Journal of Business Research 186 (2025) 114981

Contents lists available at ScienceDirect

Journal of Business Research

journal homepage: www.elsevier.com/locate/jbusres

Post hoc explanations improve consumer responses to algorithmic decisions

Mehdi Mourali [a,*], Dallas Novakowski [a], Ruth Pogacar [a], Neil Brigden [b]

[a] University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada
[b] Mount Royal University, 4825 Mount Royal Gate SW, Calgary, Alberta, Canada T3E 6K6

ARTICLE INFO

Keywords:
Automated Decisions
Algorithms
Explanations
Transparency
Fairness
Trust

ABSTRACT

Algorithms are capable of assisting with, or making, critical decisions in many areas of consumers' lives. Algorithms have consistently outperformed human decision-makers in multiple domains, and the list of cases where algorithms can make superior decisions will only grow as the technology evolves. Nevertheless, many people distrust algorithmic decisions. One concern is their lack of transparency. For instance, it is often unclear how a machine learning algorithm produces a given prediction. To address the problem, organizations have started providing post-hoc explanations of the logic behind their algorithmic decisions. However, it remains unclear to what extent explanations can improve consumer attitudes and intentions. Five experiments demonstrate that algorithmic explanations can improve perceptions of transparency, attitudes, and behavioral intentions – or they can backfire, depending on the explanation method used. The most effective explanations highlight concrete and feasible steps consumers can take to positively influence their future decision outcomes.

1. Introduction

The rapid proliferation of algorithmic decision-making in domains ranging from healthcare to finance has sparked both enthusiasm and unease among consumers. Whereas algorithms have long been valued for their superior predictive accuracy compared to human decision-makers (Dawes, 1979; Dawes, Faust, and Meehl 1989; Grove et al., 2000; Meehl, 1954) consumers often exhibit algorithm aversion, expressing distrust and negative attitudes toward algorithmic decision-making (Castelo, Bos, & Lehman, 2019; Dietvorst, Simmons, & Massey, 2015, 2018; Longoni, Bonezzi, & Morewedge, 2019; Yalcin et al., 2022).

One major concern is the algorithms' lack of transparency (Burrell, 2016; Lepri et al., 2018). For instance, machine learning algorithms are often referred to as "black boxes" because one can only see their output, not the underlying mechanisms. Such opacity poses a serious challenge to perceptions of fairness (Lepri et al., 2018; Pasquale, 2015; Selbst & Barocas, 2018), and may contribute to the erosion of trust in algorithmic decision-making (Kizilicec, 2016; Shin & Park, 2019). This has led to extensive calls for increased transparency in algorithmic decision systems (Lepri et al., 2018; Pasquale, 2015). For example, the General Data Protection Regulation GDPR in the European Union requires organizations to "provide meaningful information about the logic involved" in decisions made by their algorithms (Wachter, Mittelstadt, & Floridi, 2017). Organizations may wish to avoid providing explanations of algorithmic decisions and the associated risk of disclosing proprietary information. However, explanations of how decisions were made also hold the potential to enhance consumer trust and acceptance of algorithms and the institutions deploying them. The present research addresses the questions of whether and how providing a post-hoc explanation influences consumer perceptions of algorithmic decisions and of the organizations deploying them. It makes three key contributions to the literature. First, it demonstrates that providing post-hoc explanations of algorithmic decisions that give consumers a sense of control over their outcomes can reduce negative attitudes toward algorithmic decisions and the organizations deploying them. This is significant because companies can easily modify the design and delivery of such explanations. Second, it tests theoretically grounded predictions about when and how different types of explanations are likely to improve consumer responses and when they are likely to backfire. Third, it extends the extant literature on consumer responses to algorithms by focusing on how people react when they are the subjects of algorithmic decisions, with no option to reject an algorithm-determined outcome. This contrasts with the literature's emphasis on contexts in which consumers are merely users of algorithmic recommendations (e.g., Castelo, et al., 2019; Dietvorst et al., 2015, 2018; Leung, Paolacci, & Puntoni,

* Corresponding author.
E-mail addresses: mehdi.mourali@haskayne.ucalgary.ca (M. Mourali), dallas.novakowski1@ucalgary.ca (D. Novakowski), ruth.pogacar@haskayne.ucalgary.ca (R. Pogacar), nbrigden@mtroyal.ca (N. Brigden).

https://doi.org/10.1016/j.jbusres.2024.114981
Received 3 October 2023; Received in revised form 29 August 2024; Accepted 20 September 2024
Available online 25 September 2024

2019; Logg et al., 2020; Longoni et al., 2019; Longoni & Cian, 2022). Given the potentially high-stakes and consequential nature of many algorithmic decisions, understanding how to effectively communicate these decisions is both theoretically and substantively important.

2. Post hoc explanations and consumer response to algorithmic decisions

The rapidly expanding field of explainable AI (XAI) is concerned with developing interpretable decision-making models. One popular method is through a post-hoc explanation, which does not seek to provide a complete explanation of how a model works. Instead, it seeks to justify decisions made by the model by explaining how the model reached its conclusion (Lepri, Staiano, Sangokoya, Letouzé, & Oliver, 2017; Lipton, 2018; Montavon, Lapuschkin, Binder, Samek, & Müller, 2017). A notable advantage of post-hoc explanations is that they offer the ability to explain decisions from complex and opaque models in a way that is comprehensible to consumers, without having to make changes to the model that sacrifice predictive performance (Lipton, 2018). Importantly, regardless of how a model functions internally, the content and method of the post-hoc explanation may influence consumers' perceptions of transparency.

XAI research is focused primarily on addressing the computational challenges of making algorithms more interpretable. However, to create interpretable descriptions, one must consider the psychological underpinnings of how people process and respond to explanations (Miller, 2019; Miller, Howe, & Sonenberg, 2017; Mittelstadt, Russell, & Wachter, 2019). The psychology literature suggests that human-to-human explanations serve several functions. In addition to transferring knowledge, they play an important role in persuasion, learning, and assignment of blame (Lombrozo, 2006, 2007). This suggests that providing a convincing explanation of an algorithmic decision can not only increase transparency and understanding but can also enhance consumers' evaluations of the decision process and the organization behind it.

Binns et al. (2018) reviewed the technical literature on interpretable machine learning and derived four types of model-agnostic (i.e., applicable to any kind of ML model) explanations that are both computationally feasible and likely to meet the legal requirements of current and future regulations on transparency of algorithmic decisions (e.g. GDPR).

1. **Sensitivity**: Shows how much the value of an input variable used in the decision would have to differ in order to change the output class.
2. **Case**: Shows a case from the model's training data which is most similar to the decision being explained.
3. **Demographic**: Shows aggregate statistics on the outcome classes for people in the same demographic categories as the decision-subject, such as age, gender, income level or occupation.

3. Input Influence: Shows a list of input variables and a quantitative measure of their influence in the decision model

Input influence and demographic explanations are considered "global" explanations, whereas sensitivity and case-based explanations are considered "local" explanations. Global explanations describe how the algorithm makes decisions in general, but may not be specific enough to convey why a particular decision was made for a specific instance. In contrast, local explanations offer justifications for why a specific decision was made for a particular instance, but may not be representative of the decision-making process for other instances (Dodge et al., 2019; Wick & Thompson, 1992). Global explanations provide a broader understanding of the algorithm's decision-making process, and are particularly relevant for programmers and decision-makers, such as managers, who seek to understand the overall performance of the model or identify patterns, trends, or biases that might be present in the data. However, local explanations are more specific and targeted, and tend to be sought by end-users desiring to understand and validate the decision that was made about them (Dodge et al. 2019; Mittlestadt et al., 2019). While consumers may be indifferent to how an algorithm functions in a global sense, consumers are motivated to understand how decisions that affect them personally came to be. Because this research is focused on consumers' perceptions of algorithmic decisions made about them, our investigation focuses on the impact of local (sensitivity-based and case-based) explanation types.

Sensitivity-based explanations show how much the values of specific variables used in a decision would have to differ to change the decision outcome (e.g., "If 10 % or less of your driving took place at night, you would have qualified for the cheapest insurance tier"). Thus, sensitivity-based explanations have the potential to offer actionable information that consumers could use to change their outcomes. In contrast, case-based explanations present a case from the model's training data that is most similar to the decision being explained (e.g., "… a similar case to yours is a previous customer: She was 38 years old, with 18 years of driving experience… she was involved in one accident in the following year"; Binns et al., 2018). Both explanations provide information designed to improve transparency (i.e., understanding of how a decision was reached). However, there is reason to believe that sensitivity-based explanations would lead to more positive evaluations of algorithmic decisions and the organizations deploying them.

