Birkbeck
UNIVERSITY OF LONDON

BIROn - Birkbeck Institutional Research Online

Tešić, Marko and Hahn, Ulrike (2022) Can counterfactual explanations of AI systems' predictions skew lay users' causal intuitions about the world? If so, can we correct for that? Patterns 3 (12), p. 100635. ISSN 2666-3899.

Downloaded from: https://eprints.bbk.ac.uk/id/eprint/50231/

Usage Guidelines:
Please refer to usage guidelines at https://eprints.bbk.ac.uk/policies.html or alternatively contact lib-eprints@bbk.ac.uk.

Patterns

CellPress
OPEN ACCESS

Perspective

Can counterfactual explanations of AI systems' predictions skew lay users' causal intuitions about the world? If so, can we correct for that?

Marko Tešić[1,*] and Ulrike Hahn[1]
[1]Birkbeck, University of London, London, UK
*Correspondence: m.tesic@bbk.ac.uk
https://doi.org/10.1016/j.patter.2022.100635

THE BIGGER PICTURE Explainable artificial intelligence provides methods for bringing in transparency into black-box artificial intelligence (AI) systems. These methods produce explanations of AI systems' predictions that are aimed at increasing the understanding of the AI systems' behavior and help us to appropriately calibrate our trust in these systems. In this perspective, we explore some of the potential undesirable effects of providing explanations of AI systems to human users and ways to mitigate such effects. We start from the observation that most AI systems capture correlations and associations in data and not causal relationships. Explanations of the AI systems' predictions would make the correlations more transparent. They would not, however, make the explained relationships causal. In two experiments, we show how providing counterfactual explanations of AI systems' predictions unjustifiably changes people's beliefs about causal relationships in the real world. We also show how we may go about preventing such a change in beliefs and hope to open doors for further exploration into psychological effects of AI explanations on human recipients.

Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

Counterfactual (CF) explanations have been employed as one of the modes of explainability in explainable artificial intelligence (AI)—both to increase the transparency of AI systems and to provide recourse. Cognitive science and psychology have pointed out that people regularly use CFs to express causal relationships. Most AI systems, however, are only able to capture associations or correlations in data, so interpreting them as casual would not be justified. In this perspective, we present two experiments (total n = 364) exploring the effects of CF explanations of AI systems' predictions on lay people's causal beliefs about the real world. In Experiment 1, we found that providing CF explanations of an AI system's predictions does indeed (unjustifiably) affect people's causal beliefs regarding factors/features the AI uses and that people are more likely to view them as causal factors in the real world. Inspired by the literature on misinformation and health warning messaging, Experiment 2 tested whether we can correct for the unjustified change in causal beliefs. We found that pointing out that AI systems capture correlations and not necessarily causal relationships can attenuate the effects of CF explanations on people's causal beliefs.

INTRODUCTION

Interest in automatically generated explanations for predictive artificial intelligence (AI) systems has grown considerably in recent years.[1–6] It is argued that explanations provide transparency for what are often black-box procedures and that transparency is viewed as critical for fostering the acceptance of AI systems in real-world practice.[7–12] Explainable AI (XAI) has emerged as a field to address this need for AI systems' predictions to be followed by explanations of these predictions.

Common approaches to (post hoc) explainability of specific predictions of AI systems include feature importance,[13,14] saliency maps,[15] and example-based methods.[16] In this perspective, we focus on counterfactual (CF) explanations of specific predictions of AI systems.[12,17–24] These explanations describe changes in an AI system's inputs (features/factors) that alter the AI system's output (prediction/label) and lead to favorable outputs. CF explanations address questions such as "why A rather than B?"; for example, "Why did the AI system deny the loan rather than approve it?" An answer to this question would

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

be a CF explanation: "If Tom's salary had been at least £30k, the AI system would have offered him a loan." CF explanations not only provide us with an insight into why an AI system made a certain prediction ("deny the loan") but also what a user can do in order to flip the prediction ("offer a loan"). In other words, CF explanations may also be able to provide recourse for users.[19] Furthermore, CF explanations naturally embody contrastiveness, i.e., the ability to address the question of why this prediction instead of some other one, which is one of the attributes that people expect explanations to have.[25]

A significant body of research on CF explanations can be found in cognitive science and psychology. Some of the results of this research suggest that CF explanations often convey causal relations[26–29] and that making causal judgments often requires comparing actual and relevant CF situations.[30,31] For example, taking painkillers can have side effects such as fatigue. In a situation where a runner sprained an ankle and took painkiller *A*, which has fatigue as one of its side effects, people would judge that painkiller *A* has caused poor performance and loss of the race when they are aware of an alternative painkiller *B* without side effects. Here, people formed a CF: if the runner had taken painkiller *B*, she would not have had the side effects. However, when painkiller *B* also leads to side effects, people judge painkiller *A* to have less causal impact on the race outcome: even if the runner had taken *B*, she still would have had side effects.[32]

