Kevin Bauer | Moritz von Zahn | Oliver Hinz

Please Take Over: XAI, Delegation of Authority, and Domain Knowledge

SAFE Working Paper No. 394 | July 2023

Leibniz Institute for Financial Research SAFE

Sustainable Architecture for Finance in Europe

info@safe-frankfurt.de | www.safe-frankfurt.de

Electronic copy available at: https://ssrn.com/abstract=4512594

Please take over: XAI, delegation of authority, and domain knowledge[*]

Kevin Bauer, Moritz von Zahn, Oliver Hinz

July 15, 2023

Abstract

Recent regulatory measures such as the European Union's AI Act require artificial intelligence (AI) systems to be explainable. As such, understanding how explainability impacts human-AI interaction and pinpointing the specific circumstances and groups affected, is imperative. In this study, we devise a formal framework and conduct an empirical investigation involving real estate agents to explore the complex interplay between explainability of and delegation to AI systems. On an aggregate level, our findings indicate that real estate agents display a higher propensity to delegate apartment evaluations to an AI system when its workings are explainable, thereby surrendering control to the machine. However, at an individual level, we detect considerable heterogeneity. Agents possessing extensive domain knowledge are generally more inclined to delegate decisions to AI and minimize their effort when provided with explanations. Conversely, agents with limited domain knowledge only exhibit this behavior when explanations correspond with their preconceived notions regarding the relationship between apartment features and listing prices. Our results illustrate that the introduction of explainability in AI systems may transfer the decision-making control from humans to AI under the veil of transparency, which has notable implications for policy makers and practitioners that we discuss.

1 Introduction

In recent years, the rapid advancement and adoption of artificial intelligence (AI) systems across business domains (see, e.g., Brynjolfsson and McAfee, 2017) led to a growing need for increased transparency in AI-supported decision-making processes (Mittelstadt et al., 2016). That is because the black box nature of AI

[*] We gratefully acknowledge research support from the University of Mannheim, the Leibniz Institute for Financial Research SAFE, and the Goethe University Frankfurt. Bauer: University of Mannheim, L15, 1-6, D-68161 Mannheim, Germany. Email: kevin.bauer@uni-mannheim.de. v. Zahn and Hinz: Goethe University Frankfurt & Leibniz Institute for Financial Research SAFE, Theodor-W.-Adorno-Platz 4, D-60323 Frankfurt am Main, Germany. Email: vzahn@wiwi.uni-frankfurt.de; hinz@wiwi.uni-frankfurt.de.

Electronic copy available at: https://ssrn.com/abstract=4512594

systems frequently hinders user trust, impedes error safeguarding, and obstructs knowledge transfers from AI to human users due to the opaque transformation of inputs into outputs (see, e.g., Bauer et al., 2021).

One approach to address these pitfalls is the use of explainable AI (XAI), which aims to provide human-interpretable explanations of AI-generated outputs, facilitating a more thorough understanding of the underlying mechanisms and logic employed by AI systems (Miller, 2019; Arrieta et al., 2020). Thereby, from a business perspective, explainability may contribute to trust in and utilization of AI outputs. Policy makers have also recognized the importance of explaining AI logic from a consumer perspective, leading them to impose guidelines and requirements for the transparency, explainability, and human oversight of AI systems, see, e.g., the European Union's AI Act (Commission, 2021). Given the imperative to confront the issues of "black box" AI from both business and regulatory standpoints, it is crucial to understand the effects of explainability on human-AI interaction. This understanding is particularly vital as the efficiency of organizational decision-making processes increasingly hinges upon the interaction between humans and AI systems (see, e.g., Berente et al., 2021). Gaining a profound understanding of the specific groups of individuals and circumstances where explainability may alter various aspects of decision-making processes that involve AI systems may help anticipate and potentially mitigate any adverse repercussions associated with the (legally required) switch from AI to XAI systems.

A key facet of decision-making processes that explainability of AI systems may affect is the delegation of authority, i.e., the ceasing of decision control (Agrawal et al., 2019; Athey et al., 2020; Baird and Maruping, 2021). Following the seminal model by Aghion and Tirole (1997), there are two types of authority: formal and real authority. Formal authority pertains to individuals' formal right to make decisions and have the final say. Real authority, in contrast, refers to individuals' effective control over decisions that arises from being better informed. Importantly, formal and real authority over a decision do not necessarily rest with the same individual in organizations. Formal authority always lies with superiors who have the formal right to make decisions including the power to eventually override their subordinate's decision. However, when subordinates are better informed and do not possess (overly) conflicting objectives, it is optimal for formal authority holders to rubber-stamp subordinates' suggested decisions instead of exerting effort to become better informed. In these cases, they effectively relinquish control over the decision. Notably, the reduction of effort separates the delegation of real authority from leveraging suggestions as decision support: whereas the former is the blind rubber-stamping of the suggestion, the latter is about thoroughly assessing suggestions to combine them with one's own conclusions.

As a tangible example in the real-estate sector, consider a manager who is formally responsible to evaluate luxury apartments for her clients. To make informed evaluations, she can do the necessary research on her own, or ask her subordinate analyst to come up with suggestions. The manager always has the final say. However, if she passes on the suggested evaluations to her

Electronic copy available at: https://ssrn.com/abstract=4512594

clients without doing her own thorough research, it is effectively the analyst who provides the service, i.e., has control over the evaluation decision. The distribution of formal and real authority is pivotal because it shapes the exertion of effort at the individual level, and the flow of information, as well as agency costs at the organizational level (Aghion and Tirole, 1997).

As AI systems continue to permeate organizations, the delegation of real authority increasingly shifts to settings where formal authority holders cease the effective decision control to AI systems instead of other humans (Fügener et al., 2022). This shift can result in more efficient and data-driven decision-making. However, the delegation of real authority to AI systems may also create adverse effects including a loss of human oversight and agency in case decision-makers, instead of exerting effort themselves, blindly rubber-stamp the AI suggestion (Berente et al., 2021; Busuioc, 2021; Meske et al., 2022). At the core of our study lies the notion that AI system explainability adds a new layer of complexity to human-AI interaction that may either contribute to a more nuanced delegation of real authority to AI or foster the excessive ceasing of control to AI systems.

Understanding whether, and if so how and when, explainability affects the delegation of real authority is crucial as the distribution of decision control typically affects economic efficiencies (Dessein, 2002). According to prior research, several determinants shape individuals' delegation of real authority to a subordinate, including individuals' level of expertise (Aghion and Tirole, 1997), beliefs about the subordinate's competence (Dessein, 2002), monitoring costs and trust (Dobrajska et al., 2015), and intrinsic preferences for agency and control (Bartling et al., 2014). In the context of human-AI interaction, we conjecture that the explainability of AI systems is another factor shaping the delegation of real authority because it contributes to bridging information gaps between the agentic information systems artifact (i. e., the AI system) and the human user as proposed in a framework by Baird and Maruping (2021). Specifically, prior research studying the impact of explainability on human-AI interaction provides mixed evidence regarding trust, and performance beliefs (see, e.g., Lim et al., 2009; Szymanski et al., 2021; Zhang et al., 2020; Alufaisan et al., 2021). Considering that these factors shape the delegation of real authority in the human-human domain, it is unclear whether explainability of AI systems counteracts or promotes the delegation of effective decision control.

The study at hand aims to fill our limited understanding of how explainability influences human's inclination to relinquish decision control to AI by developing a formal framework and conducting an incentivized empirical field experiment. Drawing inspiration from theories in economics and information systems (Aghion and Tirole, 1997; Dobrajska et al., 2015; Baird and Maruping, 2021), we place particular emphasis on exploring the role of users' existing domain knowledge and associated preconceptions about how to utilize available information to make an informed decision. Specifically, we pose two research questions:

(i) How does the introduction of explainability affect individuals' propensity

Electronic copy available at: https://ssrn.com/abstract=4512594

to delegate real authority to an AI system?

(ii) Does users' level of domain knowledge influence the effects of explainability, and, if so, under what circumstances?

Our incentivized field experiment involved 153 German real estate agents tasked with evaluating various apartments in two major German cities. The study examines three different conditions: agents performing apartment valuations on their own (control condition), agents working alongside an AI system providing black box price predictions (AI condition), and agents working with an XAI system offering feature-based explanations in addition to price predictions (XAI condition). The (X)AI system utilized in our study is a machine learning model accompanied by SHAP explanations (Lundberg and Lee, 2017), which we developed based on data collected from a prominent German real estate platform. To answer our first research question, we aim to isolate the effect of explainability on real estate agents' inclination to use price predictions as their apartment valuations and reduce their effort, i.e., their ceasing of decision control. We adopt a conservative approach to defining the delegation of real authority to the (X)AI system by considering decision control ceased only when an agent reduces her cognitive effort and simply rubber-stamps the AI prediction. To answer our second research question, we gauge the agents' expertise in estimating listing prices for German apartments, as well as their measured beliefs of how various apartment features influence those prices.

Selecting a real estate appraisal setting is well-suited for addressing our research questions for multiple reasons. First, the real estate industry, much like other sectors, has increasingly adopted AI systems for a range of tasks, including property price estimation (Olick, 2021; Tchuente and Nyawa, 2022) so that examining the impact of XAI in this context offers a realistic portrayal of the challenges and opportunities encountered by professionals in the field. Second, the decision-making process in real estate appraisals typically carries significant financial consequences for clients, rendering the delegation of real authority to AI systems highly relevant. Third, real estate price evaluation is characterized by high complexity, involving numerous features and factors that influence the final price. This intricate environment provides an ideal opportunity to evaluate the effectiveness of XAI systems in delivering explanations that aid users in comprehending and assessing AI-generated predictions.

Our research aims to offer managerial implications regarding the soon-to-be legally required integration of explainability measures in certain AI-assisted decision-making processes. With our field experiment, we can shed light on the question if and how explainability can motivate individuals to delegate real authority to AI systems depending on their domain knowledge. Based on our findings, we discuss the consequences of AI system explainability for the variability of individuals' decisions, and their inclination to "fall asleep at the wheel". From a more holistic perspective, we elaborate on the relationship between explainability and organizational decision consistency, which subsequently affects economic performance as highlighted by Kahneman et al. (2021), and the role

Electronic copy available at: https://ssrn.com/abstract=4512594

of expertise among the workforce. Overall, our findings emphasize the need for careful consideration of the heterogeneous impacts of explainable AI systems on decision-making processes for different organizational groups to ensure that explainability can indeed live up to its promise of improving human-AI collaboration.

The paper proceeds as follows: Section 2 provides our paper's conceptual foundations and identifies the research gap we aim to address. Section 3 presents a formal framework that aims to guide the reader and explains our empirical study design. Section 4 showcases our results, while Section 5 discusses our findings and offer concluding remarks.

2 Conceptual foundations

In this section we lay the conceptual groundwork for our paper. We first present the two core concepts that we consider: eXplainable AI and the delegation of real authority. Subsequently, we delineate this paper's contribution to existing research.

2.1 Explainable artificial intelligence (XAI)

Explainability of intelligent systems has early on garnered considerable attention in the information systems research community. Initial investigations center on explanation facilities of expert systems (Ye and Johnson, 1995; Dhaliwal and Benbasat, 1996; Gregor and Benbasat, 1999), which generally do not employ statistical machine learning methods and lack self-learning features. Empirical studies examining the explainability of expert systems demonstrate that explanations improve performance within cooperative problem-solving settings (Gregor, 2001). Pertinent to our research questions, earlier studies on expert systems reveal that novice users predominantly use explanations for learning purposes, while experts mainly utilize them for validating conclusions (Ji-Ye Mao, 2000). Naturally, the knowledge gained from expert system explainability serves as an informative foundation for our investigation into the explainability of contemporary machine learning (ML) systems. Nonetheless, the substantially distinct nature of the knowledge embodied in prior-generation symbolic AI compared to contemporary sub-symbolic ML-based AI (see, e.g., Teodorescu et al., 2021; Berente et al., 2021) necessitates a dedicated examination of modern explainability approaches.