Research suggests that human explanations are fundamentally social and conversational in nature (Hilton, 1990; Miller 2019). As such, they ought to adhere to Grice's (1975) maxims of quality (say only what you believe is true), quantity (say only what is necessary), relation (say only what is relevant), and manner (say it nicely). People also tend to seek counterfactual explanations. They ask not why something happened, but why it happened instead of something else (Miller, 2019). Moreover, people do not expect a full detailing of all the causes of an event, only the significant ones (Miller, 2019; Zerilli et al., 2019). Indeed, one of the most fascinating findings in the psychological literature on explanations is that when people evaluate the quality of an explanation, the probability that it is actually true is less important than other pragmatic considerations such as the explanation's perceived usefulness, relevance, and simplicity (Hilton, 1996; Lombrozo, 2007; Slugoski et al., 1993). This suggests that the impact of explanations on consumer judgments will likely depend on how useful, relevant, and simple the explanation is perceived to be.

By highlighting concrete steps that consumers can take to improve their outcomes, sensitivity explanations offer relevant and useful information. In contrast, case-based explanations merely provide a list of attributes of a third party and state this person's outcome. This type of framing does little to explicitly communicate the relative impact of the listed attributes (e.g., their importance, how much they need to change), and is thus less useful to consumers seeking to improve their outcomes. Moreover, the thresholds in a sensitivity explanation provide a window into a counterfactual case: "If you did X differently, you would have gotten Y instead of Z," providing diagnostic information about the influence of the input variable.

In sum, sensitivity explanations possess the hallmarks of effective communication. They are simple, useful, and make clear comparisons to counterfactual cases. Therefore, we predict that sensitivity explanations will enhance evaluations of algorithmic decisions and the organizations deploying them. Conversely, case-based explanations are not only less useful and comparative, but they also exhibit uniqueness neglect (Longoni et al., 2019). Telling a consumer that their loan application has been rejected because they are similar to another consumer who failed to repay their loan on time may raise concerns that the algorithm has not taken the applicant's unique characteristics and circumstances into account. Such uniqueness neglect has been found to increase resistance to AI (Longoni et al., 2019). Hence, we predict that providing a case-based explanation may be equivalent to no explanation at all. Formally:

H1: Providing a sensitivity-based explanation (compared to no explanation) increases consumers' perceptions of transparency and positive attitudes (i.e., fairness, trust) towards the algorithmic decision system, as well as increasing consumers' positive intentions toward the company.

H2: Providing a case-based explanation (compared to no-explanation) does not increase consumers' perceptions of transparency and positive attitudes (i.e., fairness, trust) towards the algorithmic decision system, and does not increase consumers' positive intentions toward the company.

We reasoned that sensitivity explanations elicit more positive attitudes and behavioral intentions because they tend to provide useful feedback. They offer actionable insight into how future outcomes might be improved after the applicant is denied the best outcome (Binns et al., 2018). The effectiveness of sensitivity explanations is thus contingent on its perceived usefulness in improving future outcomes. An actionable sensitivity explanation is an explanation that highlights factors under the consumer's direct control. Because it offers useful feedback on how to improve one's outcomes, such an explanation could enhance perceptions of the algorithmic decision system. In contrast, a non-actionable sensitivity explanation highlights factors outside of the consumer's control. Thus, it offers no useful feedback on how to improve one's outcomes and may further alienate consumers by adding insult to injury (Binns et al., 2018). Such an explanation is unlikely to improve perceptions of the algorithmic system and could in fact worsen them. Thus, sensitivity explanations are more effective than case-based and no explanations only when they offer actionable information by highlighting factors under the consumer's direct control. Specifically, we predict:

H3: Actionable sensitivity-based explanations lead to more positive attitudes towards the algorithmic decision system, and increased behavioral intentions toward the company, relative to both case-based explanations and no explanation.

H4: Non-actionable sensitivity-based explanations lead to less positive attitudes, and decreased behavioral intentions, relative to actionable sensitivity-based explanations.

H5: Non-actionable sensitivity explanations lead to equal or worse attitudes, and behavioral intentions, relative to case-based explanations or no explanation at all.

So far, we have argued that effective explanations (e.g., actionable sensitivity-based explanations) should improve consumers' perceptions of algorithmic decisions and the organizations deploying them in a way that ineffective explanations (e.g., non-actionable sensitivity-based or case-based explanations) should not. However, even the most compelling explanation will only improve perceptions if consumers are motivated to understand how a particular outcome came to be. People are more likely to seek explanations for outcomes that are unexpected or unwanted (Hilton & Slugoski, 1986; McClure et al., 2003). When the outcome of an algorithmic decision is negative, people may perceive it to be more surprising, unexpected, or unfair than when it is positive. Negative outcomes can also have more severe consequences than positive outcomes. As a result, people may be more motivated to seek an explanation to understand how they can improve their chances of getting a different outcome in the future. Thus, we predict:

H6: Providing a sensitivity explanation leads to more positive attitudes and behavioral intentions relative to no explanation, but this effect is attenuated when the decision outcome is positive.

We tested these hypotheses in a series of 5 preregistered experiments featuring both hypothetical and consequential decision outcomes. In each experiment, participants learned of an algorithmic decision, received a case-based, sensitivity, or no explanation, and then evaluated the algorithm and organization using it. The experiments demonstrate consistent impacts of explanations while identifying boundary effects and moderators. Data and analysis scripts for all five studies are available here: https://osf.io/6vwj3.

4. Study 1

4.1. Method

Study 1 was designed to test H1 and H2. We aimed to have 150 respondents per condition and expected to exclude around 10 % of the responses due to lack of attention, as per the preregistration plan. We recruited 500 Canadian residents on Prolific Academic who participated in this study in exchange for monetary compensation. Participants were randomly assigned to one of 3 conditions: no explanation vs. case-based explanation vs. sensitivity explanation, in a between-subjects design. After excluding participants who did not pass the attention check, we were left with 443 participants ($N_{sensitivity} = 148$, $N_{case} = 146$, $N_{no\text{-}explanation} = 149$; 200 men, 237 women, 5 preferred not to answer, and 1 answered other; $M_{age} = 31.1$, $SD_{age} = 10.6$). Preregistration for Study 1 is found here: https://doi.org/10.17605/OSF.IO/6WHFR.

First, participants in this and the following studies read a sample scenario where a loan applicant is denied by an algorithmic decision-maker. This scenario was used to familiarize participants with the task and for attention checks. Next, participants read a scenario adapted from Binns et al. (2018), in which a car insurance company uses an algorithmic decision system to provide customers with insurance prices based on their personal characteristics and driving behaviors [see Supplementary Information]. They were then shown the profile of an applicant and informed that the applicant did not receive the best rate. This was followed by either a case-based explanation, a sensitivity explanation, or no explanation. Participants in the case-based condition were told the company provided the applicant with the following explanation:

> "This decision was based on thousands of similar cases from the past. For example, a similar case to yours is a previous customer:
> She was 38 years old, with 18 years of driving experience, drove 850 miles per month, occasionally exceeded the speed limit, and 25 % of her trips took place at night. Claire was involved in one accident in the following year."

Those in the sensitivity condition were told "the company provided the applicant with the following information about the computer's decision":

- "If 10 % or less of your driving took place at night, you would have qualified for the cheapest tier.
- If your average miles per month were 700 or less, you would have qualified for the cheapest tier."

