In this article, we show that explanations of decisions made by machine learning systems can be improved by not only explaining why a decision was made but also explaining how an individual could obtain their desired outcome. We formally define the concept of directive explanations (those that offer specific actions an individual could take to achieve their desired outcome), introduce two forms of directive explanations (directive-specific and directive-generic), and describe how these can be generated computationally. We investigate people’s preference for and perception toward directive explanations through two online studies, one quantitative and the other qualitative, each covering two domains (the credit scoring domain and the employee satisfaction domain). We find a significant preference for both forms of directive explanations compared to non-directive counterfactual explanations. However, we also find that preferences are affected by many aspects, including individual preferences and social factors. We conclude that deciding what type of explanation to provide requires information about the recipients and other contextual information. This reinforces the need for a human-centered and context-specific approach to explainable AI.

CCS Concepts: • Human-centered computing → User studies; • Computing methodologies → Artificial intelligence; Machine learning;

Additional Key Words and Phrases: Explainable AI, directive explanations, counterfactual explanations
1 INTRODUCTION

Machine learning models are increasingly playing a critical role in decision-making in various domains, such as medicine, law, and banking [3, 4, 16, 36, 40, 42]. One of the aims of explaining decisions made by or with the aid of a machine learning model is to enable recourse, that is, to help individuals understand what they could change to receive a different outcome in the future [51, 74, 76, 78]. For example, when the use of machine learning models leads to the denial of a loan application, the explanation should not only describe the reasoning that led to the decision but also help the customer understand what they could do in the future to get the loan approved [72].

Counterfactual explanations have the potential to enable recourse [76, 78]. Counterfactuals (or counterfactual states) “describe how the world would have (had) to be different for a desirable outcome to occur” [78]. However, not all counterfactuals are actionable. For example, consider a loan applicant being told that to have their loan approved, they would have had to have no prior loan defaults in the previous 5 years; this explanation does not facilitate recourse since nothing can be done to alter history. For counterfactual explanations to enable recourse, explanations should be based on actionable input features [76]. Utsun et al. [76] propose a method for generating actionable explanations or flip sets, that is, explanations with actionable features that guarantee the desired outcome. A challenge of this approach is that some features, such as education level or income, may be mutable only for some people. This problem is usually resolved by offering multiple diverse counterfactual explanations [65, 76–78] with the hope that at least one explanation is suitable for the recipient.

While multiple counterfactuals may provide some guidance as to what circumstances would result in a different outcome (e.g., a loan being approved), they do not explicitly indicate which actions may lead to this desired result; that is, they do not provide explicit recommendations on how to act [38]. Depending on the context, how to reach the counterfactual state might not be apparent to an individual [61]. In an AI planning sense [30], counterfactual explanations provide the initial state (current instance) and the goal state (the counterfactual state), resulting in the desired outcome (decision). However, the actions that would take a person from the current state to the counterfactual state are not part of the explanation. There is an assumption that each counterfactual maps to a real-world action [6, 66], but this is not always the case [38]. Furthermore, most of the prior works on counterfactual explanations have assumed a one-step decision-making process [54, 65, 76, 78].

To better support recourse, we argue that counterfactual explanations should be directive in that they should include suggestions or recommendations of the action(s) the individual could perform, that is, how to act to get to the counterfactual state. Others have echoed similar sentiments, such as those that advocate for causal models [38–40, 50, 74].

In this article, we contribute toward the goal of making explanations directive. In Section 3, we formally define the concept of directive explanation, and we present a model and implementation for generating directive explanations. This model is based on Markov Decision Processes (MDPs) [7, 60] and gives us a framework to consider a sequence of dependent actions that a person has to take to achieve recourse.
In Sections 4 through 6, we present two studies, the first quantitative and the second qualitative, to investigate participants’ preferences and opinions toward directive explanations in the domains of credit scoring and employee satisfaction.

We conducted two studies to answer two questions: (1) Which of the three types of explanation (non-directive, directive-specific, and directive-generic) is the preferred? (2) What are the reasons someone does or does not prefer directive explanations? We conducted the first study to answer the first research question and a second study to answer the second research question. We conducted these two studies on credit scoring and lending decisions and employee turnover (whether employees were likely to resign).

For each study, we designed eight scenarios, four where the decision was favorable (e.g., the loan was approved) and four not (e.g., the loan was denied). For each scenario, we provided participants with four different types of explanations. The first was non-directive, the second was directive with specific actions, the third was directive with generic action, and the fourth was clearly not sensible and served as an attention check, with us excluding any participant who did not rank this as the least preferred explanation. The non-directive explanation informed the person how the situation must change for the desired goal to be achieved but did not suggest actions to achieve this counterfactual state. For example, the participant might be told that to prevent the employee from resigning, the employee should be required to travel only a medium amount for business, but it would not be explained how this reduction in the amount of business travel could occur. Conversely, the directive-specific explanation recommended specific actions that an individual could take to reach the counterfactual state. For example, the participant might be told that to reduce the amount of business travel from high to medium, client meetings should be conducted online. The directive-generic explanation recommended a generic class of actions. Directive-generic explanations indicate the kinds of actions that could be taken to reach the counterfactual state, but only broadly so individuals still had the freedom to decide which specific actions they would want to take. Participants ranked the four explanations from most to least preferred in the first study and provided the reasons for their choice in the second study.

We ran the studies on Amazon MTurk with 65 participants. We found significant support for the two directive explanations in both domains. In the credit scoring domain, approximately 42%, 31%, and 27% of participants selected directive-specific, directive-generic, and non-directive explanations, respectively, as their most preferred explanation. For the employee satisfaction domain, distributions were 35%, 51%, and 14%, respectively, for directive-specific, directive-generic, and non-directive explanations. The key findings are:

- We find a clear preference for the two directive explanations over non-directive counterfactual explanations in both domains. The non-directive explanation was least preferred.
- Directive-specific explanations are more suited to scenarios where the outcome is unfavorable. For example, when loans were denied or an employee was likely to leave the organization, the participants preferred directive-specific explanations. This suggests that, at least in the two domains we studied, people should have an option to receive directive explanations if they wish.
- The preference for directive-generic explanation may depend on the task. We found that participants in the employee satisfaction domain strongly preferred a directive-generic explanation. This suggests that participants prefer to provide high-level guidance and avoid specific actions when they have their own ideas for solving problems.
- Non-directives may be more suitable when the outcome is favorable, and this was certainly true for the credit scoring domain.
A qualitative analysis of the reasons participants provided for their most preferred explanation revealed that the choice for explanation type depended on multiple factors, such as social factors, and whether the participants judged the directives to be *feasible* for the recipient. These results suggest that even with an efficient computational model (e.g., like our MDP-based model) to generate directive explanations, one cannot a priori decide what type of explanation to provide—one needs further information about recipients’ preferences and contextual information to generate actionable explanations, directive or non-directive. This reinforces the need for a human-centered and context-specific approach to explainable AI.

2 BACKGROUND

Machine-learning-based systems can be complex and opaque, and their use to make critical decisions depends on the degree to which these systems are interpretable, that is, how well people understand the causes of its decision-making [9, 35, 48, 51]. There are several ways of potentially making machine learning models transparent, from using intrinsic or intelligible models [64] to using post hoc methods [1, 31, 48, 53], such as counterfactual explanations [78].

2.1 Counterfactual Explanations

Wachter et al. [78] propose the use of *unconditional counterfactual explanations* for people to understand a decision, contest it, and potentially use the explanation to change the decision or outcome. Rather than discussing the internal logic of a machine learning algorithm, counterfactual explanations describe a dependency on the external facts that led to a decision [26, 78]. The notion of counterfactuals [45] can significantly assist in making machine-learning-based systems interpretable [17, 18]. We scope our discussions to a subset of machine learning models. Specifically, we consider classification problems, which are defined in Definition 2.1. While subsequent discussions are based on classification problems, our discussions and methods can be applied to other forms of machine learning models that solve regression problems.

*Definition 2.1 (Classification Problem).* A classification problem is a tuple \((f, x, y)\), where \(f\) is a machine learning model, \(x \in X\) is a feature vector describing the instance that is being classified, and \(y \in \{0, 1\}\) is the label assigned by \(f\) to \(x\).

In the context of the classification problem, a counterfactual state is a statement of how the world would have to be different for a desirable outcome to occur. Given an input feature \(x\) and the corresponding output by a machine learning model \(f\), a counterfactual explanation is a perturbation of the input, \(x\), such that a different output, \(y\), is produced by the model, \(f\). Wachter et al. [78] propose the following formulation:

\[
c = \arg\min_c \, y_{\text{loss}}(f(c), y) + |x - c|,
\]

where \(y_{\text{loss}}()\) pushes the counterfactual state \(c\) toward a different prediction than the original instance, while the second term keeps the counterfactual close to the original instance using a distance metric.

2.2 Counterfactual Explanations and Recourse

One of the aims of counterfactual explanations is to enable recourse, and recourse is broadly related to several topics in machine learning, such as inverse classification [2], strategic classification [24, 34], adversarial perturbations [28], and anchors [63].

Utsun et al. [76] propose an optimization-based approach using integer programming to evaluate a linear classification model in terms of recourse. Their method shares similarities with existing ones [47, 62, 78] but focuses on suggesting actionable changes and evaluating the feasibility and
difficulty of recourse. Their method enables one to establish whether a person could change the decision of a machine learning model through actionable input variables, and they do this by optimizing a cost function given an input \( x \). They define an action, \( a \), as a change to the value of a feature. They choose actions from a set of actionable features, \( A(\cdot) \), that is, a set of mutable or conditionally mutable features, and each action has a cost. They solve the problem of finding actions that minimize the cost.

