INSEAD The Business School for the World®

Working Paper

2020/39/MKT

Denied by an (Unexplainable) Algorithm: Teleological Explanations for Algorithmic Decisions Enhance Customer Satisfaction

Geoff Tomaino
INSEAD, Geoffrey.tomaino@insead.edu

Hisham Abdulhalim
Ben-Gurion University of the Negev, hisham88il@gmail.com

Pavel Kireyev
INSEAD, pavel.kireyev@insead.edu

Klaus Wertenbroch
INSEAD, Klaus.wertenbroch@insead.edu

Algorithmic or automated decision-making has become commonplace, with firms implementing either rule-based or statistical models to determine whether or not to provide services to customers based on their past behaviors. Policy-makers are pressed to determine if and how to require firms to explain the decisions made by their algorithms, especially in cases where the algorithms are "unexplainable," or are equivalently subject to legal or commercial confidentiality restrictions or too complex for humans to understand. We study consumer responses to goal-oriented, or "teleological," explanations, which present the purpose or objective of the algorithm without revealing its mechanism, making them candidates for explaining decisions made by "unexplainable" algorithms. In a field experiment with a technology firm and several online lab experiments, we demonstrate the effectiveness of teleological explanations and identify conditions when teleological and mechanistic explanations can be equally satisfying. Participants perceive teleological explanations as fair, even though algorithms with a fair goal may employ an unfair mechanism. Our results show that firms may benefit by offering teleological explanations for unexplainable algorithm behavior. Regulators can mitigate possible risks by educating consumers about the potential disconnect between an algorithm's goal and its mechanism.

Keywords: Consumer Psychology; Explanations; Digital Interactions; Costumer Relations; Algorithms

Denied by an (Unexplainable) Algorithm:

Teleological Explanations for Algorithmic Decisions Enhance Customer Satisfaction

πάντες ἄνθρωποι τοῦ εἰδέναι ὀρέγονται φύσει. [All men by nature desire to know.]

Aristotle, *Metaphysics*

With recent developments in artificial intelligence and automation technologies, firms can make complex decisions about customers at scale with increased speed, efficiency, and accuracy. Despite this potential, a major concern with decision-making algorithms is their “black-box” nature (Goodman & Flaxman 2017). While an algorithm may have been created to achieve a clear decision-making objective or goal, the evolving mechanism through which it attempts that goal is often “unexplainable.” That is, the algorithm itself has developed a statistical mechanism to achieve its goal that is beyond reasonable human comprehension, improving its statistical performance at the expense of its ability to transparently relate output decisions to input data. A lack of explainability may also arise because of legal or commercial confidentiality restrictions for otherwise explainable algorithms. We study customer satisfaction with explanations for classification decisions, which, from the customer’s perspective, may have resulted from a complex, possibly unexplainable algorithm or from simple if-then rules that the company did not reveal.

Using unexplainable or black-box algorithms that subject consumers to consequential decisions about differential access to opportunities (e.g., access to financial services, credit scoring, exposure to news and advertising) creates ethical, legal, and technical challenges for firms (Goldfarb & Tucker 2011; Wertenbroch 2019). Reflecting this, regulations are beginning to endow consumers with a right to an explanation of algorithmically-determined decisions. For instance, Articles 13 to 15 of the General Data Protection Regulation (GDPR; Parliament and Council of the European Union 2016) require companies to provide “meaningful information about the logic involved” in automated profiling decisions about consumers. What constitutes such meaningful information that allows firms to comply with current and emerging regulations? To what extent, and why, do consumers value receiving explanations for algorithmic decisions about them? In this research, we address whether and how firms can satisfactorily account for decisions made by these algorithms without explaining the underlying mechanism itself.

To inform the ensuing debate between regulators, firms, and consumers about the right to an explanation, we offer a conceptual framework that distinguishes between different types of explanations and examine the psychological process by which consumers evaluate these when an algorithm denies them a desired opportunity. Importantly, we examine the acceptability of explanations (Fernbach et al. 2013), not their validity (Einhorn & Hogarth 1986). We focus on how consumers perceive and assess explanations rather than the outcomes they account for.

People, from children to scientists to consumers, have an innate desire for explanations (Lear 1988). Philosophers and psychologists have long—since Aristotle—distinguished between mechanistic (or causal) and teleological explanations. A *mechanistic* explanation describes how the parts of a system interact to cause an outcome (Craik 1943; Glennan 1996). *Teleological* explanations explain an outcome in terms of its purpose, that is, what goal an outcome serves (Kelemen 1999; Lombrozo & Carey 2006; e.g., a consumer is told that they are shown an advertisement because the advertiser wants to maximize the consumer's likelihood of purchase). While there are other types of explanations, these are less relevant to consumer contexts. For example, *probabilistic* explanations identify causes based on their covariation with events, but Ahn et al. (1995) showed that people prefer mechanistic explanations to such probabilistic ones. *Formal* explanations explain an exemplar's features by referring to its category membership (Prasada 2017) but rejecting a consumer as a member of a certain category may amount to overt, unethical, and/or illegal discrimination. For these reasons, we focus on mechanistic and teleological explanations as those most relevant to negative algorithmic decisions in practice. These answer two key questions that consumers may ask to assess the fairness of being denied by an algorithm and how to reverse this outcome or prevent it in the future: "how?" (mechanistic) and "why?" (teleological). Naturally, a teleological explanation may prompt consideration of the mechanism, which achieves an outcome's purpose (Liquin & Lombrozo 2018; Lombrozo & Carey 2006; Wright 1976). For instance, stating that a customer was rejected to prevent fraud implies that there was likely some form of mechanism at play that flags fraudulent behavior. However, what is key to the context of unexplainable algorithms is that teleological explanations can be constructed independently of understanding the underlying mechanism, despite sometimes implying one.

Across four experiments, we find that teleological explanations mitigate negative consumer responses even to unexplainable algorithmic decisions. This is because teleological explanations imply that firms aim to treat consumers fairly, provided that the purpose they convey is not unfair. Satisfaction with the explanation for an algorithmic decision may thus also enhance satisfaction with the entire experience. Our primary contribution is to extend the management literature on human interactions with algorithms by examining psychological effects of explaining algorithmic decisions to customers. People often trust algorithmic judgments less than humans (Boatsman et al. 1997; Burton et al. 2019; Dietvorst et al. 2015, 2016; Highhouse 2008; Longoni et al. 2019; Yeomans et al. 2019; but see Logg et al. 2019 for when people trust algorithms more than humans), despite the long-established greater reliability of algorithms (Blattberg & Hoch 1990; Dawes et al. 1989). Unlike these researchers, we do not compare consumer *preferences* for algorithmic versus human predictions but examine consumer *reactions* to varying explanations for algorithmic decisions. Integrating theorizing from cognitive science on the nature of explanations and from behavioral science on fairness, we provide insights for consumer, marketing, technology, and decision researchers on how to manage consumer perceptions of, and reactions to, algorithmic decisions without human involvement. Understanding the psychological process involved in evaluating these different types of explanations will help firms enhance customer satisfaction when customers face negative algorithmic decisions and will help policy makers craft regulations that take account of consumer preferences.

Teleological Explanations Appease because they Convey Algorithm Fairness

Many in the AI community call on companies to offer mechanistically explainable, or "interpretable," algorithms that can reasonably be understood by practitioners and, ideally, by consumers

(Rudin, 2019). In response, fields such as "explainable artificial intelligence" (XAI) and "fair, accountable, and transparent" (FAT) machine learning have emerged to study methods to mechanistically explain black-box systems (Turek 2017; Lepri et al. 2018; Samek et al. 2019). On the other hand, many of the benefits of algorithmic decision-making may arise from deploying mechanistically unexplainable algorithms. Moreover, legal or commercial confidentiality restrictions may prevent firms from revealing the behavior of even simple algorithms.

To help firms resolve this challenge, we propose that mechanistic explanations are not the only type of explanation that can mitigate negative consumer reactions when algorithms subject them to undesirable consequential decisions (e.g., irreversible rejection decisions that deny consumers a coveted opportunity). What drives consumer reactions to these decisions is more nuanced than whether or not the underlying mechanism can be explained. People judge the quality of different types of explanations depending on what they need to infer from an explanation in a given context (Vasilyeva et al. 2017). While explanations may satisfy curiosity (Loewenstein 1994), conversational norms (Hilton 1990), or cognitive scripts that outcomes come with explanations (even if uninformative or vacuous ones; Hemmatian & Sloman 2018; Langer et al. 1978), fairness should be a key concern when algorithmic decisions subject consumers to negative outcomes (e.g., declining financial transactions). We propose that teleological explanations for such algorithmic decisions enhance customer satisfaction because they prompt customers to infer that they have been treated fairly instead of simply addressing curiosity or a norm or script.

Specifically, we hypothesize that teleological explanations of algorithmic decisions that subject customers to negative outcomes (a) can be as effective as mechanistic explanations when the outcomes are irreversible and (b) are more effective than not providing any explanation in improving customer satisfaction and acceptance, both because teleological explanations enhance customer perceptions of being treated fairly. Note that when negative outcomes are reversible, mechanistic explanations that reveal how to reverse these outcomes will be more useful and therefore should be more satisfying.

Our hypothesis rests on two arguments. First, teleological explanations imply the tacit operation of a mechanism to achieve the goal (Wright 1976). This is consistent with findings that people exhibit weaker preferences for mechanistic details in explanations that contain teleological information about the function of particular biological features (Liquin and Lombrozo's 2018). Once the goal is known, knowledge of the precise mechanism to achieve it becomes secondary. Second, to achieve the goal provided by a teleological explanation for an algorithmic decision, the underlying mechanism must be designed to consistently and transparently apply the same set of criteria to each customer (Dawes 1979). If the mechanism applied the criteria inconsistently, arbitrarily, or in a way not well understood, it could not likely achieve its goal. The consistent application of the criteria, in turn, means that each customer is treated according to a fixed reference standard. This ensures a degree of distributive justice in outcomes, provided that the reference standard is consistent with an equitable and fair goal (Kahneman et al. 1985). Consistency and clarity are also hallmarks of procedural justice perceptions in algorithmic decision-making contexts (Binns et al. 2018). Procedural justice, in turn, can mitigate cognitive and behavioral responses to negative outcomes (Hegtvedt et al. 2003; Olson & Hafer 2001; Tyler 2006). Consistent with these arguments, teleological explanations may help legitimatize controversial management decisions in organizational settings (Bobocel & Debeyer 1998; Carter et al. 2020).

To test our hypothesis, we compare participants' reactions to teleological explanations for negative algorithmic decisions about them with those to mechanistic explanations (in the form of simple if-then rules) and/or with those in a baseline condition where no specific explanation is given. One might expect mechanistic explanations to do well in such contexts, creating conservative conditions for testing our hypothesis. First, the mechanistic nature of the process that subjects participants to negative outcomes is highly salient in computerized decision-making so that mechanistic explanations might seem more appropriate and desirable by default. Second, we use simple if-then rules to subject participants to negative decisions to ensure that they can easily understand our mechanistic explanations of these rules. Such if-then rules are also widely used in industry, yet companies often do not explain these rules for reasons of legal or commercial confidentiality. Consumers can therefore not distinguish unexplainable mechanisms from those that companies could, but do not, explain. For instance, Goldman Sachs' Apple

Card uses a simple decision tree rather than a complex machine learning model to approve credit card applicants, yet the firm did not explain the structure of this decision tree even when it was accused of gender discrimination and lack of transparency in its approval process (Horwitz 2020). Our findings should therefore be applicable to both types of algorithmic decisions, those that are genuinely 'unexplainable' and those whose underlying mechanism a company does not explain for other reasons.

We test our hypothesis across four experimental studies, in which we take ratings of customer satisfaction and/or behavioral measures of customer acceptance of algorithmic decisions (e.g., whether participants accept being denied or request a costly follow-up service). In all of our experiments, outcomes are immediate, precluding any notion that a human agent directed an outcome. Experiment 1, a large-scale field experiment with customers of a technology firm, compares the effectiveness of providing a teleological explanation versus no explanation to customers whose purchase requests are declined. We then follow up with three lab experiments modeled on this field experiment. We ran these experiments in an online setting with Amazon MTurk, which lends itself to studying how consumers react to algorithmic decision-making. Experiment 2 again compares the effectiveness of a teleological explanation of a negative outcome with that of no explanation but also with that of a mechanistic explanation. It shows that teleological explanations perform as well as mechanistic explanations unless an understanding of the decision-making mechanism enables consumers to reverse the negative outcome. The study also examines whether satisfaction with the explanation translates into satisfaction with the entire experience, as research in the services marketing literature would suggest (e.g., Bitner 1990; Karatepe 2006). Experiment 3 tests the psychological process underlying the effectiveness of teleological explanations. It shows that teleological explanations perform better because they enhance participants' perceptions of being treated fairly. Experiment 4 demonstrates that the positive effect of a teleological explanation can be substitutable with that of a mechanistic explanation; yet in an important boundary condition, the effect weakens when an algorithmic mechanism is explicitly described as unexplainable and thus cannot be presumed to be fair.

Experiment 1: Appeasing Customers with a Teleological Explanation in the Field

To test our hypothesis that providing a teleological rather than no explanation enhances customer satisfaction and acceptance among customers who have been denied access to desired opportunities by an algorithm, we partnered with a technology firm that offers e-commerce and other services, and processes a large number of purchases every day. We designed a field experiment also to assess the operational implications of our hypothesis for companies that employ algorithms on a large scale. Customers usually make a customer support inquiry when an algorithm blocks a purchase; the absence of such an inquiry constitutes our measure of satisfaction with a teleological explanation for, and acceptance of, the denial. To the best of our knowledge, this study is the first to test the effectiveness of teleological explanations of algorithmic decisions in an actual marketplace setting.