With the development of modern explainability methods for ML-based AI, research on the impact of XAI on user behavior has seen a considerable resurgence (Vilone and Longo, 2021). Nascent research in this domain typically focuses on how explanations affect users' attitudes towards, understanding of, and trust in the AI system (see, e.g., Lu and Yin, 2021). These studies produce mixed evidence on the consequences of XAI. Although several studies indicate that explanations can enhance trust and positive perceptions towards the system (see, e.g., Dodge et al., 2019; Rader et al., 2018; Yang et al., 2020; Meske

Electronic copy available at: https://ssrn.com/abstract=4512594

and Bunde, 2020), other studies provide reversed evidence (see, e.g., Erlei et al., 2020; Poursabzi-Sangdeh et al., 2021). Other research in this domain studies the effect of explanations on human learning and demonstrates how explanations shape users' understanding of the world (Abdel-Karim et al., 2022; Bauer et al., 2023). While explanations clearly show potential for "machine teaching" (Abdel-Karim et al., 2022) and may thus enable knowledge transfers from AI to human users, they also evoke significant cognitive biases that can lead to harmful side effects (Bauer et al., 2023).

Following recent research in the information systems field examining XAI (see, e.g., Doshi-Velez and Kim, 2017; **?**; Bauer et al., 2023), we conceptualize XAI as methods that present in human-understandable terms why an ML-based AI system has made a prediction. In recent years, researchers have developed various XAI methods (see, e.g., Ribeiro et al., 2016; Lundberg and Lee, 2017; Koh and Liang, 2017; Lakkaraju et al., 2019) to alleviate problems associated with the black-box nature (e.g., distrust, lack of accountability and learning, and error safeguarding) while maintaining a high level of prediction accuracy (Bauer et al., 2021). A core promise of XAI is to promote effective human-AI collaboration, allowing humans to better understand, validate, and even challenge AI-generated output (Miller, 2019). Apart from aiming to improve decision-making, XAI can facilitate regulatory compliance and risk management by making AI systems more transparent and accountable, ensuring that organizations adhere to ethical and legal guidelines (Mittelstadt et al., 2016).

Our study primarily concentrates on feature-based XAI, hereinafter referred to as XAI, which communicates the contribution of each input feature to the prediction. We choose this focus for several reasons. First, such explanations are extensively employed in practice (Bhatt et al., 2020; **?**; Gramegna and Giudici, 2021). Second, they are highly intuitive and easy to comprehend, as they fulfill most criteria for human-friendly explanations (Molnar, 2020). Third, these methods are generally applicable to systems utilizing both structured and unstructured data (see, e.g., Garreau and Luxburg, 2020). Finally, these approaches can elucidate individual predictions through local explainability, which may be the only method that complies with current or forthcoming regulations (Goodman and Flaxman, 2017).

A widely acknowledged state-of-the-art XAI method is SHAP, as recognized by numerous researchers (Gramegna and Giudici, 2021; Molnar, 2020). SHAP (Lundberg and Lee, 2017) offers explanations by employing additive feature attributions; that is, linear models illustrating the numerical influence of each feature value on the overall black box prediction. Given that SHAP learns these intelligible "surrogate models" exclusively from input-prediction pairs produced by the complex model requiring explanation, it can be applied to virtually all categories of machine learning models. SHAP draws inspiration from coalitional game theory, considering input features as a group of players collaborating to yield a payoff (the prediction). SHAP calculates the marginal contribution of each player to the total payoff, using Shapley values (Shapley, 1953) and a linear model that assigns weights to instances based on coalition properties. It is worth noting that SHAP shares a close relationship with Gregor and Benbasat's (1999)

Electronic copy available at: https://ssrn.com/abstract=4512594

seminal characterization of "why and why not explanations" within the realm of knowledge-based expert systems.

2.2 Delegation of decision authority

The delegation of authority crucially shapes the behavior of individuals within organizations, and thus eventually the organizations' overall performance. Among other things, previous research has shown that authority delegation impacts the quality of decisions (Jensen and Heckling, 1995), the timeliness of decisions (Patacconi, 2009), and employees' intrinsic motivation (Benabou and Tirole, 2003). The delegation of authority inherently trades off improving decision quality by utilizing others' better information with losing control over delegated decisions which may result in agency or some intrinsic costs for the delegating individual (Holmström, 1979). One approach to conceptualize this trade-off is the distinction between formal and real authority as proposed by Aghion and Tirole (1997). Formal authority is the *right to decide*, i. e., having the final say, whereas real authority represents the *effective control over decisions* due to better information, i. e., actually choosing the action. In organizations, superiors who always hold the formal right to decide may delegate the real authority to subordinates, depending on the balance between the subordinate's competence and the superior's need for control.

The delegation of formal and real authority has been widely studied in the context of economics as well as operations and management research. Prior work in this domain has delved into the intricate dynamics that influence decision-making, addressing factors such as information asymmetry, trust, and interest alignment (Aghion and Tirole, 1997; Eisenhardt, 1989). The delegation of real authority can enhance decision-making efficiency, effort provisions, and the information flow between individuals; but also create moral hazard and overreliance issues (Dessein, 2002; Dobrajska et al., 2015; Bartling et al., 2014). Factors determining the delegation of authority from one person to another include the delegators' level of expertise (Aghion and Tirole, 1997), beliefs about the delegatee's competence (Dessein, 2002), monitoring costs and trust (Dobrajska et al., 2015), and intrinsic motives to make decisions (Bartling et al., 2014).

As AI systems increasingly permeate various industries, the delegation of real authority has evolved from a human-human context (e. g., Dominguez-Martinez et al., 2014; Dobrajska et al., 2015) to a human-AI context (Leyer and Schneider, 2019; Fügener et al., 2022; Baird and Maruping, 2021; Candrian and Scherer, 2022). Consequently, recent research has begun exploring the delegation of authority to AI by investigating when and how humans use AI advice as decision support and comparing it to settings where other humans provide this support. This research suggests that an individual's willingness to utilize AI (versus human) advice as decision support depends on factors such as trust in the system, perception of the system, and task nature (Dietvorst et al., 2015; Logg et al., 2019; Castelo et al., 2019; Shrestha et al., 2019). The studies on individual utilization of AI as decision support contribute significantly to understanding how AI impacts human decision-making. However, they do not account for sit-

Electronic copy available at: https://ssrn.com/abstract=4512594

uations where human decision-makers, while retaining formal decision-making rights, delegate real authority to AI systems, relinquishing their effective control.

In distinguishing between the delegation of real decision authority and decision support, it is imperative to recognize the differential implications for the effort expended by the advice recipient. In the case of delegation, the advice recipient effectively reduces her efforts, relinquishing control to the advice provider, thereby rendering her role largely perfunctory in nature. In other words, she assumes a passive position, often merely rubber-stamping the recommendations forwarded by the advice giver. By contrast, when advice functions as decision support, the advice recipient retains and typically increases her level of engagement in the decision-making process. Rather than simply ratifying the advice, she conscientiously assimilates it into her decision-making deliberations. This approach invariably necessitates effort from the advice receiver.

The distinction between authority delegation and decision support is crucial because it highlights the difference between a transfer of control from humans to machines reflecting radical shifts in power structures and collaborative human-AI interactions (Baird and Maruping, 2021). As an illustration, consider a scenario wherein a real estate manager, responsible for appraising a variety of apartments, has access to an AI system that provides estimations based on observable apartment features. In one approach, the manager could diligently scrutinize each apartment, treating the AI prediction as one amongst many factors – a form of decision support. In this case, the manager retains control and invests effort in reconciling the AI's advice with other relevant considerations such as her own experience with given neighborhood amenities or market dynamics. Conversely, the manager might opt to forego processing all the available information, basing the valuation solely on the AI's estimated apartment price. By doing so, the manager effectively delegates real authority to the AI system. This scenario, where the manager minimizes personal effort and relinquishes authority to the AI, mirrors a form of automated decision-making.

While these approaches reflect two distinct ways of utilizing advice, it's important to recognize that the boundary between decision support and delegation is not always clear-cut. There are scenarios where a manager might rely more heavily on the AI's advice while still incorporating some level of personal evaluation. For example, Baird and Maruping (2021) view delegation as a broad concept that applies "[..] whenever an individual or collective leverages IS artifacts to perform tasks they would otherwise have to do themselves." To gain insights into how AI system explainability may affect humans' delegation of real authority to machines, we examine individuals' propensity to use an AI system's continuous scale prediction as their final decision together with the effort they exert in the task.

2.3 Contribution to the literature

We contribute to three streams of literature in the information systems field.

Human-XAI interaction. The first and most closely related line of work explores the interplay between explainability and users' interaction with AI

Electronic copy available at: https://ssrn.com/abstract=4512594

systems. A central tenet of XAI is that by providing individuals with an explanation about why an AI system produces its prediction, users are better able to identify the cases in which AI's reasoning was incorrect so they can overrule such a prediction. However, evidence on XAI's efficacy to enable the detection of prediction errors is mixed. Some empirical studies indicate that explainability decreases users' ability to identify inaccurate predictions (Bansal et al., 2021; Poursabzi-Sangdeh et al., 2021; Buçinca et al., 2021). For instance, Szymanski et al. (2021) show how explanations with a visual component may lead to misattributed trust. Bussone et al. (2015) provide complementary evidence in a medical diagnosing context. Chen et al. (2023) conduct a think-aloud, mixed-methods study and find that feature-based explanations may reduce overall decision performance. By contrast to these studies, Zhang et al. (2020) and Alufaisan et al. (2021) do not find XAI to have an impact on trust in AI and decision performance. Some studies, such as Yang et al. (2020) and Wang and Yin (2021), even find some evidence that XAI can actually reduce users' ability to detect incorrect predictions. This capacity, however, largely depends on the properties of the decision making task (Wang and Yin, 2021). Despite the growing importance of understanding how explainability affects human-AI collaboration, research has so far been unable to reconcile these differential findings. Other potential pitfalls of explainability seem to encompass reasoning errors such as backward reasoning and confirmation bias (Chromik et al., 2021; Szymanski et al., 2021), which overall may foster user biases and even impair decision-making (Poursabzi-Sangdeh et al., 2021; Ghassemi et al., 2021). In a recent paper, Bauer et al. (2023) provide evidence that explainability may induce biases that persist over time, impair decision-making in the long term, and even spill over to related yet disparate domains. These adverse effects may occur because users heuristically evaluate explanations – e.g., narrative heuristic or availability heuristic – instead of engaging in cognitively effortful analysis (Buçinca et al., 2021).

While much of the existing literature focuses on how XAI influences user trust in and perceptions of AI systems, as well as users' ability to detect incorrect predictions, the question of whether XAI promotes users' delegation of real authority – and thus a form of control relinquishment to machines – remains unanswered. Our paper takes an initial step in this direction. On the one hand, we examine cases where users effectively let the AI make the decision for them, while always maintaining the formal right of having the final say. To the best of our knowledge, we are the first to consider this case of delegating real authority. On the other hand, we also complement prior research by emphasizing heterogeneities concerning users' levels of domain knowledge and associated preconceptions about the relationship between features and AI predictions. Examining whether, and if so, for whom and when, explainability impacts individuals' propensity to grant AI systems effective decision control is crucial for anticipating potential shifts in task allocation and effort exertion within organizations. Moreover, our novel insights into potential sources of heterogeneity regarding the impact of XAI may help reconcile the currently mixed empirical evidence presented in related studies.