Several methods provide multiple counterfactuals to people seeking recourse \([65, 76, 78]\). Offering multiple counterfactuals may ensure that at least one has actionable features for an individual. Recently, others have extended this work \([54, 59]\). Although nearest counterfactual explanations provide an understanding of the most similar set of features that result in the desired prediction, they fall short of giving explicit recommendations on how to act to realize this set of features, and this limits agency for the individual seeking recourse \([38]\). Karimi et al. \([38]\) show that current forms of counterfactuals do not translate to an optimal or feasible set of recommendations. Instead, they propose minimizing the cost of performing actions in a world governed by a set of laws captured in a structural causal model.

### 2.3 Beyond One-step Action for Recourse Using Markov Decision Process

Recently, research has been looking at moving beyond the one-step action assumption prevalent in the space of algorithmic recourse to considering the problem as a multi-step sequential decision-making problem \([14, 55, 57, 74]\).

More recently, Tsirtsis et al. \([74]\) proposed a method to find counterfactual explanations for sequential decision-making processes, modeled as discrete-time Markov Decision Process, where the state and action spaces are discrete and low-dimensional. Their method identifies counterfactual trajectories (sequence of actions) that achieve better outcomes and differ by \( k \) actions from the observed sequence. They model the transition probabilities between a pair of states, given an action, using the Gumbel-Max structural causal model \([57]\) because that delivers a desirable counterfactual stability property and reliable estimation of the counterfactual outcome.

Similar works exist in the space of reinforcement learning \([14, 50, 57]\). For example, Madumal et al. \([50]\) proposed an action influence model to relate actions to states and to explain the learned actions or policies that people readily understand, and Oberst and Sontag \([57]\) use the Gumbel-Max SCM to evaluate counterfactual policies. A few models take advantage of causal assumptions \([25, 38, 40, 43]\) but in the context of one-step action; therefore, they are different from our model and that of Tsirtsis et al. \([74]\). We differ from Tsirtsis et al. \([74]\) in that they generate counterfactual recommendations given an already observed sequence of actions, while we generate the directives (sequence of actions) without reference to any observed trajectories. However, similar to Tsirtsis et al. \([74]\), we model the problem of synthesizing directives as an MDP.

Similar to Karimi et al. \([38]\) and Tsirtsis and Gomez-Rodriguez \([75]\), we believe that actionable counterfactual explanations should provide some guidance to individuals on how to act. In other words, they should be directive. As such, as we take our first steps toward directive explanations, we conducted two online studies to investigate individuals’ perception of and preference for directive explanations relative to merely counterfactual explanations. We discuss the details of the studies and propose a conceptual model capable of generating the directives.

### 3 A MODEL FOR DIRECTIVE EXPLANATIONS

This section formally defines the concept of directive explanations and defines a model for generating directive explanations for classification problems. We focus our discussion and examples on classification, but this can also apply more broadly to regression problems.
Definition 3.1 (Directive Explanation). A directive explanation is a tuple \( de = (f, x, y, C, \Phi, Y') \), in which \( f \) is a machine learning model, \( x \in X \) is the original input vector, \( y = f(x) \) is the current class label, \( C \) is the set of possible counterfactuals such that each \( c_i \in C \) has a different class label (i.e., \( \forall c_i, c_j : i \neq j, f(c_i) \neq f(c_j), f(c_i) \neq y, f(c_j) \neq y \)), \( \Phi \) is the set of possible policies such that each \( \pi_i \in \Phi \) is a policy (a set of directives) that transitions \( x \) to \( c_i \), and \( Y' \) is the set of possible class labels with each \( y'_i = f(c_i), y'_i \in Y' \) being the outcome or class label for each counterfactual \( c_i \in C \).

Our desiderata for such an approach consists of the following. First, the model must generate a set of directives that show how to get from the factual state \( x \) to a counterfactual state, \( c_i \). Actions from \( \pi_i \) must lead from \( x \) to \( c_i \). Second, the model must capture different ways to achieve specific outcomes; that is, getting to each counterfactual state \( c_i \in C \) can be done in multiple different ways. Third, the model must capture inherent uncertainty in the outcomes of these actions in achieving outcomes. Finally, the model should also account for action costs to account for the costs that individuals may incur when trying to reach a counterfactual state using the directives, which allows us to model that some directives are more costly than others, and even to consider different costs for different individuals. To identify potential states that change the outcome, \( C \), we can use any existing counterfactual generator, e.g., \([54, 65]\).

From these desiderata, it is clear that the framework of MDPs \([60]\) is a suitable formalism for modeling this problem. This allows us to use a planning-based approach to generate a policy, \( \pi_i \), that transitions \( x \) to \( c_i \in C \). Policy \( \pi_i \in \Phi \) is the source of the directives in the directive explanations. We define a conceptual model for generating the directives below.

Definition 3.2 (Markov Decision Process \([60]\)). An MDP is a tuple \( \Pi = (S, A, P, R, \lambda) \), in which \( S \) is a set of states; \( A \) is a set of actions; \( P(s, a, s') \) is a transition function from \( S \times A \rightarrow 2^S \), which defines the probability of action \( a \) going to state \( s' \) if executed in state \( s \); \( R(s, a, s') \) is the reward received for transitions from executing action \( a \) in state \( s \) and ending up in state \( s' \); and \( \lambda \) is the discount factor.

MDPs can be conceptualized as graphs that map states with transitions (actions), along with the transition probabilities and rewards. If \( \sum_{s' \in S} P(s, a, s') > 0 \), then this means that action \( a \) is enabled in state \( s \) and will transition to one of the states \( s' \) for which \( P(s, a, s') > 0 \). The discount factor controls how much weight or importance is placed on future rewards.

Definition 3.3 (Planning Problem \([60]\)). A planning problem is a tuple \( (\Pi, I, O) \), in which \( I \in S \) is the initial state and \( O \) is the objective to be achieved. In the simplest case, a goal-directed MDP \([30]\), \( O \) is just a set of goal states, such that \( O \subset S \), but a more common objective is simply to maximize the expected discounted reward \([60]\). The task is to synthesize a policy \( \pi : S \rightarrow A \) from states to actions that starts in state \( I \) and achieves object \( O \).

To show how to apply this to directive explanation, we map Definition 3.3 to Definition 3.1. The initial state \( I = x \) such that \( f(I) = y \), and the objective \( O \) is to “reach” \( c_i \in C \), which is achieved when \( f(c_i) = y'_i \). That is, \( x \) is the initial state and \( c_i \) is one of the “goal states,” which can be modeled as receiving a reward if and only if \( f(c_i) = y'_i \). Conceptually, for each \( c_i \), we want to generate a policy of actions that transition from the initial state \( x \) to the counterfactual state \( c_i \). The solution given for the planning problem \( \pi_i \) is the set of directives. Each action \( a \) is a directive that transitions the state to a new state \( s' \), which represents the perturbed feature vector, \( x' \). For multi-class problems, a simple approach would be to generate a plan, \( \pi_i \in \Phi \), for each \( c_i \in C \) to provide to the user.

There are several ways to solve the planning problem \( \Pi \), such as using dynamic programming or model-free reinforcement learning \([30, 70, 71, 74]\). We have implemented this model using
Monte-Carlo Tree Search [11] to create an approximate policy, \( \pi \). We choose the set of actions, \( A \), such that they modify only mutable features. For each \( a \in A \), we specify exactly how the features are modified by taking directive \( a \). For example, if \( a \) is to cancel a credit card, the feature “number of credit cards” is subtracted by 1. To keep the problem representation simple, for each \( a \in A \), we enumerate multiple versions of the actions, \( a_1, \ldots, a_n \), for every possible assignment of feature values. For example, if an action \( a \) updates a feature, \( f_b \), taking on two values, then we would generate two versions of the action \( a \): \( a_1(b = 0) \) and \( a_2(b = 1) \). We binned the continuous features to use with our method (we tested the model on categorical features only). The tree’s root node is the initial feature vector, \( x \), and each edge represents a possible action. To guide the search toward the counterfactual, \( c_i \in C \), we use a multi-objective reward stated as a linear function of two objectives:

\[
    r_{s'} = (r_{decision} + r_{distance}),
\]

where

\[
    r_{decision} = \begin{cases} 
    \alpha, & f(s') = y' \\
    0, & otherwise \end{cases},
\]

\[
    r_{distance} = \begin{cases} 
    \beta, & dist(s', c) \leq \delta \\
    0, & otherwise \end{cases},
\]

\( y' = f(c) \) is the expected counterfactual outcome, \( s' \in S \) is the state reached after performing the policy \( \pi \), \( dist(s', c) \) is the Euclidean distance (\( \ell_2 \) norm), and \( \delta \) is the radius or distance threshold. The radius \( \delta \) allows us to generate multiple directives within \( \delta \) distance away from \( c \). During the rollout, Upper Confidence Bounds (UCBs) guide the selection of nodes.

For experiments, we set \( \alpha = 0.5 \), \( \beta = 0.5 \), and \( \delta = [1, 10] \) (we arrived at the \( \delta \) values empirically for each scenario to get multiple trajectories for the two types of directive explanation; from our experience, \( \delta \) is scenario- or task-dependent). The rewards were discounted by \( \gamma = 0.8 \); this value was also arrived at empirically. Finally, we chose all categorical features and associated actions, \( A \), to illustrate the directive explanations. We provide an algorithm in Appendix F.

