put less emphasis on the overall prediction (−26.8%). Explanations also evoked asymmetric changes in lay users' conceptions about the relationship between borrower traits and repayment inclinations that influence behaviors even when they do not observe explanations anymore, the average absolute change in estimated coefficients equals 61.8%; that is, explanations affect mental models. Explanation-driven effects decreased lay users' decision-making performance in our setting. Compared with opaque predictions, explanations decreased participants investment performance by 8.9% while observing them and by 9.8% even when not observing explanations anymore. Study 2 extended these results in three ways. First, we find that even expert users in a considerably more applied domain adjusted mental models by about 25 percentage points. Second, results indicate that asymmetric mental model adjustments were a manifestation of the confirmation bias because posterior beliefs resembled observed explanations about 50% more closely if explanations confirmed prior beliefs. Third, Study 2 reveals that mental model adjustments created spillover effects leading to an average absolute change in apartment price estimates for a different market by approximately 20%.

From a theoretical perspective, our results contribute to our understanding of the role of popular XAI methods in effective knowledge transfers from ML-based AI systems to human users. A key promise of modern AI systems is that the application of ML techniques will discover new knowledge from Big Data that has previously eluded even experienced experts (Berente et al. 2021, van den Broek et al. 2021). This "machine knowledge" is typically codified in the form of a complex predictive model that outperforms humans. We show that providing predictions alone is insufficient to achieve systematic knowledge transfers from AI systems to human users. In both our studies, neither laymen nor experts adapted their understanding of the relationships between features X and label Y according to "machine knowledge" when observing only opaque predictions. Merely in treatments where users also had access to explanations, they began to adapt their approach to solving the task so that it more closely matched the strategy of the AI system. Therefore, XAI methods appear to be a pivotal factor contributing to an effective channel through which AI systems can pass on their self-learned knowledge to human users. Crucially, feature-based XAI methods seem to induce an asymmetry in mental model adjustments: users adjust their beliefs more in the direction of observed explanations if they confirm rather than disconfirm their priors. This asymmetry contradicts with the updating behavior of a Bayesian observer who would neither over- nor underweight explanations conditional on them confirming or disconfirming prior beliefs. This asymmetry occurred regardless of whether we provide graphically visualized LIME or numerically represented SHAP explanations. It therefore seems as if additive feature-based explanations more generally evoke cognitive processes leading users to learn from the machine selectively. Researchers across disciplines commonly refer to such an asymmetry as confirmation bias (Yin et al. 2016). Study 2 provides consistent evidence that explanation-driven knowledge transfers from an AI to a human similarly suffer from confirmation bias as knowledge transfers in the human-to-human domain. For example, confidence in prior conceptions and their difference from the new information moderate confirmation bias (Pyszczynski and Greenberg 1987). Similar to learning from other humans, users seem unwilling to internalize potentially helpful, XAI-channeled machine knowledge if it is inconsistent with what they already, perhaps incorrectly, believe to be true. From the perspective of the Mental Models framework, individuals more frequently engage in maintaining rather than in building mental models of the relationships between features and labels. One reason for this effect could be the need to attain or maintain a high level of self-esteem (Klayman 1995), leading users to focus inappropriately on explanations that make them feel competent. In other words, they may derive

a positive intrinsic benefit from being in the right (Gilad et al. 1987). From this perspective, people may misuse the XAI as a tool to enhance their self-esteem. If left unaddressed, the asymmetric adaptation of mental models by humans may prevent modern (X)AI applications from fulfilling their promise of making humans smarter, which (ironically) may also hinder the further development of AI applications by humans.

Interpreting our results in the light of the model by Agrawal et al. (2019) yields another theoretical insight regarding the ramifications of XAI. Our results indicate that users' willingness to follow XAI predictions depends on whether the explanations conform with their mental models. One way to rationalize this behavior is that their objective function includes a component that accounts for experiencing some positive (negative) intrinsic utility when obtaining a signal that their mental model may (not) be accurate (Festinger 1962, Gilad et al. 1987, Harmon-Jones 2019). In the model by Agrawal et al. (2019), AI systems make predictions about uncertain states of the world that relate to the profitability of taking specific actions. Human users, in turn, assess the expected payoffs associated with specific actions, that is, make judgments. Our results suggest that human judgment in this model encompasses not only the material consequences of an action but also the psychological impact of receiving a signal that implicitly shows whether current mental models are correct. If explanations reveal that the AI system arrived at a prediction in a way that contradicts their held mental models, taking an action that follows this prediction effectively constitutes a signal to oneself that the current mental model is incorrect, creating psychological distress, for example, in the form of a cognitive dissonance (Harmon-Jones 2019). This mental toll may lead users not to follow the prediction in the first place. Conversely, users may follow unreliable predictions more often if the explanations are consistent with their current mental models because doing so provides a psychologically valuable self-signal that they are in the right (Gilad et al. 1987). Against this background, users' inclination to follow predictions of an XAI system, and thus their ultimate decisions and gains, is subject to greater variance than with a black-box AI. That is because users' propensity to follow predictions depends on the consistency of the explanations with their mental models.