Electronic copy available at: https://ssrn.com/abstract=4512594

Algorithm aversion and appreciation. The second stream of literature that our work complements investigates the conditions under which humans are hesitant or willing to process AI-generated advice, exhibiting either algorithm aversion or appreciation, respectively. Algorithm aversion entails a tendency for individuals to undervalue machine-generated advice in comparison to human advice, including their own, even when they acknowledge that the machine's guidance is more accurate (Grove and Meehl, 1996; Grove and Lloyd, 2006; Önkal et al., 2009; Dietvorst et al., 2015). Conversely, recent studies (see, e.g., Logg et al., 2019; Gunaratne et al., 2018) have identified instances of algorithm appreciation, where people demonstrate a preference for algorithmic advice over human counsel. Several factors influence the occurrence of algorithm aversion or appreciation, including the perceived subjectivity of the task (Yeomans et al., 2019; Castelo et al., 2019), the capacity to modify predictions (Dietvorst et al., 2018), the observation of algorithmic errors (Dietvorst et al., 2015), and the disparity between actual and expected predictive performance (Jussupow et al., 2020). Studies by Longoni et al. (2019) and Starke et al. (2022) further indicate that the perceived fairness of algorithms can impact the level of algorithm aversion or appreciation, while Gaube et al. (2021) provides evidence that high-expertise users exhibit stronger algorithm aversion than low-expertise users.

Our paper complements this literature by providing additional evidence on the heterogeneous nature of algorithm aversion and appreciation. Specifically, we contribute to the limited number of studies indicating that the explainability of AI systems may alleviate algorithm aversion, and we demonstrate how this effect depends on users' domain knowledge and associated preconceptions about how input features relate to the evaluation of apartments. Gaining a better understanding of the complex nature of algorithm aversion and the factors that may exacerbate or mitigate it is important from a business perspective because it enables organizations to harness the full potential of AI systems while ensuring that employees remain engaged in the decision-making process. This understanding can help guide the development and implementation of AI and explainability technologies that foster trust and collaboration between humans and machines, ultimately enhancing organizational effectiveness and efficiency.

Automation bias in decision-making. Finally, at a higher level, our work relates to existing literature on decision-making automation (see, e.g., McLeod Jr and Jones, 1987; Brancheau and Wetherbe, 1987; Bucklin et al., 1998; Heimbach et al., 2015). Previous work suggests that decision support systems, and, more recently, AI, can enhance decision-making by providing data-driven insights into potential future states of the world, thereby augmenting decision-making efficacy (Arnott and Pervan, 2015; Agrawal et al., 2019). However, the ready availability of these systems may inadvertently promote an excessive reliance on decision support, adversely affecting decision-making performance (see, e.g., Mosier et al., 1998; Skitka et al., 1999). This tendency, also known as automation bias, could pose considerable challenges in contexts where intelligent systems are prevalent (Goddard et al., 2012b), as it implies a relinquishment of human agency and responsibility in the final decision-making process. With the growing integration of AI in organizations, this issue becomes particularly signif-

Electronic copy available at: https://ssrn.com/abstract=4512594

icant in high-stakes domains, where it could unintentionally exacerbate machine bias (Angwin et al., 2016; Green and Chen, 2019, see, e.g.,), paradoxically even under the guise of human control over transparent algorithmic decision-making. A seminal study by Parasuraman and Riley (1997) attributes automation bias to users' excessive confidence in the infallibility of systems. Subsequent studies support these claims and further associate automation bias with the reduction in cognitive load (Mosier et al., 1998; Lyell and Coiera, 2017). Additional research suggests that automation bias may also depend on factors such as technological literacy (Jacobs et al., 2021), user expertise (Gaube et al., 2021), and task subjectivity (Yeomans et al., 2019).

Our research contributes to this literature by investigating the conditions under which explanations might inadvertently promote excessive automation of decisions, i.e., facilitating automation bias. With the advent of regulations calling for AI explainability, such as the EU's AI Act and the Algorithmic Accountability Act in the US, it is imperative to discern whether explanations enhance users' ability to accurately assess the correctness of AI predictions, or conversely, if they instigate an overreliance on AI outputs, which subsequently diminishes decision-making performance. Comprehending such unanticipated side effects of explainability could guide policy formulation and encourage best practices, thus potentially preventing XAI misuse or unintentional harm in critical decision-making processes.

3 Formal framework and empirical study design

3.1 Formal framework

In this subsection, we develop a simple theoretical framework to formally illustrate the mechanisms we consider and derive propositions to investigate. The framework is inspired by Agrawal et al. (2019)'s model. Notably, we do not intend the framework to be a complex and all-encompassing theory. Instead, our objective is to guide the reader by giving a structured overview of the mechanisms we are interested in. We present our framework in the context of our experimental design. However, it is also applicable to other decision-making processes where individuals have access to AI predictions, e.g., stock price or sales forecasting.

Following our experimental design, individual i's task is to estimate the listing price p_j of apartment j. Each real estate agent $i \in N$ observes apartment characteristics X_j and exerts cognitive effort $e_i \in [0, 1]$ to come up with an informed evaluation about the apartment price. Exerting cognitive effort is increasingly costly for an agent. Without loss of generality, let the convex function $c(e_i) = \frac{e_i^2}{\alpha_i}$ describe the relation between efforts and costs where $\alpha_i > 0$ reflects agent i's level of domain knowledge regarding apartment valuation.

With probability e the agent is able to make sense of the apartment characteristics X_j and come up with an appropriate evaluation that yields a payoff π_H. Conversely, with probability $(1 - e)$, the agent is unable to make sense

Electronic copy available at: https://ssrn.com/abstract=4512594

of the information about an apartment so she effectively makes an uninformed guess. In this case, she earns a low payoff π_L with $\pi_H > \pi_L$. Hence, in the absence of an artificial intelligence (AI) system that provides a price prediction – our control condition – agent i's maximization problem can be described as:

$$\max_{e} \quad e \cdot \pi_H - (1 - e) \cdot \pi_L - \frac{e_i^2}{\alpha_i}. \tag{1}$$

Now consider an AI system that uses the observable apartment characteristics X_j to produce a prediction for the apartment price. The model's predictions are correct with probability $E \in [0, 1]$. In other words, E effectively represents prediction accuracy. We allow agents to form subjective beliefs about the prediction accuracy $\mu(E) \in [0, 1]$. For simplicity we assume that agents only make use of the AI system, if they fail to make sense of the observable information in which case they take over the prediction as their apartment valuation, i.e., they delegate real authority.

Relying on the prediction of the system, however, comes at a cost ω reflecting an intrinsic discomfort of losing agency or control (see, e.g., Bartling et al., 2014). Following this intuition, we can augment the maximization problem (1) to account for agents' access to an AI prediction as follows:

$$\max_{e} \quad e \cdot \pi_H - (1 - e) \cdot (\mu(E)\pi_H + (1 - \mu(E))\pi_L - \omega) - \frac{e_i^2}{\alpha_i} \tag{2}$$

Notably, in this setting, e can be understood as the inverse degree of rubber-stamping the prediction of the AI system: agents will more likely use the prediction as their evaluation when they do not succeed in making sense of the information. The main notion we examine in this study is whether explainability fosters the delegation of real authority to the AI system. Following arguments from the literature, providing explanations on top of predictions affect transparency, user understanding, fairness perceptions, and trust in the AI system (see, e.g., Gregor and Benbasat, 1999; Abdul et al., 2018; Peters et al., 2020). These effects can be understood as changing both, the costs $\omega(I_{Expl})$ agents experience when they rely on the AI system, and their subjective beliefs about the accuracy of the prediction $\mu(E|I_{Expl})$, where I_{Expl} is an indicator variable that is equal to 1 if explanations accompany predictions. Computing the first order condition of equation (2) and rearranging it to depict how the optimal effort level e^* depends on payoff structures, intrinsic costs, and the subjective prediction accuracy shows that

$$e^* = \frac{(1 - \mu(E|I_{Expl})) \cdot (\pi_H - \pi_L) + \omega(I_{Expl})}{2\alpha_i}. \tag{3}$$

Our paper aims to explore whether, and if so for whom and when, explainability increases the delegation of real authority to the AI system, i.e., increases agents' likelihood to rubber-stamp the AI prediction and reduce e^*. Following the results of previous work on human-XAI interaction, we conjecture

Electronic copy available at: https://ssrn.com/abstract=4512594

that $\frac{\partial \omega}{I_{Expl}} < 0$ because explainability may provide agents with a better understanding of why the system recommends a certain appraisal, thereby enhancing their sense of control and agency. As e^* positively depends on the costs ω, an explainability-driven decrease in costs should lead to more delegation.

Proposition 1 *Explainability reduces real estate agents' intrinsic costs of relying on an AI prediction and, thereby, can increase their ceasing of effective decision control to the AI system.*

Regarding the role of explainability for subjective beliefs $\mu(E|I_{Expl})$, the direction of the effect is more complex. While an increase in subjective accuracy beliefs should translate into more delegation, it is ex-ante unclear whether explainability does actually lead to higher accuracy beliefs. On the one hand, with explainability helping agents to identify when predictions are incorrect, subjective accuracy beliefs may converge to the true accuracy of the system. In this case, the impact of explainability on the delegation of real authority depends on whether agents' prior subjective beliefs were too high or too low. On the other hand, it may be possible that agents interpret explanations in an incorrect or biased way so that they inappropriately adjust their subjective accuracy beliefs up- or downwards.

Proposition 2 *Explainability affects real estate agents' subjective beliefs about the accuracy of AI predictions and, thereby, can increase or decrease their ceasing of effective decision control to the AI system.*

3.2 Design

Overview. In our empirical study that took place in the field with actual human experts, we conducted an incentivized experiment that allowed us to (i) exogenously manipulate individuals' access to (X)AI systems, (ii) measure their domain knowledge and preconceptions regarding input features and their relationship to predictions, and (iii) maintain strict control over potential confounding factors such as prior experience with the specific XAI system or organizational structures. This approach enables us to isolate the causal effects of explainability and focus on the treatment heterogeneities of primary interest. In total, 153 real estate experts participated in the empirical study that we conducted with the help of our industry partner, the *Real Estate Association Germany (IVD).* In the study, the experts evaluated the listing price per square meter (in Euros) of various apartments, which we had previously gathered from a large online platform.[1] To elicit genuine beliefs, we compensated participants based on their decision-making performance. Due to Covid-19 restrictions, we implemented the field study as an online experiment using oTree

[1] 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"). We constructed a dataset 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 supplementary material (Table 4).

Electronic copy available at: https://ssrn.com/abstract=4512594

(Chen et al., 2016), Python, and HTML.[2] Participants were presented with ten apartment attributes to facilitate informed evaluations, without receiving intermediate feedback. In order to mitigate task complexity and prevent information overload, we fixed seven apartment features across all stages, leaving only three characteristics that varied: location (Frankfurt or Cologne), presence of a balcony (yes or no), and the green voter share in the district (below city average, city average, or above city average). We provide screenshots of the interfaces in the supplementary material. Our experimental design comprises an introductory stage and a main stage, which we describe in detail below (see Figure 1 for an overview).[3]

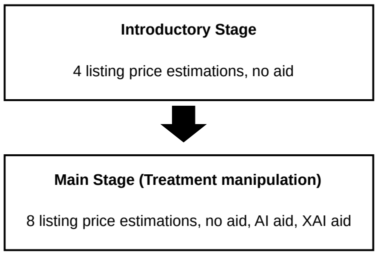

Introductory Stage
4 listing price estimations, no aid
Main Stage (Treatment manipulation)
8 listing price estimations, no aid, AI aid, XAI aid

Figure 1: Experimental stages

Introductory stage. During the introductory stage, we measured real estate agents' preconceptions regarding the impact of the three varying apartment attributes on the overall listing price. Specifically, agents evaluated four randomly selected apartments, each with different combinations of these attributes. They submitted evaluations by indicating the marginal contributions of each attribute to the price using a slider, which ranged from -2,500€ to +2,500€ in increments of 50€. We initially set the marginal contributions and overall price estimation to 0€ and the average listing price (9,600€), respectively. We calculated the final listing price evaluation for an apartment as 9,600€ plus the sum of the entered marginal contributions of the three variable features. Additionally, participants rated their confidence in both the marginal contributions and the resulting price estimation on a five-point scale.