In our implementation, while we have not considered diverse directives, there are numerous methods to measure the plan differences, and these can be used to devise a metric to compute multiple diverse directives [12, 41].

Notice that the set of actions in the policy, \( A_{pi} \subseteq A \), are directive-specific actions. That is, in the policy, \( \pi \), each action \( a \in A \) is directive-specific. In our study in Section 4, we perform post-processing on the \( \pi \) to generate directive-generic explanation. First, we generate a graph that starts with a parent or root node, \( p \). This root node simply performs the role of providing an attachment point for directive-generic explanations. Second, each directive-generic explanation, \( a_{gen} \in A_{gen} \), is connected to \( p \), and then each specific directive, \( a \in A \), is connected with its respective \( a_{gen} \). Finally, during post-processing, we simply replace \( a \) with \( a_{gen} \).

For example, assume that \{“consolidate credit cards,” “pay off credit card”\} \( \in A \), and \{“reduce credit cards”\} \( \in A_{gen} \), and \( p \) is the root node. Then we have an edge from \( p \) to “reduce credit cards.” There will be two edges from “reduce credit cards,” one to “consolidate credit cards” and the other to “pay off credit card.” If the model suggests “pay off credit card,” then this action in the directive-specific explanation is replaced with “reduce credit cards” for the directive-generic version of the explanation.

4 STUDIES

For counterfactual explanations to be directive, we argue that they must provide individuals with recommendations on how to act, as opposed to indicating only what state the individual needs to reach. We wished to know whether individuals preferred directive explanations over mere counterfactual explanations and, if so, whether they preferred specific or generic directive explanations. We conducted two studies to answer two questions: (1) Which of the three types of explanation (non-directive, directive-specific, and directive-generic) is preferred most? (2) What are the reasons someone does or does not prefer directive explanations?
We describe two studies in the following sections. We conducted the first study to answer the first research question: Which of the three types of explanation (non-directive, directive-specific, and directive-generic) is preferred the most? We ran a second study, a qualitative study, to answer the second research question: What are the reasons someone does or does not prefer directive explanations? Our studies involved an automated system explaining to an intermediary why the automated system made a particular decision, such as denying a loan. The intermediary then selected one of the four possible explanations to provide to the client. In many contexts, such as loan applications, we believe that an automated system assists people (loan officers) who assist others (customers). Therefore, this setup allows us to understand what a human considers relevant when explaining decisions to another human and provide insights from this perspective.

We conducted the two studies using scenarios designed around credit risk and employee turnover. We chose the two domains because we anticipated that most participants would be aware of the basics of both domains and, therefore, would not require training to understand the domain concepts. The other reason is that we had experience with the two domains. Finally, both domains are typical case studies in the explainable AI community.

4.1 Explanation Types

We provided participants with three explanation types: (1) non-directive, (2) directive-specific, and (3) directive-generic, as defined below. We presented only one explanation of each type for each scenario to keep the number of explanations of each type consistent across scenarios.

Explanation Type 1 - Non-directive: These were standard counterfactual explanations; that is, they specified which parts of the data would have to change to reverse a decision and to what extent they would need to change. For example, a non-directive explanation to a customer could state the maximum debt-to-income ratio needed to approve the loan. Crucially, the explanation did not include directives on achieving the required change.

Explanation Type 2 - Directive-specific: These included two components: the desired counterfactual state and a set of specific actions to help the participant reach that state. For example, it might suggest that the customer pay off their car loan to reduce the debt-to-income ratio.

Explanation Type 3 - Directive-generic: These explanations suggested a general class of actions that individuals could take to reach the desired counterfactual state without recommending a specific action. The idea was to preserve individuals’ autonomy in deciding which specific actions they want to take while still guiding their direction. For example, we might direct the customer to find strategies to reduce the total debt without giving examples of any specific strategies they could use.

4.2 Identifying Directives

To generate a list of candidate actions that we used in directive explanations, we reviewed a number of websites that provided financial advice1,2,3,4,5,6 and advice regarding improving employee

---

1https://www.experian.com/blogs/ask-experian/credit-education/debt-to-income-ratio/.
2https://www.marketwatch.com/story/try-these-creative-strategies-for-lowering-your-debt-to-income-ratio-2018-09-07.
3https://www.credit.com/blog/6-creative-ways-to-lower-your-debt-to-income-ratio-185695/.
4https://bettermoneyhabits.bankofamerica.com/en/credit/what-is-debt-to-income-ratio.
5https://www.upgrade.com/credit-health/insights/credit-utilization-ratio/.
6https://www.creditkarma.com/advice/i/how-to-lower-your-credit-card-utilization/.
job satisfaction, job involvement, managing overtime, and other Human Resource (HR)-related strategies.\textsuperscript{7,8,9,10,11}

To develop a simple model of how actions affect model features, we first identified a subset of features that were used to train machine learning models and that we believe could be observed and acted upon by decision makers. For each feature in the subset, for example, employee satisfaction or credit rating, we searched one or more of the websites listed above to identify the actions that could potentially modify them. We assume that these are the only interventions that modify the features, but realistically, there are unobserved noise variables that may influence how the features are modified \textsuperscript{38,40,74}. Furthermore, for the study, we limited the number of features each action could modify to one. For more details on the model, please see Section 3. As an alternative to planning for directives, one could learn behavior models and use those to generate candidate actions \textsuperscript{5}.

5 STUDY 1

We conducted our study in two domains, credit scoring and employee satisfaction. We trained a machine learning model to predict the outcome in each case.

For the credit scoring domain, we trained a logistic regression model to predict whether a borrower would default on a loan using the Lending Club dataset.\textsuperscript{12} The model achieved an accuracy of 85%. Similarly, for the employee satisfaction domain, we trained a logistic regression model to predict whether an employee would likely resign using an existing dataset.\textsuperscript{13} The model achieved an accuracy of 76%. To generate the counterfactual explanations, we used Russell’s \textsuperscript{65} algorithm, and we used our model to generate the directive explanations. Russell’s \textsuperscript{65} algorithm can generate many diverse counterfactual explanations. For our study, we used Russell’s \textsuperscript{65} algorithm to generate only one counterfactual, $c$, that is closest to the factual instance, $x$, with a different outcome by solving the following problem:

$$\arg\min_c \max_t \|x - c\| + \tau (f(x) - f(c)).$$  (3)

The distance function used in \textsuperscript{65} is $\ell_1$, weighted by the inverse Median Absolute Deviation ($\|\cdot\|_{1,MAD}$). The function $\tau$ maximizes the difference between the prediction of the counterfactual, $c$, and the factual point, $x$. This means that the counterfactual instance we use in our studies is the closest point to the instance we are explaining with a different outcome.

The machine learning model was used in the credit scoring domain to decide whether to approve or deny a customer’s loan application. In this domain, participants played the role of a Loan Officer. They received machine-generated explanations, and we told them their task would be to communicate the decision (approval or denial) and explain it to a customer. In the second domain, the employee satisfaction domain, the machine learning model was used to predict whether an employee is likely to resign in the near future. The participants played the role of an HR officer, who communicated the prediction to the employee’s supervisor using one of the explanations we provided. In each domain, we provided the participants with our explanations: non-directive, directive-specific, directive-generic, and an attention check question.

We designed eight scenarios in each domain (see Appendices B and C for a complete list of scenarios). Each scenario included details of a person, for example, a loan applicant (customer) or

\textsuperscript{7}https://www.savion.com/blog/effective-strategies-reduce-employee-turnover/. \hfill \textsuperscript{8}https://www.findmyshift.com/au/blog/why-overtime-working-can-harm-businesses-and-how-to-reduce-it. \hfill \textsuperscript{9}https://www.challengeconsulting.com.au/announcements/six-strategies-for-increased-job-satisfaction/. \hfill \textsuperscript{10}https://www.findmyshift.com/au/blog/why-overtime-working-can-harm-businesses-and-how-to-reduce-it. \hfill \textsuperscript{11}https://www.challengeconsulting.com.au/announcements/six-strategies-for-increased-job-satisfaction/. \hfill \textsuperscript{12}https://www.kaggle.com/husainsb/lendingclub-issued-loans#lc_loan.csv. \hfill \textsuperscript{13}https://www.kaggle.com/pavansubhash/ibm-hr-analytics-attrition-dataset.
an employee. We asked our participants to read an introductory section that included the decision (e.g., whether the loan was approved or denied or whether an employee was likely to resign) and then to rank the four explanations of the decision. The purpose of the introductory section was to avoid repeating certain pieces of information in each explanation; for example, rather than repeating the decision in each explanation, we included the decision in the introductory section. The participants were required to rank the explanations from most to least preferred to indicate which explanation they were most likely to use to communicate the decision to the individual concerned.

One out of the four possible explanations was clearly incorrect. For example, it might suggest actions that would have made the employee more likely to leave. We used this as a quality control measure; we removed any participant who did not indicate that this was the least preferred explanation in two or more scenarios. The other three explanations were non-directive, directive-specific, and directive-generic. To generate the counterfactual explanations (type 1), we used Russell’s [65] algorithm, and we used our model to generate the directive explanations (see Section 3 for more details).

5.1 Procedure

We conducted the first study using Amazon MTurk, a crowd-sourcing platform popular for human-subject experiments [15]. We designed and administered the experiments as a Qualtrics[14] survey. Before the experiments, we received ethics approval from our institution. Participants were paid USD $15 per hour for participating in the study.

Seventy-nine people participated in the study, spread over two domains: credit scoring and employee satisfaction. We recruited Masters workers, that is, workers who have consistently demonstrated a high degree of success in performing a wide range of tasks across a large number of requesters. All participants were from the United States.

