Good Explanation for Algorithmic Transparency

Joy Lu
Assistant Professor of Marketing
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
412-268-5162
tonglu@andrew.cmu.edu

Dokyun Lee
Assistant Professor of Business Analytics
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
412-268-3702
dokyun@cmu.edu

Tae Wan Kim
Associate Professor of Business Ethics
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
412-268-3703
twkim@andrew.cmu.edu

David Danks
L.L. Thurstone Professor of Philosophy and Psychology
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
412-268-8047
ddanks@cmu.edu

Good Explanation for Algorithmic Transparency

ABSTRACT

Machine learning (ML) algorithms have gained widespread usage across a variety of domains, both in providing predictions to expert users and recommending decisions to everyday users. End-users are rarely provided with an explanation of the algorithmic output, and the critical need for explanation in artificial intelligence (AI) has led to calls for algorithmic transparency, including the "right to explanation" in the EU General Data Protection Regulation (GDPR), which requires many companies to provide a meaningful explanation to involved parties. However, the research on what constitutes a meaningful or good explanation from an end-user's perspective has been limited. In this paper, we (1) develop a generalizable framework grounded in philosophy, psychology, and interpretable machine learning to investigate and define the characteristics of a good explanation for ML decisions, and (2) conduct large-scale studies to measure the impact of different factors on perceptions of understanding, fairness, and trust in the context of credit loan and credit card application decisions. The framework and studies together form a concrete guide for managers to present algorithmic prediction rationales to end-users to foster trust and adoption.

Keywords: artificial intelligence, financial decision making, consumer knowledge

INTRODUCTION

Machine learning (ML) algorithms and other forms of artificial intelligence (AI) systems are rapidly gaining important roles in everyday decisions for both consumers and managers. For example, AI-powered systems give product recommendations on e-commerce websites, while banks are increasingly using ML to make large-scale decisions, including credit scoring, fraud detection, and loan approvals.

The most successful algorithms in use today are extremely complex. Different varieties of deep neural network models (e.g., CNN, LSTM, etc.) have been shown to excel across a variety of domains, including finance, medicine, and law (LeCun, Bengio, and Hinton 2015). Ensemble models such as boosted trees (Chen and Guestrin 2016) often dominate data science competitions. These models are fundamentally "black boxes" with many hidden layers of nonlinear transformations. It can be quite difficult for anyone – including developers and end-users – to understand the underlying reasons for an algorithm's output.

There are many reported cases where black box algorithms have caused harm at scale, and where we have limited ability to determine why certain predictions were made and/or how to fix the underlying algorithm. Most famously, Angwin et al. (2016) reported that an algorithm used to guide parole and bail decisions was systematically biased against black defendants. Zech et al. (2018) warned that a black box neural network that uses x-rays to diagnose disease could be misled by spurious noise without end-users knowing why. The ability to discover, audit, and address these issues requires explanation, as well as human understanding, of the inner workings of the algorithm.

Many industry experts have pointed out the critical need for human-oriented explanation. According to an IBM survey, about 60% of 5,000 executives were concerned about explainability of AI decisions (Brenna et al. 2018) while another study of 3,000 executives identified "developing intuitive understanding of AI" to be the most important challenge (Ransbotham et al. 2017). Similar concerns are shared by researchers and policymakers. Hosanagar (2020) proposes the Algorithmic Bill of Rights to protect consumers from harmful and unintended consequences of AI, emphasizing the importance of algorithmic transparency. The EU General Data Protection Regulation (GDPR)

requires many companies to provide an ex post meaningful explanation to involved parties (e.g., users, customers, or employees), while the Equal Credit Opportunity Act demands that finance companies provide reasons for decisions to their customers.

Clearly, efforts to develop more interpretable, explainable, or intelligible algorithms comprise a key area of current research.[1] A growing number of researchers have been developing eXplainable AI or "XAI" (see Guidotti et al. 2018 for a survey). However, the definition of interpretability and desiderata for a good explanation remain elusive (Lipton 2016; Rudin 2019), and so different researchers use different, often problem- or domain-specific, definitions. More alarmingly, XAI research rarely involves systematic investigation of human responses with regard to a "good" or "satisfactory" explanation; researchers typically rely on their own intuitions even though they form a highly biased sample. To the best of our knowledge, there have not been rigorous empirical studies of how explanations for ML or AI systems do (or do not) advance consumers' interests and objectives.

The goal of this paper is to address the key question: "What is a good explanation for AI output?" We approach this question in three stages. First, we establish a definition of good explanation based on pragmatic theories of explanation in the philosophy of science, which argue that an explanation should enable recipients to advance their goals or objectives. These goals may vary across contexts, but are generally centered around increasing understanding, trust, and adoption among users. Motivated by pragmatic theories of explanation, we develop a theoretical framework for the potentially relevant features of an explanation, as well as situational and human factors, that may impact the recipient's perception of the explanation and ultimately the objectives. Second, to demonstrate how this framework can be applied, we conduct several large-scale experiments using both real-world and hypothetical financial data and decisions to investigate the impact of varying different dimensions within our framework.

For managers, our framework can be used to uncover the best way to present algorithmic pre-

[1]There is no consensus on terminology, with different researchers using the terms "explainable", "intelligible", "interpretable", and "understandable" interchangeably and/or in different ways. The focus of this paper is not to reconcile these definitions or create a new definition, but rather to understand what makes a given explanation (i.e., as provided by the system or by the researcher) "good".

diction rationales to end-users to foster trust and adoption. For AI and XAI researchers, the framework highlights important factors when devising algorithms and their explanations. For business and social science researchers, our framework can motivate more in-depth investigation of different dimensions that impact the "goodness" of explanations.

EXPLANATION *IN PHILOSOPHY OF SCIENCE*

Theories of explanation in philosophy (of science) generally divide into two groups based on whether pragmatic factors are part of what makes something a good explanation. First, traditional (*non*-pragmatic) theories argue that something is a good explanation only when it provides the correct answer to a "why-question" (see Strevens 2008). For example, an explanation for why someone buys product A instead of B might provide the causal sequence that starts with search within the product category and culminates in the final purchase. Importantly, these theories all contend that the quality of an explanation is determined by its correspondence to the aspect of the world being explained. Thus, the intended audience of a particular explanation plays no role in assessing its quality; if the audience cannot understand the explanation (e.g., a five-year-old given a technically accurate explanation of quantum mechanics), then the failing is with the audience, not the explanation.

In contrast, pragmatic theories of explanation (Achinstein 1983; Van Fraassen 1988) argue that a good explanation must provide an answer to a why-question that also enables the recipient to advance her goals or interests. If someone cannot understand a proposed explanation or if the explanation fails to provide the information that they need for their present goals, then the fault does *not* lie entirely with the recipient; rather, the explanation itself is judged to be poor (for that audience). Of course, the same explanation might be considered good for one audience but not for another. That is, pragmatic theories of explanation imply explanatory pluralism (Lipton 2008; Mantzavinos 2016; McCauley 1996), or the idea that there can be multiple good explanations for the very same event. Moreover, good explanations need not be consistent with one another.

As a concrete example, a good explanation of stock diversification for the layperson might refer to an adage of "not putting all eggs in one basket" and spreading out risks, since this enables

them to understand and predict what might happen if that basket is damaged. However, a good explanation for a portfolio manager should contain statistical theories that enable them to deal with complex market situations. The layperson's explanation is ultimately inconsistent with the portfolio manager's explanation, though they give the same predictions or recommendations in many cases. According to traditional non-pragmatic theories of explanation, the layperson's "explanation" is not actually an explanation at all since it does not truly correspond to features of the world (though it approximates them); however, according to pragmatic theories, it is a good explanation that enables the layperson to reach their personal finance goals.

As this example indicates, one surprising implication of pragmatic theories of explanation is that, in extreme cases, even falsehoods can be explanatory if they lead to overall improved ability to reach the recipient's goals (Elgin 2007).[2] This performance constraint is critical, as it addresses the concern that pragmatic theories provide an "anything goes" approach (Van Bouwel and Weber 2008). Not all explanations are equally good, since not all explanations support people's goals in the same way. Moreover, one key goal for many audiences will be to "know relevant facts about the world," in which case explanations favored by non-pragmatic approaches will likely be evaluated as good explanations by pragmatic theories as well. However, the objectively "truthful" explanation is judged to be good in this case because it supports the recipient's goals (to gain knowledge), not because of some intrinsic value of truthfulness. This is particularly relevant to XAI, where there may be a trade off between the accuracy and transparency of an algorithm, and where many methods developed to explain black-box algorithms use approximation methods (see Arrieta et al. 2020 for a review).

The present paper is focused on the ways in which XAI does (or does not) enable people to advance their goals and interests. Hence, pragmatic approaches are significantly more appealing than non-pragmatic ones. Use of a pragmatic account of explanation for XAI, however, requires that we explicate not only the values and objectives that are relevant to assessing the quality of an explanation, but also how to weigh those different values in specific cases. We contend that

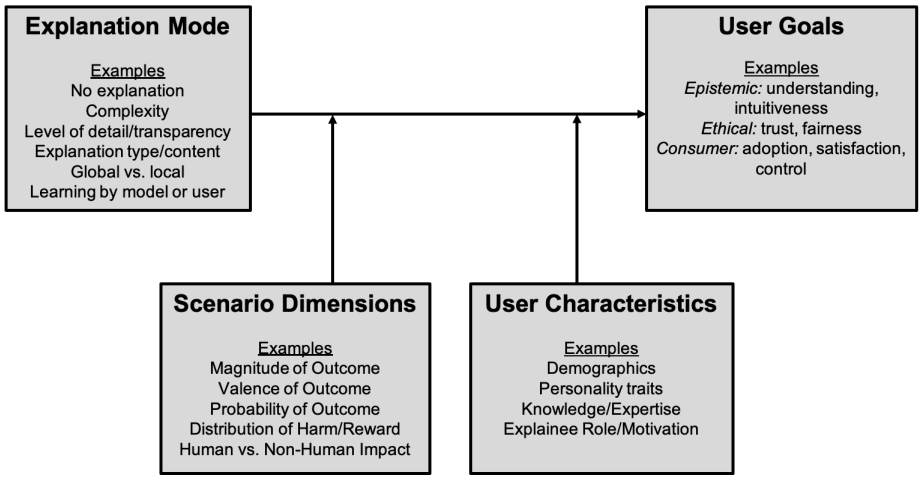
[2]Of course, falsehoods could lead to incorrect conclusions, but if those only arise in cases that the recipient does not encounter, then they are not immediately disqualifying (De Regt and Gijsbers 2016).

these values should not be legislated by philosophers, psychologists, or technologists, but should instead come from human users, including both laypeople and experts. That is, XAI should involve a participatory and deliberative process that includes the stakeholders who are impacted by the AI – developers, deployers, users, regulators, and more – to ensure that they can appropriately answer their why-questions in accordance with their values and goals. Of course, this process does not have to occur *de novo* for each new XAI deployment; lessons learned in one context can surely be transferred to new cases. Our framework and illustrative example can be understood as a first step towards identifying some of the relevant goals for a particular context and audience.

A GENERALIZABLE FRAMEWORK FOR THE DIMENSIONS OF EXPLANATION

This section develops our framework for good explanation in AI/ML settings by motivating and defining user goals, followed by different modes of explanation and dimensions that may moderate the efficacy of explanations (see Figure 1).

Figure 1: Framework for dimensions of explanation with examples of how the different dimensions may be varied for a specific use case.



Our framework is motivated by psychological, ethical, and computer science perspectives, and is by no means meant to be an exhaustive list of all possible factors that may impact how good an explanation is. We intend the framework to be a roadmap to guide users of AI/ML in thinking about how to prepare explanations for algorithmic predictions for human users. In line with this approach, Miller (2019) argues that computer scientists should draw upon the existing literature in

humanities and social sciences to properly understand multi-faceted aspects of good explanation and psychological views of how humans perceive and accept explanations.

User Goals

In psychology, a growing body of research has focused on how why-questions can or should be answered using different types of explanations. Lombrozo and Carey (2006) propose the idea of "Explanation for Export", which suggests that the function of explanations is to enable individuals to succeed in future contexts (e.g., in terms of actions, predictions, etc.), just as argued by proponents of pragmatic theories of explanation. Notably, Vasilyeva, Wilkenfeld, and Lombrozo (2017) demonstrate that the perceived "goodness" of an explanation depends specifically on its relevance to the evaluator's current task. We contend that at least three types of cognitive functions or "user goals" may be relevant to the evaluation of an explanation in business settings.

Epistemic user goals are those related to the recipient's knowledge: if she cannot understand the explanation or its implications, then the explanation cannot make a meaningful contribution to her knowledge. Even the most scientifically rigorous explanation is undesirable if users have difficulty understanding it or don't find it intuitive (Immordino-Yang and Feath 2010). To measure epistemic objectives, in some cases the recipient may only have a few moments to process the explanation, so a subjective self-reported perception of understanding and/or intuitiveness is sufficient (Cramer et al. 2008). In other cases, it may be worthwhile to measure "true" understanding using questions that test participants' ability to replicate the algorithm's predictions, consider counterfactual scenarios, or design new actions (Lage et al. 2016, Poursabzi-Sangdeh et al. 2018).

In terms of *ethical* user goals, XAI is often necessary for a company to earn customers' trust. Data subjects can reasonably place trust in data processing firms only when they are assured of a viable avenue for redress of grievances for harms or wrongs (Radin 2012). This type of "right of redress" (or to a fair trial, fair compensation, exit, etc.) typically requires explanations, which then also serve to restore trust in the company (Kim and Routledge 2020). More fundamentally, when we consider the conditions for justifiable, appropriate trust (Hardin 2002), we immediately see the key role that explanations can play in providing the necessary information to the potential trustor

(Roff and Danks 2018). In contrast, simple remediation (e.g., financial compensation for harm) does not necessarily require explanation, and is insufficient to repair trust unless accompanied by explanation of future harm prevention. Although there are many cases where algorithms actually reduce biases relative to human decision-making (Cowgill and Tucker 2019), it is also important to understand whether end-users perceive the outcomes to be fair, which may be improved via the use of appropriate explanations. Moreover, explanations can sometimes reveal other harms that have occurred or identify conditions in which harms might occur in the future, and thereby point towards ways to be more ethical.

Finally, *consumer* user goals include users' willingness-to-adopt or use the algorithm as a decision aid, or accept and/or be satisfied with the output of an algorithm. With recommender systems in particular, a firm may also want to maximize a behavioral outcome such as clickthrough rate or conversion (Tintarev and Masthoff 2015). There has been widespread resistance to algorithm adoption across a variety of domains such as forecasting and medical diagnoses (Dietvorst, Simmons, and Massey 2014; Grove and Meehl 1996), which may be related to consumers' perceived control or agency (Shaffer et al. 2013). Thus, another important consumer objective may be to give individuals a sense of control in terms of impacting the outcome generated by the system.

Explanations for AI systems can have diverse user goals beyond those we have discussed. These user goals may be correlated or there may be multiple goals within a single use-case. In our studies, we measure a variety of user goals, including both subjective and objective understanding, intuitiveness, fairness, satisfaction, ability to impact outcomes, and how ethical the use of algorithms is perceived to be in a particular domain. Importantly, we find many of these measures to be highly positively correlated, which means that providing good explanations may serve multiple goals. For example, increasing user understanding (which is mainly important for the user) may also serve to increase user trust (which is also important for the firm).

In the following subsections, we review three classes of relevant and empirically testable/measurable dimensions that may exert a significant causal impact on an explanation's ability to satisfy relevant objectives or serve as moderators, as summarized in Figure 1.

Explanation Mode

The most obvious determinant of the goodness of an explanation is the mode of the explanation itself. Nudges and choice architecture have long been studied within decision making research (see Johnson et al. 2012 for a review), with the underlying theory being that the way a choice is presented influences what a decision-maker chooses. Similarly, the way that an explanation is presented likely influences how the recipient responds.

There are many different XAI methods that may be used to present an explanation. Different forms of an explanation (e.g., text-based, visualization techniques, explanations by example, etc.) may be presented for the same model. In their taxonomy of XAI methods, Arrieta et al. (2020) make a distinction between "transparent" ML models that are understandable on their own, versus more complex models that require "post-hoc analysis". Transparent models include linear/logistic regression, decision trees, and general additive models, while non-transparent models that require post-hoc analysis include tree ensembles, support vector machines, and convolutional neural networks (CNNs). Importantly, the algorithm described by the explanation need not be the same as the actual computational system. As discussed in the previous section, many XAI techniques for post-hoc analysis depend on finding accurate linear approximations of an underlying non-linear model (Andrews et al. 1995; Montavon et al. 2018), but according to pragmatic theories of explanation they may still provide good explanations that advance user goals.