Main stage. In the main stage of our experiment, participants evaluated eight randomly selected apartments. As in the introductory stage, they observed ten

[2] See the supplementary material for details on the experimental procedures including payments, instructions, and screenshots.

[3] Note that the two stages we focus on in this paper were followed by two additional stages that we analyze in another paper (blinded for peer review). The isolated analyses of the first two stages are methodologically sound because participants were never aware of any subsequent stages of the study. As a result, later stages could by design not have affected participants' behavior in any previous stages, e.g., through evoking strategic choices that aim to optimize behavior across all stages.

Electronic copy available at: https://ssrn.com/abstract=4512594

apartment attributes, with the same three characteristics varying across apartments. Unlike the previous stage, participants directly entered their estimated listing price on a continuous scale, requiring only that the final evaluation be greater than 0. As a reference point, we reminded participants of the average listing price for an apartment in our sample.

Crucially, we introduced our between-subject treatment variation during this stage of the experiment. There were three distinct treatment variations, differing only in terms of the availability and explanation facilities of an AI system. In the *NoAid* condition, which served as our between-subject control, real estate agents submitted their evaluations without any assistance from an AI system. Participants in the *AI* condition were provided with opaque listing price predictions from a stationary, non-learning AI system, which had been trained on 4,975 collected observations.[4] In the *XAI* condition, participants not only observed predictions from the same underlying AI system but also received numerically presented SHAP values for the three variable apartment characteristics, which represented the marginal contributions to the AI's prediction in Euros. In other words, the AI and XAI conditions differed only in the provision of SHAP explanations. After evaluating all eight apartments in this stage, participants in the AI and XAI conditions completed a survey addressing their trust, degree of reliance, and perceived transparency of the AI system. At the end of the experiment, participants in all treatment conditions filled out our socio-demographic survey containing items such as their age, domain knowledge, and overconfidence.

It is important to note that since the predictions are continuous in nature, it is arguably unlikely that real estate agents' evaluations would exactly match the AI prediction unless they actually observed it in the AI and XAI conditions.[5] The continuous nature of predictions is a crucial aspect of our study. Specifically, we contend that, relative to the between-subject control condition, any increase in real estate agents' propensity to select the prediction as their evaluation that is accompanied by a reduction in effort represents an implicit relinquishing of control over the decision to the underlying AI. By comparing these patterns for the AI and XAI conditions, we are able to isolate the causal effect of presenting SHAP explanations alongside predictions.

4 Results

This section presents our findings in two steps. First, we investigate on an aggregate level whether the provision of feature-based explanations affects the tendency of real estate agents to effectively delegate decision-making to an AI system. Second, we consider the roles of domain knowledge and the alignment between beliefs about how apartment characteristics influence listing prices and

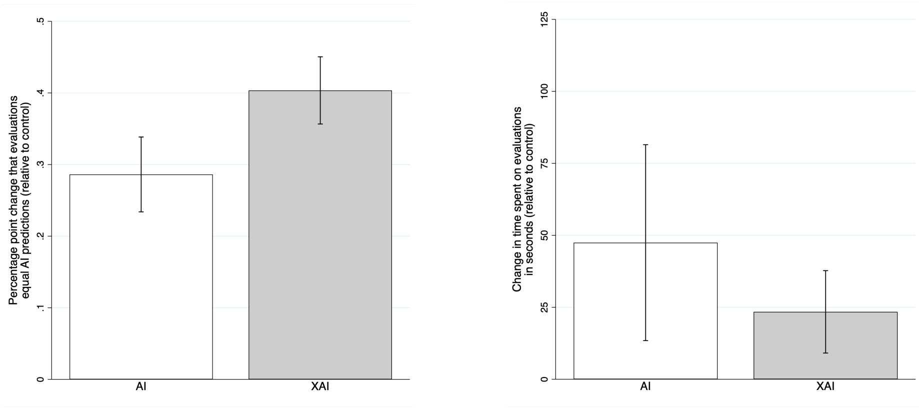
[4] The AI system is a random forest that achieves a performance of $R^2 = 0.72$ on unseen test data. See the supplementary material for additional information.

[5] In fact, the evaluation of participants in the control condition equaled the underlying prediction by chance merely on 2.8% of the cases.

Electronic copy available at: https://ssrn.com/abstract=4512594

the observed SHAP explanations. Our analyses focus on two variables: agents' likelihood to rubber-stamp the prediction of the AI system and the time they spend on apartment valuations. We always report results for the AI and XAI treatment conditions relative to the between-subject control condition, where participants had no access to any aid, allowing us to isolate the effects attributable to the provision of black box predictions and explanations.

4.1 Explainability and delegation of real authority



(a) Delegation of decision

(b) Effort provided

Figure 2: Aggregate level findings

Notes: Relative to the control condition, panel (a) depicts the change in the average probability with which evaluations equal AI predictions; panel(b) depicts the average change in the time agents spent to evaluate apartments. Whiskers represent 95% confidence intervals. We show results separately for the AI and XAI conditions.

Figure 2a illustrates the average change in the frequency of participants' apartment valuations matching the AI system predictions relative to our control condition. Different bars depict the results for our AI and XAI conditions.

The figure reveals that the provision of SHAP explanations alongside predictions increases the likelihood that real estate agents' apartment valuations equal the system's predictions. Compared to the *No Aid* control condition – where evaluations equal predictions by chance in 2.8% of the cases – the likelihood that AI participants' apartment valuations match the observed prediction rises by 28.3 percentage points. Real estate agents in the XAI condition who observed SHAP explanations in conjunction with predictions exhibit an even stronger inclination to adopt the observed prediction as their apartment valuation. Specifically, relative to the control condition, XAI participants' evaluations are 40 percentage points more likely to equal the observed prediction. In other words, the additional provision of explanations further elevated the likelihood of evaluations equalling predictions by 11.7 percentage points. The difference in the percentage point increases between the AI and XAI conditions is both

Electronic copy available at: https://ssrn.com/abstract=4512594

economically (+27.3%) and statistically highly significant ($p < 0.01, F$-test; see column (1) of Table 2).

But does this explainability-driven increase reflect a more pronounced delegation of real authority? To answer this question we inquire into agents' efforts when evaluating apartments. Utilizing the change in time real estate agents spend appraising apartments as a proxy for the change in invested efforts, we find evidence supporting this conjecture. Figure 2b presents the average change in time real estate agents took to evaluate apartments across our treatment conditions (relative to the control condition).

Our findings reveal that explainability not only increases the likelihood that apartment valuations equal predictions, but also decreases the time real estate agents devote to evaluating apartments. Compared to control participants who required on average 193 seconds, observing black-box predictions extends evaluation time by an average of 48 seconds (+24.9%). However, the average relative increase for real estate agents who observed explanations alongside predictions amounts to only 23 seconds (+11.9%). Put differently, real estate agents who observed explanations together with predictions took about 25 seconds less to evaluate apartments compared to their counterparts who observed black box predictions, which is in support of the notion that explainability entails an increase in the delegation of real authority. While the difference is economically significant (-10.4%), it is marginally statistically insignificant ($p = 0.14$, Wilcoxon rank-sum test).[6]

Considering the formal framework we derive in Section 3.1, the reader may naturally wonder whether the impact of explainability originates from a change in real estate agents' subjective accuracy beliefs or intrinsic costs of using the AI system. To provide insights into the mechanism underlying the observed effect, we compare the difference in AI and XAI participants' estimated accuracy of the AI system and their reported trust calibrations (Komiak and Benbasat, 2006) by means of regression analyses. Specifically, we examine the impact of explainability on the accuracy they believe the system to have, which we interpret as subjective accuracy beliefs, and the impact on levels of trust in the system's competence and integrity, as well as emotional trust – which we consider as inverse proxies for intrinsic costs of using the system.

Regression results reported in Table 1 indicate that explainability does not affect real estate agents' beliefs about the accuracy of predictions (column 1), however, increases parts of real estate agents' trust in the system (columns 2 and 3). Controlling for the accuracy belief and trust measures when regressing participants' likelihood to use predictions as their evaluations on treatment dummies, we find that the magnitude of the treatment effect of explainability decreases by about 20% (see Table 5 in the appendix). Interpreting these

[6] Figure 2b further indicates that showing black box predictions alone increased the time real estate agents took to evaluate apartments. This increase may depict that real estate agents tried to make sense of the prediction and spent cognitive efforts to decide whether they can rely on it. Against this background, the pure AI treatment effect on the propensity to use the prediction as their final evaluation may not represent an increase in the delegation of real authority. Instead, it may actually reflect the outcome of thoughtful consideration.

Electronic copy available at: https://ssrn.com/abstract=4512594

	(1)	(2)	(3)	(4)
Dep. variable:	Acc. belief	Comp. trust	Emo. trust	Integr. trust
Observing explanation	-2.56	0.6*	0.57**	0.4
	(3.784)	(0.328)	(0.270)	(0.273)
N	98	98	98	98
p	0.000	0.000	0.000	0.000
R^2	0.196	0.214	0.3	0.155

Table 1: Subjective accuracy beliefs and intrinsic costs.

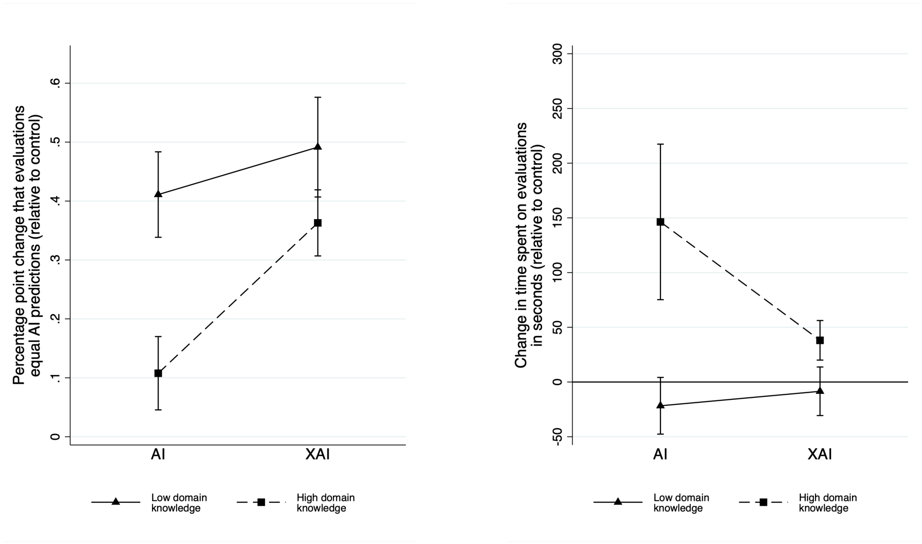
Notes: We depict results from OLS regression models with robust standard errors. In different columns, the dependent variable equals the accuracy beliefs and different trust measures following (Komiak and Benbasat, 2006). As we measured accuracy beliefs and trust only in treatments where participants actually interacted with an AI system, we can only include observations from the AI and XAI treatments. As independent variables, we include a treatment dummy for XAI and controls on agents characteristics. We denote significance levels by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

findings from the perspective of our framework suggests that explainability affects the delegation of real authority to the AI system not by increasing subjective accuracy beliefs, but reducing the intrinsic costs associated with using the system.[7]

Result 1: *Explainability fosters users' inclination to delegate the evaluation to the AI system. This effect seems not to stem from changes in subjective beliefs about the system's performance but only from the intrinsic costs of using the system.*

What remains open thus far is which individuals are enticed to delegate real authority to the AI upon receiving explanations, and under what circumstances. Comprehending the factors that moderate the explanation-driven increase in delegation can aid in devising strategies to counteract an excessive delegation of real authority to the AI. Such strategies can help mitigate concerns that explainability effectively increases automation of decisions under the disguise of transparency. We examine potential sources of heterogeneity in the next subsection.