The participants first received a plain language statement, and if they decided to continue the experiment, they were given a consent form. If the participants agreed to all items in the consent form, they were asked a few logical questions to filter out automated respondents. Then we asked the participants to provide their Amazon MTurk WorkerID and fill in the demographics questionnaire. Following this, they were allocated at random one of the two domains, credit scoring or employee satisfaction. We randomly selected six of the eight scenarios and presented these one at a time. Recall that we had four scenarios with a favorable outcome (e.g., the loan was approved) and four scenarios with an unfavorable outcome. We randomly selected three of the four scenarios with a favorable outcome and three of the four with an unfavorable outcome, giving us six scenarios. We randomized the scenarios and explanations to eliminate ordering effects. The scenarios were presented sequentially without the option of going back and changing previous answers. Participants were required to rank the four explanations from most to least preferred for each scenario.

At the end of the survey, participants were thanked and given a randomly generated code to enter into their Amazon MTurk session so they could be paid for completing the task.

5.2 Study 1 Results

In this section, we present the quantitative analysis showing that directive-specific and directive-generic explanations were preferred more than non-directive explanations. We also show that the preference was domain-dependent. In the credit scoring, participants preferred directive-specific explanations the most, while in the employee satisfaction domain, directive-generic explanations were preferred the most.

14https://www.qualtrics.com/.
15https://www.mturk.com/worker/help.
Domain 1 - Credit Scoring: Before doing the analysis, we used the attention check question to exclude participants who may not have been engaged with the task. Of the 39 participants, we excluded those who did not rank the attention check question as their last preference for two or more scenarios out of six. That is, if a participant made one error with ranking the attention check question, we discarded that ranking, keeping the other five. If, however, a participant made two or more errors, we removed the participant completely from the dataset. After elimination, we had 32 participants. All analysis presented in the following sections is based on the remaining 32 participants. The mean task completion time was 27 minutes (SD = 11 mins).

5.2.1 Participant Demographics. All participants were from the United States. Around 57% self-identified as males, 40% as females, and 3% did not state their gender. In terms of age, 32% were 25 to 34, 36% were 35 to 44, 25% were 45 to 54, and the rest were above 55 (7%). Regarding education, 18% were high school graduates, 14% had some college but no degree, 64% had an Associate’s or Bachelor’s degree, and 4% had a Doctoral degree. Regarding familiarity with the domain, 27% reported that they were slightly familiar with the loan application processes, 48% were moderately familiar, 18% were very familiar, and 7% were extremely familiar.

5.2.2 Explanation Type Preference. We provided participants with a non-directive explanation and two forms of directive explanations. Figure 1(a) shows participants’ explanation type choices for the three preferences. Directive-specific explanation was the most preferred, providing strong evidence that directive explanations are well accepted in this domain. Overall, we collected 192 rankings. Of the 192 first-preference choices, 81 (42%) were for directive-specific explanations, 51 (27%) for directive-generic explanations, and 60 (31%) for non-directive explanations. A chi-square goodness-of-fit test was performed to examine the likelihood of the participants’ choices being uniform. The likelihood of observing the data if the choices for the most preferred explanations were random is low, $\chi^2(2, N = 191) = 7.58, p < 0.02$. Similar results were obtained for the second and third preferences (see Appendix A).

5.2.3 Directive-specific Explanations Preferred for Unfavorable Decisions. We encoded the data such that we had the counts of the three types of explanations by each participant’s preference. Essentially, we represented the number of times a participant chose each explanation type over the eight scenarios. As such, for each participant, we had nine values. The first three were the counts of each explanation type the participant chose as the first preference, the next three were the counts...
of the explanation types for the second preference, and the last three for the third preference. The first-preference counts represent the number of times each participant would have given a particular explanation type to a customer.

We performed a non-parametric Friedman test of the differences between the number of times the participants chose each explanation type. We did this test for the first, second, and third preferences separately. We did not find significant differences between the number of times each participant chose an explanation type, $\chi^2(2) = 3.07, p < 0.23$, Kendall’s $W = 0.05$. This suggests that, overall, participants chose each explanation type almost equally for the eight scenarios.

We separately analyzed the participants’ preferences for scenarios where the loan was approved (favorable outcome, three scenarios) and those where the loan was denied (three scenarios). We performed a non-parametric Friedman test of the differences between the number of times each explanation type was chosen by participants when the loan was approved. We found no significant differences between explanation type choices. $\chi^2(2) = 2.58, p = 0.27$, Kendall’s $W = 0.04$. We found that non-directive explanation was chosen for $(M = 1.21, SD = 0.8)$ scenarios, directive-specific explanations for $(M = 1.0, SD = 0.8)$ scenarios, and directive-generic explanations for $(M = 0.78, SD = 0.1)$ scenarios.

We performed a non-parametric Friedman test of the differences between the number of times each explanation type was chosen by participants for scenarios when the loan was denied. We found significant differences between explanation type choices, $\chi^2(2) = 10.75, p = 0.004$, Kendall’s $W = 0.17$. We performed the Nemenyi post hoc analysis and found that directive-specific explanation was chosen for significantly more scenarios $(M = 1.53, SD = 0.9)$ than non-directive explanations $(M = 0.65, SD = 0.7, p < 0.001)$ and for moderately significantly more scenarios than directive-generic explanations $(M = 0.81, SD = 0.8, p = 0.05)$.

The above suggests that directive-specific explanation was more suitable when the decision was unfavorable.

5.2.4 Scenario and Individual Preferences Influenced Choices. The analysis so far showed that the choices were not random. To investigate which factors influenced these choices, we first examined whether the scenario influenced the preferred explanation type. We encoded the data to get the counts of each explanation type grouped by scenario for first preference.

We then examined whether we could explain the choices by a combination of scenario and individual preferences. Individual preferences were encoded as the proportion of choices for non-directive and directive-specific explanations, noting that directive-generic explanation was linearly dependent (we could compute counts of directive-generic choices given the other two). In other words, we computed the probability of the participants choosing non-directive and directive-specific explanations. We encoded the scenario effects as the average number of choices for non-directive and directive-specific explanations, that is, the probability of participants choosing non-directive and directive-specific explanations for each scenario. Using this data, we then built and compared two multinomial logit models using the \textit{mlogit} library in R.

The first model was built using directive-generic explanation as the base outcome and using only the individual preferences. We found that on average, the participant was a good predictor of which explanation type choice would be made for a given scenario ($\ell = -156.48$, McFadden $R^2 = 0.25, \chi^2 = 101.64, p < 0.001$). Then, we built a model with both the scenario effects and individual differences. We found that both the scenario and individual differences influenced the choice of explanation type ($\ell = -129.33$, McFadden $R^2 = 0.38, \chi^2 = 155.95, p < 0.001$). Also, a likelihood ratio test showed that the second model (with both scenario and individual differences) was significantly better than the first ($\chi^2(1) = 54.31, p < 0.001$).
Domain 2 - Employee Satisfaction: We used the same attention check question and criteria as in domain 1 to eliminate participants who may not have been engaged. Of the 40 participants who completed the experiment, after elimination, 33 remained. All analysis presented in the following sections is based on the remaining 33 participants. The mean task completion time was 28 minutes ($SD = 12$ mins).

5.2.5 Participant Demographics. All participants were from the United States. Around 50% self-identified as males, 48% as females, and 3% did not state their gender. In terms of age, 23% were 25 to 34, 39% were 35 to 44, 23% were 45 to 54, and the rest were above 55 (15%). Regarding education, 13% were high school graduates, 13% had some college but no degree, 65% had an Associate’s or Bachelor’s degree, and 9% had a Master’s degree. Regarding familiarity with the domain, 36% reported that they were slightly familiar with the human resource management processes, 45% were moderately familiar, 15% were very familiar, and 4% were extremely familiar.

5.2.6 Explanation Type Preference. Figure 1(b) shows participants’ explanation type choices for the three preferences. Participants chose directive-generic explanations more than directive-specific, and the non-directive explanation was least preferred, providing strong evidence that the two directive explanations are well accepted in the employee satisfaction domain. Overall, we collected 183 rankings. Of the 183 first-preference choices, 94 (51%) were of directive-generic explanations, 64 (35%) of directive-specific explanations, and 25 (14%) of non-directive explanations. A chi-square goodness-of-fit test was performed to examine the likelihood of the participants’ choices being uniform. The likelihood of observing the data if the choices for the most preferred explanations were random is low, $\chi^2(2, N = 183) = 39.25, p < 0.001$. We obtained similar results for the second and third preferences (see Appendix A).

5.2.7 Directive-generic Explanations Preferred by Most Participants. We started by encoding the data as we did for the credit scoring domain; that is, for each participant, we had nine values. The first three were the counts of each explanation type the participant chose as the first preference, the next three were the counts of the explanation types for the second preference, and the last three for the third preference. The first preference counts essentially represent the number of times each participant would have given an explanation type to an employee’s supervisor.

We performed a non-parametric Friedman test of the differences between the number of times the participants chose each explanation type. We did this test for the first, second, and third preferences separately. For the first preference, we found significant differences between explanation type choices, $\chi^2(2) = 30.07, p < 0.001, Kendall’s W = 0.47$. We performed the Nemenyi post hoc analysis and found that for the first preference, directive-generic explanation ($M = 2.98, SD = 1.2$) was chosen for significantly more scenarios than non-directive explanations ($M = 0.78, SD = 1.0, p < 0.001$), but we did not find any significant difference when it came to directive-specific explanations ($M = 2.0, SD = 1.10, p = 0.13$). The directive-specific explanations were chosen for significantly more scenarios than non-directive explanations ($p = 0.003$). We obtained similar results for the second and third preferences (see Appendix A).