We then measured perceived transparency, attitude (perceived fairness, trust in the algorithmic system), and behavioral intentions on seven-point scales anchored at 1 (*strongly disagree*) and 7 (*strongly agree*), using prompts specifically created for this project. *Transparency* ($\alpha = 0.95$) "I feel like I know exactly how the computer system made its decision"; "It is perfectly clear how the computer system made its decision"; "The decision process used by the computer system is transparent." *Perceived fairness*: "Using this kind of automated decision system is fair." *Trust* in the algorithmic decision system: "I would trust this automated decision system to determine the price I pay for insurance." *Behavioral intentions* (*std.* $\alpha = 0.96$)[1]: "If I were in the market for insurance, I would want to do business with this company"; "I would

[1] For a two-item scale, the standardized coefficient alpha (std. α) is equivalent to the Spearman-Brown coefficient and is considered a more appropriate measure of scale reliability than Cronbach's alpha (Eisinga, Grotenhuis, & Pelzer, 2013).

recommend this insurance company to a friend." (Additional exploratory measures are described in the preregistration). Perceived fairness and trust were highly correlated (r = 0.85), so we combined them into a single index of *Attitude* toward the automated decision system (*std.* α = 0.92).

4.2. Results and discussion

Explanation type had a significant effect on perceived transparency ($F(2, 440) = 121.72$, $p < 0.001$, $\eta^2 = 0.356$), attitude ($F(2, 440) = 25.48$, $p < 0.001$, $\eta^2 = 0.104$), and intentions ($F(2, 440) = 19.55$, $p < 0.001$, $\eta^2 = 0.096$). Consistent with H1, planned contrasts[2] revealed that consumers who received a sensitivity explanation reported greater perceived transparency, more positive attitude toward the automated decision system (i.e., rated the automated system to be fairer and more trustworthy), and more favorable behavioral intentions toward the insurance company than those who received a case-based explanation as well as those who received no explanation (see Table 1 and Fig. 1). Contrary to H2, case-based explanations also yielded higher perceptions of transparency, attitudes, and behavioral intentions, relative to no explanation. We reasoned that case-based explanations fail to communicate useful information for consumers seeking to improve their outcomes. Moreover, when a consumer's loan application is rejected with the explanation that they are similar to another consumer who failed to repay their loan on time, this may raise concerns that the algorithm has not adequately accounted for the consumer's unique characteristics and circumstances. Thus, we predicted that providing a case-based explanation would be no more effective than providing no explanation at all. The data in Study 1 appear inconsistent with this account. It is possible that H2 is incorrect and that providing a case-based explanation is, in fact, an effective strategy. However, as will be later shown, Studies 2, 3, and 4 consistently find that providing a case-based explanation does not improve attitudes and intentions compared to providing no explanation. Thus, while a case-based explanation may sometimes be better than no explanation, it is not advisable based on the weight of the evidence we will present. Finally, in an exploratory analysis, we tested the direct and moderating effects of gender and found no significant effects on any of our DVs.

A potential limitation of Study 1 is that it asked participants to evaluate decisions that happened to other people rather than themselves. This design feature allows us to keep the applicant profile constant across participants and conditions. However, judgments of fairness and trust may vary when reflecting on personal outcomes versus the outcomes of others. We address this limitation in the next study, in which participants are themselves rejected by an algorithm designed to assess whether they should receive a real, financial participation bonus.

Table 1
Means (and Standards Deviations) of Perceived Transparency, Attitude, and Intentions Across Experimental Conditions and Planned Contrasts – Study 1.

	Transparent	Attitude	Intention
None	3.00 (1.45)	2.67 (1.37)	2.40 (1.40)
Case	4.35 (1.54)	3.22 (1.50)	2.87 (1.52)
Sensitivity	5.61 (1.34)	3.93 (1.65)	3.58 (1.57)
Sensitivity vs. Case	7.46***	4.01***	4.07***
Sensitivity vs. None	15.60***	7.12***	6.80***
Case vs. None	8.10***	3.08**	2.70*

Note. The t-values for the planned contrasts are based on t-tests with 440 degrees of freedom. Statistical significance markers: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

[2] All analyses involving multiple comparisons use a Bonferroni correction.

5. STUDY 2

5.1. Method

Five hundred US residents, recruited on Prolific Academic, participated in this study in exchange for monetary compensation. Participants were randomly assigned to one of 3 conditions: no explanation vs. case-based explanation vs. sensitivity explanation, in a between-subjects design. After excluding those who did not pass the attention check, we were left with 493 participants ($N_{sensitivity} = 168$, $N_{case} = 161$, $N_{no\text{-}explanation} = 168$; 249 men, 236 women, 2 preferred not to answer, and 6 answered other; $M_{age} = 35.1$, $SD_{age} = 11.7$). Preregistration for Study 2 is available here: https://doi.org/10.17605/OSF.IO/9KN4J.

Upon completing an initial task, participants learned that, if eligible, they could participate in a bonus section of the study (for extra pay). They were informed that their eligibility would be determined by an automated system, based on their responses to a battery of questions, including "How many languages can you speak well enough to order coffee?"; "How many anagram puzzles have you done in the past week?"; "How many hours a week do you spend reading for pleasure?"; "How expert are you at Sudoku (the logic-based, combinatorial number puzzle game)?"; and "At what age did you complete your first Sudoku puzzle?".

After submitting their answers, all participants learned that they failed to qualify for the bonus section. Those in the no-explanation condition received the following message: "Unfortunately, the automated system determined that you are not eligible to complete this section." Those in the case-based explanation received the same message, augmented with the following explanation: "Your responses closely match those of other participants who did not provide useful data. Here are your responses and an example of one of those other participants' responses." The other participant's responses were generated by adding random noise to the participant's own responses. Those in the sensitivity-based explanation were told: "If you had completed 2 more anagrams in the past week or if you completed more than 19 Sudoku puzzles per week, you would have qualified.".

Participants then reported their perceived *transparency* (α = 0.95), *attitude* toward the algorithmic system (*std.* α = 0.88), and *behavioral intentions* (*std.* α = 0.95) on adapted versions of the seven-point scales from Study 1 (see Supplementary Information). Finally, they were debriefed, and all received bonus pay.

5.2. Results and discussion

As in study 1, explanation type had a significant effect on perceived transparency ($F(2, 490) = 77.62$, $p < 0.001$, $\eta^2 = 0.241$), attitude ($F(2, 490) = 31.98$, $p < 0.001$, $\eta^2 = 0.115$), and intentions ($F(2, 490) = 31.70$, $p < 0.001$, $\eta^2 = 0.115$). Participants who received a sensitivity explanation perceived the decision to be more transparent, expressed a more positive attitude toward the automated decision system, and reported more favorable intentions toward studies that use an automated decision system than those who received a case-based explanation, and those who received no explanation (see Table 2 and Fig. 2). Moreover, participants in the case-based explanation condition reported less favorable intentions than those in the no-explanation condition, while the case-based and no-explanation groups did not differ significantly in their attitude toward the automated decision system and their perceived transparency. Finally, the case-based explanation produced more negative behavioral intentions than no explanation at all.

Using a task in which participants were themselves rejected by an algorithm, Study 2 replicated Study 1′s finding that providing a sensitivity explanation improves perceptions of transparency, attitudes, and behavioral intentions, lending further support to H1. Unlike Study 1, Study 2 also found that a case-based explanation performed no better (and worse in terms of intentions) than no explanation at all, which is consistent with H2.

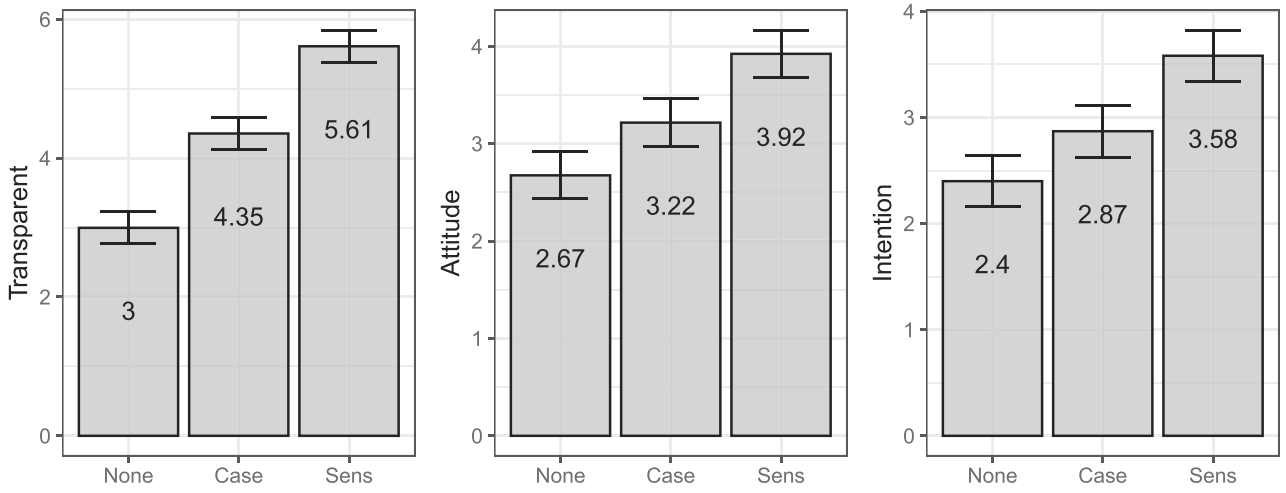

Fig. 1. Study 1: Effect of explanation on transparency, attitudes, and intentions. Error bars represent 95% confidence intervals.