Another theoretical contribution of our work is to show the potential of feature-based XAI to link different decision domains by influencing users' beliefs about the feature-label relationship. Study 2 results show that observing explanations for listing price predictions for apartments in Market A influenced the price estimation of experts in a different Market B, where the learned pattern does not exist, and they did not have access to XAI decision support. We find that listing prices estimated by experts who observed explanations differed significantly from those estimated by experts who either had no decision aid or only observed opaque predictions. This spillover effect seems to occur because of the adjustment of mental models that experts draw on in both situations. Therefore, as an unintended side effect, increasing public and private efforts to promote the use of XAI methods may extend the already significant influence of AI systems from areas where we interact with them (Rahwan et al. 2019) to areas where such systems are not in use. Feature-based XAI methods' potential to link different domains is particularly concerning given recent evidence on their susceptibility to intentional manipulation and adversarial attacks (Lipton 2018). Many modern XAI methods, including LIME and SHAP, optimize fidelity, that is, ensure that explanations accurately mimic the predictions of the black box model. However, even small perturbations of the input data (e.g., deliberate manipulation and measurement errors) can lead to considerably different explanations for identical predictions, that is, depict different feature-label relations (Ghorbani et al. 2019, Lakkaraju and Bastani 2020). The potential instability of explanations allows manipulating user behaviors. Following our results, the creation of misleading explanations may not only affect users' trust in the AI system (Lakkaraju and Bastani 2020) but also lead to an (asymmetric) adjustment of mental models that affect users' decision making beyond the XAI augmented decision at hand. Specifically, the depiction of certain feature-label relationships that are not present can evoke inappropriate mental model adjustments that, given the documented asymmetry, will cause users who already believe these patterns to be true, to feel vindicated and reinforce these beliefs. In general, the documented spillover effects may magnify the reach and impact of intentional manipulations of explanations, increasing deceiving parties' incentive to do so.

4.2. Implications

Reported results have important practical implications for organizations and policymakers. Our finding that XAI can change human thinking points to potential pitfalls for companies that want, or have to, use XAI. Consider a company that plans to implement XAI methods to explain to its employees why an AI system makes certain predictions. As Study 1 shows, providing explanations in addition to predictions may draw users' attention excessively to the explanations, to the detriment of the prediction itself. Users may place too much emphasis on individual explanations that confirm their prior beliefs, rather than adhering to the overall prediction. As a result, employees' decision-making performance for the task at hand may deteriorate, which is in line with evidence from related research (Poursabzi-Sangdeh et al. 2021). In domains where explanations are becoming a regulatory standard, managers need to take such potential downsides into account and contemplate

the ramifications of implementing explainability measures. Following our results, managers who, in the future, are obliged to put XAI methods in place, should not take these steps too lightly. From a business perspective, our documented downsides of explainability could render the continued use of AI-based decision support systems unattractive. Considering that AI systems are often deeply interwoven with business processes, this XAI-driven discontinuance may entail considerable organizational change. As a result, managers may be well advised to assess potential inconsistencies between the AI system's internal logic and employees' understanding of the task it supports before rolling out explainability measures. This puts managers in a position to evaluate the magnitude of the potential downside of explainability and use countermeasures. For example, managers may obviate confirmation bias by openly discussing explanations that conflict with employees' mental models and showcasing arguments in support of the explanation.