We separately analyzed the scenarios where an employee was more likely to stay than resign (favorable outcome) and those where the employee was predicted to leave. We performed a non-parametric Friedman test of the differences between the number of times each explanation type was chosen by participants for scenarios when the employee was not likely to leave. We found significant differences between explanation type choices, $\chi^2(2) = 2.26, p < 0.001, Kendall’s W = 0.39$. The Nemenyi post hoc analysis found that directive-generic explanation ($M = 1.81, SD = 0.9$) was chosen for significantly more scenarios than non-directive explanations ($M = 0.62, SD = 0.7, p = 0.001$) and directive-specific explanations ($M = 0.43, SD = 0.7, p < 0.001$).
We performed a non-parametric Friedman test of the differences between the number of times each explanation type was chosen by participants for scenarios when the employee was likely to leave or resign. We found significant differences between explanation type choices, $\chi^2(2) = 32.62, p < 0.001$, Kendall’s $W = 0.5$. The Nemenyi post hoc analysis found that directive-specific explanation ($M = 1.56, SD = 0.8$) was chosen for significantly more scenarios than non-directive explanations ($M = 0.16, SD = 0.4, p < 0.001$) but not directive-specific explanations ($M = 1.13, SD = 0.8, p = 0.53$).

The results show a shift in the preferred explanation type from directive-generic to directive-specific when the decision was not favorable, suggesting, like the credit scoring domain, that directive-specific explanation was more suitable when the decision was unfavorable.

6 STUDY 2

We repeated our study using almost the same procedure and a similar number of participants (ending up with 54 participants from 70 after elimination) to learn why participants chose their most preferred explanation. We added seven more scenarios, taking the total number of scenarios to 15. This time, the participants were required to rank the explanations from most to least preferred to indicate which explanation they were most likely to use to communicate the decision to the concerned individual for all 15 scenarios and provide reasons for their selection in an open-ended text box. We asked the participants to answer one open-ended question after ranking the explanations, which was: *Please provide the reason(s) for choosing the most preferred explanation over the other three explanations*. We asked this question to learn why participants chose their explanations. We include the quantitative analysis for this study in Appendix E.

We performed a thematic analysis of the participants’ reasons. However, we did the thematic analysis for the two tasks separately. First, we performed a thematic analysis for the credit scoring task. Then, to test the generalizability of the codes and themes, we ran a validation sub-study to code the reasons for employee satisfaction tasks using the codes and themes from the credit scoring task. This sub-study aimed to validate the model from the credit scoring domain, that is, to learn to what extent the codes and themes from the credit scoring domain translated to employee satisfaction.

6.1 Qualitative Analysis for Credit Scoring Task

To perform the thematic analysis, we followed the steps outlined in the existing literature on thematic analysis [10, 21, 56]. In particular, we followed Nowell et al. [56], who provide a step-by-step guide to ensure that this qualitative data analysis is precise, consistent, and exhaustive. We formed a group of three (all authors on the article), with the lead author analyzing and documenting the process, the codes, and the themes. Two other members verified the codes and themes by critically analyzing these, and through triangulation, the three researchers decided on the final list of codes and themes after multiple iterations. During coding, it became clear that it was helpful to organize the codes according to whether or not they could be used to predict the participants’ choices. We coded reasons as *non-predictive* if the participant was justifying the choice and indicated what factor the participant considered was the most important when making a choice, but we could not determine which specific explanation the participant chose based on this response. Otherwise, the code was predictive, and of the four themes, three contained predictive codes. The four themes were *Action-related*, *Language-related*, *Usefulness/practical*, and *Non-predictive*.

Figure 2 shows the themes and codes that resulted from the thematic analysis. Definitions for the codes can be found in Appendix D.
Fig. 2 Themes and codes.
**Action-related:** This theme encompassed all responses that we considered to be action-related. Most participants preferred directive explanations precisely because they explicitly told the recipient (e.g., the customer) what he or she needed to do. We saw earlier that individual preferences influenced preference for explanation type. Participants were split between the two directive explanations, and some did not want directives. Several participants preferred the directive explanation because it had multiple options. For example, P15 stated:

“This explanation provides alternatives for Amir to get a higher spending limit.”

The directive-generic explanations were meant to promote the autonomy of the individuals trying to achieve recourse. This was indeed recognized by participants choosing directive-generic explanations and summarized well by P9:

“The preferred option [directive-generic] is the most flexible in terms of how Evan can increase their income. It doesn’t limit him to just getting another job, but he can get creative with how to increase his income.”

Other participants chose directive-specific explanations because this explanation type was specific. That is, it provided clear actions for an individual to take. For example, P42 provided the following reason for choosing the directive-specific explanation:

“My first preference [directive-specific] gives her a realistic option on what she has to do. My 2nd option [directive-generic] is not bad but doesn’t seem to be as specific. The 3rd preference [non-directive] is honest but will leave the customer wondering what to do next.”

Not everyone preferred directive explanations. There were several reasons participants were not attracted to directives. Participants chose the non-directive explanation because they did not prefer to tell the recipient what to do. In these circumstances, the non-directive explanation was sufficient to indicate to the recipient when the decision would change instead of providing directives. For example, P53 stated:

“Option one [non-directive] because two [directive-specific] and three [directive-generic] are telling her what to do and will make them mad.”

We also found that participants carefully analyzed the directives when choosing the directive explanations, looking at the practical value of the suggested directives in the short term or the long term. For example, P20 provided the following reason for selecting a directive explanation (the loan was approved):

“It [directive-generic] provides reasons for the approval but also ways in which he can ensure he continues to get approved in the future.”

Knowing what one is doing right may be particularly important for business customers, who may require credit multiple times over the life of the business.

**Usefulness/practical:** This theme included all reasons that alluded to the usefulness or practicality of explanations. We included counterfactual information in all explanations. Participants found the counterfactual information useful not only to know when the decision of the ML model would change but also to understand the limits or the decision boundary. For example, in scenarios with approved loans, participants often selected explanations because the explanation had information about the decision boundary that could help customers behave to ensure approval in the future. For example, P27 mentioned that:

“The explanation I chose [directive-specific] explains why he was denied the best and what amount he could apply for and be approved.”
Several participants tried to imagine how reasonable or feasible the explanation would likely be for the recipient. For example, P43 provided the following justification for their choice:

“I picked [directive-generic] based on how feasible I thought each strategy would be.”

The above example indicates that participants were engaging in perspective-taking and trying to judge the cost of the directives suggested for the recipient.

The explainer may not always be aware of how costly or how actionable the explanation truly is. One way for the explainer to know the hidden costs is through dialogue [49, 68], that is, explicitly requesting this information. This suggests that dialogue is probably necessary when there is uncertainty around the feasibility of an actionable counterfactual explanation.

Finally, many participants did not feel the need to explain, especially when the loan application was approved or the employee was unlikely to resign. If the participant indicated that an explanation was unnecessary, they typically chose the non-directive. For example, P4:

“He got approved. He’s not looking for a long-winded explanation of why, just the simplest (if he read the explanation at all).”

**Language-related:** This theme encompassed all responses that suggested that language-related factors influenced the participant’s choice. Participants were attracted (mostly toward non-directive explanations) to simple, short, or direct explanations. We found that participants were particularly attracted to non-directive explanations in Scenario 3. In this scenario, the customer’s loan was denied because of the income, and the two directive explanations suggested that the customer could increase his income by changing his job, finding a second job, or getting a promotion. Many participants found these two explanations “condescending” or “impolite.” For example, P6 wrote:

“The first two options [directive-specific and directive-generic] feel condescending and don’t take into account Evan’s personal situation. He may not be able to increase his income. The third one [non-directive] is more matter-of-fact and doesn’t try to get into Evan’s personal life.”

We note that our suggestions in the directive explanations are very similar to the tips commonly found on financial advice websites. It appears that people may be comfortable reading this information on their own but not being “told” to do so within an explanation. As such, from an algorithmic standpoint, it appears that there may be specific attributes/features for which a non-directive explanation is a more reasonable option than telling people how to act.

**Non-predictive reasons:** Our final theme was created to cater to responses that did not predict the explanation type chosen by participants, which is why they are described as non-predictive. There were four sub-themes under the non-predictive theme: readability/informative, tone, opinion, and miscellaneous. Many participants justified their choice in terms of the clarity of the explanations or if explanations were informative. For example, P34 stated:

“This explanation [directive-generic] is clear and is easily understandable when compared to others.”

We observed that participants justified their choice based on tone, that is, how polite or friendly the explanations were, how diplomatic or professional they sounded, or how it would have made the recipient feel. For example, P26 wrote that an explanation could come out as impolite:

“Because that explanation [directive-specific] gently explains the customers the whole scenario rather then being just rude. they told if instalment is been missed for 6 months| that’s a clear point they made for customer. and customer will also know the dead ends.”

Some participants justified their choice by expressing an opinion toward an explanation:
Finally, many other codes were thin and fell under the non-predictive category; we decided to collect them under the miscellaneous sub-theme.