Table 2
Means (and Standards Deviations) of Perceived Transparency, Attitude, and Intentions Across Experimental Conditions and Planned Contrasts – Study 2.

	Transparent	Attitude	Intention
None	2.94 (1.90)	3.54 (1.68)	3.36 (1.81)
Case	3.02 (1.82)	3.29 (1.63)	2.78 (1.66)
Sensitivity	5.13 (1.74)	4.67 (1.71)	4.30 (1.79)
Sensitivity vs. Case	10.59^{***}	7.47^{***}	7.88^{***}
Sensitivity vs. None	10.99^{***}	6.17^{***}	4.89^{***}
Case vs. None	0.43	−1.32	-3.00^{**}

Note. The t-values for the planned contrasts are based on t-tests with 490 degrees of freedom. Statistical significance markers: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Studies 1 and 2 demonstrated the superior effectiveness of providing sensitivity explanations but did not address the underlying rationale. We have argued that sensitivity explanations perform especially well because, unlike other types of explanations, they tend to offer actionable insight into how future outcomes might be improved. Thus, we expect that a sensitivity explanation that lacks actionability (e.g., by highlighting factors that are not under the consumer's control) will fail to improve consumer evaluations of an algorithmic system and the company deploying it. We test these predictions in the next study.

6. Study 3

6.1. Method

Study 3 was designed to test hypotheses 3, 4, and 5. We recruited 660 US residents on Prolific Academic. Participants were randomly assigned to one of four conditions (sensitivity-actionable vs. sensitivity-non-actionable vs. case-based vs. no explanation) in a between-subjects design. After excluding participants who did not pass the attention check, we were left with 601 participants ($N_{\text{sensitivity-actionable}} = 149$, $N_{\text{sensitivity-non-actionable}} = 154$, $N_{\text{case}} = 153$, $N_{\text{no-explanation}} = 145$; 281 men, 308 women, 5 preferred not to answer, and 6 answered other; $M_{age} = 32.9$, $SD_{age} = 12.5$). Preregistration is found here: https://doi.org/10.17605/OSF.IO/W5BU9.

Participants read the same car insurance scenario used in Study 1 and were then shown the same profile of an applicant, before reading one of the three explanations, or no explanation, depending on condition. The case and sensitivity-actionable explanations were the same as in Study 1. Participants in the sensitivity-non-actionable condition were told that the applicant received the following feedback:

- If you were aged 45–55 years old, you would have qualified for the cheapest tier.
- If you had owned your home, you would have qualified for the cheapest tier.

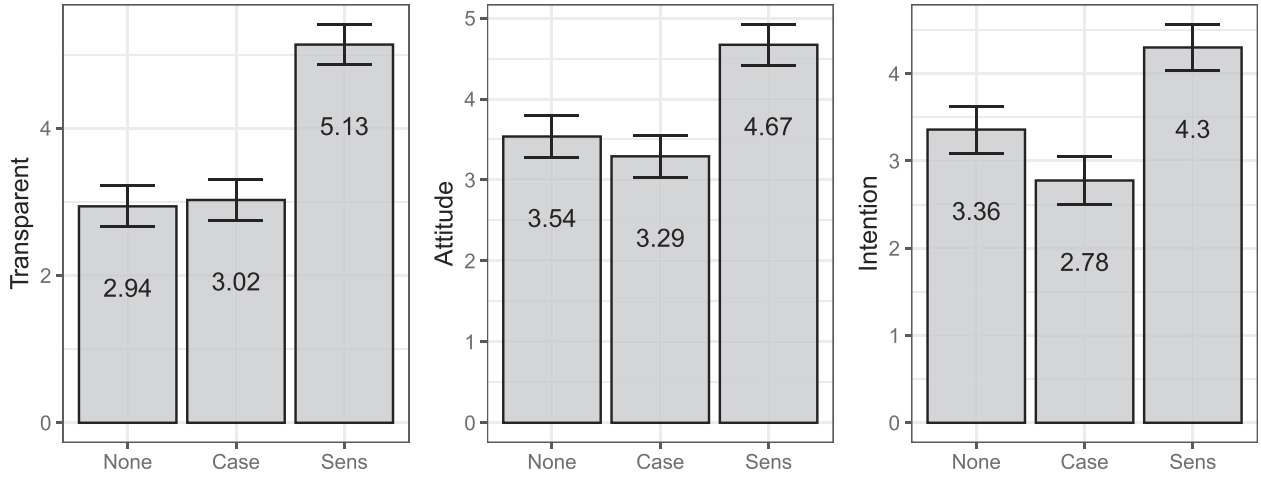

Fig. 2. Effect of explanation on transparency, attitudes, and intentions. Error bars represent 95% confidence intervals.

We then measured perceived *transparency* (α = 0.93), *attitude* toward the algorithmic system (*std.* α = 0.85), and *behavioral intentions* (*std.* α = 0.96) on the same seven-point scales used in Study 1. Finally, we measured perceived *actionability* (α = 0.85) – that is, the degree to which the information provided can be used by the applicant to get a better rate in the future – by asking "How easy or difficult would it be for [the applicant] to make changes that would alter whether she qualifies for the best rate?"; "I believe that if [the applicant] changed her behavior, she could end up being eligible for a different rate"; "[the applicant's] quoted insurance rate is due to her own behavior." (Additional exploratory measures are described in the preregistration and reported on osf.io).

6.2. Results and discussion

The sensitivity-actionable explanation was perceived to be significantly more actionable than the sensitivity-non-actionable explanation ($M_{\text{sensitivity-actionable}} = 4.72$, $SD_{\text{sensitivity-actionable}} = 1.31$ vs. $M_{\text{sensitivity-non-actionable}} = 2.40$, $SD_{\text{sensitivity-non-actionable}} = 1.39$; $t(300.81) = 14.96$, $p < 0.001$), confirming that the actionability manipulation worked as intended.

Explanation type had a significant effect on perceived transparency ($F(3, 597) = 70.76$, $p < 0.001$, $\eta^2 = 0.262$), attitude ($F(3, 597) = 14.26$, $p < 0.001$, $\eta^2 = 0.067$), and intentions ($F(3, 597) = 11.16$, $p < 0.001$, $\eta^2 = 0.053$). Consistent with H3, actionable sensitivity explanations yielded more positive attitudes and higher behavioral intentions than case-based explanations and no explanation (see Table 3 and Fig. 3). Consistent with H4, non-actionable sensitivity explanations resulted in less positive attitudes and lower intentions than actionable sensitivity explanations. Consistent with H5, attitudes and intentions were no better in the non-actionable sensitivity condition than they were in the case-based and no-explanation conditions. Surprisingly, actionable and non-actionable sensitivity explanations did differ in perceived transparency, with actionable explanations perceived as more transparent. This discrepancy in perceived transparency may be due to a halo effect; participants' positive attitudes towards actionable explanations may have colored their other judgments, including their perceived transparency, despite both conditions having the same number of features listed. The non-actionable sensitivity explanation was perceived to be equally transparent as the case explanation, and both were perceived as more transparent than no explanation.

These results demonstrate the importance of providing information that consumers can use to improve their outcomes. Without such actionable information, a "sensitivity" explanation may be no better than no explanation at all. Together, the results of Study 3 suggest that there is good reason for firms to provide actionable explanations – doing so may improve consumer attitudes and desire to recommend and patronize the business. However, the results thus far are limited to a single financial scenario. We therefore expand the range of application domains in study 4.

Table 3
Means (and Standards Deviations) of Perceived Transparency, Attitude, and Intentions Across Experimental Conditions and Planned Contrasts – Study 3.