Another pitfall for companies concerns the transfer of knowledge from AI systems to human users. As Study 2 shows, even experts can overgeneralize learned feature-label relationships that are only applicable in the context in which they interact with the system. With the confirmatory learning from explanations, existing differences in employees' initial conceptions may lead to differences in how they collaborate with and what they learn from the XAI, for example, fostering the biased weighting of certain information. From this perspective, providing explanations might decrease individual level noise in the decision-making process (Kahneman et al. 2021) because individuals' decisions become more consistent. This is in line with Fügener et al. (2021b), who find decisions to be increasingly consistent among users engaging with opaque predictions. On a more aggregate level, however, our results suggest that explained predictions may additionally foster differences in the decision-making process across subgroups of users that possess heterogeneous priors. As a consequence, the variation of decisions on a group level can grow. As pointed out by Kahneman et al. (2021), variation in decisions can substantially contribute to errors and ultimately harm business performance. Consider our previous example of loan officers. XAI may cause loan approval decisions to increasingly depend on the particular employee, with idiosyncratic mental models, assessing the applicant's creditworthiness. This increase in loan approval variation may create considerable business, legal, and reputational risks. Against this background, managers should closely monitor the introduction of XAI to identify a possible increase in decision variance. For instance, managers could complement XAI with "noise audits" and the development of "reasoned rules" (as proposed by Kahneman et al. 2021) to overcome the hidden costs of XAI-driven increases in inconsistent decision making.

From a societal perspective, our results indicate that broad, indiscriminate implementation of XAI methods may create unintended downstream ramifications. Our finding that XAI can lead users to adjust mental models in a confirmatory way and carry over learned patterns to other domains may, in an extreme case, foster discrimination and social divisions. Assume all recruiters start to collaborate with an XAI system to support hiring decisions. For example, a subgroup of recruiters may discriminate against women because they believe female applicants to be less productive on the job. If the XAI (occasionally) provides local explanations that depict being female as negative evidence for high future performance, the subgroup that statistically discriminates based on gender will readily reinforce its prior belief, that is, engage in mental model maintenance. As a result, these recruiters may become more biased and less noisy in their behavior as they hire female applicants consistently less. Given the spillover effects we find, they may even carry over their strengthened conceptions about women's productivity to other jobs, further reinforcing discriminatory patterns. Additionally, because nondiscriminating recruiters will most likely refrain from adjusting their mental model, that is, not engage in mental model building, social divisions among recruiters may develop and accumulate along the lines of gender biases. Hence, without any malicious intent, the broad use of XAI may ironically foster human discriminatory tendencies and divide social groups. Notably, with the possibility to manipulate explanations, deceiving third parties could also intentionally cause explanations to exhibit specific prediction contributions for sensitive attributes such as race, gender, or age. This effect could lead human users who already hold prejudices, stereotypes, or discriminatory tendencies to reinforce their views, which could promote certain political agendas.

4.3. Limitations and Future Research

As with any other research study, ours is not without limitations. In light of increasing regulatory requirements and private initiatives, we believe that these limitations open up fruitful avenues for future research. One limitation of our work concerns the lack of feedback on the decision outcomes and thus the performance of the AI system. In both our studies, we did not provide feedback for two reasons. First, it adds a considerable layer of complexity that impedes the measurement and interpretation of isolated explanation-driven effects on users' cognitive processes. Second, in practice, many AI-supported decisions do not yield immediate feedback, or only yield feedback for some of the predictions. Hence, users have to interact with the system without learning its prediction accuracy, at least for a certain period. Examples include hiring decisions supported by an on-the-job performance predicting AI

system, investment decisions supported by a return predicting AI system, and drug treatment decisions supported by an effectiveness predicting AI system. Consequently, explanations may alter users' situational information processing and mental models before feedback on system performance arrives. Nonetheless, we strongly encourage future research to examine the role of feedback as it may introduce unexpected dynamics in the cognitive effects we document. For instance, the (selective) reinforcement of their mental models through explanations, may lead users to be more forgiving and maintain trust in the AI system, even if they eventually see it making mistakes. In this way, the interaction between feedback and explanations might constitute a factor contributing to unwarranted algorithm appreciation (Logg et al. 2019), leading users to rely on incorrect outcomes blindly. Additionally, people's adjustments of the situational information processing and existing mental models possibly depend on the extent to which the XAI system's predictions outperform their own. If users learn that an XAI system's predictions perform considerably better than their subjective ones, the magnitude of reported confirmation biases may vary. Conversely, when users' predictions are better than the XAI, their confirmation bias might be even stronger. Future research could examine to what extent our reported effects, at the intensive margin, depend on users' perceptions about differences in their own and the XAI system's predictive performance.