6.2 Qualitative Analysis for Employee Satisfaction Task

We ran a validation study to test the generalizability of the codes and themes we had identified when coding the reasons from the credit scoring task. The goal of this study was to validate the model, that is, to see to what extent the codes and themes translated to another domain. To do this, we recruited six coders. We introduced the codebook from the first study to the six coders by having an initial 30-minute briefing where the lead author explained the goal of the task (which was to code the reasons so that we could understand why a participant chose a particular type of explanation), the existing codebook from study 1 with examples, and the procedure that the coders had to follow. Following this, the coders did a 60-minute tutorial prepared by the lead author that explained how the lead author would have coded a few examples. The tutorial also included a practice set of 10 reasons for the participant to get familiar with the codebook. We held a further 45-minute briefing to clarify any questions and go through six further examples. The participants had around 3.5 hours to code around 180 reasons. We used Qualtrics to administer the task, and the coders were compensated at AUD $50 per hour. Because we had around 360 reasons to code, we split the reasons into two groups of 180 reasons and created a separate survey for each group of 180 reasons. We randomly allocated the six coders to one of the surveys.

For each reason, we provided the coders with a simplified version of the employee profile, the participant’s selected explanation, and the two other valid explanations that the participant received. For each explanation, we included the explanation type (non-directive, directive-specific, and directive-generic) so that the coders were aware of the explanation type chosen by the participant and could use this information to code the reason better. Recall that for the credit scoring domain, we coded reasons as non-predictive if the participant was justifying the choice and indicated what factor the participant considered was the most important when making a choice, but we could not determine which specific explanation the participant chose based on this response. Otherwise, the code was predictive, and of the four themes, three contained predictive codes. We included the explanation type to help the coders follow the same process.

Following the employee profile and explanations, we provided the reason the participant provided us for their most preferred explanation. After the reason, we listed the 54 codes from study 1 as multiple-choice options; coders could choose more than one, and if none of the codes appropriately described the reason, they selected the miscellaneous:other option. Coders were also allowed to list any new codes that they felt were appropriate for the reason. The instructions to the coders were to be as granular as possible when coming up with new codes, and the lead author provided examples of how to do this during the initial meetings. The coders assessed each reason one at a time with the option of returning to previously coded reasons. However, no coder did this because of the inconvenience of clicking the back button repeatedly. We configured Qualtrics so that a coder could stop multiple times and complete the coding over multiple days. Most coders completed the task within 2 working days.

We analyzed the data for the two groups of 180 reasons separately and then combined the results of the two surveys. Our first analysis was to see the number of new codes (or themes) that were required. The six coders generated eight new codes that covered 3% of the codes. That is, 97% of reasons could be coded using the model produced in the credit risk study.

Next, we investigated the agreement at the code level. We only counted codes that two of the three coders assigned to each reason. The rationale is that it is possible for coders to choose similar
but not the same codes for a reason. For each reason, the coders had a choice of 54 codes. We took
the majority code—if two coders assign the same code to a reason, we assume it is the correct
code(s). At least two coders assigned the same code for 254/360 (70%) reasons, and we discarded
the other 30% before further analysis.

We also analyzed the agreement at the theme level. Naturally, a theme consists of multiple codes,
and coders could choose different codes within each theme. Therefore, we looked at whether the
coders agreed on the theme. Note that the coders were responsible for assigning the codes, not
the themes. At a theme level, the agreement was 91%. Overall, we observed that the codes and
themes from the credit scoring domain had good coverage (it covered 97% of the reasons from the
employee satisfaction domain).

The top two themes were Action-related (33% of codes) and Usefulness/practical (20%). The
opinion and miscellaneous themes were 17% and 16%, respectively. Finally, the lowest two were
Readability/informative and Language with 9% and 5% of the codes, respectively.

7 DISCUSSION
In this article, we proposed directive explanations, that is, explanations that give individuals di-
rectives for recourse for machine learning decisions. We assert that actionable explanations can
be improved by explicitly providing people with a single or a sequence of actions to change the
decisions. We evaluated the preference for and perception toward directive explanations over non-
directive ones through two user studies, one in the space of credit scoring and the other in em-
ployee satisfaction domains.

Our quantitative analysis indicates a strong preference for the two directive explanations. The
participants’ first and second preferences were mostly for the two directive explanations. In the
credit scoring domain, 69% chose one of the two directive explanations as their most preferred
explanation, and for the employee satisfaction domain, 86% did so. Our results suggest that the
two directive explanations complement (non-directive) counterfactual explanations [54, 59, 76, 78].
While we show that explanations should be directive, we found that participants were spread
between directive-specific and directive-generic explanations between the two domains.

Participants chose directive-specific explanations because they provided a specific solution to
help the recipient achieve recourse, particularly when the decision was not favorable (when the
loan was denied or an employee was likely to resign). For example, in the second study, one of the
participants liked that the directive-specific explanation provided specific advice:

“I chose my most preferred explanation [directive-specific] because it gets at the root of
the problem (travel) and offers up a good suggestion on how to solve that problem.”

Conversely, sometimes participants preferred directive-generic explanations because they were
perceived as providing some autonomy for people to choose their own specific course of action
to achieve recourse. This finding echoes that of Binns et al. [8], who reported that their partici-
pants thought that providing alternatives to people when the decision is not favorable was a good
idea. Generally, directive-generic explanations are most suitable when someone prefers options
or at least has or feels some sense of agency when deciding the specific course of action. For ex-
ample, a participant provided the following reasoning in study 2 for choosing a directive-generic
explanation:

“I like this reason [directive-generic] because it set clear goals for which areas need to
be improved, specially travel and job satisfaction, which is in line with her responsibility
expectation when she accepted the job. Also, it gives suggestion to achieve the goals while
allowing freedom to the supervisor to choose the means and methods.”

ACM Transactions on Interactive Intelligent Systems, Vol. 13, No. 4, Article 23. Publication date: December 2023.
We noted a higher preference for directive-generic explanations in the employee satisfaction domain. We believe that this could be due to a few reasons. First, participants were slightly more familiar with the credit domain than the employee domain (69% stated that they were between moderately and extremely familiar with the credit domain, while 57% stated that they were between moderately and extremely familiar with the employee domain). This could be why people were more comfortable suggesting directive-specific explanations in the credit domain and directive-generic explanations in the employee domain. Second, we believe that most people would have their own ideas on improving job satisfaction, which would have a lot of personal preferences. Therefore, it was potentially easier for the HR officer to leave the specific course of action that the employee’s supervisor would take to improve the job satisfaction of the concerned employee. On the other hand, recourse for credit scoring is about changing behavior to “game” the credit scoring model, with which many people would have limited experience, so more concrete advice would be appreciated.

While we saw significant support for directive explanations, around 31% and 13% of responses in the two domains were for non-directive explanations. One of the main reasons participants sometimes chose non-directive explanations was the decision; many participants suggested that when the decision is favorable, the most important information is when the decision is likely to change (counterfactual information) and not necessarily how that would happen, as one of the participants describes below:

“I like the basic and simple explanation that overtime could cause him to resign [non-directive]. I don’t think you should try to give a reason for it, just whether or not it happens.”

Various other factors potentially influenced the choice of an explanation type. In some scenarios, the choice was impacted by social factors. In one credit scoring, the directive explanation suggested that the customer change jobs, do part-time work, or try to get a promotion to increase their income (these recommendations are common on various websites that provide financial advice). For this scenario, participants were almost evenly distributed between the explanation types. However, many participants highlighted that it was condescending to tell people to change their jobs. In several scenarios in the employee satisfaction domain, we found that the participants were choosing directives based on which one makes an employee happier. For example, one of the participants wrote the following for choosing a directive-generic explanation:

“I choose my most preferred over the others because it gives the suggestion to remove his over time but would allow him to do the projects more effectively and quicker, saving the company both time and money and probably making him a happier employee.”

Socio-technical systems usually have many stakeholders. For example, credit risk assessment involves customers, data modelers, model builders, model users (such as loan officers), and others. The roles influence the relevance of different types of explanations [32, 73]. This could explain why some participants found directive explanations helpful while others did not.

The above discussions imply that it is not straightforward to select between explanation types, reinforcing that we cannot decide a priori whether non-directive or directive explanations are more suitable for all individuals in all circumstances. This finding is not limited to directives explanations. For example, Ehsan et al. [27] found that for rationale generation, participants’ requirements for the type of explanation was context-dependent; they preferred short and simple rationales to understand agents, but detailed rationales for identifying failure or unexpected behavior. Thus, the explanation type choice is influenced by individual, social, and contextual factors, and what is or is not actionable must be identified by the individual concerned [46, 58, 79].
To summarize:

- We find a clear preference for the two directive explanations over non-directive counterfactual explanations in two domains; the non-directive explanation was the least preferred explanation type.
- Directive-specific explanations are more suited to scenarios where the outcome is unfavorable. We found that in scenarios where the loan was denied or the employee was likely to leave, the participants strongly preferred directive-specific explanations. This suggests that in the two domains, explanations should be constructed so that there are options for people to receive directive explanations. We find a strong preference for it, which suggests that people will find it useful.
- The domain may influence the preference for the two directive explanations (see discussion above for a higher preference for directive-generic explanation in the employee satisfaction domain).
- Non-directives are unsuitable when the outcome is favorable for the credit scoring domain. The non-directive explanations provide decision boundaries that will be useful to continue good financial behaviors. In the employee satisfaction domain, the dominant preference for a directive-generic explanation could be because people may want to encourage positive behaviors and keep people employed for longer.

7.1 Limitations

Our studies involved an automated system explaining to an intermediary why the automated system made a particular decision, such as denying a loan. The intermediary then selected one of four possible explanations to provide to the client. In many contexts, such as loan applications, we believe that an automated system assists people (loan officers) who assist others (customers). Therefore, this setup allows us to understand what a human considers relevant when explaining decisions to another human and provide insights from this perspective. However, we do acknowledge that our study is limited to these settings.