	Transparent	Attitude	Intention
None	2.90 (1.41)	2.91 (1.39)	2.67 (1.45)
Case	4.24 (1.64)	3.02 (1.60)	2.80 (1.59)
Sens-A	5.45 (1.41)	3.95 (1.61)	3.60 (1.65)
Sens-NA	4.64 (1.65)	3.20 (1.46)	2.81 (1.51)
Sens-A vs. Sens-NA	4.58***	4.28***	4.45***
Sens-A vs. Case	6.88***	5.33***	4.48***
Sens-A vs. None	14.25***	5.90***	5.13***
Case vs. Sens-NA	−2.32	−1.06	−0.04
Case vs. None	7.51***	0.64	0.71
Sens-NA vs. None	9.81***	1.69	0.76

Note. Sens-A=Sensitivity-Actionable; Sens-NA=Sensitivity-Non-Actionable. The t-values for the planned contrasts are based on t-tests with 597 degrees of freedom. Statistical significance markers: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

7. Study 4

7.1. Method

Study 4 tests the generalizability of our findings across multiple domains. We recruited 660 US residents on Prolific. After excluding participants who did not pass the attention checks, we were left with 593 participants ($N_{\text{sensitivity-actionable}} = 149$, $N_{\text{sensitivity-non-actionable}} = 149$, $N_{\text{case}} = 144$, $N_{\text{no-explanation}} = 151$; 283 men, 300 women, 5 preferred not to answer, and 5 answered other; $M_{age} = 33.7$, $SD_{age} = 12.9$). Preregistration is found here: https://doi.org/10.17605/OSF.IO/45URB.

The study consisted of a 4 (explanation condition) x 4 (scenario) mixed design. Participants were randomly assigned to one of four between-subjects explanation conditions, as in Study 3. Then, each participant responded to four new scenarios in a within-subjects format: credit card application; parole decision; admission to a professional training program; and health insurance pricing, presented in a random order (see Supplementary Information). In each scenario, participants reported on their perceived transparency, attitude, and intentions using the same measures as before (see Table 4 for reliability indicators).

Since each participant provided four sets of ratings, we tested our hypotheses by fitting a series of linear mixed-effects models (multiple measures for each participant; one model for each outcome variable) to account for the dependencies in the data. The models were estimated using maximum likelihood. They included random intercepts for participants and scenarios, and fixed effects for explanation type.[3] To test the overall effect of explanation type (i.e., the omnibus test), each model was compared to a corresponding reduced model, which included the same random intercepts for participants and scenarios but did not include the fixed effect of explanation type. Specific contrasts were then assessed through pairwise comparisons using Bonferroni corrections for multiple comparisons. Analyses pertaining to hypothesis testing focus on the aggregated data and are summarized in Table 6. In addition, the means and standard deviations for each scenario are presented in Table 7.

7.2. Results and discussion

In all four scenarios, perceived actionability was significantly higher in the sensitivity-actionable condition than in the sensitivity-non-actionable, suggesting a successful manipulation of actionability (Table 5).

Likelihood ratio tests comparing the specified and reduced models indicated that model fit improves significantly when explanation type is present. These tests suggest a significant effect of explanation type on perceived transparency ($\chi^2 = 108.17$, $df = 3$, $P < .001$), attitude ($\chi^2 = 90.82$, $df = 3$, $p < 0.001$), and intention ($\chi^2 = 82.81$, $df = 3$, $p < 0.001$). Consistent with H3 and H4, follow-up analyses found that participants who received an actionable sensitivity explanation reported higher levels of transparency, attitude, and intentions than any other group (see Table 6 and Fig. 4). Unlike the nonsignificant differences observed in Study 3 (car insurance), non-actionable sensitivity explanations produced worse attitudes and behavioral intentions than both case and no explanation, as predicted in H5. Presumably, explanations highlighting

[3] Our analyses depart from the preregistered analyses, in which the mixed-effects models were specified to include random intercepts for participants but not for scenarios. Instead, the preregistered models included fixed effects for explanation condition, scenarios, and their interaction. The revised specification of the mixed-effects models is more in line with the study's aim, which is to assess the generalizability of the effect of explanation on attitudes and intentions across domains rather than the moderating effect of scenarios.

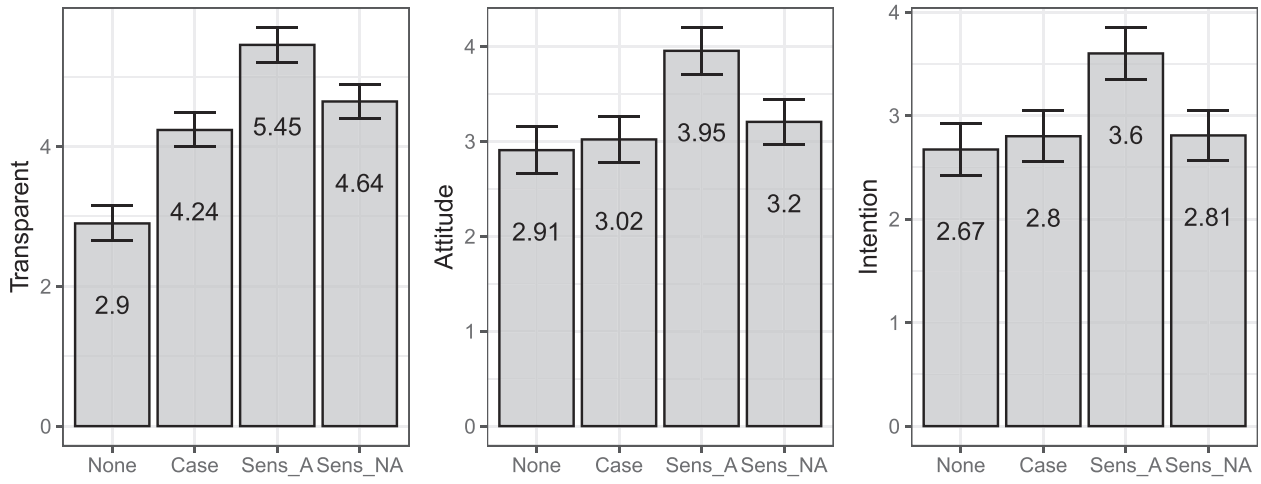

Fig 3. Effect of explanation condition on transparency, attitudes, and intentions. Error bars represent 95% confidence intervals.

Table 4
Reliability of Multi-item Scales.

Scale (reliability)	Number of Items	Credit Card	Professional Training	Parole Decision	Health Insurance
Transparent (Cronbach α)	3	0.95	0.95	0.84	0.95
Attitude (*std.* α)	2	0.95	0.89	0.93	0.93
Intention (*std.* α)	2	0.96	0.94	0.95	0.95
Actionability (Cronbach α)	3	0.86	0.84	0.86	0.87

Table 5
Perceived Actionability of Sensitivity Explanations in All Scenarios.

	Sens-A	Sens-NA	*Welsh t-test*
Credit Card	4.71 (1.32)	2.55 (1.44)	12.09^{***}
Parole Decision	5.01 (1.30)	3.01 (1.51)	10.77^{***}
Health Insurance	4.55 (1.36)	2.12 (1.41)	14.20^{***}
Professional Training	4.28 (1.39)	2.59 (1.37)	10.11^{***}

Note. Sens-A=Sensitivity-Actionable; Sens-NA=Sensitivity-Non-Actionable. Statistical significance markers: * $p < 0.5$; ** $p < 0.01$; *** $p < 0.001$.

Table 6
Means (and Standards Deviations) of Perceived Transparency, Attitude, and Intentions Across Experimental Conditions and Planned Contrasts – Study 4.

	Transparent	Attitude	Intention
None	3.50 (1.60)	3.17 (1.60)	2.79 (1.51)
Case	4.30 (1.48)	2.98 (1.63)	2.66 (1.57)
Sens-A	4.98 (1.51)	3.74 (1.71)	3.41 (1.70)
Sens-NA	4.01 (1.70)	2.53 (1.56)	2.17 (1.44)
Sens-A vs. Sens-NA	7.01^{***}	9.30^{***}	9.83^{***}
Sens-A vs. Case	4.82^{***}	5.77^{***}	5.90^{***}
Sens-A vs. None	10.64^{***}	4.38^{***}	4.95^{***}
Case vs. Sens-NA	2.13	3.44^{**}	3.85^{***}
Case vs. None	5.72^{***}	−1.45	−1.01
Sens-NA vs. None	3.62^{**}	-4.95^{***}	-4.91^{***}

Note. Sens = sensitivity. The t-values for the planned contrasts are based on t-tests with 592 degrees of freedom. Statistical significance markers: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table 7
Means (and Standards Deviations) of Perceived Transparency, Attitude, and Intentions Across Explanation Conditions in Each Scenario – Study 4.