Another limitation of our work originates from letting participants interact with local, feature-based XAI methods. We opted to use these explanations because they are already widely in use in practice and because there are arguments that feature-based explanations on an individual level are necessary to comply with (upcoming) regulatory requirements (Goodman and Flaxman 2017). Yet, there exist other forms of explanations, for example, global feature-based explanations or even example-based explanations. Although an investigation and comparison of the interplay between different forms of explanations and cognitive processes are beyond the scope of this paper, it is worthwhile for future research to explore whether, and if so why, the effects we document would change if users (additionally) obtain other forms of explanations. Consider, for instance, global explanations. Although local explanations help understand why an AI system produces a prediction on a case-by-case basis, global explanations reveal important high-level patterns and nonlinearities in the system's logic. Such global explanations effectively aggregate individual-level information for the user and help to understand the system's overall logic. By taking over this information aggregation task, global explainability could mitigate concerns about the selective processing of isolated local explanations that arguably contribute to the occurrence of confirmation bias. Additionally, the global representation may facilitate comparison and reflection processes that ultimately improves the transfer of knowledge from the AI system to the user.

4.4. Conclusion

A concluding remark is worth making. Of course, our work is not meant to be an argument, let alone a plea, against making "black box" AI systems more explainable or transparent. Instead, we comprehend our findings as a warning that the indiscriminate use of modern XAI methods as an isolated measure may lead to unintended, unforeseen problems because it creates a new channel through which AI systems can affect human behaviors across domains. The pervasive human inclination to process information in a way that confirms their preconceptions while ignoring potentially helpful yet conflicting information needs addressing if explainability is to become an effective means to combat accountability, transparency, and fairness issues without creating adverse second-order effects. For instance, one might restrict the provision of explanations of sensitive features for end users of the system and only use them to ensure the proper and unbiased functioning of the AI system during the development process. Additionally, it might be important to provide developers and data scientists with cognitive awareness trainings to make them more sensitive to their own biased mental processes.

Endnotes

[1] For anecdotal evidence of such nontraditional data use, see LendoEFL.com or https://money.cnn.com/2016/08/24/technology/lenddo-smartphone-battery-loan/index.html.

[2] On a high level, both decisions effectively constitute sequential economic transactions under uncertainty that strongly depend on trust.

[3] Explainability may enter the model of Agrawal et al. by changing the prediction reliability. Following Proposition 2, the necessity for providing explanations decreases with the users' judgment. However, the model does not consider the idea presented in our paper that explainability may also affect users' understanding of the process that determines the uncertain state of the world the AI tries to predict. One could integrate this notion into the framework by modeling that explanations affect users' judgment capabilities by influencing beliefs about underlying processes. Extending the model of Agrawal et al. in this direction may be a fruitful endeavor to better understand whether explainability modulates the relationship between prediction and judgment. However, an extension of the formal model is beyond the scope of this paper and left for future research.

[4] See the online appendix for details on the experimental procedures including payments, instructions, and screenshots.

[5] When a risk-neutral, purely self-interested investor expects that the borrower repays her with a probability of $p > 0.5$, for example, because she believes the borrower to possess altruistic, efficiency, or fairness preferences, they have a strict incentive to invest because they maximize their expected earnings. Importantly, holding such expectations about the borrower's preferences is justified and frequently observed

in sequential games: A considerable share of people does respond reciprocally in sequential exchanges if they are trusted (see Miettinen et al. 2020 for an overview).

[6] To reduce the complexity for the reader, we only report the three main stages of the experiment. Right before and after Stage II, we additionally measured participants' prior and posterior preferences to observe three borrower characteristics. We use these measures as robustness and consistency checks. We provide a detailed description of these measurements in the online appendix.

[7] The questionnaire items included in the field study were selected partly for exploratory reasons and partly motivated by previous research documenting their association with individuals' repayment behavior in investment games (Ben-Ner and Halldorsson 2010).

[8] For most traits, values and LIME values are almost perfectly correlated producing severe problems of multicollinearity (see Table 7 in the online appendix). Therefore, in our regression analyses, we only include LIME explanations for which there exists a tolerable correlation between the trait and LIME values: Openness, Agreeableness, and Conscientiousness.

[9] See Table 8 in the online appendix.

[10] These results do not allow us to isolate how explanations affect what investors consider to be a borrower's motivation to repay them or not. The change in the weighting of competitiveness could stem from a reinforced perception that competitiveness predicts a low repayment likelihood because it proxies for antisocial, income-maximizing, or relative income-maximizing motives. Although we cannot isolate investors' latent belief(s) about borrowers' motives, our results effectively show that the provision of explanations does entail a change in at least one of these perceived latent motives, that is, that XAI can change the processing of information. A similar argument applies regarding mental model adjustments outlined later.

