Received: 23 December 2022 | Accepted: 16 January 2024

DOI: 10.1002/jcpy.1416

RESEARCH ARTICLE

Unintended effects of algorithmic transparency: The mere prospect of an explanation can foster the illusion of understanding how an algorithm works

Massimiliano Ostinelli[1] | **Andrea Bonezzi**[2] | **Monika Lisjak**[3]

[1]College of Business, Florida Atlantic University, Boca Raton, Florida, USA

[2]Leonard N. Stern School of Business, New York University, New York, New York, USA

[3]W. P. Carey School of Business, Arizona State University, Tempe, Arizona, USA

Correspondence
Massimiliano Ostinelli, College of Business, Florida Atlantic University, Boca Raton, FL, USA.
Email: mostinelli@fau.edu

Abstract

This research shows that merely believing that one can access an explanation of how an algorithm works can foster an illusory sense of understanding the algorithm, even when people do not actually access and read the explanation. This effect occurs because the belief that one can access an explanation provides a feeling of empowerment that fosters an illusory sense of understanding. In turn, this illusory sense of understanding can yield unfounded reliance on algorithmic determinations. We further show that this effect is moderated by the target of an explanation and by the perceived utility of an explanation in enabling consumers to use an algorithm more effectively. From a theoretical standpoint, we offer a novel psychological account of illusory understanding based on empowerment. From a practical standpoint, we point to an unintended effect of algorithmic transparency: merely knowing that one can access an explanation for how an algorithm works may lull consumers into a false sense of understanding that yields unfounded reliance on algorithmic recommendations.

KEYWORDS

algorithmic transparency, algorithms, illusion of understanding, psychology of explanations

INTRODUCTION

Algorithms exert a pervasive influence on consumers' lives (Castelo et al., 2019; Granulo et al., 2021; Mende et al., 2019; Puntoni et al., 2021; Schmitt, 2019). However, a concern associated with the increasing diffusion of this technology is that consumers often don't understand how algorithms reach their determinations (Watson & Nations, 2019). For example, consumers generally do not know how algorithms make product recommendations (Sinha & Swearingen, 2002), pick financial investments (Tan, 2020), or diagnose diseases (Rajkomar et al., 2018). Without understanding how algorithms work, consumers cannot make informed decisions about whether to rely on algorithmic determinations. For this reason, in many domains, consumers are reluctant to rely on algorithms they do not understand (Bonezzi et al., 2022; Yeomans et al., 2019), and advocacy groups and policymakers have called for greater transparency into the inner workings of algorithms (Goodman & Flaxman, 2017; Rajkomar et al., 2018; Tan, 2020).

In response to these calls for algorithmic transparency, legislators have institutionalized a "right to an explanation," that is, a right for consumers to access explanations for how an algorithm reaches a certain determination (see Appendix A for examples). At the same time, organizations have started to provide access to explanations for how their algorithms work (see Appendix B for examples). Access to these explanations is often provided through links on companies' websites that, if clicked on, direct consumers to dedicated web pages or documents that explain how an algorithm works—e.g., hyperlinks such as "click here to learn how it works" or "an explanation is available in this document." As a result, consumers are increasingly exposed to cues that signal the opportunity to access an explanation of how an algorithm works.

Intuitively, providing the opportunity to access an explanation of how an algorithm works should benefit

Accepted by Lauren Block, Editor; Associate Editor, Rosellina Ferraro

consumers. Contrary to this intuition, we suggest that the mere belief that one can access an explanation of how an algorithm works may also have an unintended and potentially detrimental effect. We propose that merely believing that one could access an explanation of how an algorithm works—without actually accessing and reading such an explanation—may provide a feeling of empowerment that fosters an illusory sense of understanding. This illusory sense of understanding may, in turn, yield unfounded reliance on algorithmic determinations. We elaborate on the theoretical rationale for our prediction in the next section.

Our work contributes to prior research on the psychology of explanations (Dhaliwal & Benbasat, 1996; Johnson & Johnson, 1993; Lombrozo, 2006). While prior research has investigated how *reading* different types of explanations influences people's propensity to rely on algorithmic recommendation (Buchanan & Shortliffe, 1984; Demirdag & Shu, 2021; Ye & Johnson, 1995), our research offers a first attempt to explore what happens when consumers have the opportunity to access an explanation of how an algorithm works, yet they do not access and read it, a common situation that has received no attention so far. Furthermore, this research contributes to the literature on the illusion of understanding (Fisher et al., 2015; Rabb et al., 2019; Sloman & Rabb, 2016; Ward, 2021). Past research has attributed the illusion of understanding to a cognitive mechanism whereby people confuse the knowledge that others have with their own knowledge. Our research offers a novel account whereby the illusion of understanding stems from feeling empowered. Finally, from a practical standpoint, we document an unintended and potentially detrimental effect of algorithmic transparency initiatives that have implications for policymakers. Existing regulations, such as the European Union's General Data Protection Regulation (Goodman & Flaxman, 2017), require companies to inform consumers of the accessibility of an explanation of how an algorithm works. Consumers, however, can acknowledge the existence of the explanation and proceed to use a company's services without reading the explanation. Our work suggests that this practice may foster unfounded confidence in algorithmic determinations.

Theoretical background

It is well established that reading explanations for how algorithms work can increase consumers' reliance on algorithmic determinations (Buchanan & Shortliffe, 1984; Demirdag & Shu, 2021; Tomaino et al., 2020; Ye & Johnson, 1995). Based on this notion, research has investigated how different explanations influence consumers' reliance on algorithmic determinations. For instance, past research has compared explanations that focus on the reasons (i.e., the why) versus the process (i.e., the how) for a recommendation (Cotter et al., 2017), explanations that provide more versus fewer details (Narayanan et al., 2018; Yeomans et al., 2019), or explanations that provide visual versus textual information (Gedikli et al., 2014; Oduor & Wiebe, 2008).

The experimental paradigm typically used in that research entails presenting participants with an explanation of how an algorithm works and assessing their propensity to follow the algorithm's recommendations after they read such an explanation. Although this paradigm provides valuable insights into the effects of different types of explanations, it comes with a limitation: it forces respondents to read the explanation. In many situations, however, consumers may be exposed to cues indicating they have the opportunity to access an explanation but may not be directly presented with the explanation itself. This is the case, for instance, when consumers encounter links to an explanation (e.g., "click here to learn how it works") or, more generally, when they are informed of the possibility of accessing an explanation. However, providing the opportunity to access an explanation does not guarantee that consumers will actually access and read it. In fact, a considerable amount of research shows that most consumers do not read the explanations that companies provide, even when these explanations are considered important and easily accessible (Acquisti et al., 2015; Obar & Oeldorf-Hirsch, 2020). Industry reports suggest that about 90% of consumers might not read explanations made available by companies (Hart, 2019).

To date, no research has examined whether cues that signal the opportunity to access an explanation of how an algorithm works can influence consumers if they do not access and read the explanation. The objective of this work is to address this gap in the literature. To be clear, we are *not* interested in exploring when and why consumers might or might not read an explanation (for research on this question, see: Obar & Oeldorf-Hirsch, 2020; Rudolph et al., 2018). Rather, we are interested in whether merely believing that one could access an explanation of how an algorithm works can influence consumer behavior when people do not access the explanation. We use the term *prospect of an explanation* to refer to the *belief* that one has the ability to access an explanation, even if such an ability is not enacted. We propose that the prospect of an explanation can provide a feeling of empowerment that can foster an illusory sense of understanding how the algorithm works and yield unfounded reliance on algorithmic determinations.

Prospect of an explanation and empowerment

Empowerment has been defined as the process by which people are provided with means to gain control and mastery over the factors that affect their environment (Han & Broniarczyk, 2022; Kozinets et al., 2021). Put differently, empowerment is the process of promoting and

enhancing the ability to meet one's needs, solve one's problems, and mobilize the necessary resources to feel in control of one's environment (Gibson, 1991). Thus, empowerment can be summarized as the affordance of an ability to perform an action instrumental to achieving an intended outcome (Kozinets et al., 2021).

Research on empowerment has often implicitly assumed that people empowered with a certain ability to act will exercise it. For this reason, the effects of empowerment have generally been equated to the benefits associated with taking advantage of an afforded ability to act. For example, research has investigated how granting access to information reduces asymmetries between consumers and producers (Labrecque et al., 2013) under the assumption that people empowered with the ability to access information will do so.

Here, we take a different perspective and focus on the *psychological experience of empowerment* (Conger & Kanungo, 1988; Galinsky et al., 2003), that is, the feeling of control that stems from believing that one has the ability to perform an action instrumental to a certain objective, regardless of whether such ability is, in fact, enacted. Key to our research, prior literature suggests that having the ability to access information can foster such a psychological experience of empowerment (Han & Broniarczyk, 2022). We build on this notion to propose that the mere prospect of an explanation of how an algorithm works can foster a psychological experience of empowerment in that it gives people control over acquiring knowledge instrumental to making better decisions about an algorithm's recommendations. We further suggest that this feeling of empowerment can, in turn, yield an illusory sense of understanding how an algorithm works. Next, we explain the rationale for this prediction.