	None	Case	Sens-A	Sens-NA
Credit Card				
Transparent	4.03 (1.53)	4.50 (1.47)	5.39 (1.38)	3.42 (1.72)
Attitude	4.22 (1.52)	3.64 (1.73)	4.61 (1.62)	2.10 (1.41)
Intention	3.70 (1.50)	3.29 (1.73)	4.21 (1.68)	1.81 (1.27)
Parole Decision				
Transparent	3.72 (1.68)	4.55 (1.30)	5.11 (1.37)	4.22 (1.51)
Attitude	2.91 (1.65)	2.92 (1.63)	3.59 (1.69)	2.80 (1.74)
Intention	2.56 (1.62)	2.63 (1.55)	3.27 (1.68)	2.39 (1.64)
Health Insurance				
Transparent	3.46 (1.56)	4.25 (1.52)	4.51 (1.62)	4.16 (1.76)
Attitude	3.04 (1.48)	2.80 (1.62)	3.35 (1.63)	2.73 (1.57)
Intention	2.64 (1.36)	2.52 (1.58)	3.03 (1.62)	2.31 (1.44)
Professional Training				
Transparent	2.79 (1.38)	3.93 (1.54)	4.92 (1.53)	4.22 (1.67)
Attitude	2.52 (1.19)	2.56 (1.31)	3.40 (1.60)	2.47 (1.44)
Intention	2.25 (1.10)	2.20 (1.21)	3.13 (1.59)	2.16 (1.34)

Note. Sens-A=Sensitivity-Actionable; Sens-NA=Sensitivity-Non-Actionable.

uncontrollable factors not only fail to offer useful information on how to improve one's outcomes but may also backfire, frustrating consumers by adding insult to injury.

Study 4 demonstrates the robustness of the observed explanation effects across diverse algorithmic decision domains, including credit card, parole, health insurance, and professional training. In each of these domains, actionable sensitivity explanations produced the most positive perceptions of transparency, attitudes, and behavioral intentions. However, it is possible that the factors highlighted in the actionable explanation conditions are seen as not only more controllable than those highlighted in the non-actionable conditions but also as more relevant to the algorithmic decisions. Thus, the observed differences in attitudes and intentions between the actionable and non-actionable conditions may be driven more by differences in perceived relevance than perceived controllability. We investigated this possibility through a post-test (see Supplementary Information), in which participants rated the relevance of all the actionable and non-actionable factors used in the five scenarios of our studies. Perceived relevance did not vary significantly between the actionable and non-actionable conditions in three out of five scenarios (parole decision, health insurance, professional training, and car insurance). In the remaining scenarios (car insurance and credit card), the actionable factors were perceived to be more relevant than the non-actionable factors. These findings, combined with those of studies 3 and 4, suggest that although our manipulation of actionability may have impacted the explanations' perceived relevance

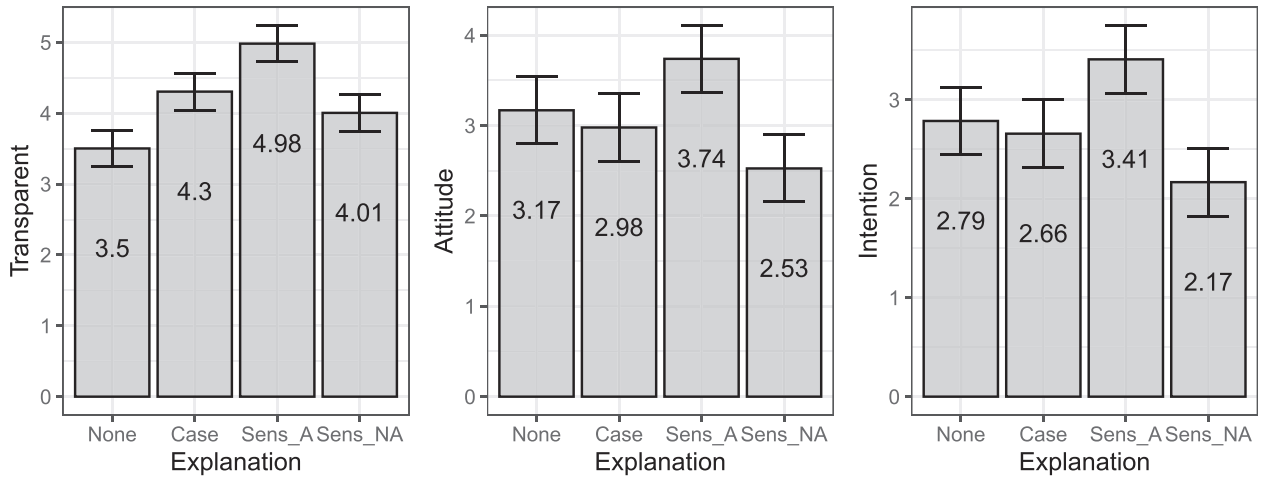

Fig. 4. Study 4: Effect of explanation condition on transparency, attitudes, and intentions across multiple scenarios. Error bars represent 95% confidence intervals.

(at least in some scenarios), the observed differences in attitudes and intentions across our studies are driven more by differences in perceived controllability than perceived relevance.

8. Study 5

Study 5 examines the moderating effect of outcome valence (H6). We argued that even the most effective explanation will only improve perceptions if consumers are motivated to understand how a particular outcome came to be. Since consumers are less likely to seek an explanation for a positive outcome than for a negative outcome, we predicted that the effect of providing the best type of explanation (i.e., actionable sensitivity explanation) on consumers' attitudes and behavioral intentions will be attenuated when the decision outcome is positive. Preregistration is found here: https://doi.org/10.17605/OSF.IO/9YSAJ.

8.1. Method

We recruited 880 US residents on Prolific Academic with the aim of obtaining 200 respondents per condition. After excluding participants who did not pass the attention check, we were left with 764 participants ($N_{\text{sensitivity-positive}} = 210$, $N_{\text{sensitivity-negative}} = 183$, $N_{\text{no-explanation-positive}} = 175$, $N_{\text{no-explanation-negative}} = 196$; 375 men, 374 women, 7 preferred not to answer, and 8 answered other; $M_{age} = 33.7$, $SD_{age} = 13.3$).

Study 5 used a 2 (no explanation vs. sensitivity explanation) x 2 (positive vs. negative outcome) between-subjects design. Participants read the same car insurance scenario used in Studies 1 and 3 and were then shown the same profile of an applicant, however in this study they were either informed that the applicant had – or had not – received the best rate. Then, participants either viewed the same sensitivity explanation used in Study 1 or no explanation. We measured perceived *transparency* ($\alpha = 0.95$), *attitude* toward the algorithmic system (*std.* $\alpha = 0.90$), and *behavioral intentions* (*std.* $\alpha = 0.96$) as before. (Additional exploratory measures are described in the preregistration and reported on osf.io).

8.2. Results and discussion

There was a significant main effect of explanation on transparency ($F(1, 760) = 240.61$, $p < 0.001$, $\eta_p^2 = 0.240$), attitude ($F(1, 760) = 52.68$, $p < 0.001$, $\eta_p^2 = 0.065$), and intention ($F(1, 760) = 54.45$, $p < 0.001$, $\eta_p^2 = 0.067$). There was also a significant main effect of outcome valence on transparency ($F(1, 760) = 112.32$, $p < 0.001$, $\eta_p^2 = 0.129$), attitude ($F(1, 760) = 145.19$, $p < 0.001$, $\eta_p^2 = 0.160$), and intention ($F(1, 760) = 144.35$, $p < 0.001$, $\eta_p^2 = 0.160$). Most importantly, there was a significant explanation × outcome valence interaction effect on transparency ($F(1, 760) = 62.49$, $p < 0.001$, $\eta_p^2 = 0.076$), attitude ($F(1, 760) = 20.21$, $p < 0.001$, $\eta_p^2 = 0.026$), and intention ($F(1, 760) = 20.77$, $p < 0.001$, $\eta_p^2 = 0.027$).