[11] Reported results are robust to excluding participants who always or never invested in our analyses, respectively, alleviating concerns that our results are driven by pure altruists or players who always choose the game-theoretically dominant strategy (see the subsection on additional robustness checks in the online appendix). Instead, our results stem from those participants whose behavior suggests that they try to invest with borrowers whom they believe will make a repayment, that is, individuals who, from a conceptual point of view, should be most inclined to learn to recognize repaying borrowers. Results 1.2 and 1.3 are equally robust to excluding these "extreme" types, warranting a similar interpretation.

[12] The significant explanation effect for *Openness* and *Extraversion* may be a consequence of participants' significantly stronger weighting of borrowers' *Competitiveness* and *Patience* and a limited capacity to process information. Specifically, XAI participants in Stage III place similarly low weight on all borrower traits but *Competitiveness*, *Agreeableness*, and *Patience*. This pattern may suggest that participants heuristically focus on the three characteristics that they themselves and the AI system deemed most relevant to the decision. As a result, they place less weight on all other traits, which for Openness led to a statistically significant effect.

[13] We show ROC curves in Figures 17 to 19 in the online appendix.

[14] Participants neither knew their own nor the AI system's performance because we did not provide intermediate feedback. Therefore, they could not see how much better or worse the system performs compared with themselves. Although unknown to participants, predictions are accurate in about 69.3% of the cases. This performance holds equally for both repaying (69.7%) and nonrepaying borrowers (67.7%). Participants in Stage I correctly invested with (non-)repaying borrowers in 66.1% (41.2%) of the cases and overall in 60.5% of the cases. Put differently, the AI system outperforms them overall (+14.5%) and especially for the identification of nonrepaying ones (+64.3%). As a result, participants could have benefited from relying on the predictions, which baseline participants did at least partially.

[15] Across Stages I and II, baseline participants' access to the AI system significantly increased the accuracy by 4.6% ($p < 0.01$, F test), whereas the recall effectively remained constant ($p < 0.82$, F test). XAI participants performance significantly decreased regarding both the accuracy (−5.3%; $p < 0.01$, F test) and recall score (−14.6%; $p < 0.01$, F test).

[16] A purely linear distinction between most competitive and other borrowers does not allow to draw conclusions about their repayment likelihood: they, respectively, made a repayment in 77.4% and 79.8% of the cases ($p = 0.85$, F test).

[17] We scraped data from a large online platform in February 2022. We collected observations for all apartments listed for sale in the seven major cities of Germany ("A-Cities") and a medium-sized eastern German city (Chemnitz). We constructed a data set consisting of eight apartment attributes and the listing price directly obtained from the platform and two additionally collected features from public statistics. We provide summary statistics in the online appendix (Table 6).

[18] We selected these three characteristics for technical reasons regarding the ML model and based on the input from our industry partner. The notion is that these characteristics together are (i) sufficiently relevant to the prediction and (ii) familiar/accessible to experts.

[19] The AI system is a random forest that achieves a performance of $R^2 = 0.72$ on unseen test data. See the online appendix for additional information.

[20] For instance, A-cities exhibit considerably higher average wages, more liberal political attitudes, and faster population growth (Cajias et al. 2020).

[21] The positive coefficient for β_5 may be related to the fact that SHAP values and overall predictions are inextricably linked. Merely observing high (low) predictions may lead to adjustments of reported beliefs upward (downward), creating a positive, however, insignificant correlation with underlying SHAP values in the data.

[22] Participants, on average, have worked in the real estate industry for 13.8 years and, on a scale from 1 to 10, report that their experience level in rating apartment listing prices is 5.7.

[23] Our main insights are robust to defining more restrictively that explanations confirm priors if the absolute distance between the prior and the observed average SHAP value is smaller than the absolute distance between the prior and 0€ and, at the same time, smaller than the absolute distance between the prior and the closest extreme, that is, ±2,500€ (see Table 25 in the online appendix).

[24] We did not include Chemnitz observations in the data to train the AI model. We conducted several analyses showing that the most important predictors for listing prices in Frankfurt and Cologne (cities in Stages I to III) differ considerably from listing price predictors in Chemnitz. Real estate experts are arguably aware of the structural differences in apartment markets.

References

Abdel-Karim BM, Pfeuffer N, Carl V, Hinz O (2022) How AI-based systems can induce reflections: The case of AI-augmented diagnostic work. *Management Inform. Systems Quart.* Forthcoming.

Abdel-Karim BM, Pfeuffer N, Rohde G, Hinz O (2020) How and what can humans learn from being in the loop? *German J. Artificial Intelligence* 34(2):199–207.