Empowerment and illusory understanding

Our prediction that empowerment can foster illusory understanding is based on the premise that the sense of understanding how something works is a subjective feeling that is often misaligned with actually knowing how something works (Alter et al., 2010; Carlson et al., 2009; Hadar et al., 2013; Rozenblit & Keil, 2002; Trout, 2002). This is because the sense of understanding how something works is a metacognitive feeling that signals whether one has sufficient knowledge to act effectively in a given situation, such as whether one has adequate knowledge to use a product or make a decision (Craik, 1967; Gopnik, 1998; Long et al., 2018; Waytz et al., 2010; Ylikoski, 2019). As such, the sense of understanding how something works doesn't necessarily hinge on knowing the inner workings of an object but rather can stem from the belief that one has the ability to interact effectively with that object. For this reason, the sense of understanding how something works can be illusory.

Consistent with this notion, scholars have theorized that when people feel empowered to interact effectively with an object, they might experience an illusory sense of understanding how that object works (Keil, 2011; Ylikoski, 2019). For example, Keil (2011) argues that knowing how to interact effectively with an object, such as a computer mouse, might elicit a "sense of mastery that might seem like explanatory understanding." That is, the feeling of control that stems from confidence in one's ability to use an object effectively can foster an illusory sense of understanding how that object works. Ylikoski (2019, p. 101) makes a similar prediction based on the fact that an objective understanding of how something works, that is, an understanding based on actual knowledge, can afford control of one's environment. Yet, because assessing one's objective understanding of how something works is often difficult, people might take their feeling of control as an indication, often misguided, of how much they understand how something works. Additional support for the notion that feeling empowered can yield an illusory sense of understanding comes from prior literature on uncertainty reduction (Edwards & Weary, 1998; Weary & Edwards, 1994, 1996). According to this literature, knowing how to interact effectively with an object reduces the perceived uncertainty surrounding that object, and reductions in uncertainty can elicit a sense of understanding of how that object works.

Building on these notions, we propose that the feeling of empowerment elicited by the mere prospect of an explanation can yield an illusory sense of understanding how an algorithm works. The prospect of an explanation empowers people in that it gives them confidence that by taking an action—accessing the explanation—they can learn more about how the algorithm works and thus make more informed decisions about the algorithm's recommendations. This empowering belief that one can take an action instrumental to using an algorithm more effectively reduces the perceived uncertainty surrounding that algorithm, fostering an illusory sense of understanding how the algorithm works. Based on these notions, we propose that feeling empowered by the prospect of an explanation can increase consumers' sense of understanding how the algorithm works *relative* to not having such access. We further propose that a higher sense of understanding can, in turn, increase consumers' propensity to rely on the algorithm's recommendations. Indeed, prior research shows that the sense of understanding how something works drives consumers' attitudes and preferences in many domains (Fernbach et al., 2019; Hadar et al., 2013; Long et al., 2018), including algorithmic decision-making (Bonezzi et al., 2022; Yeomans et al., 2019).

Although our prediction has not been empirically tested before, prior research on power offers some consistent evidence. Prior findings show that when individuals experience a feeling of empowerment, they

also feel more confident in their ability to act (Brinol et al., 2007), presumably because they feel they have a sufficient understanding of how things work (see also Maheswaran & Chaiken, 1991). For example, a field study shows that people who felt more empowered also felt more confident in their ability to make judgments and sought less advice from their colleagues (See et al., 2011), possibly because they felt they understood how to behave effectively. To this point, Wan and Rucker (2013, p. 978) argue that "high confidence leads people to infer that their knowledge is sufficient," an experience that might trigger an illusory sense of understanding.

Additional indirect evidence comes from research on illusory control (Gilovich & Douglas, 1986; Langer, 1975; Presson & Benassi, 1996). This research shows that individuals who perceive a sense of control over a game of chance tend to express greater confidence in their ability to predict the game's outcome. Feeling in control might lead to an illusion of understanding how an action (e.g., blowing on a pair of dice before throwing them) results in a particular outcome (e.g., winning). Overall, these findings suggest that empowerment can lead to an illusory sense of understanding how something works by heightening people's confidence in their ability to act effectively. Based on this discussion, we propose that,

H1. The prospect to access an explanation of how an algorithm works leads to a feeling of empowerment that results in an illusion of understanding how the algorithm works, compared to not having the prospect to access an explanation.

Our prediction is also consistent with the emerging stream of research on the community of knowledge, which asserts that people think that they understand a phenomenon simply by virtue of knowing that others in their community—generally experts—understand it (Keil, 2005; Kominsky & Keil, 2014; Rabb et al., 2019; Sloman et al., 2021; Sloman & Rabb, 2016; Ward, 2021). The community of knowledge argues that people take credit for others' knowledge because they fail to draw clear cognitive boundaries between knowledge that resides in their heads and knowledge that resides elsewhere, such as in experts' minds, books, and websites (Rabb et al., 2019; Ward, 2021). The community of knowledge effect has been attributed to a cognitive process whereby people have mental pointers to expert sources of knowledge and confuse those mental pointers for possessing the actual knowledge that those pointers index (Rabb et al., 2019; Sloman et al., 2021). Our research builds on and extends prior research by proposing a novel process based on empowerment that operates above and beyond the cognitive process proposed by the community of knowledge. Our empowerment account is especially relevant in the domain of algorithmic decision-making, where explanations can enable people to interact more effectively with the algorithm, and, for this reason, the prospect of accessing such explanations can foster a feeling of empowerment.

Our empowerment account makes unique predictions that set it apart from the theorizing put forth by the community of knowledge literature. First, our empowerment account predicts that the target of an explanation should moderate the illusion of understanding. Specifically, the illusion of understanding should be more pronounced for explanations written for the general public (i.e., laypeople) than for experts. This is because explanations written for the general public should be more empowering, as they should foster a stronger belief that one can learn how an algorithm works and use that knowledge to make more informed decisions about the algorithm's recommendations. This prediction is important both from a theoretical and from a practical standpoint. From a theoretical standpoint, this prediction sets our theorizing apart from past literature, as it would not be directly accounted for by the community of knowledge theorizing, which centers around the idea that people confuse mental pointers with potential sources of knowledge for possessing that knowledge. In particular, the community of knowledge has focused on the idea that the illusion of understanding stems from possessing mental pointers to sources of knowledge with established expertise, such as scientists (Hemmatian & Sloman, 2018; Keil, 2005; Rabb et al., 2019; Sloman et al., 2021; Sloman & Rabb, 2016). For example, Sloman and Rabb (2016, p. 1458) argue that "knowing that experts understand a phenomenon gives individuals the sense that they understand it better themselves." Thus, based on the community of knowledge account, the prospect of explanations written for laypeople, which do not reference an established source of expertise, might result in the same or even lower illusory sense of understanding as the prospect of explanations written for experts. By contrast, our theorizing predicts that the prospect of laypeople explanations will lead to a greater illusory sense of understanding. From a practical standpoint, this prediction contributes to the ongoing debate about the welfare implications of explanations designed for experts versus laypeople (Kaminski & Malgieri, 2021; Ribera & Lapedriza García, 2019). While explanations targeted to a lay audience might be seen as providing stronger protection for people's rights, our theorizing suggests that these explanations might also lead to a greater illusory sense of understanding. In sum, our empowerment account predicts that:

H2. The prospect to access an explanation of how an algorithm works leads to a greater illusion of understanding when the explanation is written for the general public than for experts.

Furthermore, our empowerment account predicts that the illusion of understanding should be moderated by the type of knowledge made accessible and, specifically, by its instrumentality. To the extent that the prospect of an explanation fosters illusory understanding by empowering people to use an algorithm's recommendations more effectively, believing that one can access knowledge that does not enable them to make more informed decisions about an algorithm's recommendations (i.e., lower instrumentality) should reduce the illusion of understanding, compared to believing that one can access knowledge that enables to make more informed decisions about the algorithm's recommendations (i.e., higher instrumentality). This is because being able to access knowledge that is not instrumental to making more informed decisions about an algorithm's recommendations should reduce the feeling of empowerment and, as such, the illusion of understanding. This prediction further sets our theorizing apart from previous literature on the community of knowledge. To the extent that the illusion of understanding is based on confusion between mental pointers to sources of knowledge and knowledge itself, whether the knowledge made accessible is or is not instrumental to using an algorithm's recommendations more effectively should be irrelevant to experiencing an illusory sense of understanding. Indeed, the literature on the community of knowledge has suggested that illusory understanding is driven by possessing mental pointers to sources of expert knowledge regardless of whether such knowledge can provide practical guidance or not (Sloman et al., 2021). Instead, we predict a greater illusory sense of understanding when accessible knowledge is believed to be instrumental in enabling people to interact more effectively with their environment. Specifically, our empowerment account predicts that:

H3. The prospect to access an explanation of how an algorithm works leads to a smaller illusion of understanding when the explanation is not instrumental to interacting more effectively with an algorithm.

OVERVIEW OF STUDIES

We begin our empirical investigation by showing that the mere prospect of an explanation can increase the sense of understanding how an algorithm works (Experiment 1) and have downstream consequences (Experiment 2). Next, we provide evidence for the prosed process via empowerment (H1) while probing several alternative explanations (Experiments 3 and 4). We then test whether the prospect of an explanation designed for laypeople leads to a greater illusion of understanding than the prospect of an explanation designed for experts (H2; Experiment 4). Finally, we provide a process-by-moderation test of our proposed psychological process by examining whether the illusion of understanding is reduced when an explanation is not instrumental to interacting more effectively with an algorithm's recommendations (H3; Experiment 5). We report data collection practices along with additional analyses in the web Appendices. Survey instruments and datasets are available at osf.io/g3k6d.