Consistent with H6, when the outcome was negative, that is when the applicant did not receive the best rate, participants in the sensitivity explanation condition perceived the decision to be more transparent, reported more positive attitudes toward the algorithmic decision system, and more favorable behavioral intentions toward the insurance company than those in the no explanation condition (see Table 8 and Fig. 5). In contrast, when the decision was positive (i.e., consumers got the best interest rate) the effect of explanation on transparency was attenuated, and the effect of explanation on attitudes and behavioral intentions was eliminated. These results suggest that choosing the optimal explanation type is essential, particularly when consumers are likely to question the outcome because it is not favorable. However, when outcomes are positive, explanations have less power to influence consumer attitudes and intentions. Thus, firms can conserve resources (e.g., user experience designers' time; follow-up mailings) by focusing post-hoc explanation efforts on consumers who have received negative outcomes.

9. General discussion

Algorithms are now widely used to automate decision-making in a variety of domains, such as finance, healthcare, education, and law enforcement. Despite offering notable gains in prediction accuracy, algorithmic decision systems are often met with skepticism and mistrust. One important obstacle to the acceptance of algorithmic decisions is their perceived opacity. People mistrust obscure decision systems when these systems impact their access to resources and opportunities.

This research examined the impact of providing different types of post-hoc explanations on consumers' perceptions of transparency, attitudes toward algorithmic decision systems, and behavioral intentions

Table 8
Means (and Standards Deviations) of Perceived Transparency, Attitude, and Intention Across Experimental Conditions and Planned Contrasts – Study 5.

	Transparent	Attitude	Intention
None/Negative	2.94 (1.48)	2.70 (1.43)	2.44 (1.46)
Sensitivity/Negative	5.26 (1.49)	3.80 (1.59)	3.59 (1.67)
None/Positive	4.54 (1.45)	4.54 (1.43)	4.34 (1.50)
Sensitivity/Positive	5.19 (1.38)	4.68 (1.43)	4.49 (1.46)
Sensitivity vs. None/Negative	15.51^{***}	7.26^{***}	7.38^{***}
Sensitivity vs. None/Positive	4.36^{***}	0.36	0.34

Note. The t-values for the planned contrasts are based on t-tests with 760 degrees of freedom. Statistical significance markers: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

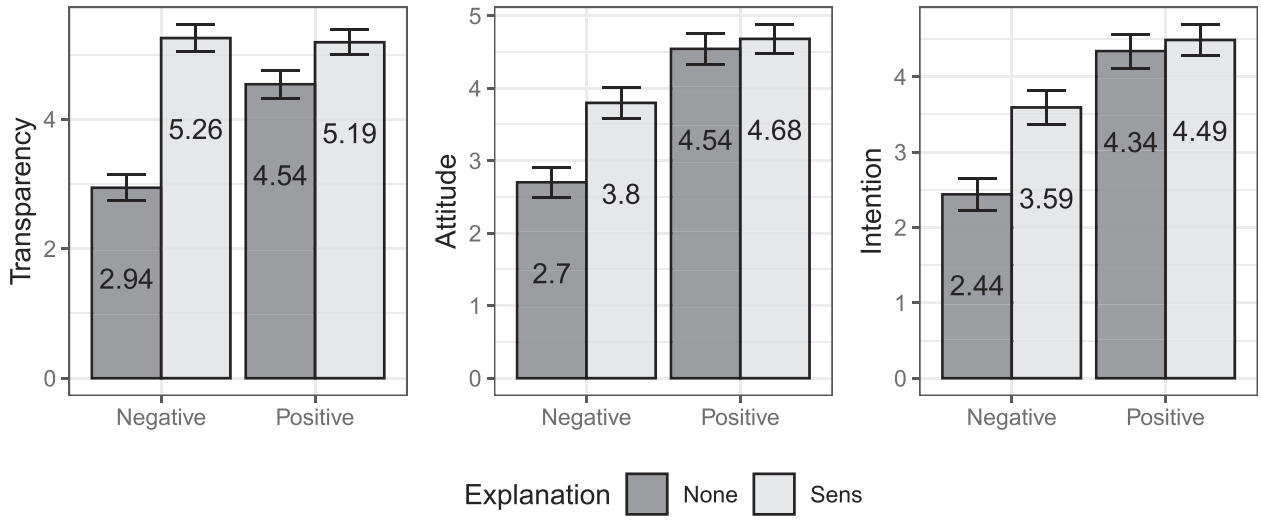

Fig. 5. Study 5: Effect of explanation and outcome valence on transparency, attitudes, and intentions. Error bars represent 95% confidence intervals.

toward the organizations deploying algorithmic systems. We found that sensitivity explanations were the most effective at improving perceptions of algorithmic fairness and trust (attitudes) and behavioral intentions, so long as the explanation offered actionable feedback and the decision outcome was unfavorable. These effects held across multiple relevant domains. However, *non*-actionable sensitivity explanations sometimes backfired, reducing attitudes and behavioral intentions.

9.1. Theoretical implications

Our research contributes to the growing literature in marketing and consumer behavior on consumer acceptance of algorithmic systems (Castelo, et al., 2019; Dietvorst et al., 2015, 2018; Leung et al., 2019; Logg et al., 2020; Longoni et al., 2019; Longoni & Cian, 2022). This line of work has documented a general aversive stance toward algorithms but has also identified several conditions under which consumers may be positively disposed toward algorithms. For instance, consumers are more likely to accept an algorithm's recommendation if they believe it to be tailored to an individual's unique case (Longoni et al., 2019), if they can modify the algorithm's forecast (Dietvorst et al., 2018), if they perceive the task to be objective (Castelo et al., 2019), or if they have a utilitarian (vs. hedonic) goal (Longoni & Cian, 2022). We extend this literature by showing that resistance to algorithmic systems can also be attenuated by providing an adequate explanation of the decision outcome. This is significant because the design and provision of explanations are controllable factors that are entirely within a company's power.

Our research also contributes to the broader literature on interpretable AI by testing theoretically-derived predictions of how people might respond to simple post-hoc explanations of outputs from complex machine learning models. Machine learning uses black-box functions for making decisions. These functions may be extremely complex and may contain millions of interdependent values, making them impossible for humans to fully comprehend. As a result, it may not be possible to offer faithful and complete explanations of decisions made by machine learning models. Post-hoc explanations circumvent these constraints. In particular, local explanations aim to justify decisions made by the model by explaining how the model behaved in a particular case rather than how it works in general. However, not all post-hoc explanations are equally effective. Building on research emphasizing the social nature of explanations, we predicted and found that sensitivity-based explanations produce the most positive evaluations of algorithmic decisions and the organizations deploying them.

9.2. Managerial implications

Organizations may wish to avoid explaining algorithmic decisions, in an effort to protect proprietary information. However, our findings demonstrate that businesses need not fear potential disclosure mandates. On the contrary, explanations hold the potential to enhance consumer trust and acceptance of algorithms, algorithmic decisions, and the institutions deploying them. These findings also point to clear recommendations for organizations seeking to offer effective explanations of algorithmic decisions. First, whenever possible, firms should present explanations in a sensitivity format. Such explanations help consumers understand how they can obtain a better outcome, rather than direct attention and frustration at the company and its algorithm.

To empower consumers to improve their future outcomes, sensitivity explanations must focus on features and variables that can be feasibly changed by consumers. For instance, a consumer's age might be a heavily-weighted feature in an algorithm's decision to qualify a person for a good health insurance rate; moreover, a person's age might have narrowly missed the threshold for a better outcome (e.g., being only one year too old). A company's desire to provide transparent and meaningful explanations may lead them to highlight the consumer's age in a post-hoc explanation. However, our research suggests that since a variable such as age is impossible to change (i.e., non-actionable), this choice of content may render the explanation ineffective, or even counterproductive. Since complex decision systems such as machine learning algorithms often consider hundreds of variables, organizations may benefit from being selective when crafting explanations. Specifically, firms should opt to include variables that are more actionable, and omit features that are less actionable, in order to improve consumer responses. Lastly, given that firms operate with limited time and resources, our results indicate that efforts to craft explanations should be prioritized for decisions with negative outcomes, as we failed to find any significant benefits to explaining decisions with positive outcomes.