Agarwal R, Dhar V (2014) Big data, data science, and analytics: The opportunity and challenge for IS research. *Inform. Systems Res.* 25(3):443–448.

Agrawal A, Gans JS, Goldfarb A (2019) Exploring the impact of artificial intelligence: Prediction vs. judgment. *Inform. Econom. Policy* 47:1–6.

Alavi M, Marakas GM, Yoo Y (2002) A comparative study of distributed learning environments on learning outcomes. *Inform. Systems Res.* 13(4):404–415.

Ban GY, El Karoui N, Lim AE (2018) Machine learning and portfolio optimization. *Management Sci.* 64(3):1136–1154.

Bauer K, Hinz O, van der Aalst W, Weinhardt C (2021) Expl(AI)n it to me: Explainable AI and information systems research. *Bus. Inform. Systems Engrg.* 63(2):79–82.

Ben-Ner A, Halldorsson F (2010) Trusting and trustworthiness: What are they, how to measure them, and what affects them. *J. Econom. Psych.* 31(1):64–79.

Berente N, Gu B, Recker J, Santhanam R (2021) Managing artificial intelligence. *Management Inform. Systems Quart.* 45(3):1433–1450.

Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity, and social history. *Games Econom. Behav.* 10(1):122–142.

Bhatt U, Xiang A, Sharma S, Weller A, Taly A, Jia Y, Ghosh J, et al. (2020) Explainable machine learning in deployment. *Proc. Conf. on Fairness, Accountability, and Transparency* (Association for Computing Machinery, New York).

Brewer WF (1987) Schemas vs. mental models in human memory. Morris P, ed. *Modelling Cognition* (John Wiley & Sons, Oxford, UK), 187–197.

Bussone A, Stumpf S, O'Sullivan D (2015) The role of explanations on trust and reliance in clinical decision support systems. *Proc. Internat. Conf. on Healthcare Informatics* (Institute of Electrical and Electronics Engineers (IEEE), New York).

Cabral TS (2021) AI and the right to explanation: Three legal bases under the GDPR. *Data Protection Artificial Intelligence* 13:29–56.

Cajias M, Freudenreich P, Freudenreich A, Schäfers W (2020) Liquidity and prices: A cluster analysis of the German residential real estate market. *J. Bus. Econom.* 90(7):1021–1056.

Case N (2018) How to become a centaur. *J. Design Sci.* https://jods.mitpress.mit.edu/pub/issue3-case/release/6?version=53b19e72-d43a-4eda-8c48-6ed3cdc03218.

Castelo N, Bos MW, Lehmann DR (2019) Task-dependent algorithm aversion. *J. Marketing Res.* 56(5):809–825.

Chatterjee S, Sarker S, Valacich JS (2015) The behavioral roots of information systems security: Exploring key factors related to unethical IT use. *J. Management Inform. Systems* 31(4):49–87.

DeGroot MH (1974) Reaching a consensus. *J. Amer. Statist. Assoc.* 69(345):118–121.

Dellermann D, Ebel P, Söllner M, Leimeister JM (2019) Hybrid intelligence. *Bus. Inform. Systems Engrg.* 61(5):637–643.

Dhaliwal JS, Benbasat I (1996) The use and effects of knowledge-based system explanations: Theoretical foundations and a framework for empirical evaluation. *Inform. Systems Res.* 7(3):342–362.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Experiment. Psych. General* 144(1):114–126.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Sci.* 64(3):1155–1170.

Dijkstra JJ (1999) User agreement with incorrect expert system advice. *Behav. Informs. Tech.* 18(6):399–411.

Dodge J, Liao QV, Zhang Y, Bellamy RK, Dugan C (2019) Explaining models: An empirical study of how explanations impact fairness judgment. *Proc. Internat. Conf. on Intelligent User Interfaces*.

Doshi-Velez F, Kim B (2017) Toward a rigorous science of interpretable machine learning. Preprint, submitted March 2, https://arxiv.org/abs/1702.08608.

Erlei A, Nekdem F, Meub L, Anand A, Gadiraju U (2020) Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. *Proc. AAAI Conf. on Human Comput. and Crowdsourcing*.

EU (2016) Regulation EU 2016/679 of the European Parliament and of the Council of 27 April 2016, article 22. *Official J. Eur. Union Law* 119:59.

EU (2021) Proposal for a regulation EU of the European Parliament and of the Council of April 21, 2021, laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. *Official J. Eur. Union Law* 119. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

Festinger L (1962) Cognitive dissonance. *Sci. Amer.* 207(4):93–106.