EXPERIMENT 1: PROSPECT OF EXPLANATION AND UNDERSTANDING

Experiment 1 tests whether the prospect of an explanation of how an algorithm works can increase consumers' sense of understanding of how an algorithm works when they do not read the explanation. We test our hypothesis in the context of credit scores. Credit scores measure a consumer's creditworthiness and are used in many important decisions, such as determining people's access to credit (e.g., amount of credit, interest rate). Credit scores are computed by algorithms that consider different pieces of information, such as payment history, credit utilization, and credit history. An accurate understanding of how credit scores are computed has implications for one's financial well-being, as people who overestimate their understanding of how credit scores are computed are more likely to be denied credit and pay higher interest rates on loans (Courchane et al., 2008). We predict that the prospect of an explanation of how an algorithm computes a credit score leads to a greater sense of understanding of how the algorithm works even though people do not access and read the explanation.

Method

Two hundred Cloud Research-approved participants (M_{age} = 40.4, SD = 12.0, 50% female) were recruited to participate in a study about VantageScore, one of the main algorithms used to compute credit scores. Participants were assigned to one of two conditions. In the control condition, people were asked to indicate their understanding of how the VantageScore algorithm works on a scale adopted from prior research (1 = do not understand at all, 7 = understand completely; Alter et al., 2010; Bonezzi et al., 2022; Fernbach et al., 2013). In the explanation prospect condition, participants were informed that a description of how the VantageScore algorithm computes a person's credit score was available, and they were presented with a link to such an explanation. Participants who clicked on the link were directed to an explanation of how the VantageScore algorithm computes a person's credit score. Participants were then asked to report their understanding of how the VantageScore algorithm works, regardless of whether or not they clicked on the link.

Results and discussion

In line with past research showing that most people do not read explanations even when easily accessible (Acquisti et al., 2015; Hart, 2019; Nissenbaum, 2011), only ten participants assigned to the explanation prospect condition clicked on the link. Since our focus is on whether the mere prospect of an explanation leads to an illusion of understanding when people do not access and read the explanation, we restricted the analysis to those participants who did not click on the link. We examined whether participants in the explanation prospect condition who did not access the explanation would indicate a higher understanding of how the VantageScore algorithm works than participants in the control condition even though, in both cases, nobody read an explanation.

As predicted, participants presented with a link to an explanation reported a greater sense of understanding of how the VantageScore algorithm works ($M = 3.75$, $SD = 1.90$) compared to participants in the control condition ($M = 2.87$, $SD = 1.71$; $t(188) = 3.37$, $p = 0.001$, $d = 0.49$), even though they did not access the explanation. This result provides preliminary evidence for our hypothesis that the mere prospect of an explanation can increase consumers' sense of understanding of how an algorithm works, even when they do not read the explanation.

While these results provide evidence for the proposed effect, participants in the explanation prospect condition who clicked on the link were excluded from the analysis. This could have led to an overestimation of the illusion of understanding effect to the extent that these individuals would have reported a lower sense of understanding had they not accessed the explanation. We address this concern in two ways. First, we conducted a robustness analysis aimed at equating the exclusions across the two conditions. Specifically, we also excluded participants from the control condition who reported the lowest sense of understanding, under the assumption that those excluded from the explanation prospect condition would have reported the lowest sense of understanding in the sample had they not accessed the explanation. This analysis replicated the conclusion presented earlier (explanation prospect: $M = 3.75$, $SD = 1.90$ vs. control: $M = 3.08$, $SD = 1.67$; $t(178) = 2.53$, $p = 0.012$, $d = 0.37$), thus limiting concerns about the differential exclusion of participants. Second, we designed experiments that manipulated the prospect of an explanation without providing participants with a link they could click on, thus avoiding any differential exclusion of participants (Experiments 3, 4, and 5).

EXPERIMENT 2: PROSPECT OF EXPLANATION, UNDERSTANDING, AND CHOICE

Experiment 2 further tests whether the prospect of an explanation can lead to an illusion of understanding how an algorithm works and, in turn, yield a greater likelihood of relying on the algorithm's recommendation. We test this hypothesis with an incentive-compatible task where participants choose between a more accurate and a less accurate robot advisor. We investigate whether having access to an explanation for the less accurate robo-advisor could sway preferences toward the less accurate robo-advisor even when participants do not access the explanation. This question is practically relevant as algorithms for which an explanation can be accessed are often less accurate than inherently inexplicable algorithms for which an explanation cannot be accessed (Gunning & Aha, 2019; Herm et al., 2023).

Method

Four hundred Cloud Research-approved participants ($M_{age} = 42.5$ years, $SD = 12.7$ years, 50% female) were randomly assigned to one of four conditions of a 2 (prospect of explanation for more accurate robo-advisor: present vs. absent) × 2 (prospect of explanation for less accurate robo-advisor: present vs. absent) between-subject design. Participants were asked to choose between two robo-advisors (A and B) used to predict the S&P 500. Participants were informed that they could receive a bonus payment if the forecast of the robo-advisor they chose turned out to be the one closest to the actual value of the S&P 500 index on a specific day.

In all conditions, participants were informed that, on average, robo-advisor A outperformed robo-advisor B. Specifically, the accuracy rate was 94% for robo-advisor A and 89% for robo-advisor B. The study further mentioned that the accuracy rate of both robo-advisors was certified by an independent company. In the prospect of an explanation present conditions, participants were informed that an explanation of how the target robo-advisor works was made available by the developer of the robo-advisor, and they were provided with a clickable link to that explanation. In the prospect of an explanation absent conditions, participants were informed that no explanation of how the target robo-advisor works was made available. Participants were then asked to choose one of the two robo-advisors. After choosing one of the two robo-advisors, participants were asked to indicate their understanding of the two robo-advisors on a 7-point scale (1 = "I understand better how robo-advisor A works" and 7 = "I understand better how robo-advisor B works"). At the end of the study, participants were asked to recall the accuracy rate of each robo-advisor: 80% correctly recalled the accuracy rate of robo-advisor A, and 78% correctly recalled the accuracy rate of robo-advisor B (McNemar's Test (χ^2 (1) = 0.49, $p = 0.48$)). Participants were also asked to indicate which of the two robo advisors they had the opportunity to access an explanation for: 79% answered correctly. The conclusions reported

next do not change if participants who fail any of the checks above are omitted from the analysis.

Results and discussion

Sixteen participants among those who had the opportunity to access an explanation did so. These participants were not included in the subsequent analysis since we are interested in whether the prospect of an explanation affected the sense of understanding of participants who did not read the explanation. A logistic regression on choice revealed a significant interaction effect (χ^2 (1) = 5.40, p = 0.020; see Table 1). When no prospect of an explanation was present for the more accurate robo-advisor (A), participants were more likely to choose the less accurate robo-advisor (B) if they had the prospect of accessing an explanation for the less accurate robo-advisor versus not, χ^2 (1) = 18.53, p < 0.001. Yet, unsurprisingly, when the prospect of an explanation for the more accurate robo-advisor was presented, participants' choice of the less accurate robo-advisor was low regardless of whether they had the prospect of accessing an explanation for that algorithm or not, χ^2 (1) = 0.08, p = 0.77.

An ANOVA on the sense of understanding revealed a significant interaction effect (F(1,380) = 5.85, p = 0.016; see Table 1). When no explanation prospect was present for the more accurate robo-advisor (A), giving participants the prospect to access an explanation for the less accurate robo-advisor (B) increased the sense of understanding this robo-advisor relative to not having the prospect to access such an explanation, t(380) = 7.39, p < 0.001, d = 1.17. This effect was attenuated, albeit still significant, when participants were given the prospect of an explanation for the more accurate robo-advisor A, t(380) = 3.88, p < 0.001, d = 0.52.

A mediation analysis (Hayes, 2017; Model 7) revealed a significant indirect effect through sense of understanding (explanation prospec $t_{\text{robo-adv. B}} \rightarrow \text{understanding}_{\text{robo-adv. B}} \rightarrow \text{choice}_{\text{robo-adv. B}}$) both when an explanation prospect for the more accurate robo-advisor (A) was absent (b = 1.453, 95% CI: 1.020, 2.051) as well as when it was present (b = 0.775, 95% CI: 0.335, 1.299). Moreover, when controlling for sense of understanding, the prospect of an explanation no longer influenced choice (Z = 0.54, p = 0.59), while sense of understanding significantly predicted choice (Z = 6.79, p < 0.001), suggesting full mediation. These findings provide evidence that the mere prospect of an explanation can foster illusory understanding and that such illusory understanding can influence subsequent behavior. We also conducted a robustness analysis with the same goal as the one described in Experiment 1 to address the differential exclusion of participants. This analysis led to the same conclusion (see Appendix D), suggesting that the differential exclusion of participants did not account for the effect. This experiment also provides important practical implications in light of the current technological developments posing a potential tradeoff between accuracy and explainability, where most accurate algorithms are often the least explainable (Gunning & Aha, 2019). The results from this experiment show that the prospect of an explanation can lull consumers into preferring suboptimal (i.e., less accurate) algorithms simply because they afford the possibility of accessing an explanation. In the next experiment, we explore the role of empowerment as an antecedent of the illusory sense of understanding triggered by the prospect of an explanation.