The company allows customers to load cash balances in their accounts and use these balances to complete purchases with third parties, either paying from their accounts in full or partially alongside a different payment method. Customers may be permitted to complete a purchase even if their account balance is negative or falls below zero as a result of the transaction. The company uses a machine learning algorithm trained on historical data to compute a purchase-specific risk score, which determines whether a purchase will be permitted even if the customer has an insufficient balance. Similarly, a purchase may be declined if it has a high risk score even if the customer has a sufficient balance. We focus on a specific subset of declined purchases, which would have been permitted with certainty if customers had met specific conditions. We can identify such cases by focusing on declined purchases for a subset of customers whom the company classifies as "elite users," a select group of customers whose purchases are always approved if the purchase amount does not exceed their account balance or a limited line of credit, regardless of the purchase-specific risk score based on other variables. Therefore, the ultimate mechanism determining the customer outcome in our sample is a combination of the following if-then rules: "If the balance in the customer's account exceeds the purchase amount, then approve the purchase," or "If the purchase amount exceeds the customer's balance by no more than their individual line of credit and the customer's risk score is sufficiently low, then approve the purchase." As a result, it is sufficient for users to update their account balance to complete a purchase if it gets declined, but these "elite users" are not aware of the specific decision rules. From their perspective, the company might be using much more complex rules or even an unexplainable algorithm to make decisions. From the company's perspective, the ultimate goal of the algorithm is to protect the financial well-being of its customers. Communicating this goal as the reason for declining purchases serves as the teleological explanation in our experimental design.

Method

We began with a sample of 16,399 purchases that were declined during an arbitrary 39-hour period in the fall of 2019. The company's algorithm would have approved these purchases if the accounts had had sufficient balances. The company's AB-testing platform allowed us to conduct our experiment only at the purchase level. For privacy, legal, and security reasons, the company only shared partial account identifiers with us. Based on these, we were able to identify 62.8% of the purchases that were made by unique accounts that appeared in the sample only once. We restricted our analyses to these purchases, after applying the experimental treatment, to ensure that no individual customer in our final data set could have experienced both the treatment and the control conditions of our experiment. This yielded a final sample of n = 10,295 customers, each with one purchase.

We employed a two-cell between-participants design. The control condition was the default provided by the company to customers, while every seventh purchase received the treatment explanation. The company restricted the maximum number of customers we were allowed to treat over the course of the experiment. We treated every seventh purchase instance to achieve that limit while ensuring the experiment ran throughout the allotted duration. Customers did not know the order in which they purchased. As a result, the condition to which a customer was assigned was independent of any characteristics that may have affected the outcome. Hence, we can treat the manipulation as causal, even after selecting only those transactions associated with a unique identifier. Additionally, we implement covariate matching in our analysis but find that it does not change any of our findings, reinforcing the orthogonality of the manipulation. In the control condition (*no* explanation given, n = 8,816), the website where customers attempted to complete their purchase stated the default message: "(The company) has blocked this purchase. (The company) blocked the purchase due to customer-related issues."[1] No further explanation was provided. In the treatment condition (*teleological* explanation given, n = 1,479), in addition to the default message, the website stated: "(The company) blocks such purchases to ensure the financial well-being of our customers."[2] After providing these notifications, we recorded whether or not customers made a customer support inquiry (either through an online support center or by directly calling the customer support hotline) and whether they ultimately completed the purchase within 90 hours after the test ended by balancing their account. These two metrics, customer inquiries and completed purchases, serve as the dependent measures in this study, with the absence of a support inquiry as a proxy for acceptance and purchase completion as the functional benefit derived in each condition.

Table 1 summarizes the characteristics of the purchases in the sample across the control and treatment groups. As "elite users," the customers in our sample have used the platform longer, tend to use it more frequently, and have higher purchase amounts than the average user. They typically purchase baskets of items across a wide variety of categories using this platform. We observe only the purchase amounts but do not have access to which items customers purchased. Purchase amounts range from 1¢ cent (typically for paying small fees) to an upper bound of US$1,000, which may correspond to a single expensive purchase or the total price of a basket of items. The company restricted the maximum purchase amount to $1,000 for the purposes of our study before the treatment was applied. The average purchase amount is about US$164. Seniority measures the number of days since the date, on which a customer

[1] We slightly adjusted the wording of the field experiment conditions in this manuscript per the company's request. The adjustments do not affect the semantic meaning of the explanations.
[2] Note that the number of customers in the treatment condition does not exactly equal one-seventh of the total number of customers ($1{,}471 = 1/7 \times 10{,}295$) because we apply the treatment before selecting only those customers who transacted once. This disproportionately removes some transactions in the control group (but to a very minor and insignificant extent).

created their account. Elite users have been with the platform for at least three months. The most senior customer in our sample has been registered with the firm for almost four years at the time of the experiment, the average customer for about two years. These experienced customers may have encountered the same message in the past if they attempted a purchase with an insufficient balance. However, given the uninformative nature of the default message, we do not expect such past experiences to significantly influence customer behavior. Throughout the 39-hour observation period, we observe fewer purchases late at night and early in the morning, as can be expected. Overall, all characteristics are similar across the treatment and control groups, suggesting that the treatment assignment was indeed independent of purchase or customer characteristics.

Table 1. Field Experiment Summary Statistics.

	N	Min	Median	Mean	Max
Control Group	8,816				
Amount (USD)		0.01	61.99	163.82	1000.00
Seniority (Days)		94	675	675	1422
Time of Purchase (Day, Hour:Minute)		1, 09:09	2, 04:19	2, 03:37	2, 23:59
Treatment Group	1,479				
Amount (USD)		0.01	70.72	164.88	1000.00
Seniority (Days)		103	687	678	1406
Time of Purchase (Day, Hour:Minute)		1, 09:12	2, 04:38	2, 03:50	2, 23:59

Results

We estimate a series of linear probability models to study the effect of the teleological explanation on our outcome variables.[3] First, we focus on customer support inquiries. We estimate a regression of the form

$$Y_i = \alpha + \beta T_i + \boldsymbol{\gamma} \boldsymbol{X_i} + \boldsymbol{\delta} \boldsymbol{H_i} + \epsilon_i \qquad (1)$$

where Y_i is equal to 0 if the customer i did not launch a customer support inquiry and 1 if the customer

[3] To enhance readability, we present linear probability models instead of logistic models because the treatment effect coefficient can be directly interpreted as the percentage change in the outcome induced by the treatment. The standard errors are valid given the large sample size of the study. Logistic specifications yielded substantively equivalent findings.

did, α is an intercept term which measures the fraction of customers who initiate an inquiry in the control condition, T_i is equal to 1 if the customer received the teleological explanation and 0 if the customer received the default explanation, β captures the treatment effect, $\boldsymbol{X_i}$ includes amount and seniority as control variables with associated coefficients $\boldsymbol{\gamma}$, $\boldsymbol{H_i}$ is a vector of dummy variable indicators for the day-hour when the transaction by customer i occurred with $\boldsymbol{\delta}$ as the associated set of fixed effects, and ϵ_i is an error term. In principle, the incorporation of controls $\boldsymbol{X_i}$ and time fixed effects $\boldsymbol{\delta}$ should not affect the treatment effect estimate β as treatment assignment was independent of any customer or purchase characteristics. We include these variables to control for their independent effects on the probability of launching a customer support inquiry.

Columns i-iii of Table 2 present estimates from a series of regressions including different sets of variables as controls. The estimate in the first column in the row labelled "Teleological Explanation" shows that providing the explanation caused a significant reduction in the number of customer support inquiries, indicating greater customer acceptance of the algorithm's decision (b = -0.074, 95% CI [-0.080, -0.068], $p < .001$). The estimate can be interpreted as a decrease of 7.4% in the number of customer support inquiries given the regression specification as a linear probability model. This estimate remains stable regardless of whether controls are incorporated in the model (columns ii-iii), consistent with the fact that the treatment was assigned independently of the characteristics. Column iv shows the results of an additional robustness check, in which each treatment observation is first matched to its "closest" control observation before performing the regression (Ho et al. 2007, 2011). Closeness is based on the nearest-neighbor distance between the two observations based on amount, seniority, and time.[4] This approach ensures that any imbalances in the distributions of the characteristics across the treatment and control groups are minimized before the regression is estimated. This approach does not significantly affect our treatment effect estimate. The standard error for the treatment effect increases as the sample

[4] The approach we implement identifies nearest neighbors based on a propensity-score generated through a logistic regression of the treatment indicator T_i on the characteristics $\boldsymbol{X_i}$ and purchase time in seconds. Each treatment observation is matched with a control observation with a similar estimated probability of falling into the treatment group based on the logistic regression.

size shrinks when we use only matched control observations.

Table 2. Customer Support Inquiries.

Dependent Variable:	i	ii	iii	iv
Teleological Explanation	-0.074***	-0.074***	-0.074***	-0.075***
	(0.003)	(0.003)	(0.003)	(0.007)
log(Amount)$\times$ 100		0.082	0.091	0.315
		(0.058)	(0.059)	(0.211)
log(Seniority)$\times$ 100		-0.052	-0.090	0.375
		(0.040)	(0.192)	(0.678)
Hour Fixed Effects	No	No	Yes	Yes
Observations	10,295	10,295	10,295	2,958
R^2	0.074	0.075	0.078	0.087

Note: ***p < .0001.

We do not report the estimate for the intercept term because each and every one of the 8,816 customers in the control condition initiated a customer support inquiry within 90 hours of the purchase blockage. As a result, the intercept is effectively fixed at 1. According to the company, such a 100% inquiry rate is expected from customers classified as elite users. The hassle cost of opening an inquiry is not very large as it involves the press of a button in the website notification of a blocked purchase or, at most, a phone call. Moreover, elite users heavily rely on this platform and may be especially concerned about losing access to their account if they do not know why their purchase was blocked. The company generally receives a very large volume of calls, and one of its business objectives is to find ways to reduce the workload for customer support. It is therefore all the more notable that providing the teleological explanation reduced the customer inquiry rate to 92.6%, a reduction of 7.4%. In total, 110 customers out of 1,479 in the treatment condition refrained from launching an inquiry.

Next, we analyze the time taken by customer support to resolve the inquiries. Table 3 shows the estimates of regressions of the resolution time on characteristics of the purchase and an indicator for whether or not the customer received a teleological explanation. The specification is the same as in Equation 1, except that the outcome variable Y_i now denotes the hours until a customer support representative has closed the inquiry. These regressions are estimated only on those cases, for which a user initiated a support inquiry; they apply to a selected subset of users and should therefore not be interpreted as causal. The estimates in column i suggest that the average inquiry resolution time was 1.94 hours lower ($p < .01$) for customers who received a teleological explanation than the average resolution time of 36.78 hours for customers in the control condition. Columns ii-iii show that purchase and customer characteristics do not significantly predict resolution times of customer support inquiries when they are included as control variables in the model. Column iv focuses only on the matched sample of transactions and reveals no significant changes in the estimates. These drops in inquiries and resolution time point to a significant potential to increase operational efficiency and save costs, or at least to create benefits for customers, although we do not have data on how many man hours they could save the company as we do not know the total number of customer support specialists or what fraction of the resolution time these specialists actually spend working to resolve the issue.

Table 3. Resolution Time for Customer Support Inquiries.

	i	ii	iii	iv
Intercept	36.779***	34.562***	36.766***	33.671***
	(0.216)	(2.574)	(2.785)	(5.319)
Teleological Explanation	-1.940*	-1.940*	-1.930*	-2.014*
	(0.590)	(0.590)	(0.590)	(0.764)
log(Amount)		-0.070	-0.091	0.019
		(0.121)	(0.121)	(0.234)
log(Seniority)		-0.391	0.369	1.215
		(0.393)	(0.394)	(0.754)
Hour Fixed Effects	No	No	Yes	Yes
Observations	10,185	10,185	10,185	2,848
R^2	0.001	0.001	0.006	0.017

Note: ***$p < .0001$, **$p < .001$, *$p < .01$. Resolution time measured in hours. Observations where the customer did not contact support are excluded.

Next, we analyze the impact of providing the teleological explanation on purchase completion. Table 4 summarizes the estimates of several regressions specified as in Equation 1, except that Y_i now equals 1 if the customer completed the purchase by updating her balance and 0 otherwise. Providing a teleological explanation not only reduced the customer support workload, it did so without reducing the purchase completion rate. Based on the estimates in column i of Table 4, without an explanation (control condition; n = 8,816), 45.1% of customers completed the purchase within 90 hours by updating their balance, compared to 46.7% among those who received the teleological explanation (treatment condition; n = 1,479). The difference between the two conditions was not significant (b = 0.016, 95% CI [-0.011 - 0.043], p = 0.25). The incorporation of control variables (columns ii-iii) and an analysis of the matched sample (column iv) did not yield any significant changes in the treatment effect estimate.

Table 4. Purchase Completion.

Dependent Variable:	i	ii	iii	iv
Intercept	0.451***	0.477***	0.174***	0.379***
	(0.005)	(0.063)	(0.066)	(0.126)
Teleological Explanation	0.016	0.016	0.016	0.005
	(0.014)	(0.014)	(0.014)	(0.018)
log(Amount)$\times$ 100		0.310	0.528*	-0.058
		(0.295)	(0.288)	(0.557)
log(Seniority)$\times$ 100		-0.602	-0.648	-2.845
		(0.961)	(0.936)	(1.785)
Hour Fixed Effects	No	No	Yes	Yes
Observations	10,295	10,295	10,295	2,958
R^2	0.000	0.000	0.052	0.054

Note: ***p < .0001, **p < .001, *p < .01. The intercept changes when we incorporate hour fixed effects as it takes on the value of the effect of an arbitrary hour in the sample, when participants appeared to have a lower completion rate. R^2 captures variation explained by the variables excluding the intercept. As these variables have limited significance, the R^2 is very low in columns i and ii and is rounded to 0.