Electronic copy available at: https://ssrn.com/abstract=4512594



(a) Delegation of decision

(b) Effort provided

Figure 3: Treatment heterogeneities

Notes: Relative to the control condition, panel (a) depicts the change in the average probability with which evaluations equal AI predictions; panel(b) depicts the average change in the time agents spent to evaluate apartments. Whiskers represent 95% confidence intervals. Lines with different symbols show results separately for real estate agents with high and low domain knowledge. We show results separately for the AI and XAI conditions.

4.2 The role of domain knowledge

Figure 3 portrays treatment heterogeneities for our variables of main interest based on real estate agents' self-reported domain knowledge in estimating apartment listing prices. We categorize agents as possessing relatively low (high) domain knowledge if their self-reported expertise is less than or equal to (greater than) the study median of 5. Similar to our aggregate-level analyses, Figures 3a and 3b present the average change in the likelihood that real estate agents' apartment valuations match the predictions, and the average change in the time required to evaluate apartments, respectively.

Figure 3a suggests that the explainability-driven increase in the likelihood of evaluations equalling predictions originates from individuals with high do-

[7] Notably, when we assess real estate agents' performance using the mean absolute difference between their evaluations and the true listing price we find that the explanation-driven increase in delegating decision authority to AI does not lead to improved evaluations. In particular, relative to those in the AI condition, participants exposed to explanations exhibit a 10.5% increase in the mean absolute error. This difference is economically, however, not statistically significant ($p = 0.25$, Wilcoxon rank-sum test). Notably, our AI system outperforms real estate agents in the control condition by approximately 8% (based on the mean absolute error), i.e., could potentially improve the agents' valuation performance. Yet, the heightened reliance on predictions accompanied by explanations appears to impair agents' performance.

Electronic copy available at: https://ssrn.com/abstract=4512594

main knowledge. Compared to the control condition, the likelihood that low and high domain knowledge agents' evaluations in the AI condition match predictions respectively rises by 41.1 and 10.7 percentage points (from 1.7% and 4.2% in the control). This observation is in line with our formal framework which suggests that domain knowledge as such reduces incentives to delegate. Examining treatment heterogeneities of explainability, we observe that providing SHAP explanations in addition to predictions further elevates the likelihood of low (high) domain knowledge agents' evaluations aligning with predictions by 8.1 percentage points (25.5 percentage points). In contrast to individuals in the AI condition, the likelihood that low and high domain knowledge agents in the XAI condition rubber-stamp predictions increases by +15.7% and +184.5%, respectively.

We evaluate the statistical significance of these patterns through regression analyses and present our findings in column (2) of Table 2. In these analyses, the dependent variable is a dummy variable indicating whether an agent's evaluation matches the AI system's prediction for a specific apartment. To account for potential confounding factors, we include additional controls such as agents' age, risk aversion, observed prediction, and overconfidence level, and cluster robust standard errors at the individual level.[8] The reported estimates confirm that explainability only influences the propensity to adopt the prediction as their evaluation for agents with high domain knowledge. Estimates in column (2) depict an economically and statistically significant treatment heterogeneity concerning domain knowledge (see β_5) and a highly significant overall treatment effect for high domain knowledge agents ($\beta_2 + \beta_5$). Due to the heterogeneous explanation treatment effect, the observed gap between low and high domain knowledge agents regarding the rubber-stamping of predictions narrows substantially from 27.8 percentage points in the AI condition to 8.2 percentage points in the XAI condition. Consequently, the evaluations of low and high domain knowledge agents become more similar under explainability.

Investigating changes in the time required by real estate agents with varying levels of domain knowledge to evaluate apartments (see Figure 3b) depicts that neither predictions nor explanations significantly affect low domain knowledge agents' effort provision. In stark contrast, real estate agents with high domain knowledge substantially increase the time spent on apartment evaluations when exposed to a black box prediction (+146 seconds). Crucially for our research questions, we find that showing SHAP values reduces the time they spend by 108 seconds. It appears that high domain knowledge agents exert extra effort to scrutinize black box predictions; however, when explanations accompany predictions, they revert to effort levels similar to those in the control condition, which amount to 193 seconds. The observation that explainability raises rubber-stamping of predictions, while simultaneously reducing the time invested to evaluate apartments underscores that the explanation-driven increase in the delegation of real authority underlies considerable heterogeneity.

Additional regression analyses (see Tables 6 and 7 in the supplementary ma-

[8] Note: our results are robust to the additional inclusion of apartment fixed effects.

Electronic copy available at: https://ssrn.com/abstract=4512594

Dep. variable: Evaluation equals AI prediction	(1)	(2)
Observing prediction (β_1)	0.252*** (0.054)	0.384*** (0.072)
Additionally observing explanation (β_2)	0.178*** (0.064)	0.092 (0.098)
High expertise		0.101* (0.060)
High expertise*Observing prediction (β_4)		-0.347*** (0.106)
High expertise*Observing explanation (β_5)		0.234* (0.131)
F-test: $\beta_1 + \beta_4$		0.627
F-test: $\beta_2 + \beta_5$		0.000
N	1,182	1,182
p	0.000	0.000
R^2	0.283	0.304

Table 2: Delegation and domain knowledge.

Notes: We depict results from random effects GLS regression models with robust standard errors clustered on the individual level and reported in parentheses. In all columns, the dependent variable is equal to one when real estate agents' evaluation coincides with the prediction and zero otherwise. As independent variables, we include treatment dummies, a dummy indicating agents' level of expertise, and their interaction effects. The control condition where agents did not observe any prediction or explanation serves as the reference category. Additionally, we include controls on agents' gender, age, level of risk aversion, academic degree, confidence in their evaluation, familiarity with AI technology, and degree of overconfidence. We denote significance levels by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

terial) suggest that the domain knowledge-driven heterogeneity of explanations can again, at least in part, be traced back to changes in the intrinsic costs of relying on the system, but not to changes in the subjective accuracy beliefs. Specifically, explainability only significantly increases the trust in the system's competence and emotional trust for high domain knowledge experts for whom we observe an increased inclination to rubber-stamp explained predictions. We do not observe any effects for agents with low domain knowledge. This finding may help clarify why numerous recent studies have uncovered limited evidence on the theorized effects of explanations on trust calibration (see, e.g., Jacobs et al., 2021). Following our results, critical individual differences such as the level of domain expertise may play a determining role in whether explainability can foster trust.

Result 2: *The effects of explainability on real authority delegation only occur for high domain knowledge individuals.*

In the final step of our analysis, we aim to better understand this heterogeneity by examining the role of the alignment between SHAP explanations and real estate agents' preconceptions about how apartment characteristics contribute to listing prices. Specifically, we ask whether the increase in delegation for high domain knowledge agents is a general phenomenon associated with the provision

Electronic copy available at: https://ssrn.com/abstract=4512594

of explanations as such regardless of the explanation-preconception alignment – e.g., due to an epistemic curiosity: a preference for knowledge that motivates them to eliminate information gaps (Litman, 2008). Alternatively, it may also be the case that explanations often corroborate high domain knowledge agents' preconceptions, allowing them to validate what they already assume to be true. Relatedly, for agents with low domain knowledge, the absence of an explanation treatment effect may result from their lack of pronounced preconceptions about how apartment features relate to listing prices. Having no clear preconceptions may render seeing explanations to validate the AI system's reasoning mute so that we do not find a treatment effect.

We investigate the role of explanation-preconception alignments using regression analyses. We repeat the analyses reported in Table 2 separately for evaluations where real estate agents' preconceptions, measured in the introductory stage, align or conflict with observed SHAP values (see Table 3).

We define agents' preconceptions and SHAP explanations for a given apartment as aligned when the average absolute difference between SHAP values and measured preconceptions is below the median of the distribution (which equals 1275 Euro). According to this definition, preconceptions of agents with low (high) domain knowledge contradict explanations in 48.2% (53.3%) of the cases.[9]

Table 3 presents the results of our regression analyses. The two main independent variables of interest are β_2 and β_5, representing the pure effect of providing explanations and the corresponding treatment difference for agents with high domain knowledge, respectively. Columns (1) and (2), and (3) and (4) respectively display results for the subsamples of apartment evaluations where agents' preconceptions about how apartment characteristics contribute to listing prices and SHAP values were highly and lowly aligned.[10]

The results suggest that individuals' response to explainability is sensitive to the explanation-preconception alignment only for real estate agents with low domain knowledge. When explanations align with agents' preconceptions about how apartment characteristics affect listing prices, the additional provision of explanations increases the likelihood that evaluations equal predictions by about 20 percentage points ($p < 0.05$, see β_2 column (2)). Notably, we find no significant treatment heterogeneities ($p = 0.42$) concerning agents' domain knowledge (see β_5). Therefore, when explanations validate individuals' preconceptions, explainability increases real estate agents' likelihood to rubber-stamp predictions,

[9] Notably, when there is alignment, the AI system's prediction and an agent's apartment valuation would naturally be more similar. However, our regression analyses include a dummy variable indicating whether agents observed explanations in addition to the prediction. Hence, we are able to isolate the effect of observing explanations that are aligned with preconceptions from the effect attributable to the alignment of evaluations and predictions. That is because the dummy variable indicating that agents observed the prediction as such captures the latter effect. In our control condition, the absolute difference in evaluations and predictions for apartments is 25.3% lower when the elicited reasoning of how apartment characteristics contribute to listing prices and SHAP explanations have a high compared to a low alignment (1021 Euro vs. 1367 Euro, respectively)

[10] Note: our results are robust to the additional inclusion of apartment fixed effects.

Electronic copy available at: https://ssrn.com/abstract=4512594

	Explanation-preconception alignment			
Dep. variable:	High		Low	
Evaluation equals AI prediction	(1)	(2)	(3)	(4)
Observing prediction (β_1)	0.294***	0.403***	0.224***	0.357***
	(0.059)	(0.077)	(0.060)	(0.082)
Observing explanation (β_2)	0.206***	0.191**	0.210***	0.088
	(0.067)	(0.093)	(0.073)	(0.109)
High expertise		0.056		0.106*
		(0.068)		(0.064)
High expertise*Observing prediction (β_4)		-0.294**		-0.345***
		(0.118)		(0.111)
High expertise*Observing explanation (β_5)		0.124		0.279**
		(0.136)		(0.142)
F-test: $\beta_1 + \beta_4$		0.214		0.876
F-test: $\beta_2 + \beta_5$		0.001		0.000
N	591	591	591	591
p	0.000	0.000	0.000	0.000
R^2	0.331	0.351	0.258	0.281

Table 3: The role of the alignment of human and AI reasoning.

Notes: We depict results from random effects GLS regression models with robust standard errors clustered on the individual level and reported in parentheses. In all columns, the dependent variable is equal to one when real estate agents' evaluation coincides with the prediction and zero otherwise. As independent variables, we include treatment dummies, a dummy indicating agents' level of expertise, and their interaction effects. The control condition where agents did not observe any prediction or explanation serves as the reference category. Additionally, we include controls on agents' gender, age, level of risk aversion, academic degree, confidence in their evaluation, familiarity with AI technology, and degree of overconfidence. We denote significance levels by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

regardless of their domain knowledge level.

In contrast, when the explanation-preconception alignment is low, explainability only increases the likelihood that appraisals equal predictions for agents with high domain knowledge (see β_5 in column (4)). Specifically, high domain knowledge agents' evaluations are 27.9 percentage points more likely to equal predictions ($p < 0.05$). For agents with low domain knowledge, however, providing explanations alongside predictions has no significant effect on their rubber-stamping of predictions ($p = 0.36$, see β_2).