EXPERIMENT 3: EMPOWERMENT VS. ALTERNATIVE EXPLANATIONS

Experiment 3 aims to provide a first test of our empowerment account of illusory understanding vis-à-vis alternative accounts. First, it aims to rule out the possibility that when the existence of an explanation is not mentioned, as in Experiments 1 and 2, consumers might infer that an algorithm is intrinsically inexplicable and, therefore, inherently more difficult to understand. To this end, we manipulate the prospect of an explanation while making all participants aware of the existence of an explanation. Second, this experiment aims to test whether empowerment drives the sense of understanding how an algorithm works above and beyond inferences about the trustworthiness and transparency of the company that provides access to the explanation. A company that provides access to an explanation might be perceived as more trustworthy and transparent, and such

TABLE 1 Means and standard deviations in Experiment 2.

Prospect of explanation			
More accurate Robo-advisor (A)	**Less accurate Robo-advisor (B)**	**Choice of less accurate Robo-advisor (B) (%)**	**Sense of understanding[a]**
No	No	7	3.82 (1.20)
No	Yes	31	5.35 (1.41)
Yes	No	7	3.04 (1.65)
Yes	Yes	9	3.86 (1.52)

[a]Higher scores indicate a greater sense of understanding for the less accurate robo-advisor (B) over the more accurate one (A).

favorable inferences about the company might then drive illusory understanding and greater intentions to rely on the algorithm's recommendations. While the prospect of an explanation can indeed lead to inferences about the trustworthiness and transparency of a company, we propose that differences in empowerment foster an illusion of understanding above and beyond these inferences. To this end, in experiment 3, we manipulated the prospect of an explanation while aiming to minimize differences in the perceived trustworthiness and transparency of a company by design. To do so, we present respondents with an algorithm from a well-known and reputable national bank. We reasoned that respondents should have well-established beliefs about this company, and thus, their perception of trustworthiness and transparency should be less affected by whether or not they can access an explanation for the algorithm.

Method

Two hundred Cloud Research-approved participants (M_{age} = 41.4 years, SD = 12.0 years, 46% female) read about a bank's proprietary robo-advisor algorithm that provides investment recommendations. Respondents were randomly assigned to one of two conditions. In the explanation prospect condition, they were informed that a description of how the robo-advisor works had been published in a scientific journal and that a copy of the article was freely available on the website of the journal, such that anyone who had access to the internet could read about the inner workings of the algorithm. In the control condition, participants were informed that a description of how the algorithm works had been published in a scientific journal and that only those subscribed to the journal could read about the algorithm's inner workings. Thus, while the prospect of an explanation differed across conditions, all participants were made aware of the existence of an explanation. To ensure that our manipulation of the prospect of an explanation was effective, at the end of the experiment, participants reported the perceived accessibility of the explanation (To what extent do you feel you could access the article that describes how the [algorithm's name] robo-advisor works, if you wanted to? 1 = I would not be able to access it; 7 = I could easily access it).

Respondents then reported the likelihood of using the bank's robo-advisor if they were looking to invest some money (1 = very unlikely; 7 = very likely). Then, they rated their understanding of how the algorithm works (I understand how the [algorithm's name] robo-advisor works), feeling of empowerment (I feel empowered to: (a) use the [algorithm's name] robo-advisor effectively, (b) deal with the recommendations of the [algorithm's name] robo-advisor, α = 0.95), trust in the company ([bank's name] is a trustworthy company; I can trust [bank's name], α = 0.98) and transparency of the company ([bank's name] is transparent, [bank's name] has nothing to hide, α = 0.93). These questions were measured on the same scale (1 = strongly disagree; 7 = strongly agree) and administered in a randomized order.

Results

We first checked whether we effectively manipulated the prospect of an explanation while keeping the perception of the company's trustworthiness and transparency constant. The results confirmed that participants in the prospect of an explanation condition reported greater perceived access to the explanation (M = 5.89, SD = 1.32) than participants in the control condition (M = 3.50, SD = 1.94; $t(198)$ = 10.19, $p < 0.001$, d = 1.44). Yet, there was no significant difference in the perception of the company's trustworthiness between the explanation prospect condition (M = 4.43, SD = 1.85) and the control condition (M = 4.45, SD = 1.66; $t(198)$ = 0.08, p = 0.94, d = −0.01). Similarly, there was no significant difference in the perception of the company's transparency between the explanation prospect condition (M = 4.12, SD = 1.92) and the control condition (M = 3.96, SD = 1.59; $t(198)$ = 0.64, p = 0.52, d = 0.09).

We then examined how the prospect of an explanation affected respondents' likelihood to use the algorithm, sense of understanding, and feeling of empowerment. Compared to the control condition, in the explanation prospect condition, respondents indicated a greater likelihood to use the algorithm ($M_{control}$ = 3.30, $SD_{control}$ = 1.81 vs. $M_{exp.\ prospect}$ = 4.09, $SD_{exp.\ prospect}$ = 1.65; $t(198)$ = 3.23, p = 0.002, d = 0.46), a greater sense of understanding how the algorithm works ($M_{control}$ = 2.98, $SD_{control}$ = 1.87 vs. $M_{exp.\ prospect}$ = 3.77, $SD_{exp.\ prospect}$ = 1.87; $t(198)$ = 2.99, p = 0.003, d = 0.42), and a greater feeling of empowerment ($M_{control}$ = 3.30, $SD_{control}$ = 1.67 vs. $M_{exp.\ prospect}$ = 4.34, $SD_{exp.\ prospect}$ = 1.71; $t(198)$ = 4.36, $p < 0.001$, d = 0.62). A serial mediation analysis (Hayes, 2017; Model 6) showed a significant indirect effect (explanation prospect → empowerment → understanding → intentions, b = 0.088, 95% CI: 0.026, 0.172), thus providing evidence for the proposed model. Results from the mediation model with trust and transparency as alternative mediators to empowerment are reported in Appendix E.

Discussion

This experiment provides preliminary evidence for our empowerment account of the illusion of understanding how an algorithm works. The increased feeling of empowerment elicited by the prospect of an explanation was associated with a greater sense of understanding, which in turn was associated with greater intentions to use the algorithm. Importantly, the results emerged when we manipulated the prospect of an explanation

while keeping constant the perceived trustworthiness and transparency of the company that provides access to an explanation, thus corroborating the idea that the prospect of an explanation can foster an illusion of understanding through empowerment, above and beyond inferences about the trustworthiness and transparency of a company. A follow-up test also showed that the manipulation used in this experiment did not yield significant differences in trust toward the algorithm (I can trust the [algorithm's name] robo-advisor to do what is good for me, The [algorithm's name] robo-advisor can be trusted to make decisions that are good for me; adapted from Brockner et al., 1997; $\alpha = 0.97$; $M_{control} = 4.09$, $SD_{control} = 1.65$ vs. $M_{exp.\ prospect} = 4.17$, $SD_{exp.\ prospect} = 1.66$; $t(198) = 0.33$, $p = 0.74$, $d = 0.05$).

This experiment also rules out the possibility that the effect stems from inferences about the inherent explainability of an algorithm when the existence of an explanation is not mentioned, as all participants were aware of the existence of an explanation. Finally, it is worth noting that in this experiment, we found an illusion of understanding even if participants were merely informed of the possibility of accessing an explanation without being provided with a link to the explanation. Notably, this finding is consistent with our theorizing that the mere *belief* that one has the possibility to access an explanation can result in a feeling of empowerment. Moreover, this finding squares with research showing that the belief of having control is often sufficient to provide a feeling of empowerment even in the absence of actual means to exercise that control in the moment (Bandura, 1988).

EXPERIMENT 4: EMPOWERMENT AND KNOWLEDGE POINTERS

Experiment 4 aims to test our empowerment account vis-à-vis the account proposed by the community of knowledge, which argues that illusory understanding stems from a cognitive process whereby people confuse mental pointers to potential sources of knowledge with possessing actual knowledge (Sloman et al., 2021, p. 6). To this end, we test a theoretically driven distinction between explanations for experts and explanations for laypeople. Our empowerment account predicts that the prospect of an explanation designed for the general public (i.e., laypeople) should result in a greater illusion of understanding than the prospect of an explanation designed for experts (H3). This is because explanations designed for the general public should be more empowering, as they should foster a stronger belief that by accessing the explanation one could learn how an algorithm works and use that knowledge to use the algorithm more effectively, compared to explanations designed for experts. In contrast, as previously discussed, the cognitive process proposed by the community of knowledge literature would suggest that, to the extent that people know that knowledge is accessible, the prospect of explanations written for laypeople might result in the same or even lower illusory sense of understanding as the prospect of explanations written for experts. We also probe our empowerment account vis-à-vis the account proposed by the community of knowledge via mediation by measuring and testing both the role of empowerment and knowledge pointers in fostering the sense of understanding how an algorithm works. Finally, we further probe the role of inferences about a company's trustworthiness and transparency by measuring and statistically controlling for differences in such perceptions across conditions.

Method

Three hundred Cloud Research-approved participants ($M_{age} = 41.1$ years, $SD = 11.7$ years; 44% female) read about an algorithm called Ai.XR that was developed to provide financial recommendations and were randomly assigned to one of three conditions: control, experts explanation prospect, and laypeople explanation prospect. In the experts explanation prospect condition, participants were informed that an explanation was made available on the website of the company that developed the algorithm and that the explanation was written so that data scientists could understand how the algorithm works. In the laypeople explanation prospect condition, participants were informed that an explanation was made available on the website of the company that developed the algorithm and that the explanation was written so that anybody could understand how the algorithm works. In the control condition, participants were informed that no explanation was made available. To ensure participants processed the text presented, an open-ended question prompted them to type key information presented before proceeding.