9.3. Limitations and future research

Our work has limitations that provide several opportunities for future research. First, in manipulating the actionability of sensitivity explanations, we distinguished between factors that are under the consumers' direct control and those that are not. While this is an important distinction, perception of actionability may also depend on how easy or difficult it is for a consumer to change a factor that is under their direct control. For example, a consumer can control how much they drive, but an explanation suggesting reducing their monthly driving by 100 miles

would undoubtedly be perceived as more actionable than one suggesting reducing it by 500 miles. Future research may benefit from a more refined operationalization of actionability.

Second, although actionable sensitivity-based explanations were consistently found to perform best at improving attitudes and intentions toward automated systems, the comparison was limited to only one alternative (case-based explanations). Many other explanation styles exist (e.g., see Binns et al., 2018). It is possible that different styles are better suited in different contexts, depending on users' goals. For instance, local explanations may well be best at justifying outcomes and highlighting strategies for change, but for users who seek to better understand the model, global explanations would be preferable.

While our findings suggest that actionable sensitivity explanations offer a promising direction for improving the acceptability of algorithmic decision systems, simply offering post-hoc sensitivity explanations may not be the optimal long-term solution for addressing issues of actual fairness in algorithmic decisions. Sensitivity explanations are necessarily selective, and thus do not faithfully reflect what the original model actually computes. This creates at least two conundrums. First, for complex decision systems that consider hundreds of variables, it is not clear how to optimally select a set of factors to present in a sensitivity explanation. Even if one desires to present counterfactuals that offer the lowest-cost action or easiest change for the consumer to make, there could be significant heterogeneity with respect to the ease of change for different consumers (Rudin, 2019). Moreover, our data shows that actionability is a key component of consumer-friendly explanations. However, this may not be the only factor. For instance, actionable explanations may also seem more relevant, and this may play a role in consumers' positive responses. Future research might tease apart the possible multi-determinants of consumers' preferences for sensitivity explanations.

Finally, deliberately or not, sensitivity explanations can be misleading, especially if the explanation model is optimized to produce acceptable justifications (Lipton, 2018). Factors that have strongly influenced a decision outcome may be omitted in favor of less important but more palatable factors. Thus, although our research focuses on the immediate impact of explanations on individual attitudes and intentions, it is important to note that increased trust and acceptance of automated systems may have broader societal implications, including legitimate concerns about algorithmic bias and its impact on social and economic inequality. In the long run, awareness of such practices would likely reduce consumer trust instead of increasing it. This has led some observers to argue against the use of black-box models in high-stake decisions (Rudin, 2019), especially if simpler more interpretable models can be used without sacrificing prediction accuracy.

Despite the potential limitations, algorithmic decision systems generally produce superior results to human decision-makers, and the use of properly designed algorithms is, therefore, often in consumers' best interests. Yet mistrust is understandable. The current research highlights the positive role that explanations can play in improving acceptance of algorithmic decisions, especially unfavorable ones, by detailing how the consumer could change their behavior to obtain a different result. This forward-looking approach is constructive and transparent, paving the way for consumers to get the decision outcomes they want and for business to improve customer relations.

Funding: This work was supported by the Social Sciences and Humanities Research Council of Canada (SSHRC) [Insight Grant # 435–2021-0185].

CRediT authorship contribution statement

Mehdi Mourali: Writing – review & editing, Writing – original draft, Project administration, Methodology, Formal analysis, Conceptualization. **Dallas Novakowski:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Ruth Pogacar:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Neil Brigden:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data and analysis scripts are available here: https://osf.io/6vwj3.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbusres.2024.114981.

References

Binns, R, Van Kleek, M., Veale, M., Lyngs U., Zhao, Jun, & Shadbolt, N. (2018), "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions," Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Paper No 377, 1-14. Doi: 10.1145/3173574.3173951.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1). https://doi.org/10.1177/2053951715622512

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research, 56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*(7), 571. https://doi.org/10.1037//0003-066x.34.7.571

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

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. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Eisinga, R., & Grotenhuis, M.t. & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health, 58*, 637–642. https://doi.org/10.1007/s00038-012-0416-3

Grice, H. P. (1975). Logic and conversation. In Speech acts (pp. 41-58). Brill. Doi: 10.1163/9789004368811_003.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19. https://doi.org/10.1037//1040-3590.12.1.19

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin, 107*(1), 65. https://doi.org/10.1037//0033-2909.107.1.65

Hilton, D. J. (1996). Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning, 2*(4), 273–308. https://doi.org/10.1080/135467896394447

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review, 93*(1), 75. https://doi.org/10.1037//0033-295x.93.1.75

Lepri, B., Nuria, O., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. *Philosophy & Technology, 31*(4), 611–627. https://doi.org/10.1007/s13347-017-0279-x

Lepri, B., Staiano, J., Sangokoya, D., Letouzé, E., & Oliver, N. (2017). The tyranny of data? the bright and dark sides of data-driven decision-making for social good. In *Transparent data mining for big and small data* (pp. 3–24). Cham: Springer. https://doi.org/10.1007/978-3-319-54024-5_1.

Leung, E., Paolacci, G., & Puntoni, S. (2019). How technology shapes identity-based consumer behavior. In *Handbook of Research on Identity Theory in Marketing*. Cheltenham, UK: Edward Elgar Publishing. https://doi.org/10.4337/9781788117739.00027.

Lipton, Z. C. (2018). The Mythos of Model Interpretability. *Queue, 16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences, 10*(10), 464–470. https://doi.org/10.1016/j.tics.2006.08.004

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55*(3), 232–257. https://doi.org/10.1016/j.cogpsych.2006.09.006

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Longoni, C., & Cian, L. (2022). Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. *Journal of Marketing, 86*(1), 91–108. https://doi.org/10.1177/0022242920957347

McClure, J. L., Sutton, R. M., & Hilton, D. J. (2003). Implicit and Explicit Processes in Social Judgments: The Role of Goal-Based Explanations. In J. P. Forgas, & K. D. Williams (Eds.), *Social Judgments: Implicit and Explicit Processes* (pp. 306–324). and William Von Hippel, Cambridge University Press. ISBN 0-521-82248-3.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. *University of Minnesota Press*. https://doi.org/10.1037/11281-000

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. arXiv preprint arXiv:1712.00547. Doi: 10.48550/arXiv.1712.00547.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In Proceedings of the conference on fairness, accountability, and transparency, 279-288. Doi: 10.1145/3287560.3287574.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition, 65*, 211–222. https://doi.org/10.1016/j.patcog.2016.11.008

Pasquale, F. (2015), The Black Box Society, in The Secret Algorithms That Control Money and Information. Cambridge, MA: Harvard University Press. Doi: 10.4159/harvard.9780674736061.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham L. Rev., 87*, 1085. https://heinonline.org/HOL/P?h=hein.journals/flr87&i=1119.

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior, 98*(September), 277–284. https://doi.org/10.1016/j.chb.2019.04.019

Slugoski, B. R., Shields, H. A., & Dawson, K. A. (1993). Relation of conditional reasoning to heuristic processing. *Personality and Social Psychology Bulletin, 19*(2), 158–166. https://doi.org/10.1177/0146167293192004

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. https://doi.org/10.1093/idpl/ipx005

Wick, M. R., & Thompson, W. B. (1992). Reconstructive expert system explanation. *Artificial Intelligence, 54*(1–2), 33–70. https://doi.org/10.1016/0004-3702(92)90087-e

Yalcin, G., Lim, S., Puntoni, S., & van Osselaer, S. M. (2022). Thumbs Up or Down: Consumer Reactions to Decisions by Algorithms Versus Humans. *Journal of Marketing Research, 4*, 696–717. https://doi.org/10.1177/00222437211070016

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology, 32*(4), 661–683. https://doi.org/10.1007/s13347-018-0330-6

Mehdi Mourali is a Professor of Marketing at the Haskayne School of Business, University of Calgary.

Dallas Novakowski is a PhD Graduate in Marketing from the Haskayne School of Business, University of Calgary.

Ruth Pogacar is an Associate Professor of Marketing at the Haskayne School of Business, University of Calgary.

Neil Brigden is an Associate Professor of Marketing at the Bissett School of Business, Mount Royal University.