Fügener A, Grahl J, Gupta A, Ketter W (2021a) Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Inform. Systems Res.* 33(2): 678–696.

Fügener A, Grahl J, Gupta A, Ketter W (2021b) Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *Management Inform. Systems Quart.* 45(3b):1527–1556.

Garreau D, Luxburg U (2020) Explaining the explainer: A first theoretical analysis of LIME. *Proc. Internat. Conf. on Artificial Intelligence and Statist*.

Ge R, Zheng Z, Tian X, Liao L (2021) Human–robot interaction: When investors adjust the usage of robo-advisors in peer-to-peer lending. *Inform. Systems Res.* 32(3):774–785.

Ghorbani A, Abid A, Zou J (2019) Interpretation of neural networks is fragile. *Proc. AAAI Conf. on Artificial Intelligence*. 33(1): 3681–3688.

Gilad B, Kaish S, Loeb PD (1987) Cognitive dissonance and utility maximization: A general framework. *J. Econom. Behav. Organ.* 8(1):61–73.

Goldstein IM, Lawrence J, Miner AS (2017) Human-machine collaboration in cancer and beyond: The centaur care model. *JAMA Oncology* 3(10):1303–1304.

Goodman B, Flaxman S (2017) European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* 38(3):50–57.

Google AI (2019) Responsible AI practices: Interpretability. Accessed March 8, 2022, https://ai.google/responsibilities/responsible-ai-practices/?category=interpretability.

Gramegna A, Giudici P (2021) SHAP and LIME: An evaluation of discriminative power in credit risk. *Frontiers Artificial Intelligence* 4:752558.

Gregor S (2006) The nature of theory in information systems. *Management Inform. Systems Quart.* 30(3):611–642.

Gregor S, Benbasat I (1999) Explanations from intelligent systems: Theoretical foundations and implications for practice. *Management Inform. Systems Quart.* 23(4):497–530.

Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ (2019) XAI—explainable artificial intelligence. *Sci. Robot* 4(37):eaay7120.

Harmon-Jones EE (2019) *Cognitive Dissonance: Reexamining a Pivotal Theory in Psychology* (American Psychological Association).

Hemmer P, Schemmer M, Vössing M, Kühl N (2021) Human-AI complementarity in hybrid intelligence systems: A structured literature review. *Proc. 28th Pacific Asia Conf. on Inform. Systems.*

Hoffman M, Kahn LB, Li D (2018) Discretion in hiring. *Quart. J. Econom.* 133(2):765–800.

Holt CA, Smith AM (2009) An update on Bayesian updating. *J. Econom. Behav. Organ.* 69(2):125–134.

Ji-Ye Mao IB (2000) The use of explanations in knowledge-based systems: Cognitive perspectives and a process-tracing analysis. *J. Management Inform. Systems* 17(2):153–179.

Johnson-Laird PN, Goodwin GP, Khemlani SS (2017) Mental models and reasoning. *The Routledge International Handbook of Thinking and Reasoning* (Routledge, Abingdon-on-Thames, UK), 346–365.

Jones NA, Ross H, Lynam T, Perez P, Leitch A (2011) Mental models: An interdisciplinary synthesis of theory and methods. *Ecological Soc.* 16(1). https://www.jstor.org/stable/26268859#metadata_info_tab_contents.

Jussupow E, Benbasat I, Heinzl A (2020) Why are we averse toward algorithms? A comprehensive literature review on algorithm aversion. *Proc. Eur. Conf. on Inform. Systems.*

Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? An investigation into physicians'

decision-making process with artificial intelligence. *Inform. Systems Res.* 32(3):713–735.

Kahneman D, Sibony O, Sunstein CR (2021) *Noise: A Flaw in Human Judgment* (Little, Brown).

Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Wortman Vaughan J (2020) Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. *Proc. CHI Conf. on Human Factors in Comput. Systems*.

Klayman J (1995) Varieties of confirmation bias. *Psych. Learning Motives* 32:385–418.

Kleinmuntz B (1990) Why we still use our heads instead of formulas: Toward an integrative approach. *Psych. Bull.* 107(3):296.

Knobloch-Westerwick S, Meng J (2009) Looking the other way: Selective exposure to attitude-consistent and counterattitudinal political information. *Comm. Res.* 36(3):426–448.

Koh PW, Liang P (2017) Understanding black-box predictions via influence functions. *Proc. Internat. Conf. on Machine Learn*.

Lakkaraju H, Bastani O (2020) "How do I fool you?" Manipulating user trust via misleading black box explanations. *Proc. AAAI/ACM Conf. on AI, Ethics, and Society*.