Participants reported their understanding of how the algorithm works (1 = do not understand at all; 7 = understand completely) and the likelihood they would rely on the algorithm's recommendations (1 = very unlikely; 7 = very likely). Then, they indicated their agreement (1 = strongly disagree; 7 = strongly agree) on measures presented in a randomized order aimed to assess feelings of empowerment (I feel empowered to learn how: (a) to use the Ai.XR algorithm effectively and (b) deal with the recommendations of the Ai.XR algorithm, $\alpha = 0.96$) trust in the company (the company that developed the Ai.XR algorithm is trustworthy, and I can trust the company that developed the Ai.XR algorithm $\alpha = 0.95$), transparency of the company (the company that developed the Ai.XR algorithm: (a) is transparent and (b) has nothing to hide $\alpha = 0.90$), and knowledge pointers. To the best of our knowledge, no previous work has empirically assessed the role of knowledge pointers on the illusion of understanding. We designed a set of items in line with the community of knowledge central tenet that individuals often

confuse what they know with what others know because they have mental pointers to external sources of knowledge. Such pointers typically involve an awareness of who possesses the knowledge required to explain a particular phenomenon (Keil, 2005; Rabb et al., 2019; Sloman et al., 2021). Based on this logic, we measured knowledge pointers with two items: I can tell who: (a) can understand how the Ai.XR algorithm works, and (b) has knowledge about how the AI.XR algorithm works, $\alpha = 0.95$.

Results

As illustrated in Table 2, ANOVA analyses revealed significant differences across the three conditions in intentions ($F(2,297) = 48.35$, $p < 0.001$), understanding ($F(2,297) = 66.45$, $p < 0.001$), empowerment ($F(2,297) = 69.27$, $p < 0.001$), knowledge pointers ($F(2,297) = 79.99$, $p < 0.001$), trust ($F(2,297) = 41.75$, $p < 0.001$), and transparency ($F(2,297) = 89.24$, $p < 0.001$). In particular, relative to the control condition, the prospect of a laypeople explanation led to greater intentions ($t(297) = 9.83$, $p < 0.001$, $d = 1.43$), greater sense of understanding ($t(297) = 11.17$, $p < 0.001$, $d = 1.65$), greater feeling of empowerment ($t(297) = 11.20$, $p < 0.001$, $d = 1.68$), greater knowledge pointers ($t(297) = 11.06$, $p < 0.001$, $d = 1.57$), greater trust ($t(297) = 9.13$, $p < 0.001$, $d = 1.27$), and greater transparency ($t(297) = 13.34$, $p < 0.001$, $d = 1.98$). The prospect of an experts explanation also led to greater intentions ($t(297) = 4.69$, $p < 0.001$, $d = 0.65$), greater sense of understanding ($t(297) = 3.06$, $p = 0.002$, $d = 0.44$), greater feeling of empowerment ($t(297) = 2.42$, $p = 0.016$, $d = 0.34$), greater knowledge pointers ($t(297) = 10.82$, $p < 0.001$, $d = 1.57$), greater trust ($t(297) = 4.16$, $p < 0.001$, $d = 0.60$), and greater transparency ($t(297) = 6.06$, $p < 0.001$, $d = 0.85$), relative to the control condition.

Moreover, key to our hypothesis, the prospect of a laypeople explanation led to greater intentions ($t(297) = 5.22$, $p < 0.001$, $d = 0.74$), greater sense of understanding ($t(297) = 8.17$, $p < 0.001$, $d = 1.09$) and greater feeling of empowerment ($t(297) = 8.83$, $p < 0.001$, $d = 1.22$), relative to the prospect of an experts explanation. Yet, the two conditions did not significantly differ on knowledge pointers ($t(297) = 0.37$, $p = 0.71$, $d = 0.05$), indicating that our manipulation did not affect participants' awareness of who had knowledge of the algorithm's inner workings. Finally, the two conditions significantly differed in terms of trust ($t(297) = 5.03$, $p < 0.001$, $d = 0.71$) and transparency ($t(297) = 7.38$, $p < 0.001$, $d = 1.00$).

To test the role of empowerment and knowledge pointers in driving the illusion of understanding and its downstream effect on intentions, we tested the model presented in Figure 1 (Hayes, 2017; model 80). Regression coefficients appear in Figure 1, with the control condition as the comparison condition in the tests of group differences (i.e., each coefficient represents the contrast test against the control condition). First, results showed an indirect effect of knowledge pointers via understanding (explanation prospect → knowledge pointers → understanding → intentions) for both the prospect of a laypeople explanation ($b = 0.150$, 95% CI: 0.056, 0.258) and the prospect of an experts explanation ($b = 0.145$, 95% CI: 0.054, 0.253). This result is in line with the theorizing of the community of knowledge, which suggests that knowing who understands how an algorithm works (i.e., a knowledge pointer) gives individuals the sense that they understand the algorithm's inner workings better themselves. Most importantly, in line with our predictions, the results show a significant positive indirect effect of empowerment via understanding (explanation prospect → empowerment → understanding → intentions) for the prospect of a laypeople explanation ($b = 0.330$, 95% CI: 0.185, 0.506) and the prospect of an experts explanation ($b = 0.070$, 95% CI: 0.011, 0.147). As indicated by the non-overlapping confidence intervals (0.185 to 0.506 vs. 0.011 to 0.147), the effect was significantly smaller for the prospect of an experts explanation than for the prospect of a laypeople explanation. This suggests that the greater effect of the prospect of a laypeople explanation over the prospect of an experts explanation was driven by the greater feeling of empowerment, thus supporting H2.

Finally, there was no indirect effect of trust via understanding (explanation prospect → trust → understanding → intentions; laypeople: $b = 0.010$, 95% CI: −0.074, 0.120; experts: $b = 0.004$, 95% CI: −0.034, 0.058) nor of transparency via understanding (explanation prospect → transparency → understanding → intentions; laypeople: $b = 0.067$, 95% CI: −0.060, 0.216; experts: $b = 0.03$, 95% CI: −0.028, 0.098). These findings provide further evidence that differences in the perceived trustworthiness and transparency of the company that provides access to an explanation cannot explain our proposed effect.

Discussion

The findings from this experiment lead to three main considerations. First, the prospect of an explanation of

TABLE 2 Means and standard deviations in Experiment 4.

	Intentions	Understanding	Empowerment	Knowledge pointers	Trust	Transparency
Laypeople	4.39 (1.51)	4.34 (1.74)	5.05 (1.61)	4.78 (1.74)	4.80 (1.46)	5.10 (1.44)
Experts	3.23 (1.64)	2.43 (1.76)	2.93 (1.85)	4.69 (1.63)	3.82 (1.32)	3.60 (1.56)
Control	2.19 (1.57)	1.72 (1.42)	2.35 (1.60)	2.19 (1.56)	3.02 (1.34)	2.38 (1.30)

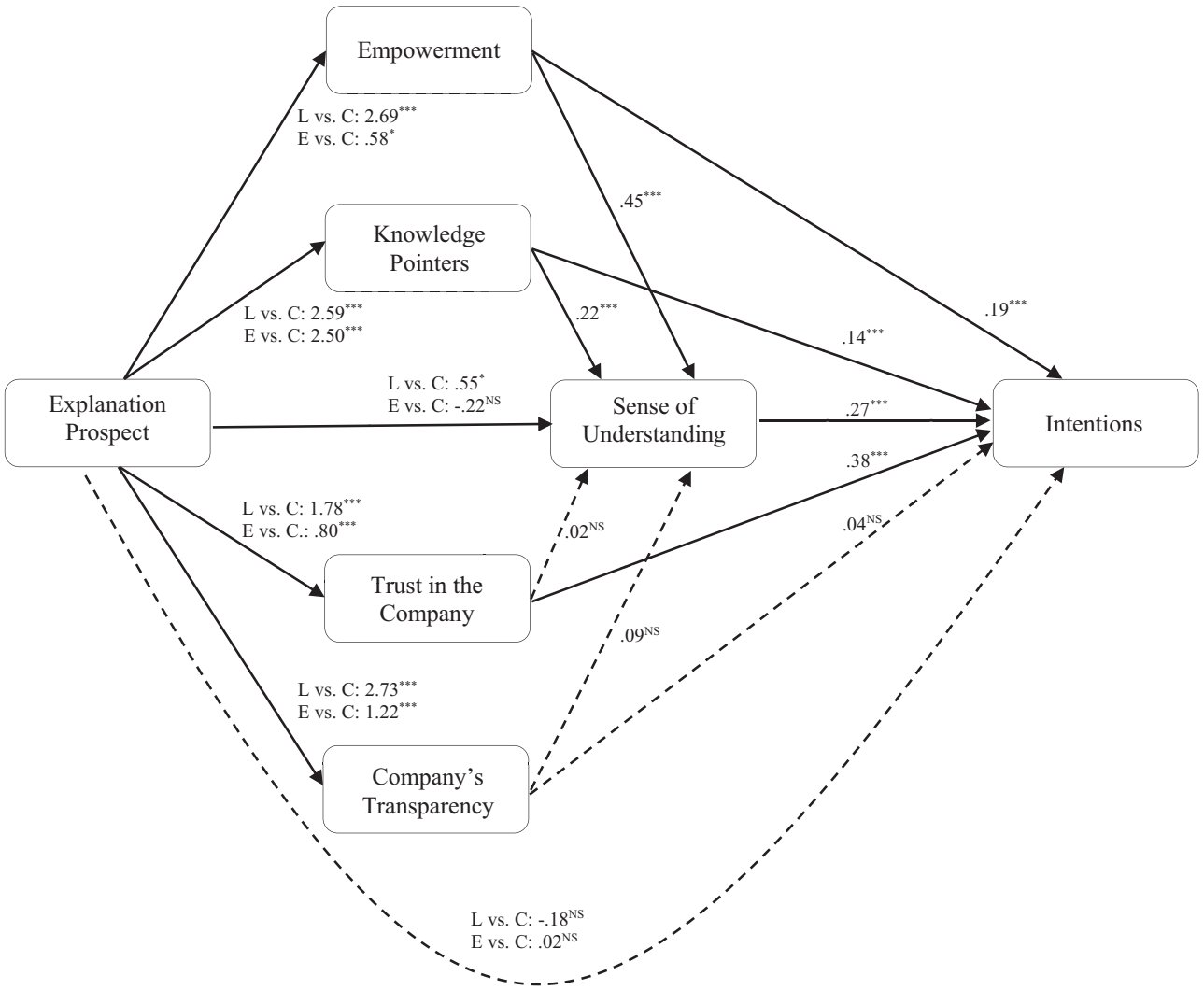

FIGURE 1 Mediation model in Experiment 4. C = control, E = Experts explanation prospect, L = Laypeople explanation prospect; *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