Within our framework, understanding the effects of the explanation mode does not mean individually testing every possible XAI method. Instead, our objective is to identify and understand the effects of some generalizable properties of explanations generated using XAI. For example, research on advice-giving has found that people over (vs. under) weight advice in difficult (vs. easy) tasks (Gino and Moore 2007), so algorithm adoption may depend upon the recipient's perceived complexity of the algorithm's task, as a result of the given explanation. The quality of an explanation can also depend on the explanation having the right level of detail (Putnam 1960). For example, in a field experiment involving peer grading within an online course, Kizilcec (2016) demonstrated that some amount of explanation of the grade calculation algorithm increases trust, but too much

transparency can actually backfire and erode trust. Explanations can also range from more global in nature by focusing on population-level aspects to more local by emphasizing the particular features of the target individual. Finally, the explanation could make reference to prior learning/adaptation by the model or only describe the system at the current moment in time. One reason why people exhibit algorithm aversion is that they believe that humans (but not machines) can learn through experience (Highhouse 2008), and so explanations that convey the adaptability of an algorithm may have an impact on the objectives. In some cases, human users themselves may be able to learn about how the algorithm works via multiple exposures or trials (Poursabzi-Sangdeh et al. 2018).

In our studies, we use three XAI methods: verbal or text-based explanation (i.e., of feature relevance), decision tree (which is a transparent model and thus generates its own explanation), and neural network (which requires post-hoc analysis methods). For the decision tree explanations, two key generalizable properties that we examine are how consistent the decision tree is with consumer expectations and the complexity of the decision tree. For the neural network explanation, we use SHAP values that indicate feature relevance, and compare local and global SHAP values. We examine how these different explanation modes impact measures of user goals relative to when no explanation is provided at all.

To reiterate, the framework in Figure 1 is designed to provide a guide to determine the factors that impact the user goals. The lists of user goals, explanation modes, and other dimensions are not to meant to be exhaustive, but to provide a starting point that can be customized to a manager's specific use case. Additionally, the underlying relationship between the dimensions and the user goals may be complex. For instance, Wang and Benbasat (2007) propose that different types of explanation (e.g., how vs. why) may enhance different trusting beliefs (e.g., competence vs. integrity). This highlights that it will not be the case that a particular explanation will always be the best for every user goal, and thus it is important to vary and test a sufficient set of dimensions and user goals for any use case.

Scenario Dimensions

The characteristics of the outcome itself, which we refer to as "scenario dimensions", might

moderate the impact of the explanation mode on the objectives. In other words, the very same sequence of words might constitute a useful explanation for one set of outcomes, but not another. A rich literature exists on how the valence, magnitude, and probability of outcomes impacts human decision-making (Van Dijk and van der Pligt 1997; Wu and Zhou 2009). For example, individuals exhibit greater sensitivity to losses versus gains (Tversky and Kahneman 1992), asymmetric updating towards positive vs. negative information (Eil and Rao 2011), information avoidance (Sweeny et al. 2010), and nonlinear probability weighting (Gonzalez and Wu 1999).

Based on these findings, the ability of an AI explanation to advance people's goals may depend on whether the outcome is beneficial vs. harmful, common vs. rare, low vs. high impact, or even affecting humans vs. non-humans. For example, a good explanation for a negative outcome might need to focus on local factors that show how the harmed individual could improve future outcomes, while a good explanation for a positive outcome might need to emphasize global factors about which groups are most likely to benefit from the AI system.

In our studies, the primary scenario dimension that we vary is whether an individual receives a positive or negative outcome, for example whether they are approved or rejected for a loan. Notably, we find that the outcome has an impact on our measures of the user goals, including how understandable and ethical individuals judge the algorithm to be. We use our measure of objective understanding (i.e., accuracy of the free response explanation and counterfactual) to demonstrate that information processing differs between individuals who receive positive versus negative outcomes.

User Characteristics

The attributes of the actual recipients of an explanation may also moderate its effectiveness. For example, researchers have proposed that algorithm aversion may depend on individual characteristics. Shaffer et al. (2013) study how patient perceptions of doctors who use decision aids depend on individual differences in locus of control and attitude towards statistics. Hoff and Bashir (2015) distinguish between factors that influence dispositional trust (e.g., culture, age) and situational trust (e.g., mood, attentional capacity) in automation.

There will almost certainly be a significant interaction between human dimensions and explanation mode. An explainee with a PhD in computer science could make use of a technically-detailed explanation of an AI system, while a non-expert would benefit from a more metaphorical explanation. The literature on how consumer knowledge and expertise shapes behavior suggests that experts value depth over breadth knowledge compared to novices (Alba and Hutchinson 1987). Thus, one testable hypothesis might be that low (vs. high) knowledge recipients value broader (vs. more specific) explanations. Importantly, there may be significant opportunities to tailor explanations to individuals if their expertise levels (and/or other characteristics) are known.

Finally, it is important to consider the role or motivation of the explainee, such as whether they are directly experiencing the AI-determined outcome vs. acting as an agent. There are documented discrepancies and similarities in how people make decisions for others vs. themselves, including less loss aversion when making gambling and social choices for others (Polman 2012), and projection when making surrogate predictions for others' end-of-life decisions (Fagerlin et al. 2001). Thus, some explanations may serve the objectives of a customer but not a manager, and vice versa. In our studies, we find that when individuals are given the role of the applicant (vs. developer), they are more satisfied with the algorithmic explanation and believe that the ability to impact the outcome is higher.

SUMMARY OF STUDIES

In the following studies, we demonstrate two concrete applications of our framework. In Studies 1A and 1B, participants are asked to consider the context of applying for a credit loan. In Study 1A, we pretested 7 possible features that might factor into a credit loan application decision to ensure that people find them to be important and generally agree on the directional impact that they should have on the decision. In addition, we examine whether individuals believe that usage of algorithms in making these decisions is ethical/unethical. In Study 1B, we present individuals with different explanations for an algorithm's decision for approving/rejecting their credit loan application, and examine their impact on various measures that may be relevant to user goals, including subjective understanding, satisfaction, and fairness. We find that participants whose hypothetical loan appli-

cations were approved rated the explanations higher across all user goal measures. We also found that the decision tree actually decreased understanding relative to no explanation in some cases, and that including both local and global neural network explanations resulted in the highest subjective understanding.

To replicate and test these findings further, in Studies 2A, 2B, and 2C, we asked participants to consider applying for a credit card. In Study 2A, we again pretested the importance and directionality of different features that would determine the outcome. In Study 2B, we developed a method of measuring objective understanding to determine whether discrepancies in subjective understanding between participants who received approve/reject outcomes resulted from differences in information processing. To explore why the decision tree seemed to decrease understanding in Study 1B, in Study 2B, we ensured that the decision tree matched expectations and presented participants with different approve/reject scenarios that went down different "branches" of the tree. In addition, we also varied whether the user played the role of the applicant or the algorithm developer. Finally, Study 2C replicated Study 2B with a sample of users from a different online panel.

STUDY 1A: PRETESTING INTUITIONS FOR CREDIT LOAN ALGORITHMS

To construct the stimuli for Studies 1A and 1B, we used the German Credit Data from the UC Irvine Machine Learning Repository (Dua and Graff 2019). This dataset consists of 1,000 individual profiles, each with 20 features and categorized as either "good" or "bad" credit risks. We selected 7 of the continuous features (see Table 1) to include in these studies (except for "Property", which was coded as categorical). Study 1A serves as a pretest of people's lay intuitions or expectations of how these 7 features should impact the probability of a credit loan application being approved or rejected. In Study 1B, we test how different explanations for algorithms trained on this data impact subjective perceptions of understanding.

Design and Method

For Study 1A, we recruited 199 participants in the United States on Amazon's Mechanical Turk (MTurk) platform. All participants were told to imagine that they were applying for credit loan

approval to purchase a new car, and that the bank used an algorithm – developed with information from 1,000 previous applications and the 7 features shown in Table 1 – to determine whether to approve or reject the loan application. In order to ensure that participants read the information given in the table, they were asked two multiple questions about the presented information that they had to respond to correctly before they could continue with the survey.

Table 1: Table of features shown to participants in Studies 1AB.

Feature Name	**Description**
1. Credit duration	How long you have to pay off the loan in months
2. Credit amount	How much the loan is for in dollars
3. Employment	Number of years you have been employed so far (0 if currently unemployed)
4. Installment rate	Monthly payments as a percentage of your disposable income
5. Property	Do you own real estate, life insurance, a car, etc.
6. Age	Age in years
7. Credit existence	How many existing loan approvals have you received from this bank in the past

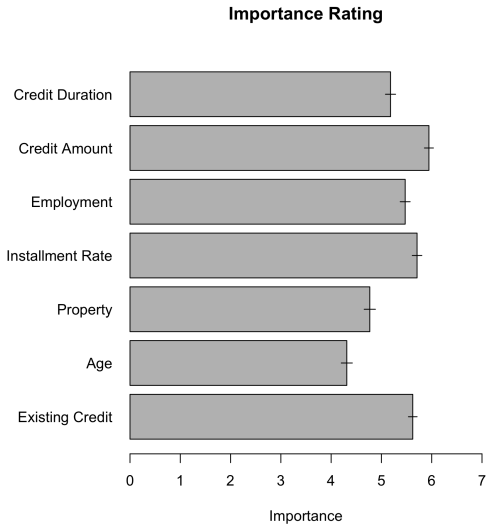
Participants were asked to rate how important they expected each factor to be when the bank's algorithm determined whether to approve or reject the loan application on a 7-point Likert scale (i.e., 1 = "Not Important At All", 4 = "Somewhat Important", 7 = "Extremely Important"). Participants were then asked to indicate whether they expected specific directional changes in each factor (e.g., an increase in credit duration) would increase, decrease, or have no effect on the probability of approval.

Finally, participants were asked to rate on a 7-point Likert scale how ethical they thought it was for the bank to use a computer algorithm to make this type of credit loan approval decision, as well as to explain in a few words why they believed it was ethical/unethical. Participants were also asked to list up to 5 additional factors (i.e., beyond the 7 given in Table 1) that they thought the bank's algorithm should take into account. Finally, participants were asked to answer several demographic questions, including gender, highest degree of education, current employment status, household income, age, and state of residence.

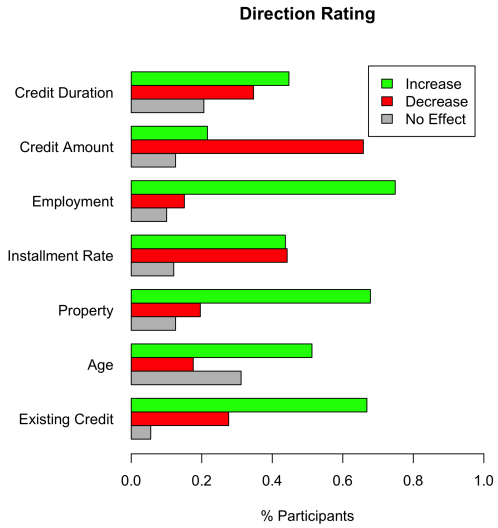
Results and Discussion

Figure 2(a) plots the average importance ratings for each of the 7 factors. All factors were rated significantly above the midpoint of 4. Figure 2(b) plots the percentage of participants who indicated that they thought a change (i.e., increase) in the factor would increase, decrease, or have no effect on the probability of approval. We see that there was a "majority" agreement on nearly all factors except for credit duration and installment rate.

Figure 2: Importance ratings and directional changes reported by participants in Study 1A.



(a) Average importance ratings



(b) Distribution of participant responses regarding how a directional change (i.e., increase) in each factor would impact the likelihood of loan approval

We also found that most participants rated the bank's use of a computer algorithm as highly ethical, significantly above the midpoint of 4 on the 7-point Likert scale (M = 5.24, SD = 1.44, $t(198) = 12.13$, $p < 0.001$). Sample responses from participants are given in Table 2. We see that those who gave a lower rating mentioned the individuality of applications, while those who gave a higher rating focused on the algorithm's ability to be free of emotion and bias. Interestingly, compared to people with lower education/schooling (e.g., no high school diploma, high school, bachelor's degree), those with higher education/schooling gave a higher ethical rating ($t(73) = 3.84$, $p < 0.001$), although there were no other demographic differences.

Table 2: Sample responses from participants explaining why they believed the use of algorithms in making bank loan decisions is ethical or unethical, organized by rating.

Rating	**Sample Responses**
1-3 (Not Ethical)	"Applicants should be looked at on a case by case basis" "Algorithm might make mistakes" "The thing is that the algorithm is going to approve or disapprove the credit based on general factors, but each person has its own history so it will be better if a human could make the decision" "Taking the time to talk to the individual about their financial situation and letting them explain certain aspects that affect their credit would help the bank make the best decision"
4 (Neutral)	"Certain things can use an algorithm (calculating how many loans they got in the past, etc) but there are some factors that should be determined by a human using human empathy"
5-7 (Ethical)	"Doing the same work that a person would do minus the emotional or trust factors. Not unethical, just math." "The bank works with actuarial type and historical record information. I think that it is ethical as those metrics do give relatable information." "It's ethical, because it's not taking things such as race and ethnicity into account, and it makes decisions fairly objectively, without emotional influence." "It is ethical as the algorithm looks at all factors objectively and will not base its decisions on any pre-conceived notions, judgments, or bias."

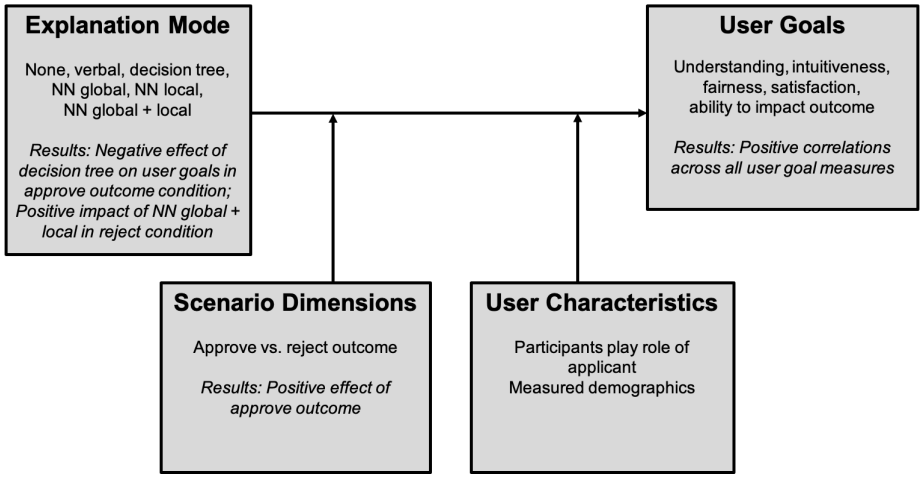
In summary, we demonstrate in Study 1A that the 7 features that we identified within the German Credit Data are rated as sufficiently important by participants in a credit loan application decision. In Study 1B, we use these 7 features to train ML algorithms to generate the explanations given to participants for different outcomes. The expected directional impact of different changes to the factors given by participants in Study 1A will allow us to identify whether any effects of the explanations are driven by how much participants' expectations match (or don't match) the algorithm. Finally, we found that most participants believe that the bank's use of an algorithm is ethical because of their perceived objectiveness of an algorithm (i.e., compared to a human).

STUDY 1B: EFFECTS OF DIFFERENT EXPLANATIONS FOR CREDIT LOAN ALGORITHMIC OUTPUT ON CUSTOMER UNDERSTANDING

The purpose of this study is to demonstrate how our framework can be applied to a common real-world setting. Specifically, we examine how consumers respond to explanations given for hy-

pothetical credit loan approval outcomes, based on actual ML models estimated using real data. As summarized in Figure 3, Study 1B serves as an example of how our framework can be applied to a real-world context in order to find the best explanation mode to maximize user goals, given a specific set of scenario dimensions and user characteristics. In the language of our framework, the explanation mode is explicitly varied in six different ways, the scenario varies by the valence of the user's outcome, and the user characteristics are specified so the explainee motivation is that of an end-user, with existing individual differences measured via questions on demographics. We measure several user goals, with the main measure of interest being subjective understanding.

Figure 3: Framework applied to a credit loan approval context in Study 1B.



Design and Method

To construct the stimuli, we used the 7 features in the German Credit Data from the Study 1A pretest (see Table 1) to train a decision tree model and a neural network (NN) model, each to predict the good/bad (i.e., approve/reject) outcome.[3] Recall that in Study 1A, all 7 features were rated to be significantly improtant (i.e., above the midpoint on a Likert scale) in a loan application decision. However, there was disagreement on the directional impact of credit duration and installment rate.