Lakkaraju H, Kamar E, Caruana R, Leskovec J (2019) Faithful and customizable explanations of black box models. *Proc. AAAI/ACM Conf. on AI, Ethics, and Society*.

Lim KH, Ward LM, Benbasat I (1997) An empirical study of computer system learning: Comparison of co-discovery and self-discovery methods. *Inform. Systems Res.* 8(3):254–272.

Lipton ZC (2018) The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16(3):31–57.

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organ. Behav. Human Decision Processes* 151:90–103.

Lu Z, Yin M (2021) Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. *Proc. CHI Conf. on Human Factors in Comput. Systems*.

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Proc. Conf. on Neural Inform. Processing Systems*.

Malle BF (2006) *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction* (MIT Press, Cambridge, MA).

Meske C, Bunde E, Schneider J, Gersch M (2022) Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Inform. Systems Management* 39(1):53–63.

Meta AI (2021) Facebook's five pillars of responsible AI. Accessed March 8, 2022, https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/.

Miettinen T, Kosfeld M, Fehr E, Weibull J (2020) Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions. *J. Econom. Behav. Organ.* 173:1–25.

Molnar C (2020) *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Accessed January 14, 2022, https://christophm.github.io/interpretable-ml-book.

Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan JW, Wallach H (2021) Manipulating and measuring model interpretability. *Proc. CHI Conf. on Human Factors in Comput. Systems*.

Pyszczynski T, Greenberg J (1987) Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. *Adv. Experiment. Soc. Psych.* 20:297–340.

Rabin M, Schrag JL (1999) First impressions matter: A model of confirmatory bias. *Quart. J. Econom.* 114(1):37–82.

Rader E, Cotter K, Cho J (2018) Explanations as mechanisms for supporting algorithmic transparency. *Proc. CHI Conf. on Human Factors in Comput. Systems*.

Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon JF, Breazeal C, Crandall JW, et al. (2019) Machine behaviour. *Nature* 568(7753):477–486.

Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?" Explaining the predictions of any classifier. *Proc. ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining*.

Rico-Juan JR, de La Paz PT (2021) Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Systems Appl.* 171:114590.

Rosenfeld A, Richardson A (2019) Explainability in human–agent systems. *Autonomic Agent Multi Agent Systems* 33(6):673–705.

Rouse WB, Morris NM (1986) On looking into the black box: Prospects and limits in the search for mental models. *Psych. Bull.* 100(3):349.

Schanke S, Burtch G, Ray G (2021) Estimating the impact of "humanizing" customer service chatbots. *Inform. Systems Res.* 32(3):736–751.

Schön DA (2017) *The Reflective Practitioner: How Professionals Think in Action* (Routledge, Abingdon-on-Thames, UK).

Senoner J, Netland T, Feuerriegel S (2021) Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing. *Management Sci.* 68(8):5704–5723.

Shapley LS (1953) A value for n-person games. *Contributions to the Theory of Games (AM-28)*, vol. II (Princeton University Press, Princeton, NJ).

Teodorescu MH, Morse L, Awwad Y, Kane GC (2021) Failures of fairness in automation require a deeper understanding of human-ML augmentation. *Management Inform. Systems Quart.* 45(3b): 1483–1499.

Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, Janda M, et al. (2020) Human–computer collaboration for skin cancer recognition. *Nature Medicine* 26(8):1229–1234.

van den Broek E, Sergeeva A, Huysman M (2021) When the machine meets the expert: An ethnography of developing AI for hiring. *Management Inform. Systems Quart.* 45(3): 1557–1580.

Vandenbosch B, Higgins C (1996) Information acquisition and mental models: An investigation into the relationship between behaviour and learning. *Inform. Systems Res.* 7(2):198–214.

Vilone G, Longo L (2021) Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inform. Fusion* 76:89–106.

Wang W, Benbasat I (2007) Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *J. Management Inform. Systems* 23(4):217–246.

Willison R, Warkentin M (2013) Beyond deterrence: An expanded view of employee computer abuse. *Management Inform. Systems Quart.* 37(1):1–20.

Yang F, Huang Z, Scholtz J, Arendt DL (2020) How do visual explanations foster end users' appropriate trust in machine learning? *Proc. Internat. Conf. on Intelligent User Interfaces*.

Yin D, Mitra S, Zhang H (2016) Research note—When do consumers value positive vs. negative reviews? An empirical investigation of confirmation bias in online word of mouth. *Inform. Systems Res.* 27(1):131–144.