Discussion

These results demonstrate the benefits of providing a teleological explanation for a negative algorithmic decision on customer satisfaction and acceptance in an actual marketplace setting. By offering customers a teleological explanation for why their purchase was rejected, the firm was able to reduce the number of customer support inquiries by 7.4% without negatively affecting purchase completion rates. In contrast, when denied customers received the uninformative default statement instead of an explanation, all 8,816 denied customers initiated an inquiry, suggesting the teleological explanation was uniquely effective beyond simply fulfilling a norm or script that a negative decision be followed by some sort of—even vacuous—statement (Hilton 1990; Langer et al. 1978). The rejection criteria applied by the algorithm implied simple if-then rules that could have easily been communicated in a mechanistic explanation. Oftentimes, though, customers are rejected by algorithms whose underlying mechanism cannot be easily explained or may even be unexplainable. Our results show that firms would do well to offer a teleological explanation instead of an uninformative or no explanation in such cases. They also show that teleological explanations need not be very detailed to be effective. As we will demonstrate in experiments 3 and 4, their efficacy hinges upon the perceived fairness of the explanation.

We do not know whether the number of customer support inquiries drops merely because customers are satisfied with the teleological explanation, or whether that satisfaction also translates into an improved overall customer experience. We examine the link between the two in Experiment 2.

Limitations. Why did we observe the slight yet statistically insignificant *increase* in completed purchases in response to the teleological explanation ("the company declines such purchases to ensure the financial well-being of its customers and merchants")? Perhaps it prompted some customers to guess that their purchase was blocked because of insufficient funds, which they then remedied by balancing their accounts. In other words, our teleological explanation may have inadvertently implied the if-then mechanisms. This would equate the teleological explanation to an indirect mechanistic explanation. That could also explain our main finding, the 7.4% drop in customer support inquiries, our measure of customer acceptance. Customers who guessed why their purchases were declined may not have felt a need to contact customer support. We test this alternative account in an experiment described in Web Appendix A but fail to find significant support for it. Specifically, in this experiment participants were unable to infer the mechanistic explanation from the teleological explanation. Moreover, Experiment 2 will show that teleological and mechanistic explanations have similar mitigating effects on customer satisfaction when customers cannot mechanistically remedy the cause of being denied an opportunity, further evidence of the effectiveness of teleological explanations.

Why do we see purchase completion rates of less than 50% in both conditions, even after customers contact customer support? When a purchase is declined, customers can either transfer funds to their accounts to complete the purchase or switch to an alternative payment means, such as a credit card.

The company only holds part of its customers' overall liquidity so that it is unlikely that purchases remain uncompleted because of liquidity constraints. Our discussions with the company indicated that almost all users eventually complete their purchases, if only with an alternative payment method, but we do not have access to these data. Our measure of purchase completion therefore captures only the effect on a user's decision to use her/his account balance and may understate overall completions. At the same time, our participants are "elite users" who rely heavily on the firm so that we may see higher purchase completion and customer support inquiry rates in our sample than among average users.

Additional findings. In addition to comparing the effect of providing a teleological explanation to that of not providing an explanation, it is also interesting to explore the impact of offering customers a mechanistic explanation. That may not be feasible where purchases are blocked by complex rules, but it is possible where purchases would have been approved with sufficient account balances. Common sense suggests that offering such simple mechanistic explanations should help consumers reverse service denials, yet many firms do not offer these explanations even when they could. An observational data set provided by the company allowed us to study the effect of providing a mechanistic explanation, "Account is restricted because of insufficient funds," on purchase completion rates and customer support inquiries. Consistent with common sense as well as past research on the instrumental value of mechanistic explanations (Lombrozo 2011), pointing customers to how they can resolve purchase blockages significantly increases completion rates and reduces customer inquiries (Web Appendix B). We next explore the effectiveness of both types of explanations under controlled experimental conditions, both when customers can, and cannot, initiate a reversal of the negative decision made by an algorithm.

Experiment 2: Teleological versus Mechanistic Explanations with(out) Instrumental Value

Our field experiment showed that giving denied customers a teleological explanation rather than no explanation increases outcome acceptance, measured by fewer customer support inquiries, without affecting purchase completion rates. Moreover, customers who received a teleological explanation and contacted customer support resolved the issue sooner. In the experiments that follow, we transition to an online laboratory setting where participants recruited through Amazon Mechanical Turk (MTurk)[5] face financial incentives and effort costs, as in our field experiment. Participants are subjected to consequential negative outcomes ostensibly determined by a Qualtrics algorithm, without any indication of human involvement. Our research hypothesis is particularly suited to be tested in such an online environment.

Experiment 2 is designed to conceptually replicate the findings of the field experiment and also to compare the effect of teleological and mechanistic explanations on satisfaction depending on whether participants can act on the explanation, that is, whether it has instrumental value. A key distinction between teleological and mechanistic explanations is their instrumentality. Mechanistic explanations often help users intervene in the process that generates outcomes and thus can be satisfying in a way that teleological explanations cannot (Craik 1943; Lombrozo 2011). However, when an explanation cannot be useful to remedy a problem, such as when a purchase opportunity is irreversibly missed when an algorithm declined the transaction (e.g., an event has sold out), we propose that people are not looking to understand the mechanism as much as wanting to be treated fairly. In such cases, we predict that teleological explanations can be as satisfying as mechanistic ones, while both will be more satisfying than no explanation. A second important factor that we manipulate in Experiment 2 is therefore the instrumental value of the explanation, by giving participants a (second) chance to act upon the explanation. We refer to this as *reversibility*. Third, we examine whether an individual's satisfaction with an explanation for a negative algorithmic decision enhances their satisfaction with the overall experience, an important possible consequence of providing teleological explanations in a business context.

Method

Participants were *n* = 863 (911 before attention-check exclusions) Amazon MTurk workers who

[5] In all online experiments for this research, individuals could only participate if they had not completed a prior experiment conducted for this paper.

received 30¢ as base compensation for their participation. The experimental design, created in Qualtrics, was a 3 (Explanation: teleological, mechanistic, no explanation) × 2 (Outcome reversibility: second chance absent, second chance present) between-participants design. Participants were randomly assigned to one of these conditions.

All participants read that, during the study, some of them would be selected to receive an extra 10¢ bonus and have fewer items to complete in the experiment. They then answered a series of visual perception questions (e.g., "Which of these two lines appears longer to you?")[6]. The last question asked which in a series of six shapes had the largest area. The shapes were designed to make it difficult to assess their areas without the help of measurement tools. Once they had picked a shape, all participants learned that they were not selected to receive the bonus and would be required to perform extra work writing descriptions of various abstract images. From the participants' perspective, the denial decision was made by an algorithm, as in our field study, as the decision was immediate, without room for (human) experimenter involvement. Depending on their experimental condition, they either did, or did not, receive an explanation for this outcome:

- **Teleological**: "This is because the researchers are interested in data from participants with a certain type of visual perception."
- **Mechanistic**: "This is because you did not select the triangle [or the arrow if the participant had selected the triangle] as appearing to have a larger area in the last question."
- **No Explanation**: In this condition participants were not given an explanation for their outcome.

Next, participants either were, or were not, informed that they had an opportunity to change their responses to the previous visual perception questions. The purpose of this second chance was to make it possible for the mechanistic explanation to have instrumental value. If participants with a second chance then picked the triangle (or the arrow if they had initially selected the triangle) in response to the last question, they learned that they would receive the bonus and would not have to perform the extra work,

6 The visual perception questions used for Experiments 2-4 are in Web Appendix C.

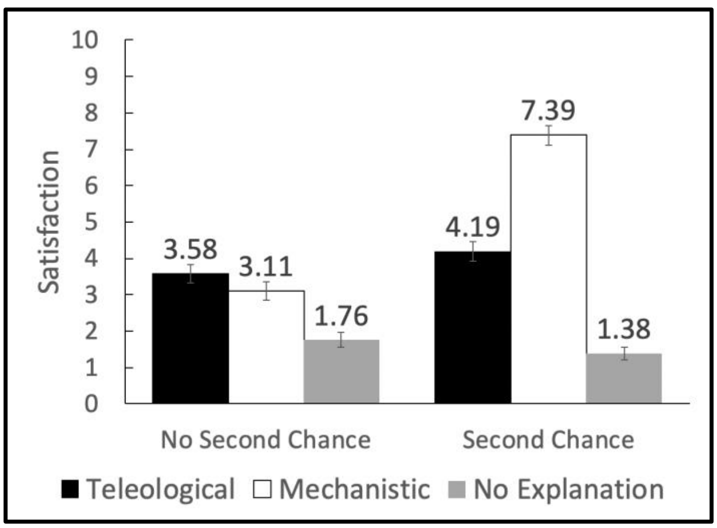
effectively reversing the negative outcome. In the mechanistic condition, 85.53% of participants successfully altered their response, whereas only 22.3% of the participants in the teleological condition changed their selected shape to receive the bonus and avoid the extra work. This suggests that our manipulation of outcome reversibility was successful, that is, that the mechanistic explanation had instrumental value for participants when they had a chance to use it to reverse the negative outcome. This also highlights that teleological and mechanistic explanations are fundamentally different – whereas the teleological explanation provides participants with the goal that their outcome serves, the mechanistic explanation directly reveals the process through which their outcome was reached.

As dependent measures, participants next rated on 11-point Likert scales how satisfying (as our main metric of participant acceptance of their explanation), reasonable (as an alternative measure of satisfaction), and useful (as a check of the instrumental value of the mechanistic explanation when there was a second chance) their assigned explanation was. Participants were shown their respective explanations again while they provided these ratings. Those in the no-explanation condition rated the statement “No explanation available” in place of an explanation. They also rated how satisfied they were with the study itself, as a proxy for overall satisfaction with consumption experiences that have negative outcomes. The scale anchors ranged from 0=“Not satisfied” to 10=“Extremely satisfied.” Then, they were asked to report their gender and age and to identify the financial value of a U.S. penny as an attention-check. Lastly, they were debriefed about the purpose of the study, notified that all of them would be receiving their base pay as well as the bonus, and thanked.

Results

Figure 1 shows the mean ratings of satisfaction across the six conditions. A 3×2 Analysis of Variance (ANOVA) revealed a significant interaction effect of type of explanation and reversibility (F(2, 857) = 52.7; $p < .001$, $\eta^2 = 0.110$). Planned contrasts revealed that, as predicted, participants in the second-chance condition rated the mechanistic explanation as more satisfying (M = 7.39, SD = 3.31) than the teleological explanation (M = 4.19, SD = 3.25; t = 9.47, p < .001, d = 0.98) or no explanation (M = 1.38, SD = 2.16; t = 18.01, p < .001, d = 2.14). Additionally, they rated the teleological explanation as more satisfying than not receiving any explanation (t = 8.23, p < .001, d = 1.02). In contrast, when participants were not given a second chance, so that the mechanistic explanation had no instrumental value, satisfaction with the teleological (M = 3.58, SD = 2.93) and the mechanistic explanations (M = 3.11, SD = 2.97) did not differ (t = 1.37, p = .17, d = 0.16). The slight directional advantage of the teleological explanation in the no-second-chance condition argues against a failure to detect an advantage of the *mechanistic* explanation simply due to insufficient statistical power. Again, also as predicted, not receiving any explanation (M = 1.76, SD = 2.47) generated less satisfaction than receiving either the teleological (t = 5.29, p < .001, d = 0.67) or the mechanistic explanations (t = 3.93, p < .001, d = 0.49).

Figure 1. Satisfaction with Explanation Types in Experiment 2.



Looking across the reversibility conditions, there was a main effect such that participants who had a second chance were more satisfied (M = 4.37, SD = 3.86) than those who did not (M = 2.84, SD = 2.90; $F(1,857) = 59.16$, p < .001, $\eta^2 = 0.065$). This is not surprising as most participants in the mechanistic condition, as well as some in the teleological condition, used the second chance to earn the bonus and avoid the extra work. Participants were marginally more satisfied with a teleological explanation when they were given a second chance (M = 4.19, SD = 3.25) than when they were not (M = 3.58, SD = 2.93; t = 1.81, p = .070, d = 0.20).[7] Most likely, that is because 22.3% of the participants who received a teleological explanation coupled with a second chance gave a response which earned them the bonus. When we exclude these participants from the analysis, there is no longer a significant difference between satisfaction with a teleological explanation when participants were given a second chance (M = 3.7, SD = 3.11) and when they were not (M = 3.58, SD = 2.93; t = .36, p = .722, d =0.04).

To examine more specifically the extent to which the difference between the teleological and mechanistic explanations depends on the instrumentality of the mechanistic explanation, we ran an additional contrast analysis with explanation satisfaction as the dependent variable and explanation type (excluding no explanation), reversibility (presence or absence of a second chance), and the interaction between the two as the independent variables. The interaction effect of explanation type and reversibility was significant, indicating that the difference between the teleological and mechanistic explanations in the second-chance conditions was significantly attenuated in the no-second-chance conditions (t = 7.66, p < .001, η^2 = 0.080).

All remaining dependent measures (explanation reasonableness, explanation usefulness, and overall satisfaction with the study) were likewise consistent with our theory (see Web Appendix D). Importantly, participants' ratings of how satisfied they were with the explanations they were given predict their satisfaction with the study overall (r = .71, p < .001.). Together, these measures show that teleological explanations can be no less satisfying than mechanistic explanations when there is no instrumental value to be derived from an explanation.