We recruited 1,205 paid participants in the United States on MTurk. Similar to Study 1A, all participants were told to imagine that they were applying for credit loan approval to purchase a new car, and that the bank used an algorithm to determine whether to approve or reject the loan

[3]See Web Appendix A for estimation methods.

application. We again included multiple choice questions that had to be answered correctly to continue with the survey to serve as attention checks.

Participants were randomly assigned to a condition in a 2 (scenario dimension: loan application approved vs. rejected) $\times$ 6 (explanation mode: none, verbal, decision tree, NN global, NN local, NN global and local) between-subjects design. From the German Credit data, we selected two existing profiles, one good and one bad, for participants in the approve and reject conditions, respectively. The information was displayed in a column entitled "Your Data" added to Table 1, with participants told to imagine that they had entered these data into their loan application. Specifically, participants across all conditions were told that the loan was for \$2800, they had been employed for 2 years, the installment rate would be 3%, they owned life insurance, they were 40 years old, and they had 1 existing loan approval from the bank. Participants in the approve condition were told that their credit duration was 6 months, while participants in the reject condition were told that their credit duration was 36 months. Thus, to minimize the effects driven by differences in the feature values, while maintaining the validity of using models trained on real data, we selected profiles that only differed by the credit duration feature (i.e., 6 months vs. 36 months).

In the no explanation condition, participants were told their approve/reject outcome with no additional information. In the verbal condition, participants were given a list of the three most important features (employment, credit duration, installment rate). In the decision tree condition, participants were shown a visual representation of the decision tree and told that the algorithm starts at the top of the tree and follows the yes/no questions downwards (see Figure 4). In the NN global condition, participants were told that the decisions were determined using a sophisticated neural network algorithm and shown a plot of SHAP *global* average values representing the average importance of each feature across customers (see Figure 5). In the NN local condition, they were instead shown the SHAP *local* values representing how the features impact chances of approval in their specific case and others with similar data features (see Figure 6). Participants in the NN global and local condition were shown both SHAP global and local values.[4]

[4]See Web Appendix B for description of XAI methods.

Figure 4: Illustration of decision tree model.

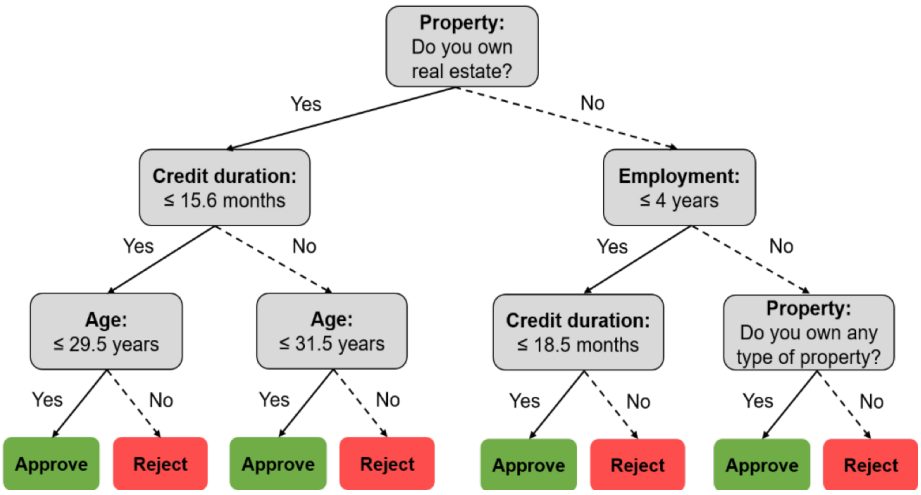


Figure 5: SHAP global average values.

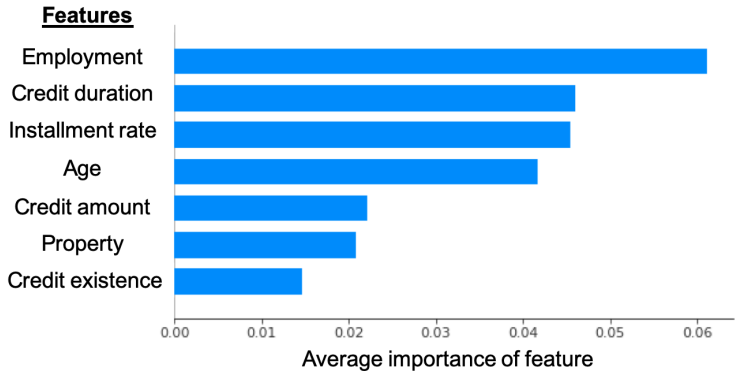
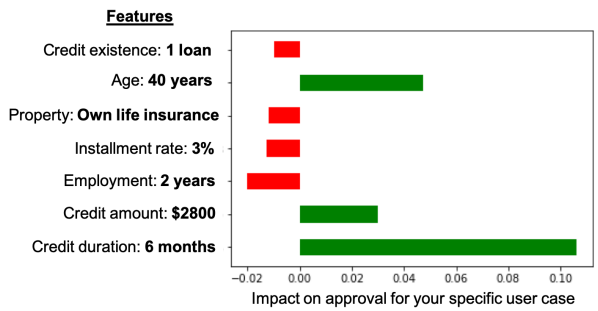
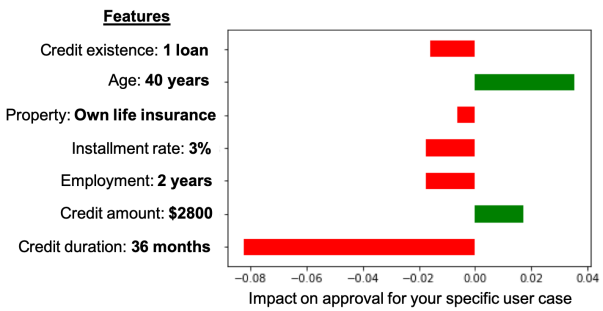


Figure 6: SHAP local values for approve vs. reject conditions, with green and red bars indicating positive and negative impact on approval, respectively.

(a) Approve Condition



(b) Reject Condition



After viewing the explanation and answering several attention and comprehension check questions, participants were asked to rate, on 7-point Likert scales, their understanding of the algorithm explanation, as well as perceived intuitiveness, fairness, satisfaction, and ability to impact the outcome. These responses serve as our measure of the key objectives for a firm implementing this algorithm. We also asked participants how complex they perceived the algorithm to be, and their familiarity with loan approval systems. Finally, participants responded to a series of demographic questions.[5]

Results and Managerial Implications

Table 3 reports the correlations between the user goal measures, and we see that all of them are significantly positively correlated. Thus, we expect the effects of the explanation modes and scenarios to be directionally similar for all measures.

Table 3: Correlations between all user goal measures.

	Understand	Intuitive	Fair	Satisfied	Impact
Understand	1				
Intuitive	0.47	1			
Fair	0.51	0.47	1		
Satisfied	0.61	0.52	0.69	1	
Impact	0.38	0.38	0.46	0.49	1

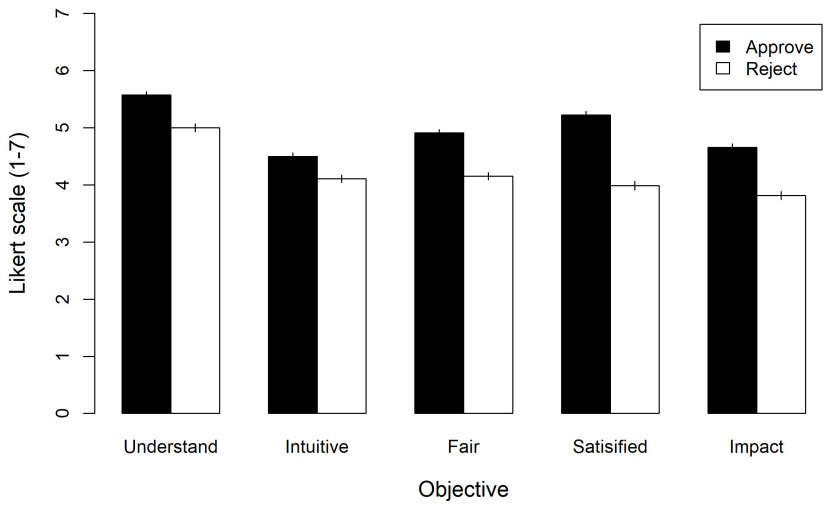
For each user goal, we conducted two-way ANOVAs with the scenario and explanation modes, and found significant main effects and interactions for nearly all variables.[6] To illustrate, in Figure 7, we compare the ratings for each objective variable between approved vs. reject conditions. In Table 4, we summarize the results from t-tests of the user goal ratings for each of the explanation modes (compared to no explanation). We highlight three key findings, along with managerial implications and directions for future research.

[5]See Web Appendix C for study instructions.

[6]See Web Appendix D for detailed statistical test results.

First, as shown in Figure 7, participants in the approve condition rate the explanation higher across all user goal measures. With positive outcomes, it is possible that recipients simply do not care about an explanation, so firms may want to focus on improving explanations for negative outcomes. Or perhaps negative outcomes decrease individuals' ability to process information, which would explain the lower ratings for understanding and intuitiveness in the reject condition and would be consistent with prior findings that people tend to focus on positive information and avoid negative information (Eil and Rao 2011; Sweeny et al. 2010). In general, perceived understanding may also reflect the recipient's affective response to good vs. bad outcomes. More involved studies with process measures (e.g., eye-tracking, clickstream) may shed light on how valence changes information processing and attention. In Study 2B, we directly test for comprehension and understanding (rather than simply "subjective" or "perceived" understanding) by having participants generate explanations for the decisions, as well as counterfactuals, which can be coded for accuracy.

Figure 7: Average explanation ratings for approve vs. reject conditions in Study 1B.



Second, as shown in Table 4(a), in the approve condition we observe a *negative* effect of the decision tree explanation for many of the objective variables. This result is surprising because intuition suggests that any explanation should improve (or at least not decrease) understanding compared to no explanation at all. In fact, shallow decision trees (like ours) are commonly used within

XAI, and are typically considered useful and intuitive by researchers and practitioners alike. One explanation for the negative effect is that consumers may expect a very complex algorithm for an important decision such as a bank loan, and the visualization of the decision tree may give the impression of an overly simplistic system. This is further supported by the positive effects of the NN global explanation, in which the algorithm was explicitly described as complex. Thus, when designing explanations, it may be important to match the explanation of the algorithm's complexity to people's expectations. This finding also highlights that, although more information should increase "true" understanding, in some cases it might decrease *perceived* understanding, which may sometimes matter more to managers. Thus, in Study 2B, we measure both subjective and objective understanding, and also vary the complexity of the decision tree.

Table 4: Summary of significant effects ($p < 0.05$) in Study 1B of different explanation modes (relative to no explanation), separated by whether participants were in the approve or reject condition.

(a) Approve Condition

Explanation Mode	**Understand**	**Intuitive**	**Fair**	**Satisfied**	**Impact**
Verbal					
Decision Tree	–	–	–	–	+
NN global	+			+	+
NN local					
NN local + global					

(b) Reject Condition

Explanation Mode	**Understand**	**Intuitive**	**Fair**	**Satisfied**	**Impact**
Verbal	+			+	
Decision Tree				+	
NN global	+			+	
NN local	+			+	
NN local + global	+	+	+	+	+

An alternative explanation is that the structure of the tree did not match participants' expectations. Specifically, as shown in Figure 2(b), participants in Study 1A were somewhat split about

whether higher credit duration should have a positive or negative effect, while in the actual estimated algorithm shown to participants in Study 1B (see Figure 4), lower credit duration is better, as determined by an algorithm estimated on real data. In addition, participants in Study 1A believed that higher age would be better, while it was better for age to be lower in the decision tree used in Study 1B. In Study 2B, we control for these effects by using only non-ambiguous features and constructing a decision tree that matches participants' expectations (as pretested in Study 2A). We also ensure that our results are robust across different applicant profiles that go down different "branches" of the tree.

Third, we see in Table 4(b) that there is an interaction in the reject condition between the effects of the NN global and NN local explanation modes. In particular, presenting each on their own does not improve intuitiveness and fairness ratings, but presenting both results in significant improvements. These results are consistent with the participant free responses from Study 1A for why they think that using algorithms to determine credit loan application decisions is ethical or unethical (see Table 2). Specifically, participants who thought algorithms were unethical mentioned individual histories, which would be addressed with local information, while those who thought algorithms were ethical mentioned fairness and objectiveness, which would be addressed with global information. In addition, only the NN global condition improved the outcome measures among participants in the approve condition, suggesting that they prefer information about the average user. Further exploration into how people process information about themselves (i.e., local information) versus the average user (i.e., global information) in the context of XAI is an important area for future work, especially as AI continues to be used both in small- and large-scale applications.

In the following set of studies, we replicate our findings from Studies 1A and 1B using a credit card application setting. In Study 2A, we pretest people's lay intuitions about an algorithm used for approving/rejecting credit card applications. In Studies 2B and 2C, we again test the effects of the approve/reject outcome and explanation mode on subjective understanding. In addition, we vary decision tree complexity and the participant's perceived role in the application process, as well as develop a method for measuring objective understanding.

STUDY 2A: PRETESTING INTUITIONS FOR CREDIT CARD APPLICATION ALGORITHMS

The purpose of Study 2A was to determine the appropriate stimuli for Studies 2A and 2B, in which we measure subjective and objective understanding of different explanations for credit card application decisions. We first determined 11 factors or pieces of information that a bank might ask for in a credit card application. 6 of the factors (credit score, delinquencies, hard inquiries, credit usage rate, income, and credit history length) were listed to be among the most important in determining credit card approval in an article posted by Credit Karma, one of the largest online credit and financial management platforms. In addition, we included 5 demographic factors (years employed, years of education, gender, age, and marital status). In Study 2A, we asked participants to indicate the importance and directional impact they expected each factor to have regarding credit card applications decisions. This allows us to ensure that the algorithmic explanations and outcomes presented in Studies 2B and 2C are consistent with general consumer expectations.

Design and Method

We recruited 200 participants in the United States on MTurk. All participants were told to imagine that they were applying for a credit card, and that the bank used an algorithm to determine whether to approve or reject their credit card application. They were then told that the algorithm was developed with information from 10,000 previous applications and used the 11 features in Table 5 to make a decision. In order to ensure that participants read the information given in the table, they were asked three multiple questions about the presented information and had to answer them correctly before proceeding further with the survey.

Similar to Study 1A, participants were asked to rate how important they expected each factor to be when the bank's algorithm determined whether to approve or reject the credit card application on a 7-point Likert scale and to indicate whether a specific change in each factor would increase, decrease, or have no effect on the probability of approval.

Finally, participants were again asked to rate how ethical they thought it was for the bank to use a computer algorithm and to explain why, to list up to 5 additional factors that should be used in the algorithm, and to answer a few demographic questions.

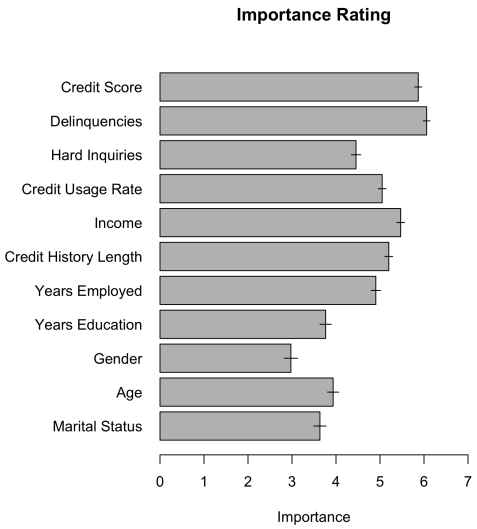
Table 5: Table of features shown to participants in Studies 2ABC.