We perform complementary analyses to examine how the explanation-preconception alignment relates to the time spent by low and high domain knowledge agents to evaluate apartments.[11] Our analyses uncover a strong positive correlation (Spearman's $\rho = 0.46, p < 0.07$) between the time that low domain knowledge agents in the XAI treatment spent evaluating apartments and the proportion of encountered apartments with misaligned preconceptions and explanations. This correlation suggests that the increased propensity of

[11] Note that we do not observe the time real estate agents spent on the evaluation of individual apartments but only on the aggregate, mainly due to privacy concerns on the part of our industry partner. As a result, we are unable to use regression analyses on the individual×apartment level when it comes to the time required to appraise apartments.

Electronic copy available at: https://ssrn.com/abstract=4512594

low domain knowledge agents to rubber-stamp predictions when preconceptions align with explanations reflects an increased delegation of real authority. Given that they require more time to evaluate apartments when preconceptions and explanations are misaligned, it seems that explanations promote scrutiny of these instances and ultimately discourage rubber-stamping predictions. For high domain knowledge agents in the XAI treatment, we do not observe a significant correlation (Spearman's $\rho = 0.09, p = 0.61$), implying that their overall reduction in evaluation time does not depend on the explanation-preconception alignment. A possible interpretation of this outcome is that explainability addresses a general need of high domain knowledge agents to be able to retrace the reasoning behind delegated decisions which, however, does not depend on the actual nature of the explanation.

Result 3: *Explainability can increase low domain knowledge individuals' delegation of real authority, however, only when explanations match their preconceptions. For high domain knowledge individuals, explainability generally fosters delegation of real authority.*

5 Discussion and conclusion

5.1 Discussion of results

This research delves into the intricate dynamics between explainability and the delegation of real authority to AI systems, with a focus on the realm of real estate valuations. Our findings indicate that agents with high domain knowledge are more inclined to delegate real authority to AI when given explanations, whereas those with low domain knowledge demonstrate increased delegation only when their preconceptions correspond with the provided explanations. Interpreting our results from the perspective of a derived framework, explainability appears to cultivate authority delegation due to a decrease in high domain knowledge agents' intrinsically experienced costs of using the system. These insights underscore the significance of acknowledging individual variations, such as domain expertise, in the design and implementation of explainable AI systems, to promote a fruitful partnership between humans and AI while averting undue dependence on AI-generated outcomes under the disguise of transparency.

Our aggregate level finding that users are more inclined to rubber-stamp AI predictions when systems are explainable implies a convergence of organizational decision-making, reducing the variability and inconsistency in judgments and decisions made by individuals, i. e., reduce noise (Kahneman et al., 2021). Following arguments by Kahneman et al. (2021), this noise reduction in decisions may streamline operations, enhance coordination among team members, and facilitate the establishment of standardized procedures, which could eventually translate into higher organizational performance. However, more standardized decision-making may also limit the diversity of perspectives and ideas, potentially stifling innovation and adaptability within organizations (Nemeth

Electronic copy available at: https://ssrn.com/abstract=4512594

and Kwan, 1987).

From a more socio-technical perspective, our results show how explainable AI systems can inadvertently lead to automation bias. Automation bias refers to the human tendency to completely and inadequately rely on recommendations made by automated systems (Parasuraman and Manzey, 2010). An increase in automation bias can entail harmful consequences (Prunkl, 2022; Banks, 2018) and potentially outweigh the gains of explainable AI. In light of potential biases and erroneous behaviors of AI systems, explainability may inadvertently cause humans in the loop to "fall asleep at the wheel" so that unfair or incorrect AI recommendations become implemented more often – even under the disguise of a human in the loop who can check the reasoning of the system.

A different lens to interpret our results is the theoretical framework on delegation to and from agentic IS artifacts (Baird and Maruping, 2021). In this framework, delegation between human users and agentic artifacts is fluid and can shift back-and-forth during interactions. Here domain knowledge may represent an endowment of the human user, whereas explainability may constitute an endowment of the agentic artifact (i. e., the AI system) that promotes delegation through improved coordination. Specifically, the agentic artifact updates the human user on the task completion by disclosing its prediction and a corresponding explanation. Thereby, the artifact both grants the ability and delegates the responsibility to intervene. The human user, however, may directly follow the prediction, effectively delegating the ultimate decision back to the agentic artifact. Our findings underline the importance of delegation as a theoretical framework for understanding new technologies such as explainable AI and enabling the "next generation of research on IS use" (Baird and Maruping, 2021).

5.2 Implications

Our results entail practical implications for managers and policymakers. For management, our results suggest at least three factors to consider when assessing how the implementation of explainability will affect the quality of decisions in the organization: workers' domain knowledge, workers' beliefs about feature-label relationships, and the degree of complementarity between the AI and workers' evaluation. For example, the AI system may outperform human workers with little complementarity and, as a consequence, the delegation of real authority to the AI system is beneficial. Here XAI would improve decisions by steering workers with high domain knowledge toward higher delegation. By contrast, XAI would have little to no effect on the delegation of workers that exhibit both low domain knowledge and little alignment in their beliefs about feature-label relationships with explanations. Hence, our results enable management to understand how XAI affects delegation within their particular workforce and, if applicable, can inform accompanying measures. Such measures could, e. g., focus on aligning the beliefs of workers with low domain knowledge and AI explanations by means of "machine teaching" (Abdel-Karim et al., 2022). Needless to say, in cases of high complementarity between AI and workers' eval-

Electronic copy available at: https://ssrn.com/abstract=4512594

uation, an increased delegation of real authority may imply overreliance and poorer decisions. Here our results suggest that introducing XAI should be met with caution. While management may still choose to implement XAI (e. g., due to upcoming regulation), they should accompany the implementation with measures promoting hybrid decision-making (Hemmer et al., 2021).

Management also needs to consider potential side effects of an explainability-driven increase in automation bias. Prior research has shown that automation bias may lead to decisions that are not in the best interest of the company, its employees, or its customers (Cummings, 2006). This can result in missed opportunities, reduced competitive advantage, and potential financial losses. Moreover, automation bias may entail negative long-term effects on the company's employees and undermine both their skills and engagement. Specifically, over time, increased automation bias means that employees have less opportunity to develop or maintain human problem-solving and decision-making skills, leading to their decline (Sheridan, 2002). As employees become more reliant on automated systems, they may also feel less responsible for their work and less engaged with their tasks, which can negatively impact job satisfaction, productivity, and employee retention (Kaplan and Haenlein, 2019). Considering the risks associated with an increase in automation bias, management should carefully monitor the side effects of XAI. Possible measures to mitigate automation bias in organizations include highlighting individual accountability of employees in the workplace (cf. Goddard et al., 2012a).

Our findings hold significant implications for policymakers currently engaged in drafting regulations for equitable and explainable AI systems. A fundamental proposition of explainable AI is the improvement of human oversight, thereby serving as a safeguard against erroneous or biased outputs (Bauer et al., 2021). However, our results illustrate that the introduction of explanations may incentivize human users to delegate more authority to AI and invest less time in checking the validity of AI outputs. Should greater explainability of AI systems indeed lead to a decrease in user engagement with AI outputs, it casts doubt on the prospect of enhanced human oversight through explainable AI. Furthermore, considering the delegation of authority to AI through the lens of automation bias, our results indicate that explainable AI (XAI) might indeed reduce accountability, responsibility, and human autonomy. Prior research has demonstrated that automation bias can result in a diffusion of responsibility, where individuals and organizations are less inclined to assume responsibility for the repercussions of their decisions (Banks, 2018). An amplified automation bias could also contribute to a decline in human autonomy and agency within the decision-making process, potentially undermining the dignity and self-determination of individuals (Prunkl, 2022). Consequently, policymakers should endeavor to extend regulations beyond merely mandating the provision of explanations, ensuring that effective human oversight is maintained, and automation bias is mitigated. For instance, more nuanced regulations could necessitate that personnel interacting with XAI systems undergo preparatory training measures to foster critical engagement with AI explanations.

Electronic copy available at: https://ssrn.com/abstract=4512594

5.3 Future research direction

Our study entails several limitations that present interesting opportunities for future research. One limitation of our study is the domain specificity that naturally arises when studying our research questions in a real-world setting. As outlined in Section 1, our setting of real estate appraisals is well-suited due to the growing role of AI in the industry, the financial consequences of evaluations, the complex relationship between numerous features and the final price, and the varying levels of domain knowledge of real estate professionals. However, there exist many different settings in which XAI systems may shape the delegation of authority and where results may vary. Other settings which we deem highly relevant due to the arguably strong impact of AI on human lives include automated hiring (van den Broek et al., 2021), medical diagnosing (Jussupow et al., 2021), and credit scoring (Khandani et al., 2010). Future research could explore these settings and test the generalizability of our findings.

A possible limitation is the choice of SHAP values to provide explanations. As described in Section 1, SHAP values represent a feature-based XAI method that is both among the most widespread in practice (Bhatt et al., 2020) and presumably necessary to comply with upcoming regulation (Goodman and Flaxman, 2017). While SHAP values seem a natural choice for our study, we acknowledge that there exist other relevant forms of explanations for AI systems, such as example-based explanations (Mittelstadt et al., 2019) or counterfactual explanations (Fernández-Loría et al., 2022). While it is not within the scope of this paper to investigate and compare the relationship between various forms of explanations and the delegation of authority, future research should examine whether and why the effects we observed would differ if users were provided with these forms.

Another limitation is the absence of feedback on the decision outcomes in our experiment. We refrain from giving feedback as it both reduces complexity to enable the clear isolation of explanation-driven effects and represents a more realistic reflection of reality (in which many AI-supported decisions do not produce immediate feedback). Examples of real-world use cases without feedback include hiring decisions that are supported by an on-the-job performance predicting AI system, investment decisions backed by a return predicting AI system, and drug treatment decisions aided by an effectiveness predicting AI system (Bauer et al., 2023). Notwithstanding these reasons, feedback – and the ability to learn from it – may have an interesting and substantial effect on the delegation of authority to AI systems, which we leave to future research to explore further.

5.4 Concluding remark

The study at hand sheds light on the complex and nuanced effects of explainability on users' willingness to delegate real authority to AI systems, and thus effectively cease control to machines. Our findings highlight that the introduction of explainability in organizations may have unforeseen, complex consequences

Electronic copy available at: https://ssrn.com/abstract=4512594

that affect different employees differently. While it is by no means our intention to advocate for keeping AI logic concealed, our research emphasizes the importance of understanding the multifaceted implications of making AI explainable by underscoring how individual differences in users' characteristics may cause them to respond differently to XAI. As AI continues to permeate various industries and explainability is increasingly required, understanding and addressing these individual-level differences will be crucial in harnessing the full potential of AI while mitigating potential risks associated with explainability.

Electronic copy available at: https://ssrn.com/abstract=4512594

References

B. M. Abdel-Karim, N. Pfeuffer, V. Carl, and O. Hinz. How AI-based systems can induce reflections: The case of ai-augmented diagnostic work. *MIS Quarterly*, page forthcoming, 2022.

A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–18, 2018.

P. Aghion and J. Tirole. Formal and real authority in organizations. *Journal of Political Economy*, 105(1):1–29, 1997.

A. Agrawal, J. S. Gans, and A. Goldfarb. Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy*, 47:1–6, 2019.

Y. Alufaisan, L. R. Marusich, J. Z. Bakdash, Y. Zhou, and M. Kantarcioglu. Does explainable artificial intelligence improve human decision-making? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6618–6626, 2021.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, 23:77–91, 2016.

D. Arnott and G. Pervan. A critical analysis of decision support systems research. *Formulating Research Methods for Information Systems: Volume 2*, pages 127–168, 2015.