[7] There was a significant impact of reversibility in the mechanistic condition (t = 12.79, p <.001, d = 1.36) but not in the no explanation condition (t = 1.09, p = .277, d = 0.16).

Discussion

Replicating the main result of the field experiment, this experiment showed that an explanation for why an algorithm denies someone a desirable opportunity is more satisfying than no explanation. When providing an explanation has no instrumental value for the customer (i.e., the outcome cannot be remedied), a teleological explanation can be no less satisfying than a mechanistic explanation even if the teleological explanation contains no explicit indication of an explainable or fair mechanism that the algorithm uses. That both are also seen as similarly reasonable suggests that teleological explanations of negative outcomes may convey fairness that can make them equally satisfying in algorithmic decision contexts (Bolton, Keh, & Alba 2010). Moreover, satisfaction with the explanation predicts satisfaction with the study, suggesting that teleological explanations can also enhance customer satisfaction with the overall transaction experience. Hence, companies need not provide mechanistic explanations when algorithms irreversibly deny customers desirable opportunities as long as they provide teleological explanations that do not appear unreasonable or unfair.

Is it possible that the teleological explanation performed so well because it actually conveyed just as much mechanistic content as the mechanistic explanation? For example, might our participants have interpreted the teleological explanation mechanistically instead (e.g., "I was rejected because the algorithm determined that I do not have the type of visual perception, for which the researchers are looking."). While teleological explanations may imply the operation of a mechanism (Liquin & Lombrozo 2018; Lombrozo & Carey 2006; Wright 1976), we find no evidence that the teleological explanation provided as much mechanistic content as the mechanistic explanation. In the second chance condition, participants who were given a teleological explanation were much less likely than those given the mechanistic explanation to correctly adjust their responses in the visual perception task.

Experiment 3: Teleological Explanations Convey Fairness

Why are teleological explanations for being denied by an algorithm acceptable? Experiment 3 directly tests the psychological process underlying the beneficial effect of teleological explanations. We hypothesized that teleological explanations for negative outcomes enhance satisfaction in algorithmic decision-making settings because they convey that customers are being treated fairly, according to a reference standard (Kahneman et al. 1986). They imply that the algorithm has been designed not to make arbitrary and inequitable decisions but to consistently apply the same objective function to all consumers (Dawes 1979), whether they provide information about the specific underlying mechanism or not. A potential alternative hypothesis is that teleological explanations are satisfying merely because they fulfill conversational norms or scripts to be given an explanation for a negative outcome (Hilton 1990; Langer et al. 1978), analogous to GDPR establishing a norm of offering explanations. To test these competing hypotheses, we do two things. First, we manipulate the fairness conveyed by teleological explanations to demonstrate that a teleological explanation that conveys an inequitable, unfair objective is less satisfying than one conveying a neutral objective. We compare both to a third condition where no explanation is provided to test for an effect of conversational norms. Second, we show that fairness perceptions mediate the effect of providing a teleological versus no explanation on participant satisfaction (Hayes 2017).

Methodology

Participants were *n* = 582 (603 before attention check exclusions) Amazon MTurk workers who received 30¢ as base compensation for their participation. They were randomly assigned to one of three conditions in a one-factor (Explanation: neutral teleological, unfair teleological, no explanation), between-participants design created in Qualtrics. The neutral teleological explanation was the same as the teleological explanation we had used in Experiment 2.

The procedure was similar to that in Experiment 2. Participants performed several visual perception tasks involving the assessment of geometric figures and were subsequently confronted with an unfavorable outcome, learning that they had missed an opportunity to avoid answering additional questions. Again, the outcome was ostensibly determined by the Qualtrics algorithm, without any indication of human involvement. Unlike in Experiment 2, there was no mention of a bonus that could be withheld. This was to test the sensitivity of our predictions by varying the severity of the consequences of the denial decision. In our previous experiments, participants either had a purchase (Experiment 1) or money and extra effort (Experiment 2) on the line. Here, the negative outcome was the inconvenience of responding to a few extra items to complete the study, similar to many real-world experiences where algorithmic decisions inconvenience consumers (e.g., requiring them to contact customer support). While such inconveniences may seem negligible and undeserving of explanations, we find that there are still benefits to explaining such outcomes.

Participants were then given one of three explanations corresponding to their assigned condition, none of which had any instrumental value for participants as in Experiment 2:

- **Neutral teleological**: "This is because the researchers are interested in data from participants with a certain type of visual perception."
- **Unfair teleological**: "This is because the researchers want to collect more data from certain participants while minimizing the costs of the study."
- **No Explanation**: In this condition participants were not given an explanation for their outcome.

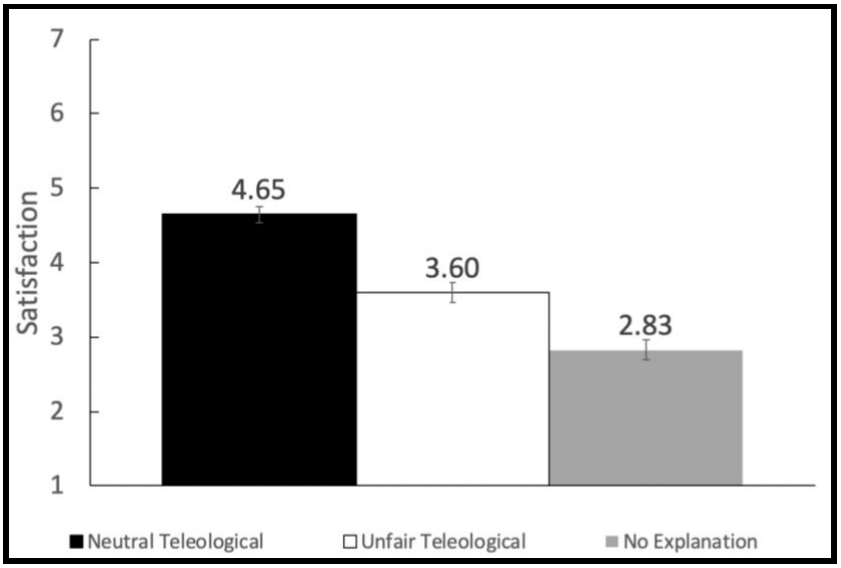
The unfair teleological explanation was designed to violate participants' reference norms of fair exchanges between experimenters and participants, akin to violations of the dual entitlement principle (Kahneman et al. 1985). As the dependent measure, participants then rated their satisfaction with the explanation, which the Qualtrics software had given them to explain why they had to answer additional questions. As in Experiment 2, participants were shown their respective explanations again while they provided these ratings. Those in the no-explanation condition rated the statement "No explanation available" in place of an explanation. Adopting additional measures of perceived fairness from Bolton et al. (2010), participants subsequently also rated how fair, reasonable, and just they felt each explanation was. All of these measures were on 7-point Likert scales, anchored by “Not satisfied / Extremely satisfied,” “Unfair / Fair,” “Not at all just / Just,” and “Unreasonable / Reasonable.” Lastly, participants were debriefed, completed an attention check, reported their age and gender, and were thanked.

Results

We averaged the ratings of the three items to create a fairness index and checked that they were internally consistent (Cronbach’s $\alpha = .95$). As predicted, a one-way ANOVA revealed a significant effect of the explanation type on the fairness index ($F(2, 579) = 75.53$; $p < .001$, $\eta^2 = 0.207$). Planned contrasts revealed that participants found the neutral teleological explanation fairer ($M = 5.15$, SD = 1.51) than the unfair teleological explanation ($M = 3.76$, SD = 1.82; $t = 8.22$, $p < .001$, $d = 0.84$) and also fairer than not receiving an explanation ($M = 3.14$, SD = 1.71; $t = 11.94$, $p < .001$, $d = 1.25$). Getting the unfair teleological explanation was also seen as fairer than not receiving any explanation ($t = 3.60$, $p < .001$, $d = 0.35$). That may be because even the unfair teleological explanation at least satisfies the conversational norm of offering an explanation.

Also, as predicted, the same pattern held for satisfaction ratings ($F(2, 579) = 56.04$; $p < .001$, $\eta^2 = 0.162$). Figure 2 shows the mean ratings of satisfaction across the three conditions. As predicted, participants who had received a neutral teleological explanation for their unfavorable outcome were more satisfied with that explanation ($M = 4.65$, SD = 1.58) than those who received an unfair teleological explanation ($M = 3.60$, SD = 1.82; $t = 6.07$, $p < .001$, $d = 0.62$) and those who received no explanation ($M = 2.83$, SD = 1.77; $t = 10.51$, $p < .001$, $d = 1.09$). Receiving the unfair teleological explanation was also more satisfying than not receiving an explanation ($t = 4.13$, $p < .001$, $d = 0.43$).

Figure 2. Satisfaction with Explanation Types in Experiment 3.



As a test of mediation in Experiment 3, we employed Barron & Kenny's (1986) mediation approach. First, we examined whether the higher satisfaction with the neutral teleological explanation relative to the unfair teleological explanation is mediated by fairness perceptions. We began by regressing explanation satisfaction on condition (= 0 if unfair, = 1 if neutral), which yielded significant results (b = 1.05, 95% CI [0.72, 1.39], p < .001). We also regressed fairness on condition, which yielded significant results (b = 1.39, 95% CI [1.06, 1.72], p < .001). Lastly, we ran a regression with explanation satisfaction as the dependent measure and with both condition and fairness as independent variables. There was a significant relationship between explanation satisfaction and fairness (b = 0.83, 95% CI [0.77, 0.89], p < .001) but not with explanation satisfaction and condition (b = -0.09, 95% CI [-0.31, 0.12], p = .392). This is consistent with full mediation.

Second, we examined whether the higher satisfaction with the neutral teleological explanation relative to no explanation is also mediated by fairness perceptions. We began by regressing explanation satisfaction on condition (= 0 no explanation, = 1 if neutral), which yielded significant results (b = 1.82, 95% CI [1.49, 2.15], p < .001). We also regressed fairness on condition, which yielded significant results (b = 2.02, 95% CI [1.70, 2.33], p < .001). Lastly, we ran a regression with explanation satisfaction as the dependent measure and with both condition and fairness as independent variables. There was a significant relationship between explanation satisfaction and fairness (*b* = 0.81, 95% CI [0.74, 0.87], *p* < .001) but not with explanation satisfaction and condition (*b* = 0.19, 95% CI [-0.05, 0.44], *p* = .122). This is also consistent with full mediation.[8]

Discussion

Experiment 3 shows that teleological explanations can be satisfying in an algorithmic decision-making setting because individuals view them as fair. First, we showed this by demonstrating that explanation fairness mediates the relationship between an explanation and the satisfaction derived from it. Second, by including an unfair teleological explanation condition, we showed that conversational norms or scripts to explain negative outcomes cannot by themselves account for why teleological explanations are satisfying, controlling for this alternative account of the beneficial effect of teleological explanations (Hilton 1990; Langer et al. 1978). That is because the neutral and unfair teleological explanations both conveyed a similar volume and depth of information, yet the neutral teleological explanation was more satisfying. Nevertheless, conversational norms may still have had a residual effect as our participants were more satisfied with the unfair teleological explanation than with no explanation.

Experiment 4: Unexplainable Mechanisms Undermine Teleological Explanations

The purpose of Experiment 4 was to demonstrate that teleological explanations can be used in the absence of mechanistic explainability. Beyond this, it was designed to provide further evidence that teleological explanations appease consumers in an algorithmic decision-making setting by conveying fairness, enhancing satisfaction in the face of negative outcomes. We test these propositions in two ways.

[8] As a robustness check, we reversed the places of fairness and satisfaction in our mediation analyses for Experiment 3 but did not find evidence in favor of this specification (Web Appendix E).

First, if teleological and mechanistic explanations similarly convey fairness, because a teleological explanation implies the tacit operation of a mechanism, then adding a fair mechanistic explanation to a neutral teleological explanation should have no incremental impact on fairness perceptions. However, if teleological explanations convey fairness differently from mechanistic explanations, then a statement that offers both types of explanations should generate greater satisfaction than one that offers just one. Second, if teleological explanations are more satisfying because they convey fairness, even in subjecting consumers to negative outcomes, then adding a mechanistic (quasi-) explanation that explicitly describes the procedure as an unexplainable black-box algorithm should undermine their perceived fairness and ability to enhance satisfaction. That is because consumers cannot presume that an unexplainable or otherwise non-verifiable, complex mechanism ensures a non-arbitrary, fair decision. To illustrate, research in machine learning that has highlighted that algorithms with fair goals can employ unfair mechanisms to achieve these goals (Kleinberg et al. 2018a, b; Obermeyer et al. 2019). We demonstrate how customers may be satisfied with fair goals, even though no evidence of a fair mechanism is provided, as long as they are not given a reason to question that fairness.

Method

Participants were n = 586 (605 before attention-check exclusions) Amazon MTurk workers who received 30¢ for their participation. They were randomly assigned to one of three conditions in a one-factor (Explanation: neutral teleological, explainable mechanistic, unexplainable mechanistic), between-participants design. The neutral teleological explanation was the same we used in Experiments 2 and 3. It was provided in all three conditions, either alone or coupled with an explainable or an unexplainable mechanistic explanation.

The procedure was again similar to that in Experiments 2 and 3. Participants performed several visual perception tasks involving the assessment of geometric figures and were then told they had missed an opportunity to be exempted from extra work. Again, this negative outcome was ostensibly determined by an algorithm, without any indication of human involvement. They were informed that they would have to answer similar visual perception questions and were given one of the following explanations:

- **Neutral teleological**: "The researchers are interested in data from participants with a certain type of visual perception."
- **Explainable mechanistic:** "The researchers are interested in data from participants with a certain type of visual perception. We use an algorithm to select these participants and it chooses those who picked shape X [always one that participants had not selected] as having a larger area."
- **Unexplainable mechanistic:** "The researchers are interested in data from participants with a certain type of visual perception. We select these participants using a complicated black-box algorithm which cannot be explained."