Feature Name	**Description**
1. Credit score	Number ranging from 300-850 that depicts credit-worthiness, based on credit history (open accounts, total levels of debt, repayment history)
2. Delinquencies	Number of credit accounts that are currently past due
3. Hard inquiries	Number of years you have been employed so far (0 if currently unemployed)
4. Credit card utilization rate	Percentage of your available credit that you are using (between 0% and 100%)
5. Income	Year income in dollars
6. Credit history length	Number of years you have had and made monthly payments on existing credit cards
7. Employment	Number of years that you have worked for your current employer (0 if currently unemployed)
8. Education	Number of years of education
9. Gender	Male or female
10. Age	Age in years
11. Marital status	Married or single

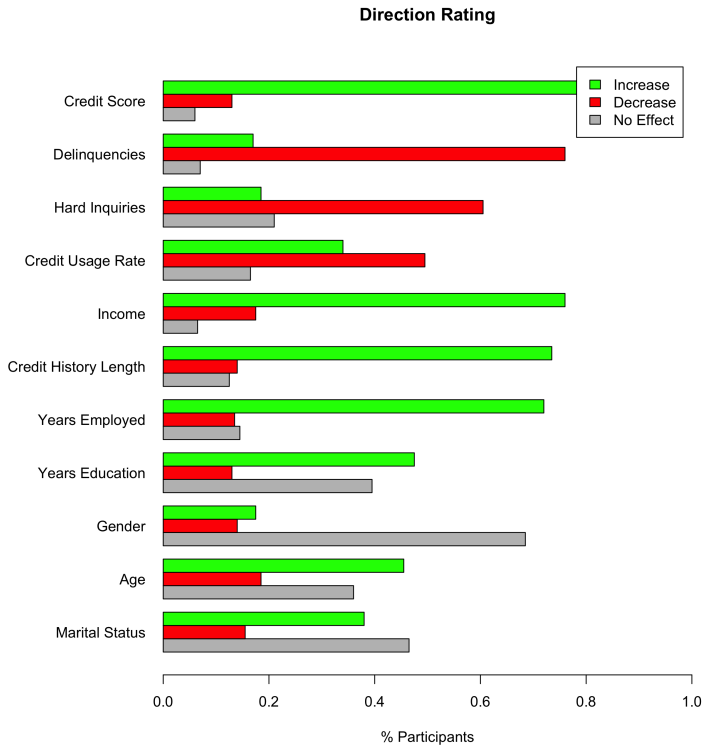
Results and Discussion

Figure 8 plots the average importance ratings and directionality for the 11 features. We see that education, gender, age, and marital status are rated as relatively unimportant, and thus we exclude them from Studies 2B and 2C. We also exclude the credit usage rate feature because, although it was rated high on importance, there was large disagreement on whether an increase in credit usage rate would have a positive or negative impact. In total, this leaves us 6 features to use in the following studies: credit score, delinquencies, hard inquiries, income, credit history length, and employment.

Figure 8: Importance ratings and directional changes reported by participants in Study 2A.



(a) Average importance ratings



(b) Distribution of participant responses regarding how a directional change (i.e., increase) in each feature would impact credit card approval. The exceptions were gender and marital status, which were expressed as binary changes (i.e., "if the applicant was female rather than male" and "if the applicant were married rather than single", respectively).

STUDY 2B: EFFECTS OF DIFFERENT EXPLANATIONS FOR CREDIT CARD ALGORITHMIC OUTPUT

The design of Study 2B allowed us address three key objectives. First, unlike in Study 1B where we used real data to estimate the algorithms and generate the stimuli, in Study 2B we directly used the responses from Study 2A to generate algorithmic explanations and outcomes that aligned with general consumer expectations. Specifically, the decision trees shown to participants in Study 2B include only features that were deemed to be important by participants in Study 2A, with branchings that are directionally consistent with their expectations. By controlling for consumer expectations and also including trees of varying complexity, we can test whether the lack of improvement in understanding with the decision tree explanation in Study 1B was due to the decision tree not matching expectations or the decision tree visualization giving participants the feeling that the algorithm was not complex enough.

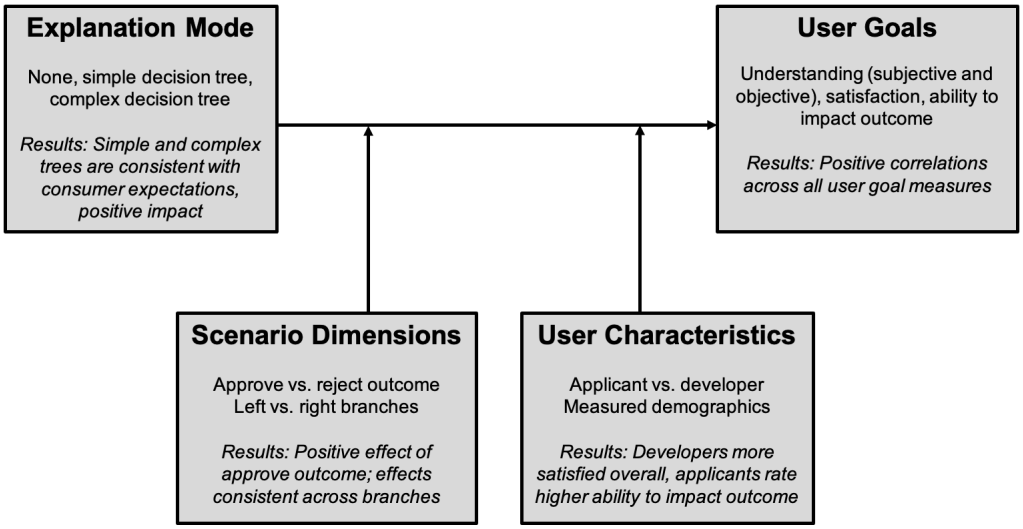
Second, we measure objective understanding of algorithmic explanations by developing a method based on coding information from participants' free responses. We compare objective understanding to participants' self-reported subjective understanding to determine whether the measures are positively correlated, which would indicate that individuals can accurately self-assess their own understanding. In addition, this would lend some support to the hypothesis that differences in understanding across different conditions in our study (e.g., positive vs. negative outcomes) are due to differences in information processing.

Third, unlike in Study 1B where we did not experimentally manipulate the user characteristics, in Study 2B, participants are told to either imagine themselves to be the credit card applicant or the developer of the algorithm. This allows us to test for whether or not being the recipient of the outcome moderates the effects of the explanation mode on understanding.

In summary, as illustrated in Figure 9, we vary the explanation mode by comparing no explanation, a simple decision tree, and a complex decision tree. The user goals that we measure are understanding (both subjective and objective), satisfaction, and perceived ability to impact the outcome. We then test whether the effects of the explanation mode on user goals are moderated by

whether participants were approved vs. rejected or in the left vs. right branches (i.e., scenario dimensions), as well as by whether participants played the role of the applicant or developer (i.e., user characteristics).

Figure 9: Framework for dimensions of explanation for Study 2B.



Design and Method

We recruited 2,411 paid participants in the United States on MTurk. Participants were told about a bank that uses a computer algorithm to determine credit card application decisions with the information in Table 5. Note that participants in Study 2B were only shown a table with the 6 important/non-ambiguous features determined in Study 2A: credit score, delinquencies, hard inquiries, income, credit history length, and employment.

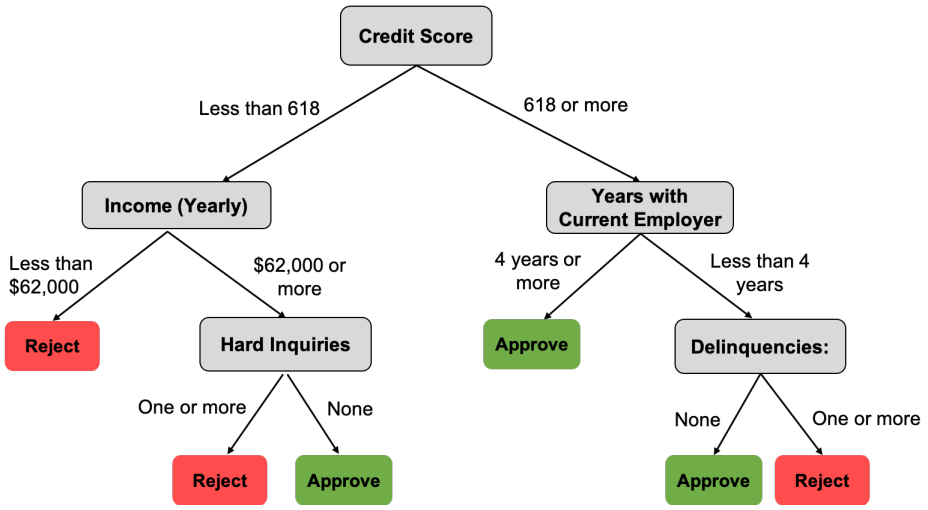
Participants were randomly assigned to a condition in a 2 (scenario: credit card application approved vs. rejected) $\times$ 3 (explanation mode: none, simple decision tree, complex decision tree) $\times$ 2 (explainee motivation: applicant vs. developer) $\times$ 2 (decision tree branching: right vs. left) between-subjects design.

Participants in the applicant condition were asked to imagine that they themselves were applying for a new credit card at a national bank. Their hypothetical information was presented in a "Your Data" column in a modified version of Table 5 that contained only the 6 features of interest. To reinforce their understanding of their role, participants were asked to enter this information into

appropriate boxes on the screen. Participants in the developer condition were told to imagine that they were working as a software developer at a national bank, and that their job was to develop a computer algorithm to help the bank determine whether to approve or reject credit card applications based on the 6 main features, and then told to consider a hypothetical customer who had submitted their application with their information presented in an "Applicant Data" column. We again included multiple choice questions that had to be answered correctly to continue with the survey to serve as attention checks.

Participants were told that the credit card application was either approved or rejected. They were then given either no explanation, shown a simple decision tree (see Figure 10), or shown a complex decision tree (see Figure 11). We used the importance and directionality ratings from participants in Study 2A to generate decision trees that were consistent with their expectations.[7] For example, in Study 2A, the delinquencies feature was rated as relatively important, with more delinquencies expected to decrease the chances of approval. Consistent with this, delinquencies determine reject vs. approve in the bottom-right node of the simple decision tree, and reject vs. continue to the next branch in the complex decision tree.

Figure 10: Simple decision tree shown to participants in Study 2B.



[7]The average importance and directionality ratings of the 6 features were used as a guide to determine the magnitude and sign of the coefficients within a logistic regression to simulate data. This simulated data was then used to estimate decision trees of varying complexity/depth, with minor adjustments made to ensure minimal differences between scenarios.

Figure 11: Complex decision tree shown to participants in Study 2B.

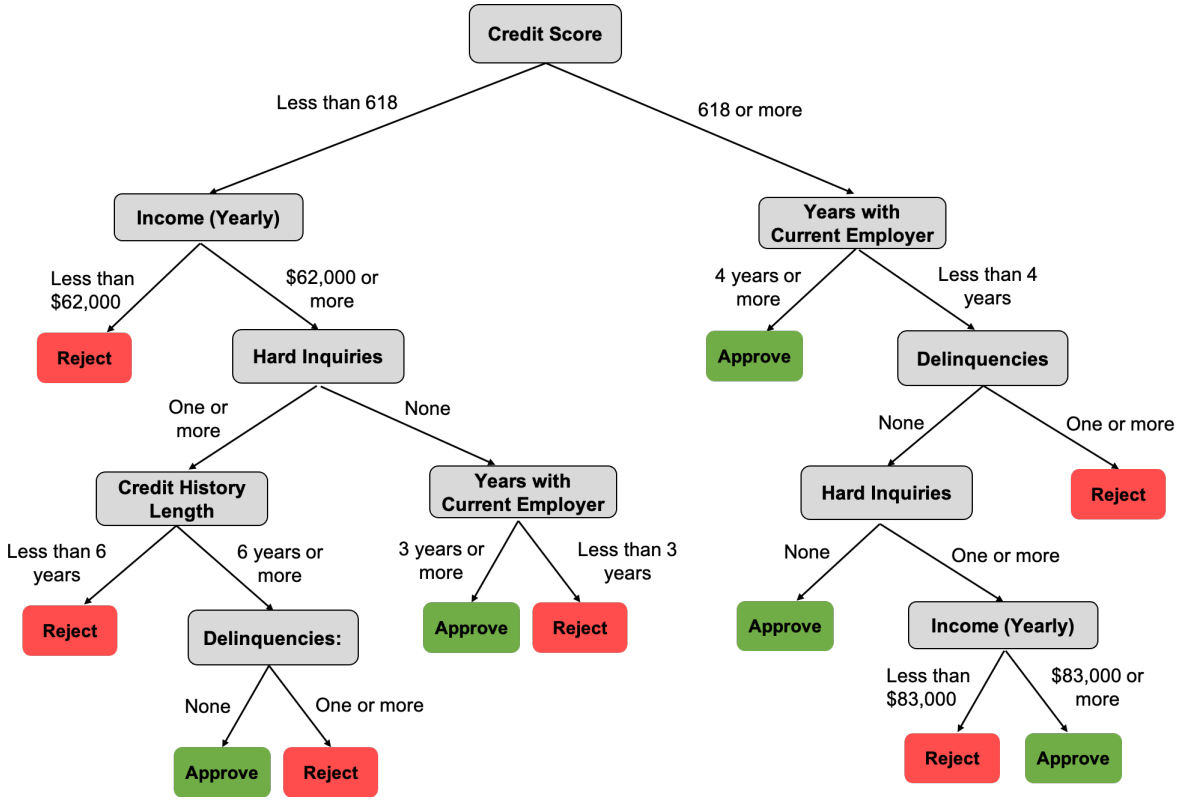


Table 6: Different scenarios shown to participants in Study 2B.

Feature Name	**Positive (Right Branch)**	**Negative (Right Branch)**	**Positive (Left Branch)**	**Negative (Left Branch)**
1. Credit score	650	650	580	580
2. Delinquencies	None	Two	None	None
3. Hard inquiries	None	None	None	Two
4. Income	$75,000	$75,000	$75,000	$75,000
5. Credit history length	4 years	4 years	4 years	4 years
6. Employment	3 years	3 years	3 years	3 years

In order to ensure the robustness of our main results to specific scenario information, we selected four hypothetical profiles, with two resulting in an approval outcome and two resulting in a reject outcome using either the simple or complex decision tree, as shown in Table 6. For the two "right branch" scenarios, the credit score is 650, and for the two "left branch" scenarios the credit score is 580. Within the positive and negative scenarios that go down the right branch of the decision

trees, the data only differs by the delinquencies feature. Within the positive and negative scenarios that go down the left branch, the data only differs by the hard inquiries feature. Thus, we generally control for feature values by comparing positive vs. negative conditions within branches, but also test for robustness with scenarios that go down different branches.

Participants were then asked to describe in their own words why the application was approved or rejected, and also to describe one or more factors in the profile that, if changed, would result in the opposite outcome for the credit card application (i.e., an alternative profile or counterfactual). By coding these responses for accuracy, we can obtain a measure of participants' objective understanding.

Participants were also asked to rate on 7-point Likert scales their (subjective) understanding of the algorithm's explanation, how sophisticated/complex they thought the algorithm was, how satisfied they were with the decision, and the perceived ability of an applicant to impact their chances of obtaining the credit card.

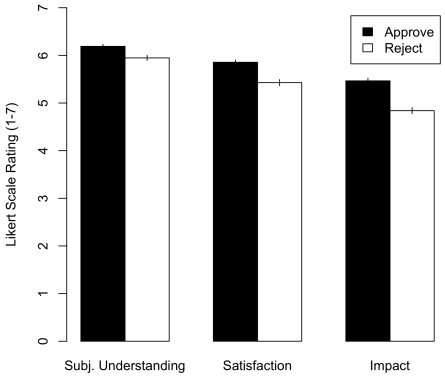
We also asked participants how ethical they think it is for a bank to use a computer algorithm to make this type of credit card approval decision on a 7-point Likert scale, and also to explain why they thought it was ethical or unethical in a free response question. Finally, they answered a few demographic questions.

Results and Discussion

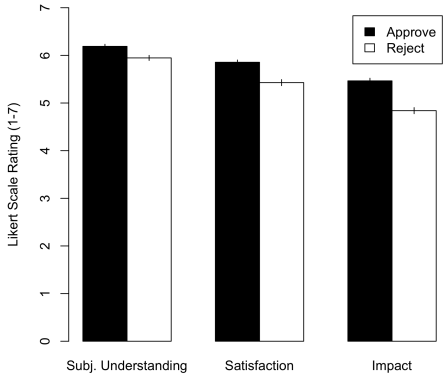
First, we replicate our findings from Study 1B that participants in the approve condition rate the explanation higher across all objectives, as shown in Figure 12, which compares the average ratings for positive/negative outcomes for the left/right branches. Thus, when faced with a positive outcome participants across both applicant and developer roles rate the algorithmic explanation higher on subjective understanding, satisfaction, and ability to impact the outcome. It is surprising that the positive/negative outcome impacts the user goal measures for participants in the developer role, which can be interpreted as our manipulation not being strong enough or individuals finding it difficult not to be impacted by the reject/accept language, even when it is not meant to represent their own outcome.

Figure 12: Average explanation ratings for approve vs. reject conditions in Study 2B.

(a) Right Branch



(b) Left Branch



Second, as shown in Table 7, the simple and complex tree explanations (relative to no explanation) significantly improve self-reported subjective understanding and satisfaction in nearly all conditions. Thus, we mitigate the negative effects of the decision tree explanation that we found among participants in Study 1B who were in the approve condition.

Table 7: Summary of significant effects ($p < 0.05$) of different explanation modes (relative to no explanation), separated by whether participants were in the approve or reject condition in Study 2B.