We noted limitations in terms of the context that we explored. In the credit scoring domain, participants felt that explanations were of no value when loans were approved. However, we do not believe this holds in all contexts. For example, if we had told the participants that the customer was a business customer who regularly applies for loans, this may have elicited a different response from these participants; for someone who applies for loans regularly, knowing why a loan was approved is useful as it indicates what they should do next time they apply for a loan.

To have confidence that directive explanations were useful in different domains, we conducted studies in credit scoring and human resource spaces. However, we need further studies in other domains to fully understand the implications of directive explanations.

We were also limited by the data collection method, as we could not run this in a lab setting due to social isolation restrictions resulting from the COVID-19 pandemic. Had we run it in a lab setting, there were many instances where we would have asked follow-up questions to the participants. As such, the input provided by the participants through the two open-ended questions could be improved if we had the opportunity to clarify the responses.

Furthermore, all participants in our studies were from the United States, and we could potentially observe a different result if we recruited participants from different countries. Several factors, such as cultural values, may influence preferences [20]. For example, users from different backgrounds responded differently to robot recommendations (Asian participants changed their decisions more than US-based participants when collaborating with robots) [80]. Therefore, it is likely that users outside of the United States may have different explanation type preferences.
7.2 Future Work

The results of our present study indicate support for both non-directive and directive explanations. First, we identified that preferences for directive vs. non-directive explanations depend on multiple factors. Further work is required to clarify why these factors matter and how they influence the selection of the explanation types across domains. Such exploration could include studying preferences from a different perspective, such as from the perspective of the loan applicant or the employee’s supervisor.

Further work is needed to understand the effectiveness of directive explanations. Our results show a clear preference for directive explanations. The next step will be to show how effectively they improve actionability. Our scenario design does not consider the cost of changing an attribute or the feasibility of the actions, and we found that participants reflected on this and it surfaced in the thematic analysis. Future work should explore scenario framing to control cost and feasibility and study the implications on preferences.

The actions we used in our MDP model were sourced from multiple public websites to get good coverage of the types of recommendations that could be included in the directive explanations. Future work could look at other ways to gather appropriate actions, such as from experts or crowdsourcing.

Efficient models are needed to generate directive explanations. Recently, Karimi et al. [38] proposed using structural causal models as one option. Madumal et al. [50] also showed that people may better understand models that employ a causal lens to generate explanations. Future work could also involve generating and evaluating diverse directives [41] and comparing MDP-based models to structural causal models [22, 33, 37, 52, 74].

While we have not considered diverse directives, there are numerous methods to measure the plan differences, and we could use these to devise a metric to compute multiple diverse directives [12, 41]. Moreover, we could use the rewards computed by the model to inform the user of the model’s preferences over these directives to make the selection easier for the user.

Another avenue for increasing diversity is by considering multiple counterfactuals. In recent work, Dandl et al. [23] proposed the Multi-Objective Counterfactuals (MOC) method and used multi-objective optimization to find a diverse set of counterfactuals with different tradeoffs between the proposed objectives. We could also combine the method in [23] with the one proposed by [13], which uses counterfactual constraints to search for a limited but more desirable set of counterfactuals. Once we have the diverse set of counterfactuals, we could use our model to generate directives for each and present these to the user as options with the hope that this further increases the actionability of directive explanations. This approach may also be relevant for multiclass problems, especially when the user may have preferences for multiple different outcomes (classes).

We could consider ways to personalize explanations. Research suggests providing multiple non-directive explanations in the hope that one of them will be actionable for the recipient [65, 77, 78]. Our results show that not all individuals wish to receive multiple explanations. At the same time, knowing the cost of action for an individual is also important—some of our participants were thinking about this, so an automated system should also consider this. One way to establish the cost of a certain action is through an interaction with individuals (see, e.g., [68]). Through dialogue, we can identify the actions individuals are more comfortable with and, therefore, better personalize the explanation to the individual’s preferences and circumstances. We could also explore asking individuals their preferences over feature values and constraining the counterfactuals to satisfy these constraints, as suggested in [67]. This approach does require individuals to divulge personal information [77], but the benefit is that they may be able to receive a more tailored and better explanation.
In recent work, [13] proposes using counterfactual constraints and distance measures to study the robustness of machine learning models across each feature. In the credit scoring domain, they showed that their method generated counterfactual explanations that allow designers to understand the robustness of machine learning models. Future work could explore the different distance measures and their impact on the model we use to generate the directives.

Finally, we could extend our work by explaining why the model believes the directives are likely to help the users achieve their goals. There is growing literature in the space of explainable planning [19, 29, 44, 69] that we could leverage concerning explaining why the suggested directive is more likely to help users achieve their goals over other possibilities.

8 CONCLUSION

We formally defined and investigated directive explanations in this article. These explanations provide individuals directives for recourse of machine learning decisions, that is, inform people on how to act. The pursuit of our goal to investigate people’s perception toward directive explanations leads us to some interesting findings. Although we demonstrated significant support for directive explanations, we conclude that we cannot always please all people. Explanation preference is subjective and depends on multiple factors; thus, we cannot generically determine the most suitable type of explanation. This reinforces the call to take a human-centered and situation-specific approach to explainable AI, especially when looking at ways of making explanations actionable.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access 6 (2018), 52138–52160.

[2] Charu C. Aggarwal, Chen Chen, and Jiawei Han. 2010. The inverse classification problem. Journal of Computer Science and Technology 25, 3 (2010), 458–468.

[3] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. 2018. Interpretable machine learning in healthcare. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB’18). Association for Computing Machinery, New York, NY, 559–560.

[4] Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. 2020. Explanation in AI and law: Past, present and future. Artif. Intell. 289 (Dec. 2020), 103387.

[5] Nikola Banovic, Anqi Wang, Yanfeng Jin, Christie Chang, Julian Ramos, Anind Dey, and Jennifer Mankoff. 2017. Leveraging human routine models to detect and generate human behaviors. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI’17). Association for Computing Machinery, New York, NY, 6683–6694.

[6] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT’20). Association for Computing Machinery, New York, NY, 80–89.

[7] Richard Bellman. 1957. A Markovian decision process. Journal of Mathematics and Mechanics 6, 5 (1957), 679–684.

[8] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "It’s reducing a human being to a percentage": Perceptions of justice in algorithmic decisions. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI’18). Association for Computing Machinery, New York, NY, 1–14.

[9] Biranand Cotton. 2017. Explanation and justification in machine learning: A survey. In IJCAI-17 Workshop on Explainable AI (XAI), Vol. 8. cs.columbia.edu, 8–13.

[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. Qualitative Research in Psychology 3, 2 (2006), 77–101.

[11] Cameron B. Browne, Edward Powley, Daniel Whitehouse, Simon M. Lucas, Peter I. Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. 2012. A survey of Monte Carlo tree search methods. IEEE Trans. Comput. Intell. AI Games 4, 1 (March 2012), 1–43.

[12] Daniel Bryce. 2014. Landmark-based plan distance measures for diverse planning. ICAPS 24 (May 2014), 56–64.

[13] Andreas C. Bueff, Mateusz Cytryński, Raffaella Calabrese, Matthew Jones, John Roberts, Jonathon Moore, and Iain Brown. 2022. Machine learning interpretability for a stress scenario generation in credit scoring based on counterfactuals. Expert Syst. Appl. 202 (Sept. 2022), 117271.

[14] Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. 2018. Woulda, coulda, shoulda: Counterfactually-guided policy search. (Nov. 2018). arXiv:1811.06272 [cs.LG]
[15] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon’s Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* 6, 1 (Jan. 2011), 3–5.

[16] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. 2020. Explainable AI in fintech risk management. *Front. Artif. Intell.* 3 (April 2020), 26.

[17] Ruth M. J. Byrne. 2016. Counterfactual thought. *Annual Review of Psychology* 67 (2016), 135–157.

[18] Ruth M. J. Byrne. 2019. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19 Macao, 10-16 August 2019)*, ijcai.org, 6276–6282.

[19] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. 2017. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. (Jan. 2017). arXiv:1701.08317 [cs.AI]

[20] Larissa Chazette and Kurt Schneider. 2020. Explainability as a non-functional requirement: Challenges and recommendations. *Requirements Engineering* 25, 4 (Dec. 2020), 493–514.

[21] Victoria Clarke and Virginia Braun. 2014. Thematic analysis. (2014), 1947–1952. https://doi.org/10.1007/978-1-4614-5583-7_311

[22] Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. 2020. Causal modeling for fairness in dynamical systems. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 2185–2195.

[23] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. 2020. Multi-objective counterfactual explanations. In *Parallel Problem Solving from Nature (PPSN XVI)*. Springer International Publishing, 448–469.

[24] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. 2018. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation (EC’18 Ithaca, NY, USA, June 18-22, 2018)*, Éva Tardos, Edith Elkind, and Rakesh Vohra (Eds.). ACM, 55–70. https://doi.org/10.1145/3219166.3219193

[25] Tri Dung Duong, Qian Li, and Guandong Xu. 2021. Prototype-based counterfactual explanation for causal classification. (May 2021). arXiv:2105.00703 [cs.LG]

[26] Lilian Edwards and Michael Veale. 2017. Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.* 16 (2017), 18.

[27] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated rationale generation: A technique for explainable AI and its effects on user perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI’19)*. Association for Computing Machinery, New York, NY, 263–274.

[28] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2018. Analysis of classifiers’ robustness to adversarial perturbations. *Machine Learning* 107, 3 (2018), 481–508.