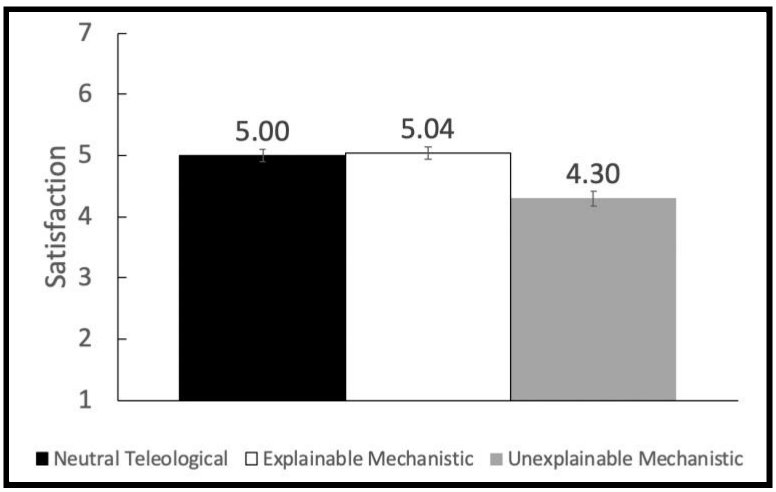
The dependent measures were the same as in Experiment 3. After having read the explanation for why they had been selected to do additional work, participants first rated how satisfied they were with the explanation and subsequently also rated the fairness, reasonableness, and justness of the explanation, all on 7-point Likert scales. Lastly, participants were debriefed, completed an attention check, filled out demographic information, and were thanked.

Results

We again averaged the ratings of the three items to create a fairness index (Cronbach's α = .91). As predicted, a one-way ANOVA revealed a significant effect of the explanation type on the fairness index ($F(2, 583) = 19.2$; $p < .001$, $\eta^2 = 0.062$). Planned contrasts revealed that participants in the unexplainable mechanistic condition rated their explanation as less fair ($M = 4.65$, SD = 1.58) than those in the explainable mechanistic ($M = 5.47$, SD = 1.34; $t = 5.60$, $p < .001$, $d = 0.55$) and neutral teleological condition ($M = 5.39$, SD = 1.31; $t = 5.14$, $p < .001$, $d = 0.51$), whereas fairness ratings between the latter two conditions did not differ (t = 0.55, p = .581, d = 0.06).

Also as predicted, the same pattern held for satisfaction ratings ($F(2, 583) = 13.97$; $p < .001$, η^2 = 0.046). Figure 3 shows the mean ratings of satisfaction across the three conditions. As predicted, participants in the unexplainable mechanistic condition were less satisfied with their explanation (M = 4.30, SD = 1.72) than those in the explainable mechanistic condition (M = 5.04, SD = 1.40; t = 4.67, p < .001, d = 0.48) and those in the neutral teleological condition (M = 5.00, SD = 1.54; t = 4.51, p < .001, d = 0.44), whereas satisfaction ratings between the latter two conditions did not differ (t = 0.24, p = .814, d = 0.02).

Figure 3. Satisfaction with Explanation Types in Experiment 4.



As in Experiment 3, we then tested whether the perceived fairness of the explanations mediated their effect on satisfaction ratings (Barron & Kenny 1986). First, we examined whether the higher satisfaction with the neutral teleological explanation relative to the unexplainable mechanism explanation is mediated by fairness perceptions. We began by regressing explanation satisfaction on condition (= 0 if unexplainable, = 1 if neutral), which yielded significant results (b = 0.71, 95% CI [0.39, 1.03], p<.001). We also regressed fairness on condition, which yielded significant results (b = 0.73, 95% CI [0.45, 1.02], p < .001). Lastly, we regressed explanation satisfaction on both condition and fairness. There was a significant relationship between explanation satisfaction and fairness (*b* = 0.79, 95% CI [0.71, 0.87], $p < .001$) but not with explanation satisfaction and condition (*b* = 0.13, 95% CI [-0.11, 0.37], $p = .275$). This is consistent with full mediation.

Second, we examined whether the higher satisfaction with the explainable mechanism explanation relative to the unexplainable mechanism explanation is also mediated by fairness perceptions. We began by regressing explanation satisfaction on condition (= 0 unexplainable, = 1 if explainable), which yielded significant results (*b* = 0.75, 95% CI [0.43, 1.06], $p < .001$). We also regressed fairness on condition, which yielded significant results (*b* = 0.81, 95% CI [0.52, 1.11], $p < .001$). Lastly, we ran a regression with explanation satisfaction as the dependent measure and with both condition and fairness as independent variables. There was a significant relationship between explanation satisfaction and fairness (*b* = 0.73, 95% CI [0.65, 0.81], $p < .001$) but not with explanation satisfaction and condition (*b* = 0.15, 95% CI [-0.09, 0.39], $p = .212$). This is also consistent with full mediation.[9]

Discussion

Experiment 4 demonstrates that individuals' satisfaction with explanations for a negative outcome can be enhanced similarly when the explanation entails a goal or mechanism. That is because a teleological explanation conveys as much fairness in the decision as does explaining the underlying mechanism. In particular, we found that pairing a fair mechanistic explanation with a fair teleological explanation was as effective as providing the teleological explanation alone. In other words, a teleological explanation is sufficient to convey fairness and boost satisfaction, without directly describing the process mechanistically. However, Experiment 4 also shows that a teleological explanation is seen as less satisfying when it is clear that the mechanism behind the outcome is unexplainable. These findings suggest that teleological explanations can be satisfying when a mechanistic explanation cannot be offered,

[9] As a robustness check, we reversed the places of fairness and satisfaction in our mediation analyses for Experiment 4 but did not find evidence in favor of this specification (Web Appendix F).

such as in a black-box algorithm setting, but will be less satisfying if the unexplainable nature of the mechanism is explicitly noted.

General Discussion

In this research, we provide converging evidence from a field and three (MTurk) lab experiments that teleological explanations convey fairness and thus help appease consumers whom algorithmic decisions subject to differential access to opportunities, such as blocking a desired financial transaction. Additional studies and analyses are presented in the Web Appendix. These positive effects on customer satisfaction with explanations occur despite regulators' and researchers' key concern that the fairness of an algorithm's underlying mechanism may be unverifiable. In our large-scale field experiment, we showed in an actual marketplace setting that teleological explanations can be more satisfying to customers than an uninformative description of a negative outcome, which the firm had previously provided by default (Experiment 1). In particular, the field experiment illustrated the potential for greater operational efficiencies as the teleological explanation reduced the number of customer inquiries that follow when customers are denied access without receiving an explanation they can understand. We then replicated the finding that teleological explanations are more satisfying than no explanations in an online experimental context and also demonstrated that they can be as satisfying as mechanistic explanations (Experiment 2) because teleological explanations are perceived as fair (Experiment 3). Furthermore, our participants perceived a teleological explanation for a negative outcome as equally fair, whether or not it was coupled with a fair mechanistic explanation, but as less fair than when they learned that the underlying mechanism was unexplainable (Experiment 4). Our key theoretical contribution is to demonstrate that providing consumers with an explanation of an algorithm's objective can be sufficient to enhance satisfaction in the face of algorithmic decisions that deny consumers desired opportunities when a firm does not, or cannot, provide a mechanistic explanation. This positive effect of teleological explanations, which is mediated by perceived fairness, weakens when consumers are aware that the underlying algorithm is unexplainable.

Theoretical implications and future research

Our framework integrates work in cognitive science on the nature of explanations with work in behavioral decision research and social psychology on fairness perceptions to help management, marketing, technology, and decision researchers understand how consumers react to decisions made about them by algorithms. We narrowed the scope of our research to explanations of negative outcomes. However, explanations can have different effects depending on whether the outcome they account for is negative or positive (Colquitt & Chertkoff 2002; Shaw et al. 2003). Future research might address which types of explanations, if any, are needed to enhance consumer satisfaction when algorithms decide to grant consumers access to desired opportunities. Customers who get what they want may not require much explaining.

Besides teleological and mechanistic explanations, other types of explanations may also create important dynamics in an algorithmic decision context. For example, *formal* explanations account for properties of an exemplar by referring to the category of which it is a member (Prasada 2017). In an algorithmic decision setting, an example of a formal explanation could be "you are shown this ad because this is an advertising revenue-based website." This type of explanation may satisfy consumers who want explanations out of curiosity rather than to determine the fairness of their outcome or to use the explanation instrumentally. Curiosity is an important driver of the need for explanations (Liquin & Lombrozo 2020). There could also be circumstances under which teleological explanations have less potential to be considered fair. For example, when consumers are highly skeptical of firms' honesty or goals (Forehand & Grier 2003), they may require a verifiable mechanistic account, consistent with our participants' reactions to the black-box explanation in Experiment 4. Future work could explore the effects of providing alternative explanation types when those may be feasible as well as additional boundary conditions of the effect of teleological explanations on customer satisfaction.

Managerial implications

Firms that rely on algorithms to provide differential access to their services without explaining their decisions to their customers face potentially costly operational consequences, as illustrated by the need to field a large number of customer inquiries in our field experiment. They also risk damaging their brands (Carmon et al. 2019). For example, financial service providers often fail to explain their algorithmic decisions when they reject or cancel credit card or other financial transactions. When customers inquire, firms often respond by vacuously declaring the decision the result of firm policy. Customers likely conclude that these providers are not reliable, do not treat their customers fairly and respectfully, and, adding insult to injury, even mock their intelligence and dignity by merely pretending to offer an explanation. Customer-centric firms can mitigate such negative attributions by automatically providing an explanation of the mechanism underlying the algorithmic decision. When the algorithm is unexplainable, or when legal or commercial considerations prevent it, they cannot. Consider a firm that employs an unexplainable algorithm designed to flag fraudulent consumers and run them through additional screening. This firm would be unable to offer flagged consumers a true mechanistic explanation of why they were selected. Given the inconvenience and insinuation associated with being flagged, well-intentioned consumers may find the firm’s decision unacceptable without a fair explanation, affecting their future relationship with the firm (Xia et al. 2004). Our research suggests that firms can enhance customer satisfaction in the face of algorithmic decisions that subject customers to such negative outcomes if they offer teleological explanations, which imply that customers are treated fairly, without revealing the unexplainability of the algorithm. Such teleological explanations may also allow firms to address the regulatory challenges noted at the outset, fulfilling requirements to explain automated decisions to their customers even when the decisions are made by unexplainable algorithms.

Policy implications

Satisfying customers and regulators is not the only challenge in automating decision-making about customers. Is it ethical for firms to provide merely teleological explanations even when providing mechanistic explanations is also feasible or when they deploy algorithms with unexplainable mechanisms? These are pressing questions for policy makers. Ethical treatment of customers may require that firms explain the mechanism underlying an algorithmic decision or, as the case may be, that they note its unexplainability, even if a teleological explanation of a fair goal were otherwise sufficient to merely mollify their customers. That is because teleological explanations of seemingly fair objectives may belie unfair underlying mechanisms. Research shows that algorithms with fair goals, such as predicting successful college applicants, reducing crime, or improving health and well-being, may exhibit significant racial or gender biases (Kleinberg et al. 2018a, b; Obermeyer et al. 2019). For example, Amazon discovered that its experimental hiring system was not rating candidates for software developer jobs in a gender-neutral way (Dastin 2018). Apple's credit card appeared to offer lower credit limits to women (Horwitz 2020). Such cases, which involve important personal consequences, suggest ethical limitations to merely relying on teleological explanations of algorithms, be they unexplainable or not, as those objectives may still be undermined by unfair mechanisms. Another ethical concern is the specificity with which a teleological explanation describes the objective that the algorithm pursues. While providing a teleological explanation may convey a sense of fairness to consumers, as we have shown, some teleological explanations may be so vague that they deceptively placate consumers' desire for fairness without actually explaining the algorithm's objective. Defining criteria for companies to provide teleological explanations in an ethical manner is thus a key challenge for policymakers.

Our results show that teleological explanations of fair goals may offer companies a consumer-centric solution to regulatory requirements to explain decisions by (un)explainable algorithms. Yet, this does not relieve firms of an ethical obligation to track whether an algorithm actually achieves its fair goal. Our work can help policy makers form a more comprehensive view of the types of explanations that companies can use and highlight the potential need to guard against teleological explanations that customers may falsely believe imply overall fairness.

References

Ahn, W K, Kalish C W, Medin D L, Gelman S A (1995) The role of covariation versus mechanism information in causal attribution. *Cognition*. 54(3): 299-352.

Binns R, Kleek M V, Veale M, Lyngs U, Zhao J, Shadbolt N (2018) It's reducing a human being to a percentage: Perceptions of justice in algorithmic decisions. *Proc. 2018 CHI Conf. Hum. Fact. Comp. Sys*.

Bitner M J (1990) Evaluating service encounters: the effects of physical surroundings and employee responses. *J. Marketing*. 54 (2): 69-82.

Blattberg R C, Hoch S J (1990) Database models and managerial intuition: 50% model + 50% manager. *Manage. Sci*. 36(8): 887-1009.

Boatsman J, Moeckel C, Pei B K W (1997) The effects of decision consequences on auditors' reliance on decision aids in audit planning. *Organ. Behav. Hum. Dec*. 71(2): 211–47.

Bobocel D R, Debeyer M (1998) Explaining controversial organizational decisions: To legitimize the means or the ends?. *Soc. Justice. Res*. 11(1): 21–40.