(a) Approve, Right Branch

Explanation Mode	Understanding	Satisfaction	Impact
Simple Tree	+	+	
Complex Tree			

(b) Reject, Right Branch

Explanation Mode	Understanding	Satisfaction	Impact
Simple Tree	+	+	
Complex Tree	+	+	

(c) Approve, Left Branch

Explanation Mode	Understanding	Satisfaction	Impact
Simple Tree	+	+	
Complex Tree	+	+	

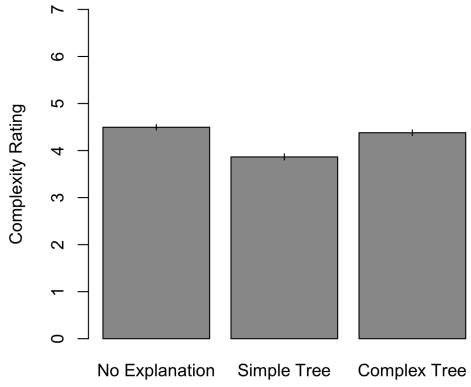
(d) Reject, Left Branch

Explanation Mode	Understanding	Satisfaction	Impact
Simple Tree	+	+	
Complex Tree	+	+	

However, it is important to note that there is no significant difference in the user goal measures between the simple tree and complex tree explanations, except for the positive right branch, where the complex tree did not improve understanding and satisfaction significantly. Overall, while we have shown that decision trees can improve the user goal measures, there are not significant gains to using more complex trees. In fact, in some cases, such as the approve, right branch condition in our study, the simple tree actually results in higher understanding and satisfaction than the complex tree.

In Figure 13 we plot the average complexity ratings for the no explanation, simple tree, and complex tree conditions. We see that the simple tree falls significantly below people's expectations with no explanation, while the complex tree is roughly equal in complexity. Thus, we find evidence against the explanation that individuals do not like an algorithm that seems too simple, and instead conclude that the negative effect of the decision tree in Study 1B was more likely due to the tree branchings not matching consumer expectations.

Figure 13: Average complexity ratings for different explanation modes in Study 2B.



To further explore the underlying mechanism of positive outcome participants reporting higher subjective understanding compared to negative outcome participants, we measured objective understanding using the answers to the free response questions. Recall that all participants in Study 2B were asked to explain in their own words why the application was accepted/rejected, as well as

to generate a counterfactual or alternative profile that would result in the opposite outcome. Table 8 gives several sample responses from participants who were in the simple tree and applicant conditions. We see that correct responses may vary depending on whether the scenario went down the right versus left branches of the decision tree.

Table 8: Sample explanation and counterfactual responses from participants across four outcome and branch conditions in Study 2A (simple tree, applicant role condition).

Condition	**Explanation**	**Counterfactual**
Approved, Right Branch	"The customer had a credit score over of over 618. Since their length of employment was less than 4 years, their delinquencies (0) was also checked. Thus the customer was approved." *(high score)*	"Credit score of 600 with no employment and many delinquencies." *(high score)*
	"I was approved because I have 0 delinquencies" *(low score)*	"If they would have had one or more hard inquiries or any delinquncies, they would have failed." *(low score)*
Rejected, Right Branch	"With a credit score of 650, the algorithm next examined the length of employment. Because that value was 3 years, it checked delinquencies. Because there had been two (the "one or more" branch), the applicant was rejected." *(high score)*	"1st option: Since this customer meets the 62K or more income requirement, all she would need to do is wait until her hard inquiries fall off. 2nd option: If she improves her credit score beyond 618 and stays with her employer for one more year she would be approved." *(high score)*
	"Income Is Too Low. The income required for a credit card varies by credit card issuer" *(low score)*	"The profile would change if the income was higher" *(low score)*
Approved, Left Branch	"The customer has a credit score below the threshold, however their income and lack of hard inquiries means they are approved" *(high score)*	"The person was approved because of employment years" *low score)*
	"My application would have been rejected based on my relatively low credit score if wither my income was below the threshold of 62000 or if over that I had one or more hard inquiries on my credit" *(high score)*	"If they had been late in their payments, resulting in delinquencies, they would be rejected" *(low score)*
Rejected, Left Branch	"I was rejected because my credit score was less than 618, my yearly income was 75000 but I have had 2 hard inquiries which results in rejection" *(high score)*	"I was probably rejected for not making enough, too many inquiries and delinquencies." *(low score)*
	"If the customer had the same credit score of 580 with a yearly income of $75,000 and no hard inquiries, the customer would be approved rather than rejected." *(high score)*	"If I had no delinquencies & had worked for my 4 or more years for my current employer." *(low score)*

For each of the 6 features, participant's responses received one point for correctly mentioning a feature that was directly relevant to the decision or correctly omitting a feature that was not directly relevant. Which features were considered relevant or irrelevant depended on the tree and branch. For example, a participant in the simple tree, right branch condition would receive points for mentioning credit score, employment, and delinquencies, as well as points for *not* mentioning income, credit history length, and hard inquiries. Thus, we assigned a score out of 6 points for each participant's explanation and counterfactual responses. Note that we do not compute a feature score for participants in the no explanation condition. As illustrated in the examples in Table 8, some responses may have been correct but scored lower because they didn't discuss as many features.

We find a significant positive correlation between the self-reported subjective understanding rating and the feature score of the explanation response ($r = 0.20$, $t(1603) = 8.34$, $p < 0.001$), as well as a positive correlation between subjective understanding and the feature score of the counterfactual response ($r = 0.12$, $t(1603) = 4.68$, $p < 0.001$).[8] This suggests that although our measures of subjective and objective understanding are not exactly equivalent, they are related and participants are generally able to accurately evaluate their own understanding.

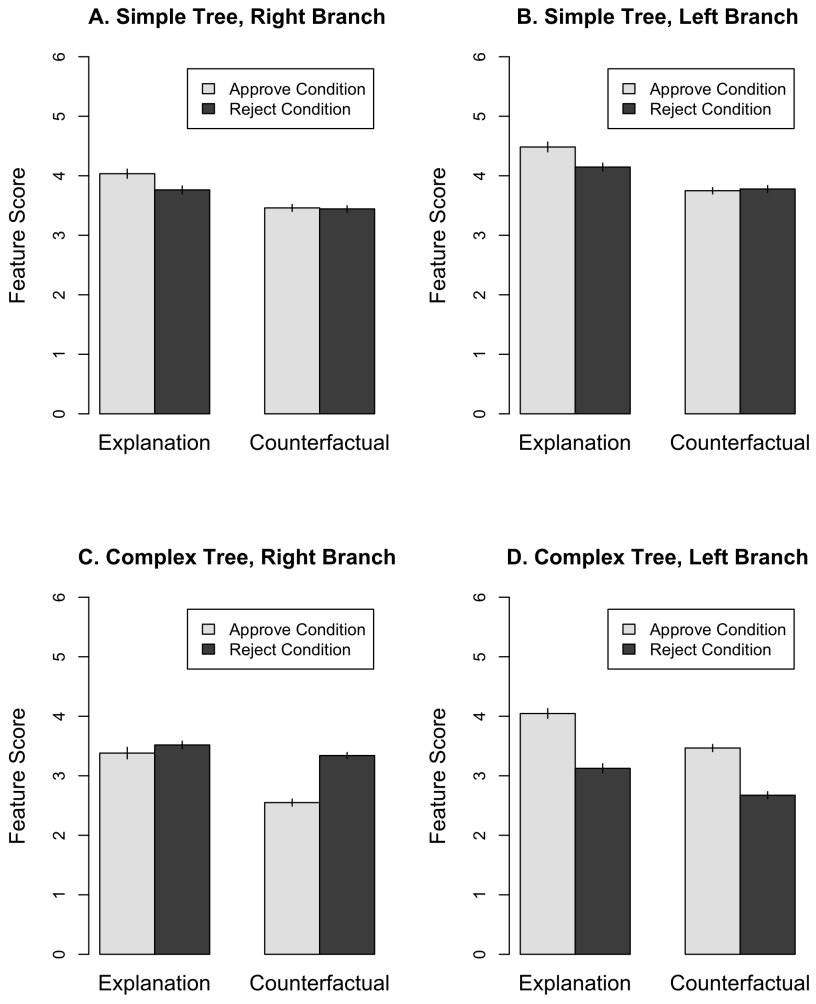
Figure 14 plots comparisons of the feature scores across the different scenarios. In panels A and B we compare the feature scores for the explanation and counterfactual responses among participants who saw the simple decision tree explanation. We see that for both the right and left branch scenarios, participants in the approve condition mentioned/omitted more features correctly in their response explanations compared to participants in the reject condition. In other words, it appears that participants in the approve condition considered more of the information and features overall. In contrast, there was no difference in feature scores for the counterfactual between the approve and reject conditions, indicating that participants in both conditions were able to consider roughly the same number of features that could be changed to generate the opposite outcome.

The results for the complex tree are less straightforward. For the right branch (panel C), there was no difference in the feature score for the explanation between conditions, but people in the reject

[8] We obtain similar results using different coding systems, such as excluding penalties for omitted features, different weightings, etc.

condition mentioned more correct features in the counterfactual response. For the left branch (panel D), participants in the approve condition mentioned more correct features for both the explanation and the counterfactual responses. These results indicate that as the complexity of a decision tree grows, it becomes more difficult to determine an individual's understanding of how the decisions are made since the algorithm's decisions can be attributed to many more features, and there are many more possible counterfactual scenarios.

Figure 14: Feature scores for explanation and counterfactual responses across different conditions in Study 2B.



Finally, we also examine whether the role of the user impacted any of the user goals. Overall, we find that developers were more satisfied compared to applicants with the algorithmic explanations

($t(2205.7) = 2.37$, $p = 0.018$), while applicants thought that the ability to impact the outcome was higher ($t(2174.3) = 2.54$, $p = 0.011$). However, there were no significant interactions between the user role and the other manipulations. Thus, it appears in our context that the strong effects of positive vs. negative outcomes and explanation modes apply to both applicants and developers. In future work, the distinction between applicant and developer roles may be heightened by surveying actual developers or by incentivizing participants based on prediction accuracy of the algorithm, which is more closely aligned with typical developer goals.

In summary, in Study 2B, by carefully constructing decision trees with branchings that match consumer expectations, we demonstrate that decision trees can help increase feelings of understanding and satisfaction relative to no explanation. In addition, we analyze free responses from participants explaining how the decision was made and possible counterfactual scenarios, and demonstrate that the objective "correctness" of the explanations are positively correlated with subjective feelings of understanding. We replicate our finding from Study 1B that positive outcomes lead to greater subjective understanding, and find evidence that this may be related to individuals being able to mention more features correctly in the approve condition. One limitation is that with more complex trees, it becomes much more difficult for individuals to attribute decisions to different features and more difficult for researchers to determine objective understanding (i.e., accuracy of participant responses). Finally, we only found small differences in the user goal measures between applicant and developer conditions, with no interactions with other conditions. To fully understand these differences, it may be necessary to use a stronger manipulation or incentivize different user goals.

STUDY 2C: REPLICATION WITH QUALTRICS PANEL

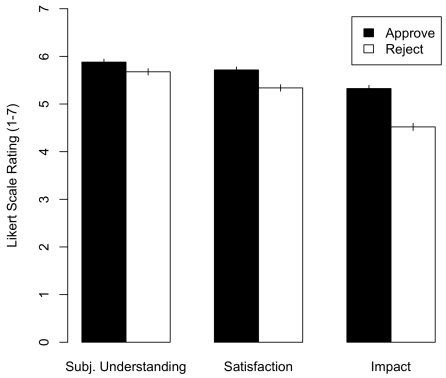
The purpose of Study 2C was to replicate our findings from Study 2B with a different sample of participants. We recruited 2,230 paid participants in the United States using the Qualtrics online panel. The design of Study 2B was identical to Study 2C, except for the demographic questions being placed at the beginning of the survey rather than at the end in order to ensure a sample with demographics roughly representative of the US population. In addition, Qualtrics manually screened out participants who completed the survey too quickly or gave nonsense answers for the

free response questions. Again, participants were randomly assigned to a condition in a 2 (scenario: credit card application approved vs. rejected) $\times$ 3 (explanation mode: none, simple decision tree, complex decision tree) $\times$ 2 (explainee motivation: applicant vs. developer) $\times$ 2 (decision tree branching: right vs. left) between-subjects design.

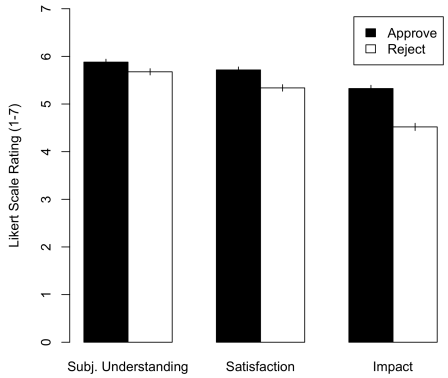
As shown in Figure 15, we replicate our findings from Studies 1B and 2B and find that those in the approve condition tend to rate the user outcome measures higher (i.e., subjective understanding, satisfaction, and ability to impact outcome). To measure objective understanding, we again code each explanation and counterfactual response by assigning them feature scores based on whether each of the 6 features was correctly mentioned or omitted. We again find a significant positive correlation between the subjective understanding rating and the feature score of the explanation response ($r = 0.24$, $t(1471) = 9.49$, $p < 0.001$), as well as the feature score of the counterfactual response ($r = 0.13, t(1471) = 5.18, p < 0.001$). Thus, our measures of subjective and objective understanding are roughly consistent and participants are able to self-assess their depth of comprehension of the algorithm.

Figure 15: Average explanation ratings for approve vs. reject conditions in Study 2C.

(a) Right Branch



(b) Left Branch



Overall, the MTurk participants in Study 2B gave higher ratings of subjective understanding ($M = 6.03$, $SD = 1.22$) compared to the Qualtrics participants in Study 2C ($M = 5.76$, $SD = 1.51$, Welch's $t(4288.4) = 6.55$, $p < 0.001$), with similar differences for satisfaction and ability

to impact the outcome. Consistent with this, MTurk participants exhibited higher feature scores for their explanation responses ($M = 3.81$, $SD = 1.18$) compared to the Qualtrics participants ($M = 3.64$, $SD = 1.17$, Welch's $t(3057.7) = 4.15$, $p < 0.001$). MTurk participants also exhibited higher feature scores for their counterfactual responses ($M = 3.30$, $SD = 0.92$) compared to the Qualtrics participants ($M = 3.19$, $SD = 0.95$, Welch's $t(3031.6) = 4.15$, $p < 0.001$). However, besides these differences, the remaining results were largely the same between the two samples.

In Table 9, we see that the simple and complex tree explanations again significantly improve subjective understanding and satisfaction. Again we find that the degree to which the decision trees improve user outcome measures depends on the scenario (i.e., approve vs. reject, right vs. left branch), and that more complex trees are not necessarily better than simple trees.

Table 9: Summary of significant effects ($p < 0.05$) of different explanation modes (relative to no explanation), separated by whether participants were in the approve or reject condition in Study 2C.

(a) Approve, Right Branch

Explanation Mode	Understanding	Satisfaction	Impact
Simple Tree	+	+	
Complex Tree			

(b) Reject, Right Branch

Explanation Mode	Understanding	Satisfaction	Impact
Simple Tree	+		
Complex Tree	+	+	

(c) Approve, Left Branch

Explanation Mode	Understanding	Satisfaction	Impact
Simple Tree	+	+	
Complex Tree		+	

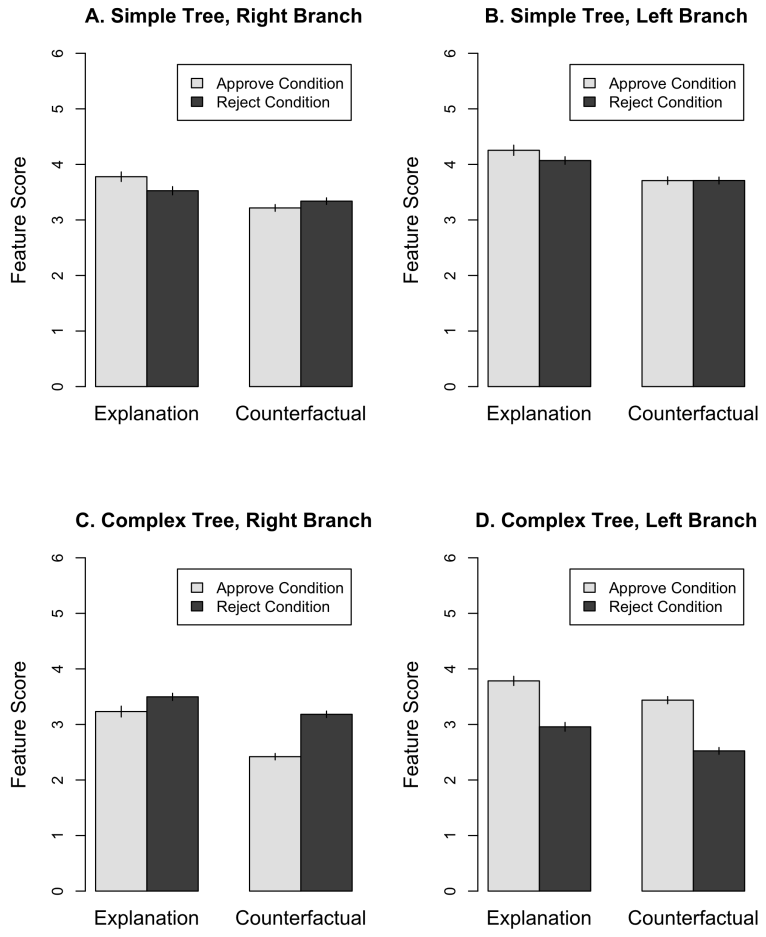
(d) Reject, Left Branch

Explanation Mode	Understanding	Satisfaction	Impact
Simple Tree	+	+	
Complex Tree	+	+	

Finally, Figure 16 plots the feature scores for the simple and complex trees across the different scenario conditions. The results closely replicate our findings from Study 2B (see Figure 14). For the simple tree (panels A and B), approved participants have higher feature scores for the explanation response, indicating that they are able to consider more information compared to rejected

participants. For the counterfactual response, approved and rejected participated consider roughly the same amount of information. For the complex tree (panels C and D), participants in the reject condition score higher for the right branch, while participants in the approve condition score higher for the left branch. In other words, objective understanding becomes more highly dependent on the specific branchings as decision trees become more complex.