how an algorithm works leads to a greater illusion of understanding and reliance on algorithmic recommendations when the explanation is written for the general public rather than for experts. This finding sets our theorizing apart from past literature and contributes to the ongoing debate about the welfare implications of explanations designed for experts versus laypeople, a point we return to in the General Discussion. Second, we find a mediation path via knowledge pointers, as predicted by the community of knowledge account, but most importantly, we find that our proposed process via empowerment operates above and beyond such a path. Thus, these findings show that the process we propose is qualitatively different from the process proposed by the community of knowledge. Finally, these findings further corroborate the idea that the prospect of an explanation can foster an illusion of understanding through empowerment, above and beyond inferences about the trustworthiness and transparency of a company.

EXPERIMENT 5: MODERATION BY EXPLANATION INSTRUMENTALITY

The objective of Experiment 5 is to further test our proposed empowerment mechanism via moderation-of-process (Spencer et al., 2005) by directly varying the feeling of empowerment orthogonally to the prospect of accessing an explanation. We manipulated empowerment via explanation instrumentality, that is, whether an explanation can enable users to interact more effectively with an algorithm (i.e., higher instrumentality) or not (i.e., lower instrumentality). We reasoned that explanations that do not enable users to interact more effectively

with an algorithm should be less empowering than explanations that enable users to interact more effectively with an algorithm, thus leading to a smaller illusion of understanding.

Method

Four hundred Cloud Research-approved participants (M_{age}=42.9, SD=12.7, 51% female) were informed that the objective of the study was to estimate the demand for a new service called Clothing Affinity and were randomly assigned to one of four conditions in a 2 (explanation prospect: present vs. absent)×2 (explanation instrumentality: baseline vs. low) between-subject design. The prospect of an explanation was manipulated as in previous experiments by informing participants that an explanation was either accessible on the website of the company that developed the algorithm or not. Explanation instrumentality was manipulated as follows. In the low-instrumentality condition, participants were informed that having an explanation for how the algorithm works would not affect their ability to make better decisions about the algorithm's recommendations. Participants in the baseline conditions were not provided with such information. Respondents then reported their understanding of how the algorithm works (1=do not understand at all; 7=understand completely), their intentions to rely on the recommendation of the algorithm (1=very likely; 7=very unlikely), and their feeling of empowerment (I feel empowered to learn how: (a) to use the algorithm effectively, (b) to deal with the recommendations of the algorithm; 1=not at all; 7=completely; α=0.92).

Results

An ANOVA analysis revealed an interaction on empowerment (F(1,396)=17.85, p<0.001). Means and standard deviations are reported in Table 3. In the baseline conditions, explanation prospect led to greater empowerment (t(396)=8.93, p<0.001, d=1.29). However, this effect was reduced when participants were informed that the explanation was not instrumental to making better decisions about the algorithm's recommendations (t(396)=2.90, p=0.004, d=0.41), thus suggesting that our manipulation was effective. Most importantly, there was a significant interaction on sense of understanding (F(1,396)=4.63, p=0.032) and intentions (F(1,396)=12.74, p<0.001). In the baseline conditions, the prospect of explanation had a greater effect on understanding (t(396)=7.35, p<0.001, d=1.08) and intentions (t(396)=5.53, p<0.001, d=0.78) than in the low-explanation instrumentality conditions (understanding: t(396)=4.25, p<0.001, d=0.58; intentions: t(396)=0.45, p=0.65, d=0.06).

In line with our hypothesis, a mediation analysis (Hayes, 2017; model 83) showed that the conditional indirect effect (explanation prospect → empowerment → understanding → intentions) in the baseline conditions (b=0.220; 95% CI=0.088; 0.372) was significantly greater than in the low-instrumentality condition (b=0.072, 95% CI: 0.017; 0.150), as indicated by the index of moderated mediation (b=0.148, 95% CI: 0.050; 0.280). Regression coefficients are reported in Figure 2.

Discussion

Experiment 5 provides moderation-of-process evidence that the feeling of empowerment triggered by the prospect of an explanation drives illusory understanding. This experiment shows that reducing the perception that an explanation is instrumental in enhancing user-algorithm interactions and, in particular, in enabling people to make more informed decisions about the recommendation of an algorithm reduces the feeling of empowerment and, as such, the illusion of understanding. Thus, this experiment provides further evidence for the role of empowerment as an antecedent of the illusion of understanding triggered by the prospect of an explanation.

GENERAL DISCUSSION

Five experiments show that the mere prospect of an explanation of how an algorithm works can foster an illusory sense of understanding the algorithm, even if people do not access and read the explanation. This effect emerged across different domains where algorithms are widely used and different operationalizations of the prospect to access an explanation. Our results suggest that this effect occurs because the mere prospect of an explanation provides a feeling of empowerment that fosters an illusory sense of understanding. In turn, this

TABLE 3 Means and standard deviations in Experiment 5.

Explanation Prospect	Explanation instrumentality	Intentions	Understanding	Empowerment
No	Baseline	2.91 (1.59)	2.33 (1.78)	2.73 (1.74)
Yes	Baseline	4.13 (1.55)	4.18 (1.64)	4.91 (1.65)
No	Low	2.97 (1.61)	2.48 (1.83)	2.94 (1.83)
Yes	Low	3.07 (1.49)	3.57 (1.91)	3.66 (1.69)

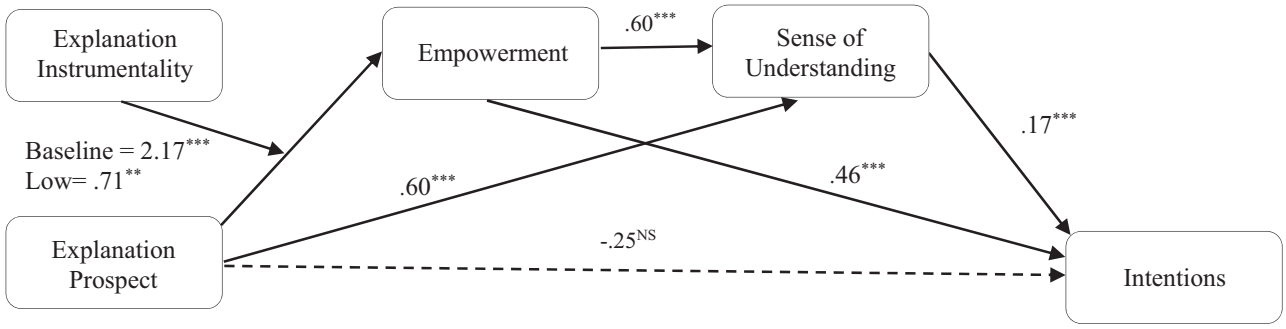

FIGURE 2 Moderated mediation via explanation instrumentality in Experiment 5. *$p<0.05$, **$p<0.01$, ***$p<0.001$.

illusory sense of understanding can yield unfounded reliance on algorithmic determinations.

Theoretical contribution

Our work contributes to prior research in several ways. First, it offers a novel conceptual lens to understand the psychology of explanations. Past research has mainly looked at explanations as means to convey knowledge (Dhaliwal & Benbasat, 1996; Johnson & Johnson, 1993; Lombrozo, 2006). As such, past research has been primarily interested in investigating how different types of explanations might be more or less effective at fostering understanding by conveying knowledge (Buchanan & Shortliffe, 1984; Ye & Johnson, 1995). We contribute to this stream of research by offering a novel account of how explanations influence understanding. We show that merely knowing one can access an explanation can foster an illusory sense of understanding by providing a feeling of empowerment, above and beyond conveying actual knowledge about how something works.

Our work also contributes to the emerging stream of research on the community of knowledge, which asserts that people often think that they understand a phenomenon simply by virtue of knowing that expert sources understand it (Kominsky & Keil, 2014; Rabb et al., 2019; Sloman & Rabb, 2016; Ward, 2021). The community of knowledge suggests that people rely on a transactive memory system in which knowledge is stored not only in one's memory but also in external repositories that one can access, such as experts' minds, books, or technological devices (Rabb et al., 2019; Ward, 2021). People often confuse having mental pointers to those external sources of knowledge with actually possessing the knowledge those pointers index (Rabb et al., 2019), and, as a result, they experience an illusion of understanding. We contribute to this literature by documenting an additional process based on empowerment that can foster illusory understanding. Our empowerment account is especially relevant in the domain of algorithmic decision-making, where explanations are expected to enable people to make more informed decisions about an algorithm's recommendations, and for this reason, the prospect of accessing such explanations can foster a feeling of empowerment.