Bolton L E, Keh H T, Alba J W (2010) How do price fairness perceptions differ across culture? *J. Marketing Res*. 47(3): 564–576.

Burton J W, Stein M K, Jensen T B (2019) A systematic review of algorithm aversion in augmented decision making. *J. Behav. Decis. Making*.

Carmon Z, Schrift R, Wertenbroch K, Yang H (2019) "Designing AI systems that customers won't hate" *MIT Sloan Manage. Rev*. December, Reprint :61315.

Carter A B, Bobocel D R, Brockner J (2020) When to explain why or how it happened: Tailoring accounts to fit observers' construal level. *J. Exp. Psychol: Applied*. 26(1): 158–170.

Colquitt J A, Chertkoff J M (2002) Explaining injustice: The interactive effect of explanation and outcome on fairness perceptions and task motivation. *J. Manage*. 28(5): 591–610.

Craik K J W (1943) *The Nature of Explanation* (Cambridge University Press, Cambridge).

Dawes R M (1979) The robust beauty of improper linear models in decision making. *Am. Psychol*. 34(7): 571–582.

Dawes R M, Faust D, Meehl P E (1989) Clinical versus actuarial judgment. *Science*. 243(4899): 1668-1674.

Dastin J (2018) Amazon scraps secret AI recruiting tool that showed bias against women. Retrieved August 10, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Dietvorst B J, Simmons J P, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Exp. Psychol: General*. 144(1): 114.

Dietvorst B J, Simmons J P, Massey C (2016) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Manage. Sci*. 64(3):1155-1170.

Einhorn H J, Hogarth R M (1986) Judging probable cause. *Psychol. Bull*. 99(1): 3-19.

Fernbach P M, Sloman S A, St. Louis R, Shube J N (2013). Explanation friends and foes: How mechanistic detail undermines understanding and preference. *J. Consum. Res*. 39: 1115– 1131.

Forehand M, Grier S (2003). When is honesty the best policy? The effect of stated company intent on consumer skepticism. *J. Consum. Psychol*. 13(3): 349–356.

Glennan S S (1996) Mechanisms and the nature of causation. *Erkenntnis*. 44: 49–71.

Goldfarb A, Tucker C E (2011) Privacy regulation and online advertising. *Manage. Sci*. 57(1): 57-71.

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

Hayes A F (2017) *Introduction to Mediation, Moderation, and Conditional Process Analysis, Second Edition: A Regression-Based Approach* (Guilford Publications, New York).

Hegtvedt K A, Clay-Warner J, Johnson C (2003) The social context of responses to injustice: Considering the indirect and direct effects of group-level factors. *Soc. Justice. Res*. 16(4): 343-66.

Hemmatian B, Sloman S A (2018) Community appeal: Explanation without information. *J. Exp. Psychol: General*. 147(11), 1677-1712.

Highhouse S (2008) Stubborn reliance on intuition and subjectivity in employee selection. *Ind. Organ. Psychol.-US*. 1(3): 333–342.

Hilton D J (1990) Conversational processes and causal explanation. *Pol. Psychol. Bull*. 107(1): 65–81.

Ho D E, Imai K, King G, Stuart E A (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal*. 15(3): 199-236.

Ho D E, Imai K, King G, Stuart E A (2011) MatchIt: nonparametric preprocessing for parametric causal inference. *J. Stat. Softw.* 42(8): 1-28.

Horwitz J (2020) Goldman explains Apple Card algorithmic rejections, including bankruptcies. Retrieved July 07, https://venturebeat.com/2020/07/02/goldman-explains-apple-card-algorithmic-rejections-including-bankruptcies/.

Kahneman D, Knetsch J L, Thaler R (1986) Fairness as a constraint on profit seeking: Entitlements in the market. *Am. Econ. Rev*. 76(4): 728-741.

Karatepe O M (2006) Customer complaints and organizational responses: the effects of complainants' perceptions of justice on satisfaction and loyalty. *Int. J. Hosp. Manag*. 25(1): 69-90.

Kelemen D (1999) Why are rocks pointy? Children's preference for teleological explanations of the natural world. *Dev. Psychol*. 35(6): 1440–1452.

Kim T, Barasz K, John L K (2019) Why Am I Seeing This Ad? The Effect of Ad Transparency on Ad Effectiveness. *J. Consum. Res*. *45*(5): 906–932.

Kleinberg J, Ludwig J, Mullainathan S, Rambachan A (2018a) Algorithmic fairness. *AEA Papers and Proceedings*. 108: 22-27.

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2018b) Human decisions and machine predictions. *Q. J. Econ*. 133(1): 237-293.

Langer E J, Blank A, Chanowitz B (1978) The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *J. Pers. Soc. Psychol.* 36(6): 635-642.

Lear J (1988) *Aristotle: The Desire to Understand* (Cambridge University Press, New York).

Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P (2018) Fair, transparent, and accountable algorithmic decision-making processes. *Philos. Tech*. 31(4): 611-627.

Leung E, Paolacci G, Puntoni S (2018) Man versus machine: Resisting automation in identity-based consumer behavior. *J. Marketing Res*. 55(6): 818-831.

Liquin E G, Lombrozo T (2018) Structure-function fit underlies the evaluation of teleological explanations. *Cognitive Psychol*. 107: 22-43.

Liquin E G, Lombrozo T (2020) A functional approach to explanation-seeking curiosity. *Cognitive Psychol*. Forthcoming.

Loewenstein G (1994) The psychology of curiosity: A review and reinterpretation. *Cognitive Psychol.* 116(1): 75-98.

Logg J M, Minson J A, Moore D A (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organ. Behav. Hum. Dec*. 151: 90–103.

Lombrozo T, Carey S (2006) Functional explanation and the function of explanation. *Cognition*. 99(2): 167–204.

Lombrozo T (2011) The instrumental value of explanations. *Philos. Compass*. 6(8): 539–551.

Longoni C, Bonezzi A, Morewedge C K (2019) Resistance to medical artificial intelligence. *J. Consum. Res*. 46 (December): 629–650.

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 366(6464): 447-453.

Olson J M, Hafer C L (2001) Tolerance of personal deprivation. In J. T. Jost & B. Major (Eds.), *The Psychology of Legitimacy: Emerging Perspectives on Ideology, Justice, and Intergroup Relations* (p. 157–175) (Cambridge University Press, New York).

Parliament and Council of the European Union (2016). *General Data Protection Regulation*: https://eurlex.europa.eu/legalcontent/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG

Prasada S (2017) The scope of formal explanation. *Psychon. B. Rev*. 24 (April): 1478–1487.

Radford A, Wu J, Child R, Luan D, Amodie D, Sutskever A (2019) Language models are unsupervised multitask learners. *OpenAI Blog*. 1.8: 9.

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell*. 1: 206-215.

Samek W, Montavon G, Vedaldi A, Hansen L K, Müller K R (2019) *Explainable AI: interpreting, explaining and visualizing deep learning. LNAI*. Vol. 11700 (Springer Nature, Basel).

Shaw J C, Wild E, Colquitt J A (2003) To justify or excuse?: A meta-analytic review of the effects of explanations. *J. Appl. Psychol.* 88(3): 444–458.

Turek M (2017) Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA).

Tyler T R (2006) Psychological perspectives on legitimacy and legitimation. *Annu. Rev. Psychol*. 57: 375-400.

Vasilyeva N, Wilkenfeld D, Lombrozo T (2017) Contextual utility affects the perceived quality of explanations. *Psychon. B. Rev*. 24 (April): 1436–1450.

Wertenbroch K (2019) "From the editor: A manifesto for research on automation in marketing and consumer behavior". *J. Mark. Behav*. 4 (1): 1-10.

Wright L (1976) *Teleological Explanations: An Etiological Analysis of Goals and Functions* (Univ. of California Press, Berkeley).

Xia L, Monroe K B, Cox J L (2004) The price is unfair! A conceptual framework of price fairness perceptions. *J. Marketing*. 68(4): 1–15.

Yeomans M, Shah A, Mullainathan S, Kleinberg J (2017) Making sense of recommendations. *J. Behav. Decis. Making*. 32(4): 403-414.

Web Appendix

Web Appendix A – Experiment 5: Teleological Explanation Does Not Reveal Mechanism

Experiment 1 not only demonstrated a significant decrease in the number of customer support inquiries by customers in the teleological condition, we also found a small directional increase in purchase completions by customers assigned to the teleological condition. This raises the question of whether the teleological explanation inadvertently provided information to customers that allowed them to guess the mechanistic if-then rule, by which their purchases had been denied (i.e., because of an insufficient account balance). To rule out this alternative account for the drop in the number of support inquiries, we ran a follow-up experiment.

We recruited 455 participants on Amazon Mechanical Turk (MTurk) who indicated that they had experience purchasing goods online. In a one-factor (Explanation: uninformative, teleological, mechanistic) between-participants design, we presented these participants with a brief description of a hypothetical e-commerce company similar to that of Experiment 1. We asked participants to imagine that the hypothetical company had rejected an attempted purchase on the company's website and that the company had given them the following explanation for the rejection, depending on their experimental condition:

- **Uninformative**:[10]

  "(The company) has blocked this purchase. (The company) blocked the purchase due to customer-related issues."

- **Teleological**:

  "(The company) has blocked this purchase. (The company) blocked the purchase due to

[10] We slightly adjusted the wording of these conditions in this manuscript per the company's request. The adjustments do not affect the semantic meaning of the explanations.

customer-related issues. (The company) blocks such purchases to ensure the financial wellbeing of our customers."

- **Mechanistic**:

  "(The company) has blocked this purchase. (The company) blocked the purchase due to customer-related issues. Account is restricted because of insufficient funds."

The uninformative message was the same as the default uninformative message provided by the company in Experiment 1.

Participants were then asked to generate and briefly write down their own mechanistic explanation, for why they thought the transaction had been rejected, and to select which of the six categories given in Table A1 most resembled the mechanistic reason they had provided. The Table summarizes the results of six linear probability models, which we estimated to determine the effect of the experimental conditions on the probability that participants selected each of the six reasons. For each model, we regressed an indicator of the mechanistic reason, which participants had selected, on indicators of which type of explanation they had been given in their respective experimental condition (teleological or mechanistic; the intercept denotes the "uninformative" condition) offered to participants. The parameter estimate for the teleological explanation in the column labeled "Insufficient Funds" is not statistically significant (b = -0.05, 95% CI [-0.15, 0.05], p = .371), indicating that participants were not more likely to infer that insufficient funds were the reason for the rejection from the teleological explanation (49.6%) relative to the uninformative control condition (54.2%). That is, the teleological reason, which we had also provided in in Experiment 1, did not provide any information to participants that their transactions had been denied because of insufficient funds in their accounts, ruling out this alternative account for the drop in the number of support inquiries in Experiment 1.

Table A1. Effects of Explanation Type of Inferred Reason for Denial.

Dependent Variable:	Insufficient Funds	Terms Violation	Suspicious Activity	Incorrect Currency	Merchant Error	Spending Limit
Intercept	0.542***	0.105***	0.386***	0.098***	0.327***	0.124***
	(0.036)	(0.025)	(0.037)	(0.021)	(0.035)	(0.027)
Mechanistic	0.354***	-0.021	-0.199**	-0.053*	-0.172**	-0.008
	(0.051)	(0.035)	(0.052)	(0.030)	(0.049)	(0.037)
Teleological	-0.046	0.031	0.193**	-0.023	-0.021	-0.002
	(0.051)	(0.036)	(0.053)	(0.030)	(0.050)	(0.038)
Observations	455	455	455	455	455	455
R^2	0.141	0.005	0.108	0.007	0.031	0.000

Note: ***p < .0001, **p < .001, *p < .01.

Additionally, we performed a qualitative text analysis of the words used by participants in their free response guess of the mechanistic reason. Table A2 summarizes the frequencies of the top words used by participants in the different conditions, excluding a host of common but uninformative words such as "I" and "the." Naturally, participants used the smallest number of unique words (315) in the mechanistic condition, and the most common words account for a larger share of all words used than in the other conditions. This suggests that participants in the mechanistic condition were the most certain of the mechanistic reason as they offered more consistent responses. Participants generated a larger number of unique words in the uninformative (442) and the teleological (465) conditions. The total number of words and the distributions of the most common words are similar across these two conditions, suggesting that participants were not more certain about the mechanistic reason in the teleological condition than in the uninformative condition.

Table A2. Word Frequency by Condition.

Explanation Type	Uninformative	%	Teleological	%	Mechanistic	%
	account	3.5	account	3.7	account	9.4
	funds	3.1	maybe	2.5	enough	5.2
	enough	2.7	funds	2.4	money	4.9
	money	2.0	may	2.3	funds	4.0
	insufficient	1.7	enough	2.1	balance	3.3
	card	1.6	transaction	2.0	negative	3.3
	may	1.6	money	1.7	bank	1.5
	maybe	1.6	insufficient	1.6	may	1.5
	information	1.2	payment	1.3	insufficient	1.3
	payment	1.2	wallet	1.2	wallet	1.1
Unique Words	442		465		315	

These results suggest that the slight increase in completed transactions and the decrease in customer support inquiries, which we observed in Experiment 1, did not occur because participants were able to guess the mechanistic reason from the teleological explanation. To maximize the similarity between the participants in this post-test and those in Experiment 1, we selected participants with experience using e-commerce services for the post-test. We find that experience with such services does not appear to lead to an increased chance of guessing the mechanistic reason from the teleological explanation.