Figure 16: Feature scores for explanation and counterfactual responses across different conditions in Study 2C.



POTENTIAL IMPROVEMENTS TO USER OUTCOMES

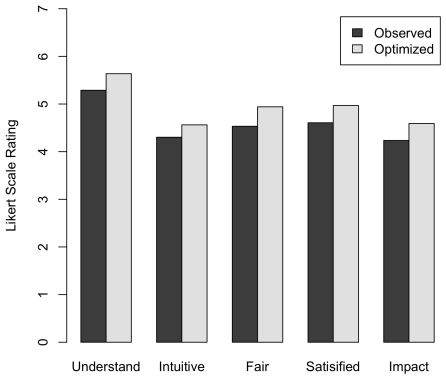
As demonstrated in the previous studies, different explanations maximize different user goals depending on the scenario dimensions and human characteristics. For example, in Study 1B, we find that for subjective understanding the NN global explanation works best for individuals whose credit

loans were approved, while the NN global and local explanation worked best for individuals whose credit loans were rejected. In Studies 2B and 2C, we found that whether the simple or complex decision tree worked best depended on whether the credit card application was approved/rejected and whether the customer profile went down the left/right branch of the tree.

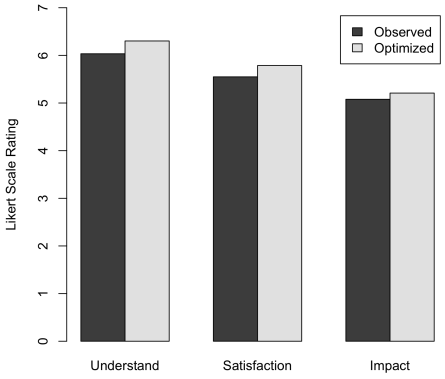
Depending on the availability of prior customer data, there is an opportunity for firms to customize explanations to different users to maximize their goals. To illustrate the potential improvement, within the different study conditions, we calculated what the average user goal measure would have been if they had seen the "best" explanation for their specific scenario condition. Figure 17 plots the comparison between the observed and "optimized" user goal measures.

Figure 17: Comparison of observed and optimized user goal ratings.

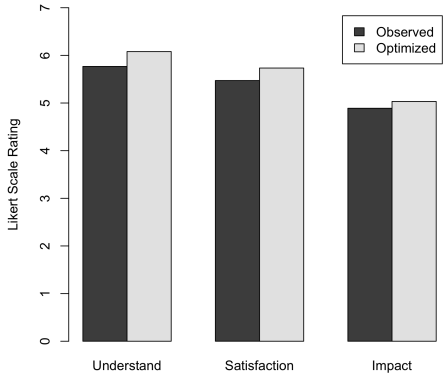
(a) Study 1B



(b) Study 2B



(c) Study 2C



The best explanation was determined by looking at which explanation resulted in the highest average rating for a particular user goal measure among participants within the scenario condition. For example, in Study 1B, among participants in the approve condition, the NN global explanation was rated highest on average compared to the other explanations in terms of subjective understanding, and so to compute an optimized understanding rating, we set the understanding rating of all participants in the approve condition to the average within the NN global condition. In contrast, among participants in the reject condition, the NN global and local explanation was the best, and so we set the optimized understanding rating for the reject condition participants to the average within the NN global and local condition. We repeat this process for all of the user goal measures and find that they result in a 6-9% improvement in the average values, as shown in Figure 17(a).

For Studies 2B and 2C, we computed the best explanation within each accept/reject and right/left branch scenarios, and also separately for the applicant/developer conditions. For example, for participants in the approve, right branch, applicant condition, the simple decision tree resulted in the highest subjective understanding rating. In contrast, the complex decision tree had the highest rating among participants in the corresponding reject condition. Additionally, the best explanation mode may differ depending on the user goal of interest. For example, although the complex tree was best among participants in the reject, right branch, applicant condition in terms of subjective understanding, the simple tree was actually rated higher in terms of the perceived ability to impact the outcome. As shown in Figure 17(b)-(c), this optimization process results in a 3-5% improvement in the various user goal measures.

GENERAL DISCUSSION

In summary, we present a framework for testing the impact of different dimensions of an explanation given to an end-user that is guided and justified by pragmatic theories of explanation. As shown in Figure 1, the explanation mode may have an impact on user goals, with scenario dimensions and user characteristics being potential moderators. Our studies offer an initial test among real human participants of how perceptions of understanding, intuitiveness, and fairness may be impacted by the dimensions of an explanation, and also demonstrate how the framework can be

applied to specific settings. Our findings include both intuitive and counter-intuitive results, and open up several avenues for further testing. This approach is generalizable and empirical applications of our framework can inform researchers on the factors to consider in XAI development, and managers on tailoring explanations for customers that maximize objectives such as user trust and adoption.

Although the primary user goals of interest in our studies fall under the epistemic category (i.e., subjective and objective understanding), we find that other important user goals such as satisfaction and perceived fairness are all highly positively correlated with subjective feelings of understanding. This suggests that by elucidating the inner workings of algorithmic decision making to consumers, firms can potentially foster trust and adoption among users. In addition, an important area of future research lies in whether different explanation modes may enable individuals to make better decisions, for example, by combining their understanding of the algorithm and their own judgment.

One consistent finding across both the credit loan and credit card studies is that positive vs. negative outcomes seem to impact the user goal ratings. We find some evidence that the higher understanding among participants in the approve conditions may be due to the greater ease with which they may generate reasons for why the loan was approved, compared to those in the reject condition. In addition, we find that user outcomes are also sensitive to other profile characteristics, such as which branch of the decision tree the profile lies. Together, our findings indicate that the scenario dimensions play an important role in determining how to provide good explanations to users. Another possibility to explore in future research is how scenario dimensions may actually change the user's goals. For example, a customer who is rejected for a loan is likely much more motivated to understand why the decision was made compared to a customer who was approved, since the rejected customer may take some action in the future to improve their chances of obtaining a loan.

In this paper, we specifically highlight two XAI methods when varying the explanation mode within our framework: decision trees and SHAP. As discussed at the beginning, decision trees fall under transparent models, while more complex models like neural networks need post-hoc analysis

such as SHAP. Additionally, decision trees may actually be used as a method of post-hoc analysis. Of course, many post-hoc XAI methods are meant to be approximations, and there has been some debate about the accuracy/transparency trade off. In our studies, although we did not explicitly vary accuracy, we did vary the complexity of the decision tree. In most circumstances, a more complex decision tree might be more accurate (especially if it were used as the post-hoc explanation of a more sophisticated non-transparent model). However, it is not always the case that a complex tree results in the highest understanding. Thus, our framework may be used by firms to understand the right degree of explanation that is accurate enough (i.e., from an ethical perspective), but also maximizes user goals. In addition, we found that user understanding and satisfaction with the decision tree explanation were highly dependent on whether the tree branchings matched consumer expectations (i.e., regarding the importance and directional impact of the various features), as well as the particular scenario that defined the user's profile. Thus, there is potential for even more detailed customization of explanations provided to users that may help maximize their goals.

XAI is a rapidly growing field that will play several crucial roles in the future development of human-technology interactions. For developers, XAI provides a way to peek inside the black box of increasingly complex algorithms. For managers and consumers, XAI will be critical in navigating the challenges of ML/AI in areas such as consumer fairness and privacy. For social science researchers, XAI provides a rich avenue for understanding the relationship between explanations and consumer knowledge and decision-making. Our framework for understanding what makes for a good explanation for AI output and empirical studies within the context of financial decisions will hopefully encourage additional research at the intersection of XAI and social science.

REFERENCES

Achinstein P (1983). *The Nature of Explanation*. Oxford University Press.

Alba, Joseph W., and J. Wesley Hutchinson (1987). "Dimensions of consumer expertise." *Journal of Consumer Research*, 13(4): 411-454.

Andrews Robert, Joachim Diederich J, and Alan B. Tickle (1995). "Survey and critique of techniques for extracting rules from trained artificial neural networks." *Knowledge-Based Systems*, 8(6):373-89.

Angwin, Julia (2016). "Making algorithms accountable." *The New York Times*.

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio GIl-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion* 58: 82-115.

Brenna, Francesco, Giorgio Danesi, Glenn Finch, Brian Goehring and Manish Goyal (2018). "Shifting toward enterprise-grade AI: Resolving data and skills gaps to realize value." *IBM Institute for Business Value.*

Chen, Tianqi, and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system." In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794.

Cowgill, Bo, and Catherine E. Tucker (2019). "Economics, fairness and algorithmic bias." *Working Paper.*

Cramer, Henriette, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga (2008). "The effects of transparency on trust in and acceptance of a content-based art recommender." *User Modeling and User-Adapted Interaction*, 18(5): 455-496.

De Regt, Henk W., and Victor Gijsbers (2016). "How false theories can yield genuine understanding." In *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, pp. 66-91, Routledge.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey (2014). "Algorithm aversion: People erroneously avoid algorithms after seeing them err." *Journal of Experimental Psychology: General*, 144(1): 114-126.

Dua, Dheeru, and Casey Graff (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Eil, David, and Justin M. Rao (2011). "The good news-bad news effect: Asymmetric processing of objective information about yourself." *American Economic Journal: Microeconomics*, 3(2): 114-138.

Elgin, Catherine (2007). "Understanding and the facts." *Philosophical Studies*, 132(1): 33-42.

Fagerlin, Angela, Peter H. Ditto, Joseph H. Danks, and Renate M. Houts (2001). "Projection in surrogate decisions about life-sustaining medical treatments." *Health Psychology*, 20(3): 166-175.

Gino, Francesca, and Don A. Moore (2007). "Effects of task difficulty on use of advice." *Journal of Behavioral Decision Making*, 20(1): 21-35.

Gonzalez, Richard, and George Wu (1999). "On the shape of the probability weighting function." *Cognitive Psychology*, 38(1): 129-66.

Grove, William M., and Paul E. Meehl (1996). "Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy." *Psychology, Public Policy, and Law*, 2(2): 293-323.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi (2019). "A survey of methods for explaining black box models." *ACM computing surveys (CSUR)*, 51(5): 1-42.

Hardin R (2002). *Trust and trustworthiness*. Russell Sage Foundation.

Highhouse, Scott (2008). "Stubborn reliance on intuition and subjectivity in employee selection." *Industrial and Organizational Psychology*, 1(3): 333-342.

Hoff, Kevin Anthony, and Masooda Bashir (2015). "Trust in automation: Integrating empirical evidence on factors that influence trust." *Human Factors*, 57(3): 407-34.

Hosanagar K (2020). *A Human's Guide to Machine Intelligence: How Algorithms are Shaping Our Lives and how We Can Stay in Control*. Penguin Books.

Immordino-Yang, Mary Helen, and Matthias Faeth (2010). "The role of emotion and skilled intuition in learning." *Mind, Brain, and Education: Neuroscience Implications for the Classroom*, pp. 69-83.

Johnson, Eric J., Suzanne B. Shu, Benedict GC Dellaert, Craig Fox, Daniel G. Goldstein, Gerald Häubl, Richard P. Larrick, John W. Payne, Ellen Peters, David Schkade, Brian Wanskin, and Elke U. Weber (2012). "Beyond nudges: Tools of a choice architecture." *Marketing Letters*, 23(2): 487-504.

Kim, Tae Wan, and Bryan Routledge (2020). "Why a right to an explanation of algorithmic decision-making should exist: A trust-based approach." *Business Ethics Quarterly, Forthcoming*.

Kizilcec, René F (2016). "How much information? Effects of transparency on trust in an algorithmic interface." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems 2016 May 7*, pp. 2390-2395.

Lage, Isaac, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez (2016). "An evaluation of the human-interpretability of explanation." *arXiv preprint arXiv:1902.00006.*

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning (2015)." *Nature*, 521(7553): 436-44.

Lipton Peter (2008). CP Laws, reduction and explanatory pluralism. *Being Reduced: New Essays on Reduction, Explanation, and Causation*, pp. 116-125.

Lipton, Zachary C. (2018). "The mythos of model interpretability." *Queue*, 16(3): 31-57.

Lombrozo, Tania, and Susan Carey (2006). "Functional explanation and the function of explanation." *Cognition*, 99(2): 167-204.

Mantzavinos, Chrysostomos (2016). *Explanatory Pluralism*, Cambridge University Press.

McCauley, Robert N. (1996). "Explanatory pluralism and the coevolution of theories in science." *The Churchlands and Their Critics*, pp. 17-47.

Miller, Tim (2019). "Explanation in artificial intelligence: Insights from the social sciences." *Artificial Intelligence*, 267: 1-38.

Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller (2018). "Methods for interpreting and understanding deep neural networks." *Digital Signal Processing*. 73:1-5.

Polman, Evan (2012). "Self–other decision making and loss aversion." *Organizational Behavior and Human Decision Processes*, 119(2): 141-50.

Poursabzi-Sangdeh, Forough, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach (2018). "Manipulating and measuring model interpretability." *arXiv preprint arXiv:1802.07810.*

Putnam, Hilary (1960)." Minds and Machines." In *Dimensions of Mind: A Symposium*, Edited by: Hook S., pp. 138–164, New York: Collier.

Radin, Margaret Jane (2012). *Boilerplate: The Fine Print, Vanishing Rights, and the Rule of Law*. Princeton University Press.

Ransbotham, Sam, David Kiron, Philipp Gerbert, and Martin Reeves (2017). "Reshaping business with artificial intelligence: Closing the gap between ambition and action." *MIT Sloan Management*

Review, 59(1).

Roff, Heather M., and David Danks (2018). " "Trust but Verify": The difficulty of trusting autonomous weapons systems." *Journal of Military Ethics*, 17(1): 2-20.

Rudin, Cynthia (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence*, 1(5): 206-215.

Shaffer, Victoria A., C. Adam Probst, Edgar C. Merkle, Hal R. Arkes, and Mitchell A. Medow (2013). "Why do patients derogate physicians who use a computer-based diagnostic support system?" *Medical Decision Making*, 33(1): 108-18.

Strevens M (2008). *Depth: An Account of Scientific Explanation*. Harvard University Press.

Sweeny, Kate, Darya Melnyk, Wendi Miller, and James A. Shepperd (2010). "Information avoidance: Who, what, when, and why." *Review of General Psychology*, 14(4): 340-53.

Tintarev, Nava, and Judith Masthoff (2015). "Explaining recommendations: Design and evaluation." In *Recommender Systems Handbook*, pp. 353-382, Springer, Boston, MA.

Tversky, Amos, and Daniel Kahneman (1992). "Advances in prospect theory: Cumulative representation of uncertainty." *Journal of Risk and Uncertainty*, 5(4): 297-323.

Van Bouwel, Jeroen, and Erik Weber (2008). "A Pragmatist Defense of Non-Relativistic Explanatory Pluralism in History and Social Science." *History and Theory*, 47(2): 168-82.

Van Dijk, Wilco W., and Joop Van der Pligt (1997)." The impact of probability and magnitude of outcome on disappointment and elation." *Organizational Behavior and Human Decision Processes*, 69(3): 277-284.