Finally, our work contributes to the broader literature on consumers' receptivity to algorithms (Castelo, 2023; Castelo et al., 2019; Dietvorst et al., 2015; Granulo et al., 2021; Longoni et al., 2019; Reich et al., 2022). Prior research suggests that the black-box nature of algorithms can make people reluctant to rely on algorithmic determinations (Bonezzi et al., 2022). Several scholars have discussed ways to open the black box and provide knowledge that helps consumers understand how algorithms work, thus increasing their receptivity to algorithmic determinations (Narayanan et al., 2018; Swartout, 1986; Ye & Johnson, 1995). This important topic has recently sparked some particularly interesting research in the domain of consumer research (Tomaino et al., 2020). We add to this stream of research by showing that the mere potential to access an explanation of how an algorithm works can foster unfounded confidence in algorithmic determinations, even when consumers do not read an explanation.

Practical implications

Our work has practical implications for policymakers tasked with creating and enforcing algorithmic transparency policies. Existing regulations, such as the European Union's GDPR (e.g., Goodman & Flaxman, 2017), require companies to inform consumers of the accessibility of an explanation of how an algorithm works. Consumers, however, can acknowledge the existence of the explanation and proceed to use a company's services without actually reading the explanation. Our findings show that this practice may have unintended consequences, as it may foster an illusory sense of understanding that can yield unfounded confidence in algorithmic determinations. This can have important implications, especially in light of recent evidence showing that algorithms can be biased (O'neil, 2016; Schwemmer et al., 2020) and even yield discriminatory determinations. If making explanations more accessible lulls consumers into blindly trusting algorithmic determinations, then policies aimed at

making explanations more accessible might have the unintended consequence of increasing consumers' acceptance of determinations that might, in fact, be unfair or discriminatory, with the risk of legitimizing discriminatory practices (Bonezzi & Ostinelli, 2021).

These adverse effects may be mitigated in contexts where companies and public agencies are interested in implementing solutions aimed to help consumers overcome their resistance to algorithms that can, in fact, enhance their well-being. Prominent examples include sectors such as financial recommendations and medical diagnoses. Algorithms can be affordable and effective tools to optimize financial planning based on individual goals and risk profiles. Similarly, algorithms can deliver affordable medical care at scale. Consumers' reluctance to embrace these algorithms can potentially result in missed financial and healthcare opportunities. In such contexts, the prospect of an explanation may increase consumers' propensity to use and benefit from algorithms.

Furthermore, our work contributes to the ongoing debate on how to implement transparency initiatives. The finding that the prospect of laypeople explanations fosters a greater illusion of understanding than the prospect of expert explanations is important in light of the current debates on how to implement algorithmic transparency. For example, the EU law on algorithmic transparency requires that explanations about the inner workings of algorithms must be comprehended by lay consumers (Grochowski et al., 2021). Our work suggests that decisions to prioritize explanations designed to be understood by consumers might need to consider the potentially detrimental effect that might ensue when consumers do not read these explanations yet develop an illusory sense of understanding that makes them more likely to rely on algorithms they actually do not understand.

Future research directions

The finding that the mere prospect of an explanation for how an algorithm works can increase reliance on algorithmic determinations opens opportunities for future research both from a theoretical and from a practical standpoint. First, there may be mechanisms other than the one we documented through which the mere prospect of an explanation can influence consumers' reliance on algorithmic determinations. For example, the prospect of an explanation could act like a signal that experts can scrutinize the algorithm, thus assuring consumers that the algorithm should work as intended. Such a signal might increase reliance on the algorithm, even when people do not access the explanation, without necessarily fostering empowerment and illusory understanding. This alternative mechanism might be particularly relevant when consumers have limited cognitive resources and are thus more likely to be influenced by peripheral cues (Petty et al., 1983). A comprehensive investigation of the mechanisms through which the mere prospect of an explanation can influence consumers' judgments and decisions, along with the identification of additional boundary conditions of the phenomenon, awaits further research.

From a practical standpoint, our work offers opportunities for future research to speak to current public policy debates. We focus our discussion on two directions that seem especially promising: (a) exploring the effect of the prospect of explanations with respect to information disclosure and (b) exploring the effects of the prospect of different types of explanations. First, future research could examine whether the effect we investigated generalizes to another domain that has important consumer welfare implications: information disclosure (Melzner et al., 2023). Algorithms often require consumer data to personalize one's browsing experience or to provide personalized services and recommendations. In many cases, consumers must grant permission to collect such data. These permission requests can be accompanied by an option to access an explanation of how algorithms function, specifically, what type of information they collect and for what purpose. Our theorizing suggests that the prospect of accessing an explanation of how an algorithm uses personal information might lead to an illusory sense of understanding that can, in turn, make consumers more likely to blindly authorize the collection of personal data. We explored this possibility in an experiment (Experiment 6, Appendix F) that provides preliminary evidence for this hypothesis.

Second, future research might explore the effect of the prospect of different types of explanations. In this work, we considered explanations that can be accessed before one uses an algorithm. These are referred to as *ex-ante* explanations (Wachter et al., 2017). Yet, in other cases, explanations are made available, often upon request, only after an algorithm has made a determination. These are referred to as *ex-post* explanations. The distinction between ex-ante and ex-post explanations has fueled a debate about the benefits provided by each type of explanation (Hamon et al., 2021; Selbst & Powles, 2017; Wachter et al., 2017). Our theorizing could add to this debate by suggesting that, while consumers might have access to an explanation in both cases, the prospect of accessing an ex-ante explanation could lead to a greater illusory sense of understanding than the prospect of accessing an ex-post explanation. This is because limiting access to an explanation to after an algorithm has made a determination constrains people's sense of control and, thus, their feeling of empowerment. According to our theorizing, this lower feeling of empowerment should translate into a lower illusion of understanding. We explored this possibility in an experiment (Experiment 7, Appendix G) that provides preliminary evidence for this hypothesis.

Overall, these preliminary findings highlight the need to systematically explore how to implement algorithmic transparency in a way that limits unintended consequences.

DATA AVAILABILITY STATEMENT

Survey instruments and datasets are available at osf.io/g3k6d.

ORCID

Massimiliano Ostinelli https://orcid.org/0000-0002-9515-9131

REFERENCES

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, *347*(6221), 509–514.

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology*, *99*(3), 436–451.

Bandura, A. (1988). Self-efficacy conception of anxiety. *Anxiety Research*, *1*(2), 77–98.

Bonezzi, A., & Ostinelli, M. (2021). Can algorithms legitimize discrimination? *Journal of Experimental Psychology: Applied*, *27*(2), 447–459.

Bonezzi, A., Ostinelli, M., & Melzner, J. (2022). The human black-box: The illusion of understanding human better than algorithmic decision-making. *Journal of Experimental Psychology: General*, *151*, 2250–2258.

Brinol, P., Petty, R. E., Valle, C., Rucker, D. D., & Becerra, A. (2007). The effects of message recipients' power before and after persuasion: A self-validation analysis. *Journal of Personality and Social Psychology*, *93*(6), 1040–1053.

Brockner, J., Siegel, P. A., Daly, J. P., Tyler, T., & Martin, C. (1997). When trust matters: The moderating effect of outcome favorability. *Administrative Science Quarterly*, *42*, 558–583.

Buchanan, B. G., & Shortliffe, E. H. (1984). *Rule-based expert systems: The MYCIN experiments of the Stanford heuristic programming project*. Addison-Wesley Longman Publishing.

Carlson, J. P., Vincent, L. H., Hardesty, D. M., & Bearden, W. O. (2009). Objective and subjective knowledge relationships: A quantitative analysis of consumer research findings. *Journal of Consumer Research*, *35*(5), 864–876.

Castelo, N. (2023). Perceived corruption reduces algorithm aversion. *Journal of Consumer Psychology*. https://doi.org/10.1002/jcpy.1373

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

Conger, J. A., & Kanungo, R. N. (1988). The empowerment process: Integrating theory and practice. *Academy of Management Review*, *13*(3), 471–482.

Cotter, K., Cho, J., & Rader, E. (2017). *Explaining the news feed algorithm: An analysis of the" News Feed FYI" blog*. Paper presented at the Proceedings of The 2017 Chi Conference Extended Abstracts on Human Factors in Computing Systems.

Courchane, M., Gailey, A., & Zorn, P. (2008). Consumer credit literacy: What Price perception? *Journal of Economics and Business*, *60*(1–2), 125–138.

Craik, K. J. W. (1967). *The nature of explanation* (Vol. *445*). CUP Archive.

Demirdag, I., & Shu, S. B. (2021). *Insights into the black box: Input Explainability drives consumer satisfaction in the digital world*. ULCA Working Paper.

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

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

Edwards, J. A., & Weary, G. (1998). Antecedents of causal uncertainty and perceived control: A prospective study. *European Journal of Personality*, *12*(2), 135–148.

Fernbach, P. M., Light, N., Scott, S. E., Inbar, Y., & Rozin, P. (2019). Extreme opponents of genetically modified foods know the least but think they know the most. *Nature Human Behaviour*, *3*(3), 251–256.

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*, *24*(6), 939–946.