[29] Maria Fox, Derek Long, and Daniele Magazzeni. 2017. Explainable planning. (Sept. 2017). arXiv:1709.10256 [cs.AI]

[30] Hector Geffner and Blai Bonet. 2013. A concise introduction to models and methods for automated planning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 8, 1 (2013), 1–141.

[31] Riccardo Guidotti, 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 (2019), 93.

[32] Mark Hall, Daniel Harborne, Vedran Galetic, Santiago Quintana-Amate, Alistair Nottle, and Alun Preece. 2019. A systematic method to understand requirements for explainable AI (XAI) systems. In *Proceedings of the IJCAI Workshop on Xplanable Artificial Intelligence (XAI’19)*, Vol. 11. dais-ita.org.

[33] Joseph Y. Halpern and Judea Pearl. 2020. Causes and explanations: A structural-model approach. Part I: Causes. *Br. J. Philos. Sci.* (2020).

[34] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS’16)*. Association for Computing Machinery, New York, NY, 111–122.

[35] Robert R. Hoffman and Gary Klein. 2017. Explaining explanation, part 1: Theoretical foundations. *IEEE Intelligent Systems* 32, 3 (2017), 68–73.

[36] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. 2017. What do we need to build expensive, yet high-quality, data? *Perspect. Psychol. Sci.* 6, 1 (Jan. 2011), 3–5.

[37] Mark Hopkins and Judea Pearl. 2007. Causality and counterfactuals in the situation calculus. *J. Logic Comput.* 17, 5 (Oct. 2007), 939–953.

[38] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic recourse: From counterfactual explanations to interventions. (2021), 353–362. https://doi.org/10.1145/3442188.3445899

[39] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. 2020. Model-agnostic counterfactual explanations for consequential decisions. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, 895–905.
Directive Explanations for Actionable Explainability in Machine Learning Applications 23:25

[40] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. 2020. Algorithmic recourse under imperfect causal knowledge: A probabilistic approach. [June 2020]. arXiv:2006.06831 [cs.LG]

[41] Michael Katz and Shirin Sohrabi. 2020. Reshaping diverse planning. AAAI 34, 06 (April 2020), 9892–9899.

[42] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. Q. J. Econ. 133, 1 (Feb. 2018), 237–293.

[43] Gunnar König, Timo Freiesleben, and Moritz Grosse-Wentrup. 2021. A causal perspective on meaningful and robust algorithmic recourse. [July 2021]. arXiv:2107.07853 [stat.ML]

[44] Benjamin Krarup, Michael Cashmore, Daniele Magazzeni, and Tim Miller. 2019. Model-based contrastive explanations for explainable planning. In ICAPS 2019 Workshop on Explainable AI Planning (XAI’19). AAAI Press, 9.

[45] David Lewis. 2013. Counterfactuals. John Wiley & Sons.

[46] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing design practices for explainable AI user experiences. arXiv preprint arXiv:2001.02478 (2020).

[47] Brian Y. Lim and Anind K. Dey. 2009. Assessing demand for intelligibility in context-aware applications. In Ubiquitous Computing, 11th International Conference (UbiComp’09), Proceedings (ACM International Conference Proceeding Series), Sumi Halal, Hans Gellersen, and Sunny Consolvo (Eds.). ACM, 195–204. https://doi.org/10.1145/1620545.1620576

[48] Zachary C. Lipton. 2018. The mythos of model interpretability. Commun. ACM 61, 10 (2018), 36–43. https://doi.org/10.1145/3233231

[49] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A grounded interaction protocol for explainable artificial intelligence. In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS’19). International Foundation for Autonomous Agents and Multiagent Systems, 1033–1041.

[50] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2020. Explainable reinforcement learning through a causal lens. (2020), 2493–2500. https://aaai.org/ojs/index.php/AAAI/article/view/5631

[51] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267 (2019), 1–38.

[52] Tim Miller. 2021. Contrastive explanation: A structural-model approach. Knowl. Eng. Rev. 36 (2021), e14.

[53] Christoph Molnar. 2020. Interpretable Machine Learning. Lulu.com.

[54] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through divergent counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 607–617.

[55] Philip Naumann and Eirini Ntoutsi. 2021. Consequence-aware sequential counterfactual generation. [April 2021]. arXiv:2104.05592 [cs.LG]

[56] Lorelli S. Nowell, Jill M. Norris, Deborah E. White, and Nancy J. Moules. 2017. Thematic analysis: Striving to meet the trustworthiness criteria. International Journal of Qualitative Methods 16, 1 (2017), 1609406917733847.

[57] Michael Oberstad and David Sontag. 2019. Counterfactual off-policy evaluation with gumbel-max structural causal models. In International Conference on Machine Learning (ICML’19). proceedings.mlrr.press, 4881–4890.

[58] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI’21, Article 237). Association for Computing Machinery, New York, NY, 1–52.

[59] Rafael Poyiadzi, Kacper Sokol, Raúl Santos-Rodríguez, Tijl De Bie, and Peter A. Flach. 2020. FACE: Feasible and actionable counterfactual explanations. In AAAI/ACM Conference on AI, Ethics, and Society (AI4ES’20, New York, NY, USA, February 7-8, 2020), Annette N. Markham, Julia Powles, Toby Walsh, and Anne L. Washington (Eds.). ACM, 344–350. https://doi.org/10.1145/3375627.3375850

[60] Martin L. Puterman. 2014. Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons.

[61] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI’18). Association for Computing Machinery, New York, NY, 1–13.

[62] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why should I trust you?”. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (Eds.). ACM, 1135–1144. https://doi.org/10.1145/2939672.2939778

[63] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI’18), the 30th innovative Applications of Artificial Intelligence (IAAI’18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18, New Orleans, Louisiana, USA, February 2–7, 2018), Sheila A. McIlraith and Killian Q. Weinberger (Eds.). AAAI Press, 1527–1535. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982.

[64] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1, 5 (2019), 206–215.

ACM Transactions on Interactive Intelligent Systems, Vol. 13, No. 4, Article 23. Publication date: December 2023.
[65] Chris Russell. 2019. Efficient search for diverse coherent explanations. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT’19, Atlanta, GA, USA, January 29-31, 2019), Danah Boyd and Jamie H. Morgenstern (Eds.). ACM, 20–28. https://doi.org/10.1145/3287560.3287569

[66] Andrew D. Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. Fordham L. Rev. 87 (2018), 1055.

[67] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2019. CERTIFAI: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. (May 2019). arXiv:1905.07857 [cs.LG]

[68] Kacper Sokol and Peter A. Flach. 2020. One explanation does not fit all: The promise of interactive explanations for machine learning transparency. CoRR abs/2001.09734 (2020). arXiv:2001.09734 https://arxiv.org/abs/2001.09734.

[69] Sarath Sreedharan, Anagha Kulkarni, and Subbarao Kambhampati. 2022. Explainable human–AI interaction: A planning perspective. Synthesis Lectures on Artificial Intelligence and Machine Learning 16, 1 (Jan. 2022), 1–184.

[70] Biplav Srivastava, Tuan Anh Nguyen, Alfonso Gerevini, Subbarao Kambhampati, Minh Binh Do, and Ivan Serina. 2007. Domain independent approaches for finding diverse plans. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI’07, Hyderabad, India, January 6-12, 2007), Manuela M. Veloso (Ed.). 2016–2022. http://ijcai.org/Proceedings/07/Papers/325.pdf.

[71] Richard S. Sutton and Andrew G. Barto. 2018. Reinforcement Learning: An Introduction (2nd ed.). MIT Press.

[72] Winnie F. Taylor. 1980. Meeting the Equal Credit Opportunity Act’s specificity requirement: Judgmental and statistical scoring systems. Buff. L. Rev. 29 (1980), 73.

[73] Richard Tomsett, Dave Braines, Dan Harborne, Alun D. Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. CoRR abs/1806.07552 (2018). arXiv:1806.07552 http://arxiv.org/abs/1806.07552.

[74] Stratis Tsirtsis, Abir De, and Manuel Gomez-Rodriguez. 2021. Counterfactual explanations in sequential decision making under uncertainty. (July 2021). arXiv:2107.02776 [cs.LG]

[75] Stratis Tsirtsis and Manuel Gomez-Rodriguez. 2020. Decisions, counterfactual explanations and strategic behavior. (Feb. 2020). arXiv:2002.04333 [cs.LG]

[76] Berk Utsun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT’19, Atlanta, GA, USA, January 29-31, 2019), Danah Boyd and Jamie H. Morgenstern (Eds.). ACM, 10–19. https://doi.org/10.1145/3287560.3287566

[77] Suresh Venkatasubramanian and Mark Alfano. 2020. The philosophical basis of algorithmic recourse. In Conference on Fairness, Accountability, and Transparency (FAccT’20, Barcelona, Spain, January 27-30, 2020), Mireille Hildebrandt, Carlos Castillo, Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna (Eds.). ACM, 284–293. https://doi.org/10.1145/3351095.3372876

[78] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. J. L. & Tech. 31 (2017), 841.

[79] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing theory-driven user-centric explainable AI. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI’19, New York, NY, USA, May 2019), Glasgow, 1–15.

[80] Lin Wang, Pei-Luen Patrick Rau, Vanessa Evers, Benjamin Krisper Robinson, and Pamela Hinds. 2010. When in Rome: The role of culture & context in adherence to robot recommendations. In 2010 5th ACM/IEEE International Conference on Human-robot Interaction (HRI’10), ieeeexplore.ieee.org, 359–366.

Received 21 February 2022; revised 13 September 2022; accepted 17 December 2022