Web Appendix B – Observational Analysis of a Mechanistic Explanation in the Field Setting

We obtained an observational dataset with n = 49,999 observations from the e-commerce company, through which we had run Experiment 1, to study the impact of providing mechanistic explanations on purchase completions and customer support inquiries in our field setting. In contrast to Experiment 1, this analysis focuses on customers from different geographic regions who may use the platform less frequently than the elite users selected in Experiment 1. The accounts are more junior and make larger purchases as this sample may also include small business accounts in addition to individual customers. The company assured us that for this group of accounts with declined purchases, the purchases would have gone through if they had amended their balance. However, we were unable to obtain additional information on how these accounts were selected or from which geographic regions they originated because of privacy restrictions. The company provided the following mechanistic explanation to customers who made 375 transactions that were declined during a 24-hour period: “(The company) has blocked this purchase. (The company) blocked the purchase due to customer-related issues. Account is restricted because of insufficient funds.” All other declined transactions in the dataset during this 24-hour period received the same default message as in Experiment 1. The assignment of the mechanistic explanation appeared to be evenly spaced throughout the day, such that the time intervals between blocked purchases that received the mechanistic explanation would be identical. However, the company did not track a control group but rather provided us with a large set of blocked purchases that received the default explanation from the same day, which may have included other types of users as the distribution of the characteristics we observed differed across the two groups. We used nearest-neighbor covariate matching (Ho et al. 2007, 2011) based on transaction time, amount, and account seniority to identify one comparable observation from the set of purchases receiving the default explanation for each one of the blocked purchases receiving a mechanistic explanation, which yielded a final sample of *N*=750 transactions. Our findings may not have a causal interpretation as we may be omitting key variables that affected the treatment assignment and also the outcome and should be interpreted subject to this constraint. We use the first few digits of account identifiers provided to us by the company to confirm that at most three customers in the resulting sample may have purchased twice, and the remaining purchases corresponded to unique customers. Given this small number, we retain all purchases in the analysis. Table B1 summarizes the characteristics of purchases in the observational study. Overall, the distribution of characteristics appears stable across the control and treatment groups suggesting that our matching procedure effectively identified the most similar observations in the control group for each treatment observation.

Table B1. Summary Statistics for Observational Analysis.

	N	Min	Median	Mean	Max
Control Group	375				
Amount (USD)		90.80	544.70	546.97	990.90
Seniority (Days)		91	246	353	1013
Time of Purchase (Hour:Minute)		01:06	11:49	11:54	23:56
Treatment Group	375				
Amount (USD)		91.80	539.30	541.07	989.60
Seniority (Days)		100	241	353	1000
Time of Purchase (Hour:Minute)		01:02	12:26	12:10	23:57

We estimate the regression specification in Equation 1 from the main text on the matched sample. Table B2 shows estimates of the impact of the mechanistic explanation on purchase completion, and Table B3 shows estimates of the impact on customer support inquiries, both within 44 hours of the blocked purchase. In both tables, the first column shows the estimates for the matched sample without any additional controls, the second column introduces amount and seniority as controls, and the third column adds hourly fixed effects. We find that the additional controls do not change the findings and focus on the results in the first column of each table. In Table B2, we see that 47.2% of all purchases are completed in the default condition and 92% are completed in the mechanistic explanation condition. The difference between the two conditions is statistically significant (b = 0.448, 95% CI [-0.391, -0.505], p < 0.001). In Table B3, we see that 74.9% of the blocked purchases lead to a customer support inquiry in the default condition, and this number falls to 38.6% in the mechanistic explanation condition. Again, the difference across conditions is statistically significant (b = -0.363, 95% CI [-0.430, -0.296], p < 0.001). These results point to the potentially strong effect of the mechanistic explanation on both increasing purchase completion and reducing company workload, subject to the assumption that we adequately control for differences in characteristics across the two groups. This effect is largely expected, as mechanistic explanations are known to be instrumental and provide very clear guidance to customers on how they can reverse the algorithm's decision.

Table B2. Purchase Completion.

Dependent Variable:	i	ii	iii
Intercept	0.472***	0.293	0.198
	(0.021)	(0.195)	(0.216)
Mechanistic Explanation	0.448***	0.448***	0.448***
	(0.029)	(0.029)	(0.030)
log(Amount)		-0.014	-0.011
		(0.025)	(0.025)
log(Seniority)		0.047^^	0.049^^
		(0.022)	(0.023)
Hour Fixed Effects	No	No	Yes
Observations	750	750	750
R^2	0.237	0.242	0.255

Note: ^*p* < .1, ^^*p* < .05, **p* < .01, ***p* < .001, ****p* < .0001.

Table B3. Customer Support Inquiries.

Dependent Variable:	i	ii	iii
Intercept	0.749***	0.085	0.031
	(0.024)	(0.225)	(0.246)
Mechanistic Explanation	-0.363***	-0.363***	-0.361***
	(0.034)	(0.034)	(0.034)
log(Amount)		0.007	0.019
		(0.028)	(0.029)
log(Seniority)		0.022	0.027
		(0.025)	(0.026)
Hour Fixed Effects	No	No	Yes
Observations	750	750	750
R^2	0.134	0.135	0.167

Note: ^*p* < .1, ^^*p* < .05, **p* < .01, ***p* < .001, ****p* < .0001.

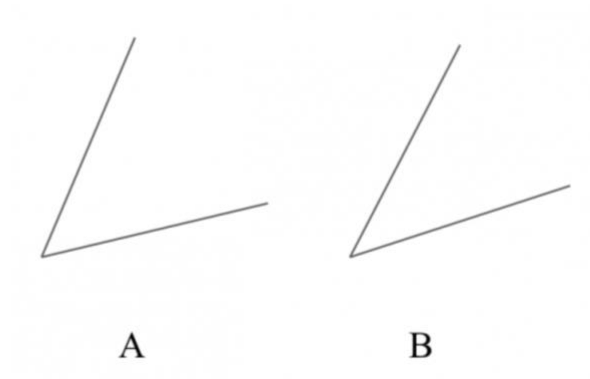
We expect that fewer customers initiate a customer support inquiry in this data set compared to Experiment 1 for several reasons. First, this sample includes less frequent users who may worry less about not being able to execute a transaction in the future compared to top users. Second, the dependent variables were recorded within 44 hours of the final blocked purchase in this case, compared to 90 hours in Experiment 1. Finally, users from different geographic regions with varying user interfaces are included in this dataset, making it potentially more difficult for some users to open a support inquiry. Overall, this observational analysis suggests that a mechanistic explanation may have a very strong effect on both purchase completion and customer support inquiries, supporting the findings of past research on the instrumentality of mechanistic explanations (Lombrozo 2011).

Web Appendix C – Visual Perception Tasks: Experiments 2 -4

Experiments 2-3

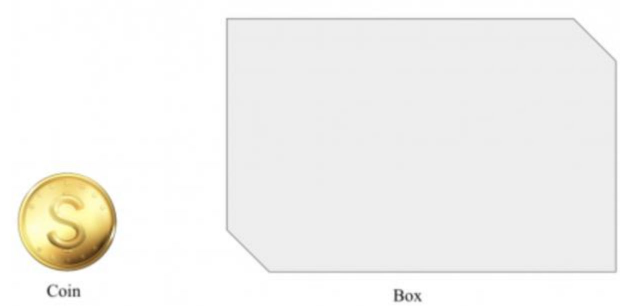
Item 1

Which of these two angles appears larger to you?



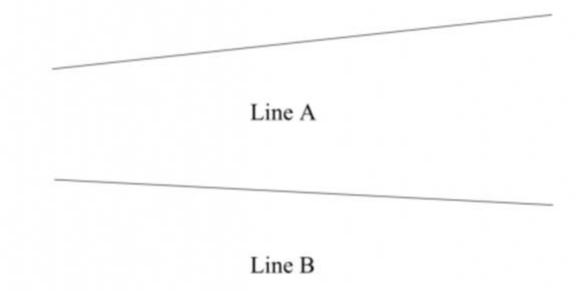
Item 2

How many of these coins do you think could fit into this box?



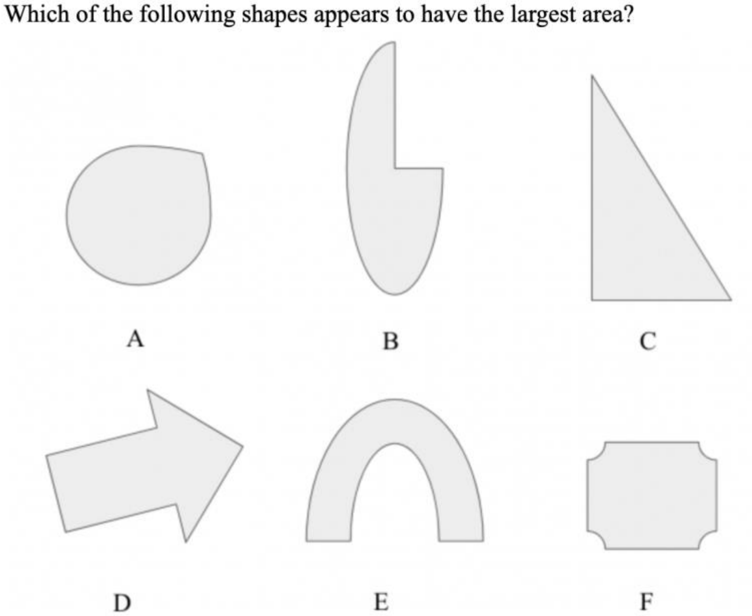
Item 3

Which of these two lines appears longer to you?



Item 4

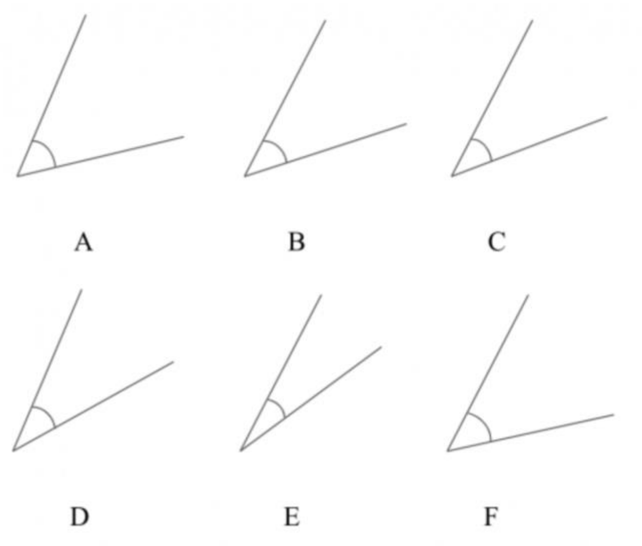
Which of the following shapes appears to have the largest area?



Experiment 4

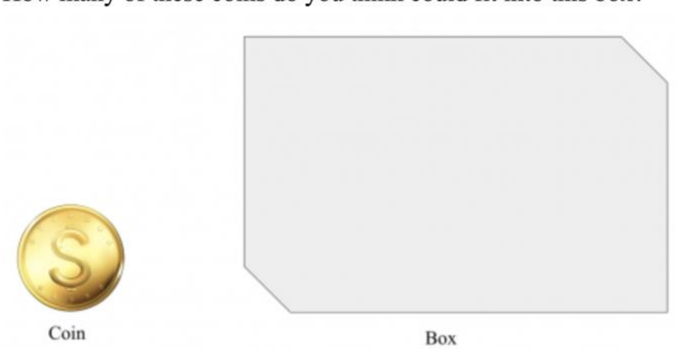
Item 1

Which of these angles appears largest to you?



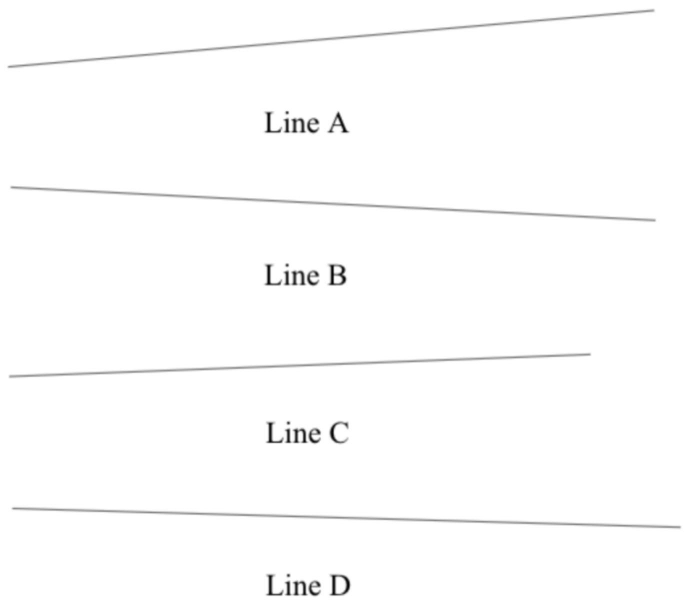
Item 2

How many of these coins do you think could fit into this box?



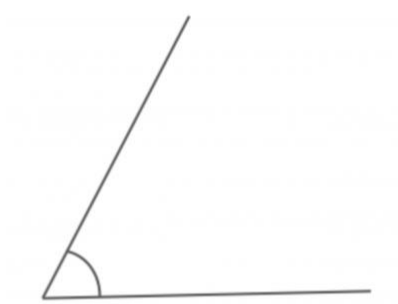
Item 3

Which of these lines appears longest to you?