Van Fraassen, Bas (1988)." The pragmatic theory of explanation." *Theories of Explanation*, pp. 135-55.

Vasilyeva, Nadya, Daniel Wilkenfeld, and Tania Lombrozo (2017). "Contextual utility affects the perceived quality of explanations." *Psychonomic Bulletin & Review*, 24(5): 1436-1450.

Wang, Weiquan, and Izak Benbasat (2007). "Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs." *Journal of Management Information Systems*, 23(4): 217-246.

Wu, Yan, and Xiaolin Zhou (2009). "The P300 and reward valence, magnitude, and expectancy in outcome evaluation." *Brain Research*, 1286:114-122.

Zech, John R., Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric K. Oermann (2018). "Confounding variables can degrade generalization performance of radiological

deep learning models." *arXiv preprint arXiv:1807.00431.*

Web Appendix For
"Good Explanation for Algorithmic Transparency"

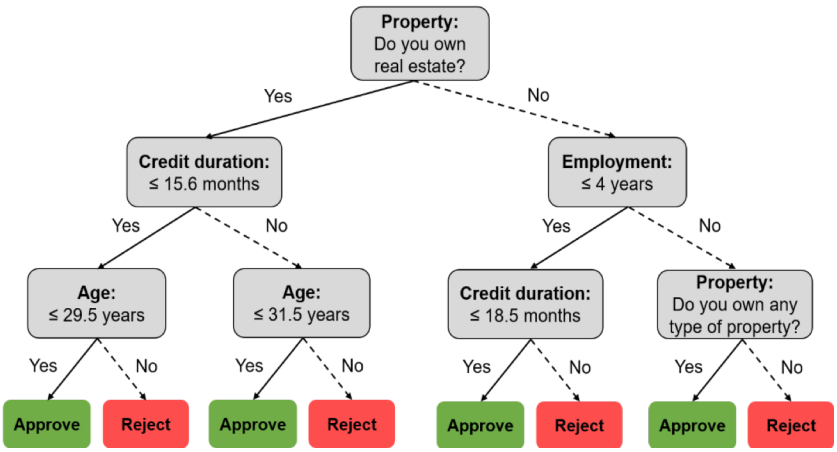
Web Appendix A: Machine Learning Algorithms

Decision Tree

Decision trees are a type of nonparametric supervised learning algorithm that can be used for both classification and regression problems. In our case, we use it for classification (i.e., "Good" or "Bad" credit risk). Decision trees, along with simple rule-set approaches, are one of the most human-interpretable or transparent algorithms when the size (i.e., depth) of the tree is not too large. Briefly described, the tree algorithm splits data feature space into subregions to maximize the purity of training data labels (i.e., each region contains largely one class). The splitting nodes of the decision tree are generated based on the principle of largest entropy (a measure of impurity) reduction. For formal introductions and more details, please see Bishop (2006) and Hastie et al. (2005).

To implement a decision tree with the German Credit Data, we utilize the python package Scikit-Learn (Pedregosa et al. 2011). Figure A1 shows an example of a decision tree trained on the data. Each node splits the data into subregions. A complete path from the top root node to the bottom leaf node corresponds to a rule that is easily understandable to humans.

Figure A1: Illustration of decision tree model



Neural Network

Neural networks are the most advanced black box algorithms in use today. A neural network algorithm consists of atomic computation units called neurons. A neuron takes a vector of input, linearly combines it much like linear regression, followed by a nonlinear activation function such as sigmoid. Output is then sent to the next neuron. A network of neurons like these form a neural network to learn a nonparametric function given training data (i.e., a dataset with independent

variables and outcome labels). "Learning" in neural network is casted as an optimization problem that is solved through stochastic gradient descent and network weights are adjusted according to Backpropagation algorithm, which is an applied chain rule. There are many different architectures of neural networks for different types of data ranging from simple tabular data, time series data, to unstructured image and text data. Neural networks have been shown to excel in all types of supervised learning problems and are one of the most widely used types of algorithms in the industry today (LeCun et al. 2015; Goodfellow et al. 2016).

To implement a neural network with the German Credit Data, we utilize Tensorflow (Abadi et al. 2016), a deep learning framework released by Google. In particular, we use a simple fully-connected neural net with 2 hidden layers of size 20 using ReLU activation functions and softmax output layer. For formal introductions and more details, please see Goodfellow et al. (2016).

Web Appendix B: XAI Methods

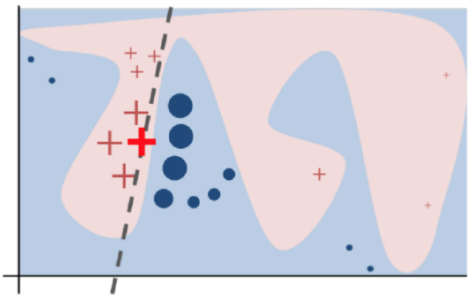
We describe the XAI algorithms used to provide explanations/rationales for black box neural network algorithms. Focusing on our case for explaining algorithmic predictions, one stream in XAI called "explainable machine learning" is most relevant (Rudin 2019). The definition of explainable machine learning is given as follows: *Given a black box predictor B and a training dataset $D = \{X, Y\}$, the explainable machine learning algorithm takes as an input a black box B and a dataset D, and returns a transparent predictor T with requirements that (1) T replicates the predictions of the black box B with high fidelity, and (2) T offers human-understandable rationale for each prediction either at the instance-level or model-average level. T may be a shallow tree, small set of rules, or linear regression with not too many explanatory variables.*

To maximize relevance and impact, we selected two of the most widely-used methods in both industry and academia (e.g., Guidotti et al. 2018; Hall et al. 2017) to construct the stimuli in our study: LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations).

LIME (Local Interpretable Model-Agnostic Explanations)

LIME (Ribeiro et al. 2016) leverages simple linear models, found to be generally more understandable, to explain any complex black box models. LIME approximates any complex black box model using a surrogate interpretable linear model (given input and output prediction) and uses this simpler model to provide explanation for a given data point at the local level. The data for training the simple explanation model is drawn from the neighborhood of the given instance to be explained by perturbation. Figure A2 shows an example of a complex nonlinear model discriminant line approximated by a simple linear model at the local level.

Figure B1: Toy example to present the intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful. Figure is from the original LIME paper.



Intuitively, LIME solves an optimization problem that balances local fidelity and complexity (i.e., the inverse of interpretability) of the interpretable surrogate model. Mathematically, the ex-

planation produced by LIME can be expressed as follows:

$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{1}$$

In Equation (1), x is the instance to be explained, G is the family of all possible explanation models (e.g., all possible linear regression models), and g is the best one among this family that minimizes the prediction loss L of g compared to the prediction of the original model f, plus the model complexity $\Omega(g)$. The lower the model complexity $\Omega(g)$, the easier it is to understand the rationale for prediction made by the model g. For example in the context of linear models, $\Omega(g)$ may be the number of coefficients. In the context of decision trees, $\Omega(g)$ may be the depth of the tree. π_x defines the proximity of the neighborhood around instance x and how much the loss function weighs a particular set of data points bootstrapped by perturbing instance x.

SHAP (SHaply Additive exPlanations)

SHAP (Lundberg and Lee 2017) is a general framework that unifies multiple local interpretability methods, including LIME. SHAP finds the feature importance of input variables for a given data point. This approach is based on Shapley Values in game theory, which is a method designed for fairly distributing the reward among the players in a cooperative game based on their contribution. In the SHAP context, each attribute in the data are the "players" and the outcome variable Y is the "reward". One step of SHAP utilizes LIME for obtaining local feature importance. For obtaining global feature importance, for each variable, we can average the SHAP feature importance values over all data points.

Web Appendix C: Study Instructions

Imagine that you are planning to purchase a new car. In order to do so, you are applying for a credit loan approval from a regional bank. Suppose that this bank determines whether to approve or reject a loan application using a computer algorithm. This algorithm uses 7 different "features" or pieces of information about the customer as the input to make an approve/reject decision. To develop the algorithm, the bank utilized information from 1000 previous loan applicants.

The following table gives the name and a brief description of each of the 7 features. The last column of the table indicates where your data would be entered when applying for a loan. Please take a few moments to scan through the list of features and their descriptions.

Now imagine that you have input your data (for this hypothetical loan application scenario), as it appears in the table below. Please take a moment to scan through your data inputs.

Feature Name	Description	Your Data
1. Credit duration	How long you have to pay off the loan in months	**6 months (accept)** **36 months (reject)**
2. Credit amount	How much the loan is for in dollars	**$2800**
3. Employment	Number of years you have been employed so far (0 if currently unemployed)	**2 years**
4. Installment rate	Monthly payments as a percentage of your disposable income	**3%**
5. Property	Do you own real estate, life insurance, a car, etc.	**Own life insurance**
6. Age	Age in years	**40 years**
7. Credit existence	How many existing loan approvals have you received from this bank in the past	**1 loan**

Now imagine that based on your data, the computer algorithm that the bank uses decides to **approve** [**reject**] your loan application. The algorithm was developed using artificial intelligence and historical user data from 1000 previous loan applicants.

No Explanation Condition

No additional information

Verbal Explanation Condition

The three main features that impacted the decision to approve [reject] your loan application were the following:

- Employment
- Credit duration
- Installment rate

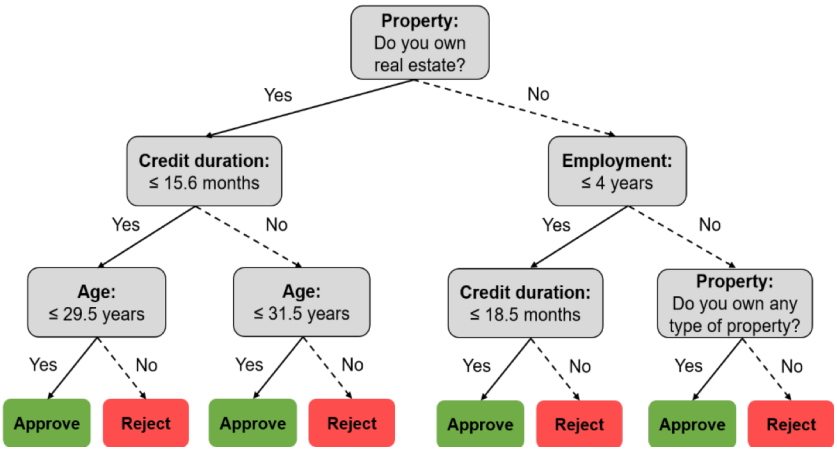
Decision Tree Condition

To help explain how the algorithm worked when making the decision to approve your loan application, the bank has provided the following visual of the "decision tree" model that the algorithm uses. For each loan application, the algorithm first starts at the top of the tree and asks a series of yes/no questions, and follows the tree all the way down to the bottom to make an approve/reject decision based on the user's data.

For example, the first question is about the property feature and asks whether or not you (the user) owns real estate. If yes, then the algorithm moves to the next question on the left (credit duration). If no, then the algorithm moves to the next question on the right (employment).

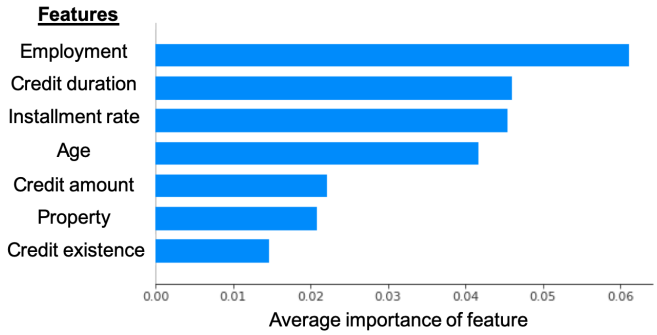
Based on your data, the algorithm followed the following steps:

- Property: Do you own real estate? **No (you own life insurance)** Employment: Less than or equal to 4 years? **Yes (2 years)**
- Credit duration: Less than or equal to 18.5 months? **Yes (6 months) [No (36 months)]**
- **Final Decision: Approve [Reject]**



NN Global Condition

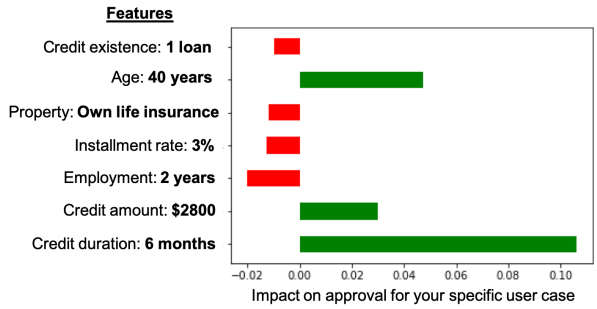
Specifically, the approve/reject decisions are determined using a sophisticated and complex neural network algorithm. To help explain how the algorithm worked when making the decision to approve your loan application, the bank has provided the following visual that shows the average importance of each of the features in the decision to approve/reject a user's loan application (i.e., the absolute value of the Shapley value). In the graph, the larger the blue bar, the more important the feature is **on average** across different customers.



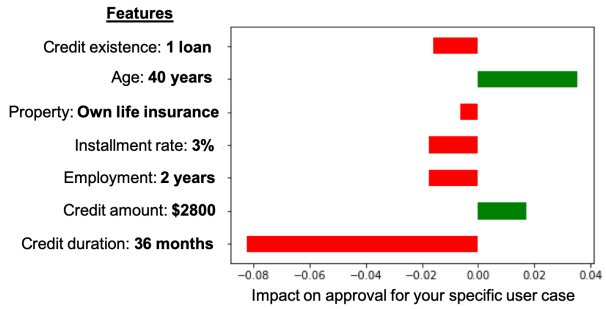
NN Local Condition

Specifically, the approve/reject decisions are determined using a sophisticated and complex neural network algorithm. To help explain how the algorithm worked when making the decision to approve your loan application, the bank has provided the following visual that shows how the different features impact the chance of approval for you and other users like yourself with similar data features. A red bar indicates that the feature had a *negative* impact on approval, while a green indicates that the feature had a *positive* impact on approval. The length of the bars indicates the magnitude of impact.

(a) Approve Condition

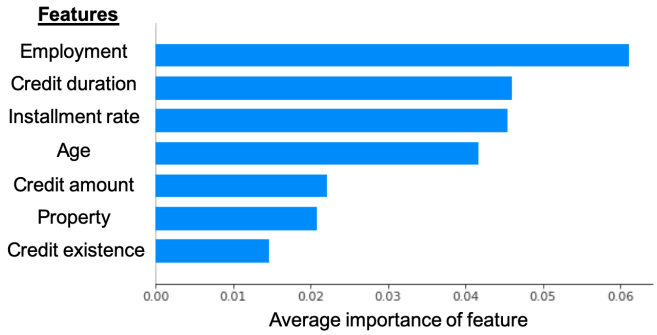


(b) Reject Condition



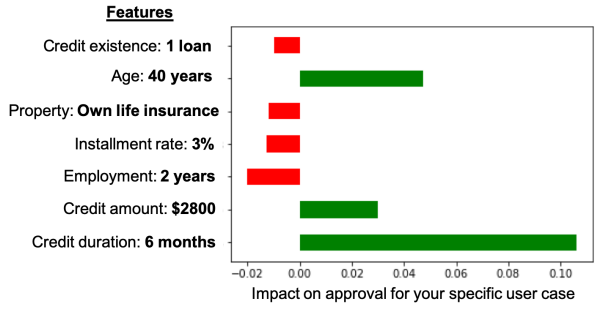
NN Global and Local Condition

Specifically, the approve/reject decisions are determined using a sophisticated and complex neural network algorithm. To help explain how the algorithm worked when making the decision to approve your loan application, the bank has provided the following visual that shows the average importance of each of the features in the decision to approve/reject a user's loan application (i.e., the absolute value of the Shapley value). In the graph, the larger the blue bar, the more important the feature is **on average** across different customers.

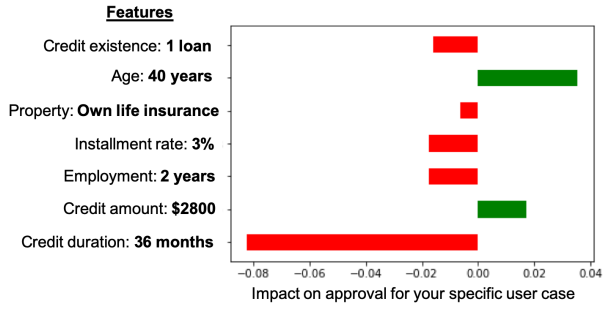


The bank has also has provided the following visual that shows how the different features impact the chance of approval for you and other users like yourself with similar data features. A red bar indicates that the feature had a *negative* impact on approval, while a green indicates that the feature had a *positive* impact on approval. The length of the bars indicates the magnitude of impact.