Fisher, M., Goddu, M., & Keil, F. (2015). Searching for explanations: How the internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General*, *144*(3), 674–687.

Galinsky, A. D., Gruenfeld, D. H., & Magee, J. C. (2003). From power to action. *Journal of Personality and Social Psychology*, *85*(3), 453–466.

Gedikli, F., Jannach, D., & Ge, M. (2014). How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, *72*(4), 367–382.

Gibson, C. H. (1991). A concept analysis of empowerment. *Journal of Advanced Nursing*, *16*(3), 354–361.

Gilovich, T., & Douglas, C. (1986). Biased evaluations of randomly determined gambling outcomes. *Journal of Experimental Social Psychology*, *22*(3), 228–241.

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

Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines*, *8*(1), 101–118.

Granulo, A., Fuchs, C., & Puntoni, S. (2021). Preference for human (vs. robotic) labor is stronger in symbolic consumption contexts. *Journal of Consumer Psychology*, *31*(1), 72–80.

Grochowski, M., Jabłonowska, A., Lagioia, F., & Sartor, G. (2021). Algorithmic transparency and explainability for EU consumer protection: Unwrapping the regulatory premises. *Critical Analysis of Law*, *8*(1), 43–63.

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, *40*(2), 44–58.

Hadar, L., Sood, S., & Fox, C. R. (2013). Subjective knowledge in consumer financial decisions. *Journal of Marketing Research*, *50*(3), 303–316.

Hamon, R., Junklewitz, H., Malgieri, G., Hert, P. D., Beslay, L., & Sanchez, I. (2021). *Impossible Explanations? Beyond explainable AI in the GDPR from a COVID-19 use case scenario*. Paper presented at the Proceedings of the 2021 ACM conference on fairness, accountability, and transparency.

Han, J., & Broniarczyk, S. (2022). The complexities of consumer empowerment in the modern consumption environment. *Current Opinion in Psychology*, *46*, 101333.

Hart, K. (2019). Privacy policies are read by an aging few. https://www.axios.com/few-people-read-privacy-policies-survey-fec3a29e-2e3a-4767-a05c-2cacdcbaecc8.html

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Publications.

Hemmatian, B., & Sloman, S. A. (2018). Community appeal: Explanation without information. *Journal of Experimental Psychology: General*, *147*(11), 1677–1712.

Herm, L.-V., Heinrich, K., Wanner, J., & Janiesch, C. (2023). Stop ordering machine learning algorithms by their explainability! A

user-centered investigation of performance and explainability. *International Journal of Information Management*, *69*, 102538.
Johnson, H., & Johnson, P. (1993). *Explanation facilities and interactive systems*. Paper presented at the Proceedings of the 1st International Conference on Intelligent User Interfaces.
Kaminski, M. E., & Malgieri, G. (2021). Algorithmic impact assessments under the GDPR: Producing multi-layered explanations. *International Data Privacy Law*, *11*, 125–144.
Keil, F. (2005). Doubt, deference, and deliberation: Understanding and using the division of cognitive labor. *Oxford Studies in Epistemology*, *1*, 143–166.
Keil, F. (2011). The problem of partial understanding. *Current Trends in LSP Research: Aims and Methods Series: Linguistic Insights*, *44*, 251–276.
Kominsky, J., & Keil, F. (2014). Overestimation of knowledge about word meanings: The "misplaced meaning" effect. *Cognitive Science*, *38*(8), 1604–1633.
Kozinets, R. V., Ferreira, D. A., & Chimenti, P. (2021). How do platforms empower consumers? Insights from the affordances and constraints of Reclame Aqui. *Journal of Consumer Research*, *48*(3), 428–455.
Labrecque, L. I., Vor Dem Esche, J., Mathwick, C., Novak, T. P., & Hofacker, C. F. (2013). Consumer power: Evolution in the digital age. *Journal of Interactive Marketing*, *27*(4), 257–269.
Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, *32*(2), 311–328.
Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470.
Long, A. R., Fernbach, P. M., & De Langhe, B. (2018). Circle of incompetence: Sense of understanding as an improper guide to investment risk. *Journal of Marketing Research*, *55*(4), 474–488.
Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650.
Maheswaran, D., & Chaiken, S. (1991). Promoting systematic processing in low-motivation settings: Effect of incongruent information on processing and judgment. *Journal of Personality and Social Psychology*, *61*(1), 13–25.
Melzner, J., Bonezzi, A., & Meyvis, T. (2023). Information disclosure in the era of voice technology. *Journal of Marketing*, *87*(4), 491–509.
Mende, M., Scott, M. L., van Doorn, J., Grewal, D., & Shanks, I. (2019). Service robots rising: How humanoid robots influence service experiences and elicit compensatory consumer responses. *Journal of Marketing Research*, *56*(4), 535–556.
Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., & Doshi-Velez, F. (2018). How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*.
Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus*, *140*(4), 32–48.
Obar, J. A., & Oeldorf-Hirsch, A. (2020). The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, *23*(1), 128–147.
Oduor, K. F., & Wiebe, E. N. (2008). *The effects of automated decision algorithm modality and transparency on reported trust and task performance*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
O'neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
Petty, R. E., Cacioppo, J. T., & Schumann, D. (1983). Central and peripheral routes to advertising effectiveness: The moderating role of involvement. *Journal of Consumer Research*, *10*(2), 135–146.
Presson, P. K., & Benassi, V. A. (1996). Illusion of control: A meta-analytic review. *Journal of Social Behavior and Personality*, *11*(3), 493.
Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and artificial intelligence: An experiential perspective. *Journal of Marketing*, *85*(1), 131–151.
Rabb, N., Fernbach, P. M., & Sloman, S. A. (2019). Individual representation in a community of knowledge. *Trends in Cognitive Sciences*, *23*(10), 891–902.
Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., … Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, *1*(1), 18.
Reich, T., Kaju, A., & Maglio, S. J. (2022). How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology*, *33*(2), 285–302.
Ribera, M., & Lapedriza García, À. (2019). *Can we do better explanations? A proposal of user-centered explainable AI*. Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA.
Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562.
Rudolph, M., Feth, D., & Polst, S. (2018). *Why users ignore privacy policies—A survey and intention model for explaining user privacy behavior*. International Conference on Human-Computer Interaction, pp. 587–598.
Schmitt, B. (2019). From atoms to bits and back: A research curation on digital technology and agenda for future research. *Journal of Consumer Research*, *46*(4), 825–832.
Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M., & Lockhart, J. W. (2020). Diagnosing gender bias in image recognition systems. *Socius*, *6*, 2378023120967171.
See, K. E., Morrison, E. W., Rothman, N. B., & Soll, J. B. (2011). The detrimental effects of power on confidence, advice taking, and accuracy. *Organizational Behavior and Human Decision Processes*, *116*(2), 272–285.
Selbst, A., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, *7*(4), 233–242.
Sinha, R., & Swearingen, K. (2002). *The role of transparency in recommender systems*. CHI'02 extended abstracts on Human factors in computing systems (pp. 830 831).
Sloman, S. A., Patterson, R., & Barbey, A. K. (2021). Cognitive neuroscience meets the community of knowledge. *Frontiers in Systems Neuroscience*, *15*, 120.
Sloman, S. A., & Rabb, N. (2016). Your understanding is my understanding: Evidence for a community of knowledge. *Psychological Science*, *27*(11), 1451–1460.
Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, *89*(6), 845–851.
Swartout, W. R. (1986). Knowledge needed for expert system explanation. *Future Computing Systems*, *1*(2), 99–114.
Tan, G. K. S. (2020). Robo-advisors and the financialization of lay investors. *Geoforum*, *117*, 46–60.
Tomaino, G., Abdulhalim, H., Kireyev, P., & Wertenbroch, K. (2020). Denied by an (Unexplainable) Algorithm: Teleological Explanations for Algorithmic Decisions Enhance Customer Satisfaction. *INSEAD Working Paper No. 2022/07/MKT*.
Trout, J. D. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science*, *69*(2), 212–233.
Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76–99.
Wan, E. W., & Rucker, D. D. (2013). Confidence and construal framing: When confidence increases versus decreases information processing. *Journal of Consumer Research*, *39*(5), 977–992.

Ward, A. F. (2021). People mistake the internet's knowledge for their own. *Proceedings of the National Academy of Sciences*, *118*(43), e2105061118.

Watson, H. J., & Nations, C. (2019). Addressing the growing need for algorithmic transparency. *Communications of the Association for Information Systems*, *45*(1), 26–510.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., & Cacioppo, J. T. (2010). Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, *99*(3), 410–435.

Weary, G., & Edwards, J. A. (1994). Individual differences in causal uncertainty. *Journal of Personality and Social Psychology*, *67*(2), 308–318.

Weary, G., & Edwards, J. A. (1996). Causal-uncertainty beliefs and related goal structures. In M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition* (pp. 148–181). The Guilford Press.

Ye, L. R., & Johnson, P. E. (1995). The impact of explanation facilities on user acceptance of expert systems advice. *MIS Quarterly*, *19*(2), 157–172.

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

Ylikoski, P. (2019). The illusion of depth of understanding in science. In H. de Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific understanding: Philosophical perspectives* (pp. 100–119). Pittsburgh University Press.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ostinelli, M., Bonezzi, A., & Lisjak, M. (2025). Unintended effects of algorithmic transparency: The mere prospect of an explanation can foster the illusion of understanding how an algorithm works. *Journal of Consumer Psychology*, *35*, 203–219. https://doi.org/10.1002/jcpy.1416