A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

S. C. Athey, K. A. Bryan, and J. S. Gans. The allocation of decision authority to human and artificial intelligence. In *AEA Papers and Proceedings*, volume 110, pages 80–84. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2020.

A. Baird and L. M. Maruping. The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly*, 45(1), 2021.

J. Banks. The human touch: practical and ethical implications of putting ai and robotics to work for patients. *IEEE pulse*, 9(3):15–18, 2018.

Electronic copy available at: https://ssrn.com/abstract=4512594

G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *CHI Conference on Human Factors in Computing Systems*, 2021.

B. Bartling, E. Fehr, and H. Herz. The intrinsic value of decision rights. *Econometrica*, 82(6):2005–2039, 2014.

K. Bauer, O. Hinz, W. van der Aalst, and C. Weinhardt. Expl(AI)n it to me–explainable AI and information systems research. *Business & Information Systems Engineering*, 63(2):79–82, 2021.

K. Bauer, M. von Zahn, and O. Hinz. Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research*, forthcoming, 2023.

R. Benabou and J. Tirole. Intrinsic and extrinsic motivation. *The review of economic studies*, 70(3):489–520, 2003.

N. Berente, B. Gu, J. Recker, and R. Santhanam. Managing artificial intelligence. *MIS Quarterly*, 45(3):1433–1450, 2021.

U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley. Explainable machine learning in deployment. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

J. C. Brancheau and J. C. Wetherbe. Key issues in information systems management. *MIS quarterly*, pages 23–45, 1987.

E. Brynjolfsson and A. McAfee. The business of artificial intelligence. *Harvard Business Review*, 6, 2017.

Z. Buçinca, M. B. Malaya, and K. Z. Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

R. Bucklin, D. Lehmann, and J. Little. From decision support to decision automation: A 2020 vision. *Marketing Letters*, 9:235–246, 1998.

A. Bussone, S. Stumpf, and D. O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *International Conference on Healthcare Informatics*, 2015.

M. Busuioc. Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review*, 81(5):825–836, 2021.

C. Candrian and A. Scherer. Rise of the machines: Delegating decisions to autonomous ai. *Computers in Human Behavior*, 134:107308, 2022.

Electronic copy available at: https://ssrn.com/abstract=4512594

N. Castelo, M. W. Bos, and D. R. Lehmann. Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825, 2019.

D. L. Chen, M. Schonger, and C. Wickens. otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, 2016.

V. Chen, Q. V. Liao, J. W. Vaughan, and G. Bansal. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *arXiv preprint arXiv:2301.07255*, 2023.

M. Chromik, M. Eiband, F. Buchner, A. Krüger, and A. Butz. I think i get your point, ai! the illusion of explanatory depth in explainable ai. In *26th International Conference on Intelligent User Interfaces*, pages 307–317, 2021.

E. Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Technical Report COM/2021/206 final, European Commission, 2021.

M. L. Cummings. Automation and accountability in decision support system interface design. 2006.

W. Dessein. Authority and communication in organizations. *The Review of Economic Studies*, 69(4):811–838, 2002.

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

B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015.

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

M. Dobrajska, S. Billinger, and S. Karim. Delegation within hierarchies: How information processing and knowledge characteristics influence the allocation of formal and real decision authority. *Organization Science*, 26(3):687–704, 2015.

J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan. Explaining models: an empirical study of how explanations impact fairness judgment. In *International conference on Intelligent User Interfaces*, 2019.

S. Dominguez-Martinez, R. Sloof, and F. A. von Siemens. Monitored by your friends, not your foes: Strategic ignorance and the delegation of real authority. *Games and Economic Behavior*, 85:289–305, 2014.

Electronic copy available at: https://ssrn.com/abstract=4512594

F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. In *arXiv:1702.08608*, 2017.

K. M. Eisenhardt. Agency theory: An assessment and review. *Academy of Management Review*, 14(1):57–74, 1989.

A. Erlei, F. Nekdem, L. Meub, A. Anand, and U. Gadiraju. Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. In *AAAI Conference on Human Computation and Crowdsourcing*, 2020.

C. Fernández-Loría, F. Provost, and X. Han. Explaining data-driven decisions made by ai systems: The counterfactual approach. *MIS Quarterly*, 46:1635–1660, 2022.

A. Fügener, J. Grahl, A. Gupta, and W. Ketter. Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research*, 33(2):678–696, 2022.

D. Garreau and U. Luxburg. Explaining the explainer: A first theoretical analysis of lime. In *International Conference on Artificial Intelligence and Statistics*, 2020.

S. Gaube, H. Suresh, M. Raue, A. Merritt, S. J. Berkowitz, E. Lermer, J. F. Coughlin, J. V. Guttag, E. Colak, and M. Ghassemi. Do as ai say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*, 4(1):31, 2021.

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

K. Goddard, A. Roudsari, and J. C. Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012a.

K. Goddard, A. Roudsari, and J. C. Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012b.

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

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

B. Green and Y. Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3 (CSCW):1–24, 2019.

Electronic copy available at: https://ssrn.com/abstract=4512594

S. Gregor. Explanations from knowledge-based systems and cooperative problem solving: an empirical study. *International Journal of Human-Computer Studies*, 54(1):81–105, 2001.

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

W. M. Grove and M. Lloyd. Meehl's contribution to clinical versus statistical prediction. *Journal of Abnormal Psychology*, 115(2):192, 2006.

W. M. Grove and P. E. Meehl. Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, 2(2): 293, 1996.

J. Gunaratne, L. Zalmanson, and O. Nov. The persuasive power of algorithmic and crowdsourced advice. *Journal of Management Information Systems*, 35 (4):1092–1120, 2018.

I. Heimbach, D. S. Kostyra, and O. Hinz. Marketing automation. *Business & Information Systems Engineering*, 57:129–133, 2015.

P. Hemmer, M. Schemmer, M. Vössing, and N. Kühl. Human-AI complementarity in hybrid intelligence systems: A structured literature review. In *Pacific Asia Conference on Information Systems (PACIS)*, 2021.

B. Holmström. Moral hazard and observability. *The Bell journal of economics*, pages 74–91, 1979.

M. Jacobs, M. F. Pradier, T. H. McCoy Jr, R. H. Perlis, F. Doshi-Velez, and K. Z. Gajos. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry*, 11(1):108, 2021.

M. C. Jensen and W. H. Heckling. Specific and general knowledge, and organizational structure. *Journal of Applied Corporate Finance*, 8(2):4–18, 1995.

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

E. Jussupow, I. Benbasat, and A. Heinzl. Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. In *European Conference on Information Systems (ECIS)*, 2020.

E. Jussupow, K. Spohrer, A. Heinzl, and J. Gawlitza. Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research*, 32(3):713–735, 2021.

Electronic copy available at: https://ssrn.com/abstract=4512594

D. Kahneman, O. Sibony, and C. R. Sunstein. *Noise: A flaw in human judgment.* Little, Brown, 2021.

A. Kaplan and M. Haenlein. Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. *Business horizons*, 62(1):15–25, 2019.

A. E. Khandani, A. J. Kim, and A. W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*, 2017.

S. Y. Komiak and I. Benbasat. The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly*, (4):941–960, 2006.

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

M. Leyer and S. Schneider. Me, you or ai? how do we feel about delegation. In *ECIS*, 2019.

B. Y. Lim, A. K. Dey, and D. Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2119–2128, 2009.

J. A. Litman. Interest and deprivation factors of epistemic curiosity. *Personality and individual differences*, 44(7):1585–1595, 2008.

J. M. Logg, J. A. Minson, and D. A. Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.

C. Longoni, A. Bonezzi, and C. K. Morewedge. Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4):629–650, 2019.

Z. Lu and M. Yin. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *CHI Conference on Human Factors in Computing Systems*, 2021.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Conference on Neural Information Processing Systems (NIPS)*, 2017.

D. Lyell and E. Coiera. Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, 24(2):423–431, 2017.

Electronic copy available at: https://ssrn.com/abstract=4512594

R. McLeod Jr and J. W. Jones. A framework for office automation. *MIS Quarterly*, (1):87–104, 1987.

C. Meske and E. Bunde. Transparency and trust in human-ai-interaction: The role of model-agnostic explanations in computer vision-based decision support. In *Artificial Intelligence in HCI: First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*, pages 54–69. Springer, 2020.

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

T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

B. Mittelstadt, C. Russell, and S. Wachter. Explaining explanations in ai. In *Conference on Fairness, Accountability, and Transparency (FAT)*, 2019.

B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679, 2016.

C. Molnar. *Interpretable machine learning: A Guide for Making Black Box Models Explainable*. 2020.

K. L. Mosier, L. J. Skitka, S. Heers, and M. Burdick. Automation bias: Decision making and performance in high-tech cockpits. *The International journal of aviation psychology*, 8(1):47–63, 1998.

C. J. Nemeth and J. L. Kwan. Minority influence, divergent thinking and detection of correct solutions. *Journal of Applied Social Psychology*, 17(9): 788–799, 1987.

D. Olick. Artificial intelligence is taking over real estate. `https://www.cnbc.com/2021/09/17/what-artificial-intelligence-means-for-homebuyers-real-est` 2021. Accessed: 2023-04-21.

D. Önkal, P. Goodwin, M. Thomson, S. Gönül, and A. Pollock. The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, 22(4):390–409, 2009.

R. Parasuraman and D. H. Manzey. Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3):381–410, 2010.

R. Parasuraman and V. Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

A. Patacconi. Coordination and delay in hierarchies. *The RAND Journal of Economics*, 40(1):190–208, 2009.

Electronic copy available at: https://ssrn.com/abstract=4512594

F. Peters, L. Pumplun, and P. Buxmann. Opening the black box: Consumer's willingness to pay for transparency of intelligent systems. 2020.

F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach. Manipulating and measuring model interpretability. In *CHI Conference on Human Factors in Computing Systems*, 2021.

C. Prunkl. Human autonomy in the age of artificial intelligence. *Nature Machine Intelligence*, 4(2):99–101, 2022.

E. Rader, K. Cotter, and J. Cho. Explanations as mechanisms for supporting algorithmic transparency. In *CHI Conference on Human Factors in Computing Systems*, 2018.

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

L. S. Shapley. A value for n-person games. In *Contributions to the Theory of Games (AM-28), Volume II.* Princeton University Press, 1953.

T. B. Sheridan. *Humans and automation: System design and research issues*, volume 280. Human Factors and Ergonomics Society Santa Monica, CA, 2002.

Y. R. Shrestha, S. M. Ben-Menahem, and G. Von Krogh. Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4):66–83, 2019.

L. J. Skitka, K. L. Mosier, and M. Burdick. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006, 1999.

C. Starke, J. Baleis, B. Keller, and F. Marcinkowski. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, 9(2):20539517221115189, 2022.

M. Szymanski, M. Millecamp, and K. Verbert. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*, pages 109–119, 2021.

D. Tchuente and S. Nyawa. Real estate price estimation in french cities using geocoding and machine learning. *Annals of Operations Research*, pages 571—-608, 2022.

M. H. Teodorescu, L. Morse, Y. Awwad, and G. C. Kane. Failures of fairness in automation require a deeper understanding of human-ml augmentation. *MIS Quarterly*, 45(3b):1483–1499, 2021.

Electronic copy available at: https://ssrn.com/abstract=4512594

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

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

X. Wang and M. Yin. Are explanations helpful? a comparative study of the effects of explanations in AI-assisted decision-making. In *International Conference on Intelligent User Interfaces*, 2021.

F. Yang, Z. Huang, J. Scholtz, and D. L. Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *International Conference on Intelligent User Interfaces*, 2020.

L. R. Ye and P. E. Johnson. The impact of explanation facilities on user acceptance of expert systems advice. *Mis Quarterly*, (2):157–172, 1995.