(a) Approve Condition



(b) Reject Condition



Dependent variables

We would now like you to answer the following questions about the decision. *(Note that in Qualtrics, participants respond on a scale from 1-7).*

- How **understandable** do you find the above explanation?
- How **complex** do you perceive the algorithm to be?

- How **intuitive** do you perceive the algorithm to be?
- How **fair** do you find the algorithm to be?
- How **satisfied** are you with the explanation of the decision?
- How would you rate your ability to **improve your chances** of obtaining this loan approval?
- How **familiar** are you with loan approval systems?

Web Appendix D: Study Analysis and Results

Table D1: Summary of two-way ANOVA analyses for all dependent variables (Study 1B).

	Scenario Dimensions	**Explanation Mode**	**Interaction**
Understanding	$F(1,1193) = 41.80$, $p < 0.001$	$F(5,1193) = 7.48$, $p < 0.001$	$F(5,1193) = 2.11$, $p = 0.061$
Complexity	$F(1,1193) = 2.15$, $p = 0.143$	$F(5,1193) = 2.15$, $p = 0.008$	$F(5,1193) = 2.11$, $p = 0.061$
Intuitiveness	$F(1,1193) = 18.69$, $p < 0.001$	$F(5,1193) = 3.43$, $p = 0.004$	$F(5,1193) = 2.31$, $p = 0.043$
Fairness	$F(1,1193) = 83.30$, $p < 0.001$	$F(1,1193) = 5.38$, $p < 0.001$	$F(5,1193) = 2.77$, $p = 0.017$
Satisfied	$F(1,1193) = 177.67$, $p < 0.001$	$F(1,1193) = 7.07$, $p < 0.001$	$F(5,1193) = 5.39$, $p < 0.001$
Impact	$F(1,1193) = 78.26$, $p < 0.001$	$F(1,1193) = 72.65$, $p < 0.001$	$F(5,1193) = 2.32$, < 0.001
Familarity	$F(1,1193) = 0.07.$ $p = 0.796$	$F(1,1193) = 1.83$, $p < 0.876$	$F(5,1193) = 0.36$, $p = 0.876$

Table D2: Summary statistics for objective variables across participants in the **approve** condition, along with t-statistics and p-values from t-tests comparing each explanation mode to the no explanation baseline (Study 1B).

Explanation Mode	**Understanding**	**Intuitiveness**	**Fairness**	**Satisfied**	**Impact**
None (Baseline)	*M=5.39, SD=1.51*	*M=4.41, SD=1.56*	*M=5.06, SD=1.25*	*M=5.20, SD=1.50*	*M=4.43, SD=1.70*
Verbal	M=5.76, SD=1.23 t=1.89, p=0.060	M=4.68, SD=1.44 t=1.31, p=0.191	M=4.96, SD=1.31 t=0.55, p=0.583	M=5.25, SD=1.43 t=0.35, p=0.811	M=4.82, SD=1.45 t=1.78, p=0.076
Decision Tree	M=4.97, SD=1.50 t=2.00, p=0.046	M=3.88, SD=1.69 t=2.30, p=0.023	M=4.29, SD=1.51 t=3.91, p=0.000	M=4.52, SD=1.61 t=3.10, p=0.002	M=4.17, SD=1.70 t=1.06, p=0.291
Neural Network (global)	M=5.94, SD=1.43 t=2.76, p=0.006	M=4.79, SD=1.34 t=1.89, p=0.060	M=5.30, SD=1.26 t=1.35, p=0.180	M=5.51, SD=1.26 t=2.213, p=0.035	M=4.99, SD=1.58 t=2.44, p=0.016
Neural Network (local)	M=5.72, SD=1.43 t=1.54, p=0.124	M=4.50, SD=1.67 t=0.39, 0.697	M=4.94, SD=1.26 t=0.68, p=0.499	M=5.41, SD=1.45 t=1.03, p=0.302	M=4.80, SD=1.56 t=2.05, p=0.042
Neural Network (global + local)	M=5.65, SD=1.14 t=1.367, p=0.174	M=4.71, SD=1.47 t=1.44, p=0.151	M=4.93, SD=1.41 t=0.69, p=0.493	M=5.37, SD=1.30 t=0.69, p=0.493	M=4.62, SD=1.54 t=0.87, p=0.386

Table D3: Summary statistics for objective variables across participants in the **reject** condition, along with t-statistics and p-values from t-tests comparing each explanation mode to the no explanation baseline (Study 1B).

Explanation Mode	Understanding	Intuitiveness	Fairness	Satisfied	Impact
None (Baseline)	*M=4.45, SD=1.94*	*M=3.89, SD=1.47*	*M=3.98, SD=1.44*	*M=3.20, SD=1.71*	*M=3.61, SD=1.40*
Verbal	M=5.93, SD=1.80 t=2.23, p=0.028	M=4.00, SD=1.59 t=0.51, p=0.612	M=4.04, SD=1.63 t=0.28, p=0.783	M=3.80, SD=1.89 t=2.35, p=0.020	M=3.73, SD=1.78 t=0.50, p=0.618
Decision Tree	M=4.83, SD=1.61 t=1.53, p=0.127	M=4.07, SD=1.59 t=0.828, p=0.409	M=3.86, SD=1.59 t=0.56, p=0.574	M=4.12, SD=1.81 t=3.61, p<0.001	M=3.78, SD=1.83 t=0.72, p=0.472
Neural Network (global)	M=5.12, SD=1.65 t=2.66, p=0.008	M=4.10, SD=1.52 t=0.99, p=0.325	M=4.14, SD=1.46 t=0.77, p=0.441	M=4.21, SD=1.58 t=4.36, p<0.001	M=3.81, SD=1.66 t=0.83, p=0.354
Neural Network (local)	M=5.26, SD=1.63 t=3.26, p=0.001	M=4.26, SD=1.80 t=1.62, p=0.107	M=4.30, SD=1.61 t=1.51, p=0.132	M=4.27, SD=1.87 t=4.27, p<0.001	M=3.76, SD=1.82 t=0.66, p=0.509
Neural Network (global + local)	M=5.33, SD=1.49 t=3.64, p<0.001	M=4.33, SD=1.51 t=2.10, p=0.037	M=4.59, SD=1.63 t=2.79, p=0.006	M=4.32, SD=1.75 t=4.59, p<0.001	M=4.19, SD=1.74 t=2.59, p=0.010

Table D4: Summary of two-way ANOVA analyses for all dependent variables for participants in the **right branch** condition (Study 2B). Note that we collapse across applicant/developer conditions.

	Scenario Dimensions	Explanation Mode	Interaction
Understanding	F(1,1192) = 13.36, p<0.001	F(2,1192) = 39.64, p<0.001	F(2,1192) = 8.53, p<0.001
Complexity	F(1,1192) = 0.72, p = 0.396	F(2,1192) = 16.12, p<0.001	F(2,1192) = 2.60, p = 0.075
Satisfaction	F(1,1192) = 33.88, p<0.001	F(2,1192) = 22.61, p<0.001	F(2,1192) = 10.07, p<0.001
Impact	F(1,1192) = 59.83, p<0.001	F(2,1192) = 0.98, p = 0.374	F(2,1192) = 0.34, p = 0.710
Ethical	F(1,1192) = 9.00, p = 0.003	F(2,1192) = 0.03, p =0.973	F(2,1192) = 3.72, p = 0.025

Table D5: Summary of two-way ANOVA analyses for all dependent variables for participants in the **left branch** condition (Study 2B). Note that we collapse across applicant/developer conditions.

	Scenario Dimensions	Explanation Mode	Interaction
Understanding	F(1,1207) = 39.99, p < 0.001	F(2,1207) = 48.37, p < 0.001	F(2,1207) = 10.45, p < 0.001
Complexity	F(1,1207) = 1.10, p = 0.294	F(2,1207) = 16.13, p < 0.001	F(2,1207) = 0.36, p=0.698
Satisfaction	F(1,1207) = 91.98, p < 0.001	F(2,1207) = 27.76, p < 0.001	F(2,1207) = 4.95, p = 0.007
Impact	F(1,1207) = 73.07, p < 0.001	F(2,1207) = 0.081, p = 0.922	F(2,1207) = 0.03, p = 0.971
Ethical	F(1,1207) = 20.35, p < 0.001	F(2,1207) = 5.79, p = 0.003	F(2,1207) = 6.14, p = 0.002

Table D6: Summary statistics for objective variables across participants in the **right branch, approve** condition, along with t-statistics and p-values from t-tests comparing each explanation mode to the no explanation baseline (Study 2B).

Explanation Mode	Understanding	Satisfied	Impact
None (Baseline)	*M=5.98, SD=1.18*	*M=5.73, SD=1.29*	*M=5.57, SD=1.31*
Simple Tree	M=6.43, SD=0.87 t=4.33, p<0.001	M=6.02, SD=1.10 t=2.36, p=0.014	M=5.50, SD=1.39 t=0.52, p=0.604
Complex Tree	M=6.16, SD=1.11 t=1.55, p=0.120	M=5.82, SD=1.13 t=0.80, p=0.423	M=5.34, SD=1.42 t=1.68, p=0.094

Table D7: Summary statistics for objective variables across participants in the **right branch, reject** condition (Study 2B).

Explanation Mode	Understanding	Satisfied	Impact
None (Baseline)	*M=5.43, SD=1.50*	*M=4.80, SD=1.76*	*M=4.82, SD=1.56*
Simple Tree	M=6.15, SD=1.20 t=5.30, p<0.001	M=5.69, SD=1.43 t=5.59 p<0.001	M=4.81, SD=1.49 t=0.10 p=0.921
Complex Tree	M=6.28, SD=0.97 t=6.77, p<0.001	M=5.73, SD=1.37 t=5.96 p<0.001	M=4.77, SD=1.73 t=0.35, p=0.725

Table D8: Summary statistics for objective variables across participants in the **left branch, approve** condition (Study 2B).

Explanation Mode	Understanding	Satisfied	Impact
None (Baseline)	*M=5.96, SD=1.17*	*M=5.61, SD=1.33*	*M=5.35, SD=1.20*
Simple Tree	M=6.36, SD=0.98 t=3.79, p<0.001	M=5.98, SD=1.10 t=3.07, p=0.002	M=5.40, SD=1.38 t=0.31, p=0.755
Complex Tree	M=6.31, SD=0.97 t=3.26, p=0.001	M=5.97, SD=1.10 t=3.07, p=0.003	M=5.41, SD=1.38 t=0.40, p=0.690

Table D9: Summary statistics for objective variables across participants in the **left branch, reject** condition (Study 2B).

Explanation Mode	Understanding	Satisfied	Impact
None (Baseline)	*M=5.10, SD=1.66*	*M=4.51, SD=1.75*	*M=4.64, SD=1.58*
Simple Tree	M=6.12, SD=1.11 t=7.27, p<0.001	M=5.31, SD=1.57 t=4.81, p<0.001	M=4.68, SD=1.64 t=0.26, p=0.795
Complex Tree	M=6.13, SD=1.04 t=7.57, p<0.001	M=5.46, SD=1.29 t=6.23, p<0.001	M=4.65, SD=1.70 t=0.06, p=0.956

Table D10: Summary of two-way ANOVA analyses for all dependent variables for participants in the **right** branch condition (Study 2C). Note that we collapse across applicant/developer conditions.

	Scenario Dimensions	**Explanation Mode**	**Interaction**
Understanding	F(1,1116) = 2.57, p = 0.109	F(2,1116) = 10.04, p $<$ 0.001	F(2,1116) = 8.27, p $<$ 0.001
Complexity	F(1,1116) = 1.17, p = 0.279	F(2,1116) = 8.79, p $<$ 0.001	F(2,1116) 2.82, p = 0.060
Satisfaction	F(1,1116) = 17.96, p $<$ 0.001	F(2,1116) = 9.36, p $<$ 0.001	F(2,1116) = 7.01, p $<$ 0.001
Impact	F(1,1116) = 70.69, p $<$ 0.001	F(2,1116) = 0.37, p=0.689	F(2,1116) = 0.412, p = 0.662
Ethical	F(1,1116) = 3.00, p = 0.083	F(2,1116) = 0.22, p = 0.799	F(2,1116) = 0.31, p = 0.737

Table D11: Summary of two-way ANOVA analyses for all dependent variables for participants in the **left** branch condition (Study 2C). Note that we collapse across applicant/developer conditions.

	Scenario Dimensions	**Explanation Mode**	**Interaction**
Understanding	F(1,1102) = 15.52, p $<$ 0.001	F(2,1102) = 17.13, p $<$ 0.001	F(2,1102) = 0.43, p = 0.649
Complexity	F(1,1102) = 1.26, p = 0.263	F(2,1102) = 11.72, p $<$ 0.001	F(2,1102) = 1.53, p = 0.217
Satisfaction	F(1,1102) = 37.53, p $<$ 0.001	F(2,1102) = 8.72, p $<$ 0.001	F(2,1102) = 0.17, p = 0.846
Impact	F(1,1102) = 72.54, p $<$ 0.001	F(2,1102) = 0.08, p = 0.923	F(2,1102) = 2.14, p =0.118
Ethical	F(1,1102) = 10.47, p = 0.001	F(2,1102) = 0.128, p = 0.880	F(2,1102) = 0.54, p = 0.582

Table D12: Summary statistics for objective variables across participants in the **right branch, approve** condition (Study 2C).

Explanation Mode	**Understanding**	**Satisfied**	**Impact**
None (Baseline)	*M=5.74, SD=1.48*	*M=5.51, SD=1.58*	*M=5.30, SD=1.55*
Simple Tree	M=6.22, SD=1.30 t=3.40, p$<$0.001	M=6.04, SD=1.28 t=3.66, p$<$0.001	M=5.36, SD=1.66 t=0.38, p=0.704
Complex Tree	M=5.69, SD=1.55 t=0.35, p=0.724	M=5.62, SD=1.52 t=0.72, p=0.474	M=5.31, SD=1.66 t=0.09, p=0.928

Table D13: Summary statistics for objective variables across participants in the **right branch, reject** condition (Study 2C).

Explanation Mode	**Understanding**	**Satisfied**	**Impact**
None (Baseline)	*M=5.34, SD=1.52*	*M=5.02, SD=1.58*	*M=4.39, SD=1.60*
Simple Tree	M=5.82, SD=1.63 t=2.94, p=0.003	M=5.34, SD=1.75 t=1.88, p=0.062	M=4.43, SD=1.91 t=0.24, p=0.807
Complex Tree	M=6.06, SD=1.42 t=4.64, p$<$0.001	M=5.73, SD=1.44 t=4.54, p$<$0.001	M=4.59, SD=1.83 t=1.13, p=0.261

Table D14: Summary statistics for objective variables across participants in the **left branch, approve** condition (Study 2C).

Explanation Mode	Understanding	Satisfied	Impact
None (Baseline)	*M=5.62, SD=1.45*	*M=5.46, SD=1.49*	*M=5.32, SD=1.18*
Simple Tree	M=6.17, SD=1.33 t=3.78, p<0.001	M=5.82, SD=1.42 t=2.41, p=0.016	M=5.18, SD=1.71 t=0.92, p=0.356
Complex Tree	M=5.93, SD=1.50 t=1.96, p=0.051	M=5.80, SD=1.40 t=2.23, p=0.026	M=5.39, SD=1.49 t=0.44, p=0.662

Table D15: Summary statistics for objective variables across participants in the **left branch, reject** condition (Study 2C).

Explanation Mode	Understanding	Satisfied	Impact
None (Baseline)	*M=5.74, SD=1.48*	*M=4.81, SD=1.52*	*M=4.46, SD=1.57*
Simple Tree	M=5.87, SD=1.52 t=4.48, p<0.001	M=5.29, SD=1.78 t=2.79, p=0.006	M=4.62, SD=1.93 t=0.88, p=0.379
Complex Tree	M=5.63, SD=1.54 t=2.97, p=0.003	M=5.26, SD=1.59 t=2.79, p=0.006	M=4.33, SD=1.72 t=0.73, p=0.463

Web Appendix References

Bishop CM (2006). *Pattern Recognition and Machine Learning*. Springer Science + Business Media.

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M (2016). Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265-283.

Goodfellow I, Bengio Y, Courville A (2016). *Deep Learning*. MIT press.

Hall P, Gill N, Kurka M, Phan W (2017). Machine Learning Interpretability with H2O Driverless AI.

Hastie T, Tibshirani R, Friedman J, Franklin J (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer* 27(2):83-5.

LeCun Y, Bengio Y, Hinton G (2015). Deep learning. *Nature* 521(7553):436-44.

Lundberg SM, Lee SI (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765-4774.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825-30.

Ribeiro MT, Singh S, Guestrin C (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135-1144. ACM.

Rudin C (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5):206-215.