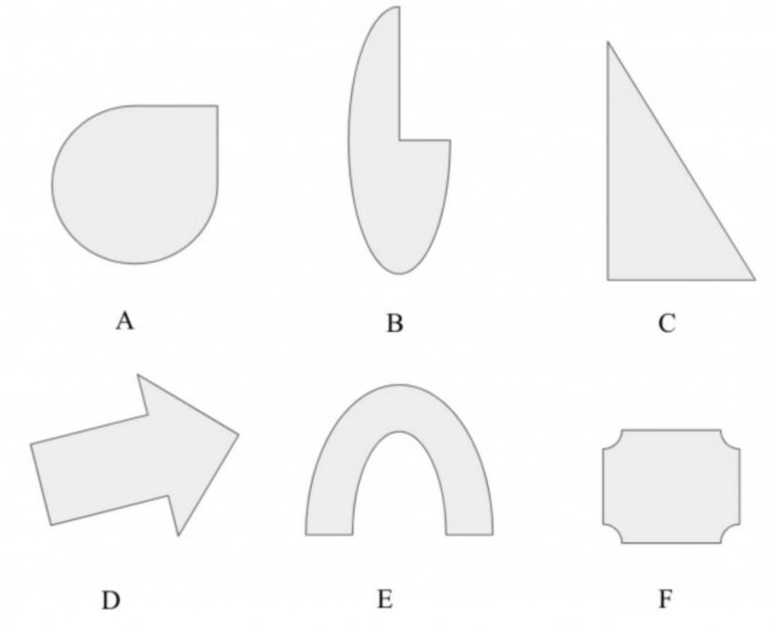
Item 4

How many degrees does this angle appear to be?



Item 5

Which of the following shapes appears to have the largest area?



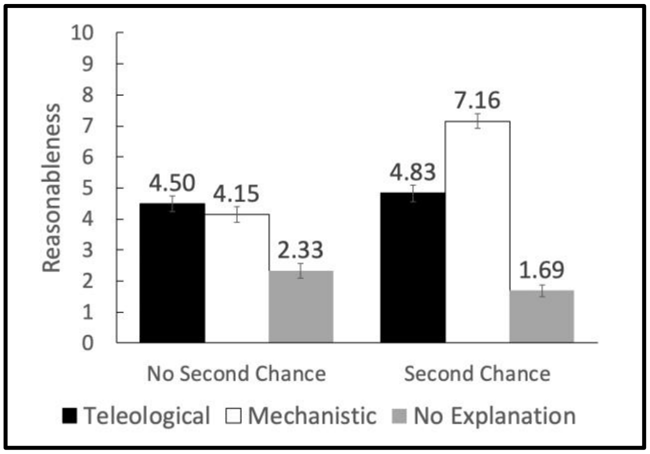
Web Appendix D – Analyses of the Additional Dependent Measures in Experiment 2

This appendix provides analyses of the additional dependent measures we took in Experiment 2. We report results for each measure depending on explanation type and on the reversibility of the negative outcome, that is, whether participants had a second chance to undo the negative outcome.

Ratings of the Reasonableness of Each Explanation

Mean ratings of the reasonableness of each explanation depending on the reversibility of the negative outcome are given in Figure C1. A 3×2 ANOVA of the effects of explanation type and reversibility revealed a significant interaction ($F(2, 857) = 31.34$, $p < .001$, $\eta^2 = 0.068$).

Figure C1. Explanation Reasonableness.



No-Second-Chance Condition. Planned contrasts revealed that participants in the teleological explanation condition did not provide different reasonableness ratings (M = 4.50, SD = 2.96) than those in the mechanistic explanation condition (M = 4.15, SD = 3.03; t = 1.05, p = .295, d = 0.12) but thought their explanation was more reasonable than those in the no explanation condition (M = 2.33, SD = 2.70; t = 6.32, p < .001, d = 0.76). Participants in the mechanistic condition also rated the explanation as more reasonable than those in the no explanation condition (t = 5.27, p < .001, d = 0.63).

Second-Chance Condition. Participants in the mechanistic explanation condition rated their explanation as more reasonable (M = 7.16, SD = 2.99) than those in the teleological explanation condition (M = 4.83, SD = 3.15; t = 6.88, p < .001, d = 0.76) and those in the no explanation condition (M = 1.69, SD = 2.37; t = 16.39, p < .001, d = 2.02). Additionally, participants in the teleological explanation condition rated their explanation as more reasonable than those in the no explanation condition (t = 9.21, p < .001, d = 1.13).

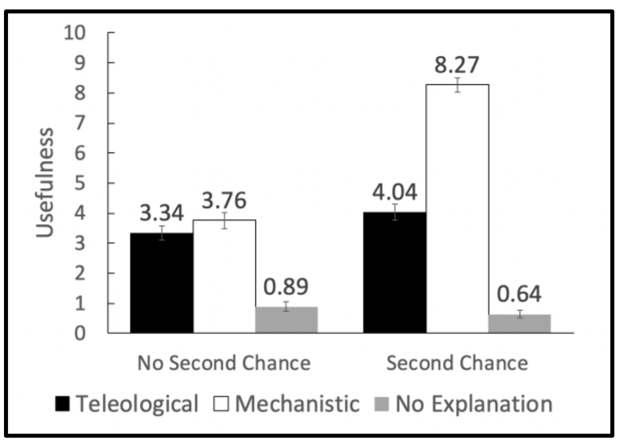
Impact of Second Chance. Planned contrasts revealed that there was no significant impact of participants being given a second chance on explanation reasonableness perceptions in the teleological condition (t = 0.98, p = .327, d = 0.11). There was, however, a significant positive impact in the mechanistic condition (t = 9.00, p < .001, d = 1.00) and a marginally significant negative impact in the no explanation condition (t = 1.86, p = .063, d = 0.25).

Interaction. Contrasts additionally revealed an interaction between explanation type (excluding the no-explanation condition) and reversibility, indicating that the difference in explanation reasonableness between the teleological and mechanistic conditions seen in the second-chance condition was significantly attenuated in the no-second-chance condition (t = 5.60; p < .001, η^2 = 0.047).

Ratings of the Usefulness of Each Explanation

Mean ratings of the usefulness of each explanation depending on the reversibility of the negative outcome are given in Figure C2. A 3×2 ANOVA of the effects of explanation type and reversibility revealed a significant interaction ($F(2, 857)$ = 64.21; p < .001, η^2 = 0.130).

Figure C2. Explanation Usefulness.



No-Second-Chance Condition. Planned contrasts revealed that participants in the teleological explanation condition did not provide different usefulness ratings (M = 3.34, SD = 2.83) than those in the mechanistic explanation condition (M = 3.76, SD = 3.31; t = 1.32, p = .189, d = 0.13) but thought their explanation was more useful than those in the no explanation condition (M = 0.89, SD = 1.90; t = 7.67, p < .001, d = 1.01). Participants in the mechanistic condition also rated the explanation as more useful than those in the no explanation condition (t = 8.94, p < .001, d = 1.06).

Second-Chance Condition. Participants in the mechanistic explanation condition rated their explanation as more useful (M = 8.27, SD = 2.87) than those in the teleological explanation condition (M = 4.04, SD = 3.11; t = 13.45, p < .001, d = 1.41) and those in the no explanation condition (M = .64, SD = 1.59; t = 24.54, p < .001, d = 3.27). Additionally, participants in the teleological explanation condition rated their explanation as more useful than those in the no explanation condition (t = 10.67, p < .001, d = 1.38).

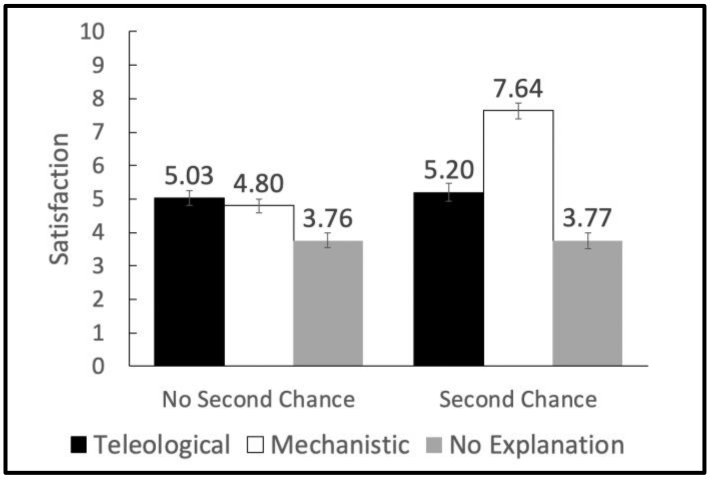
Impact of Second Chance. Planned contrasts revealed that participants rated the teleological explanation as more useful when they had a second chance compared to when they did not (t = 2.18, p = .029, d = 0.23). As with our measure of satisfaction with the study, this effect is probably driven by those participants who happened to successfully adjust their responses in the teleological condition when they were given a second chance and then attributed their success to the teleological explanation. There was also a significant positive impact in the mechanistic condition (t = 14.47, p < .001, d = 1.46). There was no significant impact in the no explanation condition (t = .77, p = .442, d = 0.14).

Interaction. Contrasts additionally revealed an interaction between explanation type (excluding the no-explanation condition) and reversibility, indicating that the difference in explanation usefulness ratings between the teleological and mechanistic conditions seen in the second-chance condition was significantly attenuated in the no-second-chance condition t = 8.58; p < .001, η^2 = 0.091).

Ratings of Satisfaction with the Study

Mean ratings of the satisfaction with the study depending on explanation type and reversibility of the negative outcome are given in Figure C3. A 3×2 ANOVA of the effects of explanation type and reversibility revealed a significant interaction (F(2, 857) = 24.65; p < .001, η^2 = 0.054).

Figure C3. Study Satisfaction.



No-Second-Chance Condition - Participants in the teleological explanation condition did not provide different study satisfaction ratings (M = 5.03, SD = 2.53) than those in the mechanistic explanation condition (M = 4.80, SD = 2.50; t = 0.73, p = .463, d = 0.09) but thought the study was more satisfying than those in the no explanation condition (M = 3.76, SD = 2.61; t = 3.92, p < .001, d = 0.50). Participants in the mechanistic condition also rated the study as more satisfying than those in the no explanation condition (t = 3.19, p = .001, d = 0.41).

Second-Chance Condition - Participants in the mechanistic explanation condition rated the study as more satisfying (M = 7.64, SD = 2.83) than those in the teleological explanation condition (M = 5.20, SD = 3.02; t = 7.60, p < .001, d = 0.83) and those in the no explanation condition (M = 3.77, SD = 2.87; t = 12.22, p < .001, d = 1.36). Additionally, participants in the teleological explanation condition rated the study as more satisfying than those in the no explanation condition (t = 4.43, p < .001, d = 0.49).

Impact of Second Chance. Planned contrasts revealed that there was no significant impact of participants being given a second chance on study satisfaction in the teleological condition (t = 0.52, p = .606, d = 0.06) and also not in the no explanation condition (t = 0.03, p = .976, d = 0.004). There was, however, a significant positive impact in the mechanistic condition (t = 8.93, p < .001, d = 1.06).

Interaction. Contrasts additionally revealed an interaction between explanation type (excluding the no-explanation condition) and reversibility, indicating that the difference in study satisfaction ratings between the teleological and mechanistic conditions seen in the second-chance condition was significantly attenuated in the no-second-chance condition $t = 5.89$; $p < .001$, $\eta^2 = 0.057$).

Web Appendix E – Additional Analyses of Mediation by Fairness in Experiment 3

We conducted the same mediation analyses from Experiment 3 (Barron & Kenny 1986) but with satisfaction as the mediating variable and fairness as the dependent variable. The first two steps in this approach are regressing the dependent variable on the independent variable and the mediating variable on the independent variable. Since these regressions are already reported in the main text, we only report here the final regressions of the dependent variable (fairness) on the independent variable (explanation type), controlling for the mediating variable (satisfaction). When comparing the neutral teleological to the unfair teleological explanation (= 0 if unfair, = 1 if neutral) we found significant coefficients for both explanation ($b = 0.56$, 95% CI [0.35, 0.76], $p < .001$) and satisfaction ($b = 0.79$, 95% CI [0.73, 0.85], $p < .001$). When comparing the neutral teleological explanation to no explanation (= 0 if no explanation, = 1 if neutral) we found significant coefficients for both explanation ($b = 0.66$, 95% CI [0.43, 0.88], $p < .001$) and satisfaction ($b = 0.75$, 95% CI [0.69, 0.81], $p < .001$). Neither of these results are consistent with full mediation.

Web Appendix F – Additional Analyses of Mediation by Fairness in Experiment 4

We again used the mediation approach from Barron & Kenny (1986) to test whether a reversal of our main model holds with satisfaction as the mediating variable and fairness as the dependent variable. As with Experiment 3, we focus here on the final regression in this mediation approach of the dependent variable (fairness) on the independent variable (explanation type) controlling for the mediating variable (satisfaction). When comparing the neutral teleological to the unexplainable mechanism explanation (= 0 if unexplainable, = 1 if neutral) we found significant coefficients for both explanation (b = 0.29, 95% CI [0.08, 0.50], p = .007) and satisfaction (b = 0.62, 95% CI [0.56, 0.69], p < .001). When comparing the explainable mechanism explanation to the unexplainable mechanism explanation (= 0 if unexplainable, = 1 if explainable) we found significant coefficients for both explanation (b = 0.33, 95% CI [0.11, 0.56], p = .003) and satisfaction (b = 0.64, 95% CI [0.57, 0.71], p < .001). Neither of these results are consistent with full mediation.

Appendix References

Baron R M, Kenny D A (1986) The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *J. Pers. Soc. Psychol.* 51: 1173-1182.

Ho Daniel E, Imai K, King G, Stuart E A (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal*. 15(3): 199-236.

Ho Daniel E, Imai K, King G, Stuart E A (2011) MatchIt: nonparametric preprocessing for parametric causal inference. *J. Stat. Softw.* 42(8): 1-28.

Lombrozo T (2011) The Instrumental Value of Explanations. *Philos. Compass*. 6(8): 539-551.