M. Yeomans, A. Shah, S. Mullainathan, and J. Kleinberg. Making sense of recommendations. *Journal of Behavioral Decision Making*, 32(4):403–414, 2019.

Y. Zhang, Q. V. Liao, and R. K. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

Electronic copy available at: https://ssrn.com/abstract=4512594

6 Appendix

Information on real estate data and the AI system

We obtained the dataset by crawling apartments listed on a large online platform in February 2022. Specifically, we considered apartments listed for sale in the seven major cities of Germany (“A-cities”) and scraped multiple different attributes reflecting the number of rooms in the apartment or whether it has a balcony. We disregarded apartments for which the information on one or several attributes was missing. In order to characterize the location of the apartment within the city, we joined third-party data from public statistics: the share of voters for the German green party and the unemployment rate. Both attributes are captured on the level of districts and, subsequently, bagged to lower, mid, and upper third within the respective city. For example, if an apartment in Berlin is in the low third for unemployment, then it is located in a district for which the unemployment rate is below the average unemployment rate in Berlin. We further treat the top 0.5% of apartments with regard to the listing price as outliers and exclude them from our data. The final, preprocessed dataset comprises 5090 apartments and is described in Table 4.

Continuous attributes	average	standard dev	0.25 quantile	median	0.75 quantile
Listing price/m^2 [€]:	7158.55	3217.37	4500.0	6500.0	8500.0
Construction [year]:	1971.18	43.07	1937.0	1972.0	2018.0
Nmbr of rooms:	2.72	1.25	2.0	3.0	3.0
Floor (storey):	1.80	2.56	0.0	1.0	3.0
Ordinal attributes			lower third	mid third	higher third
Unemployment			44.7 %	30.8 %	24.6 %
Green party electorate			39.1 %	25.8 %	35.1 %
Binary attributes				Yes	No
Basement				68.1 %	31.9 %
Elevator				45.3 %	54.7 %
Balcony				60.1 %	39.9 %
Garden				21.5 %	78.5 %
Multicat. attributes					Distribution (shares)
City					Berlin (39.2 %), Hamburg (19.4 %), Munich (16.1 %) Cologne (8.9 %), Frankfurt (7.0 %), Stuttgart (4.8 %) Dusseldorf (4.7 %)

Table 4: Descriptive statistics of real-estate data.

Notes: We scraped the data from a large real-estate platform in Germany and joined the ordinal attributes (unemployment and green party electorate) by drawing from public statistics. We considered the seven major cities in Germany (“A-Cities”). We excluded real-estate for which the price or any of the remaining attributes were not listed. This left us with 5090 observations.

We randomly split the data into different sets for training (95%) and testing (5%) of our AI system, following common conventions. Moreover, we ensure that the apartments directly featured in our experiment fall into the test set.

Our AI system is based on a random forest. To yield a prediction, the random forest averages across the predictions of multiple, randomized decision trees. In our case, the random forest predicts the listing price per square meter based on

Electronic copy available at: https://ssrn.com/abstract=4512594

the remaining 10 attributes as predictors. We determine the hyperparameters for the random forest by applying a grid search in a 5-fold cross-validation on the training set. Subsequently, we assess the performance of our AI system based on the test data ($\text{R}^2 = 0.72$).

Our explanations are based on SHAP values. We compute SHAP values for all predictors using the tree implementation of the SHAP value method. As a result, for each of the 8 apartments featured in the experimental main stage, we yield both the predicted listing price per square meter and the contribution of each of the 10 predictors.

Additional analyses on trust and domain knowledge

Dep. variable: Evaluation equals prediction	(1)	(2)
Observing explanation (b2)	0.223***	0.18***
	(0.062)	(0.049)
Accuracy belief		0.005**
		(0.002)
Comp. trust		0.031
		(0.037)
Emo. trust		0.08***
		(0.027)
Integr. trust		-0.034
		(0.022)
N	782	782
p	0.000	0.000
R^2	0.284	0.468

Table 5: Trust calibrations and domain knowledge.

Notes: We depict results from random effects GLS regression models with robust standard errors clustered on the individual level and reported in parentheses. As we measured trust only in treatments where participants actually interacted with an AI system, we can only include observations from the AI and XAI treatments. In all columns, the dependent variable is equal to one when real estate agents' evaluation coincides with the prediction an zero otherwise. As independent variables, we include a treatment dummy, subjective accuracy beliefs, and the three trust measures. We include controls on agents' gender, age, level of risk aversion, academic degree, confidence in their evaluation, familiarity with AI technology, and degree of overconfidence. We denote significance levels by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Electronic copy available at: https://ssrn.com/abstract=4512594

	(1)	(2)	(3)	(4)
Dep. variable:	Acc. belief	Comp. trust	Emo. trust	Integr. trust
Observing explanation (β_1)	0.548	0.337	0.289	0.448
	(4.982)	(0.393)	(0.432)	(0.398)
High expertise (β_2)	1.018	-0.673	-1.217*	0.396
	(5.708)	(0.521)	(0.636)	(0.564)
High expertise*Observing explanation (β_3)	-5.329	0.602	0.922	-0.225
	(7.293)	(0.587)	(0.684)	(0.633)
F-test: $\beta_1 + \beta_3$	0.41	0.04	0.03	0.65
N		98	98	98
p		0.000	0.000	0.000
R^2	0.183	0.297	0.228	0.141

Table 6: Trust calibrations and domain knowledge.

Notes: We depict results from OLS regression models with robust standard errors. In different columns, the dependent variable equals accuracy beliefs and different trust measures following Komiak and Benbasat (2006). As we measured trust only in treatments where participants actually interacted with an AI system, we can only include observations from the AI and XAI treatments. As independent variables, we include a treatment dummy for XAI, a dummy indicating agents' level of expertise, and the interaction effect. We include controls on agents' gender, age, level of risk aversion, academic degree, confidence in their evaluation, familiarity with AI technology, and degree of overconfidence. We denote significance levels by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 6 reports results from OLS regression analyses where participants' subjective accuracy and reported levels of emotional and cognitive trust in the AI system serve as dependent variables. In columns (1), (2), (3), and (4), we respectively show results for subjective accuracy beliefs, cognitive trust in competence, emotional trust, and cognitive trust in integrity measure (Komiak and Benbasat, 2006). As independent variables, we use a treatment dummy for whether or not participants observed explanations on top of predictions, a variable indicating that they possess a high level of domain knowledge, and a corresponding interaction term. We report robust standard errors in parentheses.

Reported results suggest that real estate agents with a high level of domain knowledge exhibited lower trust in the competence and lower emotional trust in a black box AI system than their low domain knowledge counterparts (see β_2 in columns (2) and (3)). This difference may originate from the inability to determine the inner logic that the AI system applies to arrive at a certain output. We further find that the additional provision of explanations on top of predictions has a significantly positive impact on these two trust measures (see $\beta_1 + \beta_3$), increasing high expertise agents' trust in the AI to levels that are insignificantly different from agents with low domain knowledge (see F-tests). Notably, we do not find a significant XAI treatment effect for agents with low domain knowledge, emphasizing the heterogeneous nature of XAI on trust calibrations in our study.

Table 7 shows results from regression analyses, where the dependent variable is a dummy indicating that a real estate agent's evaluation of an apartment equaled the prediction of our AI system. As independent variables of main in-

Electronic copy available at: https://ssrn.com/abstract=4512594

Dep. variable: Evaluation equals prediction	(1)	(2)
Observing explanation (β_1)	0.101	0.077
	(0.089)	(0.071)
High expertise	-0.241**	-0.118
	(0.107)	(0.087)
High expertise*Observing explanation (β_3)	0.281**	0.219**
	(0.133)	(0.103)
Acc. belief		0.006***
		(0.002)
Comp. trust		0.023
		(0.038)
Emo. trust		0.076***
		(0.029)
Integr. trust		-0.031
		(0.022)
N	782	782
p	0.000	0.000
R^2	0.304	0.477

Table 7: Delegation of evaluations and domain knowledge.

Notes: We depict results from random effects GLS regression models with robust standard errors clustered on the individual level and reported in parentheses. As we measured trust only in treatments where participants actually interacted with an AI system, we can only include observations from the AI and XAI treatments. In all columns, the dependent variable is equal to one when real estate agents' evaluation coincides with the prediction and zero otherwise. As independent variables, we include treatment dummies, a dummy indicating agents' level of expertise, their interaction effect, subjective accuracy beliefs, and the three trust measures. We include controls on agents' gender, age, level of risk aversion, academic degree, confidence in their evaluation, familiarity with AI technology, and degree of overconfidence. We denote significance levels by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

terest, we include treatment dummies, a variable indicating that they possess a high level of domain knowledge, and a corresponding interaction term. We report robust standard errors that we cluster at the individual level in parentheses. Our study's control condition where real estate agents did not observe a prediction serve as the reference category. Columns (1) and (2) merely differ regarding the inclusion of the additional control variables measuring subjective accuracy beliefs and trust in the AI system. We include these measures in column (2) to examine whether the XAI-driven increase in high expertise agents' trust in the AI system (see Table 6) can, at least partially, account for the increase in their reliance on the prediction, i.e., we effectively test for a mediation effect.

A comparison of columns (1) and (2) reveals that the inclusion of the trust measures causes the coefficient β_3 to become considerably smaller in magnitude (-22.1%). Considering that both the coefficient for subjective accuracy beliefs and for emotional trust are statistically significant and increase for high expertise agents through explainability, these results support the notion that the XAI-driven increase in high expertise agents' overreliance on predictions is driven by an increase in these two trust dimensions – at least partially.

Electronic copy available at: https://ssrn.com/abstract=4512594

Recent Issues

No. 393	Michael Kosfeld, Zahra Sharafi	The Preference Survey Module: Evidence on Social Preferences from Tehran
No. 392	Christian Mücke	Bank Dividend Restrictions and Banks' Institutional Investors
No. 391	Carmelo Latino, Loriana Pelizzon, Max Riedel	How to Green the European Auto ABS Market? A Literature Survey
No. 390	Kamelia Kosekova, Angela Maddaloni, Melina Papoutsi, Fabiano Schivardi	Firm-Bank Relationships: A Cross-Country Comparison
No. 389	Stefan Goldbach, Philipp Harms, Axel Jochem, Volker Nitsch, Alfons J. Weichenrieder	Retained Earnings and Foreign Portfolio Ownership: Implications for the Current Account Debate
No. 388	Gill Segal, Ivan Shaliastovich	Uncertainty, Risk, and Capital Growth
No. 387	Michele Costola, Katia Vozian	Pricing Climate Transition Risk: Evidence from European Corporate CDS
No. 386	Alperen Afşin Gözlügöl, Wolf-Georg Ringe	Net-Zero Transition and Divestments of Carbon-Intensive Assets
No. 385	Angela Maddaloni	Liquidity Support and Distress Resilience in Bank-Affiliated Mutual Funds
No. 384	Julian Greth, Alperen Afşin Gözlügöl, Tobias Tröger	The Oscillating Domains of Public and Private Markets
No. 383	Satyajit Dutt, Jan Wedigo Radermacher	Age, Wealth, and the MPC in Europe - A Supervised Machine Learning Approach
No. 382	Jan Wedigo Radermacher	Mamma Mia! Revealing Hidden Heterogeneity by PCA-Biplot - MPC Puzzle for Italy's Elderly Poor
No. 381	Olga Goldfayn, Nathanael Vellekoop	The Role of Personality Traits in Household Loan Expectations and Borrowing Constraints
No. 380	Alexander Hillert, Alexandra Niessen-Ruenzi, Stefan Ruenzi	Mutual Fund Shareholder Letters: Flows, Performance, and Managerial Behavior
No. 379	Irene Mecatti, Tobias Tröger	Who Should Hold Bail-Inable Debt and How can Regulators Police Holding Restrictions Effectively?

Electronic copy available at: https://ssrn.com/abstract=4512594