AI systems are typically predictive in nature and are capturing associations and correlations in data, not causal processes that generated the data. More specifically, in most applications of AI systems, we use data $\mathbf{X}$ and Y to estimate a function f, which in turn is used to generate predictions $\widehat{Y}$ for new instances. No underlying theoretical causal model for function f is assumed. Moreover, f is not expected to adequately capture the underlying (causal) processes or real-world mechanisms that generated the data used for training and estimation. It is thus entirely possible that explanations for predictions $\widehat{Y}$ that comprise of changes in features $\mathbf{X}$ have no clear causal connection (when, for example, $\mathbf{X}$ contains heavily engineered features) or have an anti-causal relationship, where Y is a cause of some $\mathbf{X}$. Furthermore, due to regularization, it is possible that some of the actually causal $\mathbf{X}$ are left out or that their impact on estimating $\widehat{Y}$ is reduced.[33] One should then be careful when using AI systems and explanations of their predictions not to misinterpret AI systems in a causal manner and to be wary of their limits.[34,35] This, however, may be easier said than done, particularly in the case of CF explanations of AI systems' predictions. CF explanations normally assume that the change in feature values maps onto the actions in the real world. This implies that CF explanations ought to incorporate a causal mechanism that would allow the person receiving the explanation to meaningfully intervene in the real world.[36] Some work on incorporating the real-world causal relationships in CF explanations has been done;[19,37] however, the vast majority of CF explanation generation algorithms do not account for the causal structure of the world.[24]

If people naturally associate CF with causal reasoning, as is suggested by the psychological and cognitive science research, then they may be especially prone to slipping into causal interpretations of AI system results when they are presented with CF explanations. As a consequence, they may form an (unjustified) mental model of the causal structure of the world or the underlying processes that generated the data. In other words, it may lead the recipients of CF explanations to form disingenuous and over-attributive perspectives with respect to these systems. Recent empirical work suggests that CF explanations do promote causal interpretations of features/factors used by an AI system,[38] making plausible the claim that people may indeed form an over-attributive perspective regarding AI system predictions when coupled with CF explanations.

In this perspective, we test the possibility that CF explanations may lead lay people into believing that relations captured by AI systems are causal in the real world. We report two experiments. The first investigated if lay people are more likely to form causal beliefs about the factors/features AI systems are using when these are presented with CF explanations. The second experiment explores a possible means to prevent lay people from forming inadvertent causal beliefs due to CF explanations.

EXPERIMENT 1

The aim of this experiment was to explore lay people's causal beliefs after having received a prediction made by an AI system, which is then supplemented with a CF explanation. The main hypothesis is that people's causal beliefs about the world will be (unjustifiably) affected by CF explanations of an AI system's predictions. More specifically, we hypothesize that people will erroneously hold beliefs that the features an AI has used to make predictions are more causal when a CF explanation of the AI system's prediction is provided compared with when the prediction of an AI system is presented without a CF explanation and compared with a baseline (where no AI system or its predictions are mentioned).

As AI systems are predictive in nature, one might argue that the above hypothesized effect may be due to lay people conflating the prediction/predictive power of AI systems with causation. The second hypothesis is aimed at testing this possibility. More specifically, we hypothesize that knowing an AI system is using certain feature A to predict label B and knowing what the predictions are *will* change people's *expectation* as to how good a predictor feature A is with respect B compared with the baseline. Crucially, however, we hypothesize that additionally knowing an explanation for that prediction *will not* further change people's *expectation* as to how good a predictor A is. This finding would imply that any change in *causal beliefs* would be due to the presence of a CF explanation of AI predictions and cannot be accounted for by a change in the *expectation* of how good the features the AI system uses are in predicting the label. Experiment 1 tested both hypotheses.

Methods

Participants

A total of 93 participants (n_{female} = 74, one participant identified as neither male nor female, M_{age} = 37.2, SD = 13.3) were recruited from Prolific Academic (www.prolific.ac). All participants were native English speakers residing in the UK or Ireland whose approval ratings were 95% or higher. They gave informed consent and were paid £6.24 an hour for partaking in the study, which took on average 8.1 min to complete. Both Experiments 1 and 2 were approved by the Department of Psychological

Sciences, Birkbeck, University of London Ethics Committee (reference 2021074).

Design

Participants were randomly assigned to one of three between-participant groups: the control/baseline group, where participants were asked about their intuitions regarding how certain factors/features influence salary without mentioning AI systems or explanations of AI systems (n = 30); the AI prediction group, where participants were told about the AI system and the features it uses as well as what the prediction is (n = 31); and the AI explanation group, where they were told about the AI system, the features it uses, and what prediction is and received a CF explanation of the prediction for each feature (n = 32).

The experiment had three dependent variables: expectation, confidence, and action. The expectation dependent variable measured people's beliefs regarding how well features predict the label. The confidence dependent variable measured how confident participants were in their expectation estimates. The main reason for including the confidence dependent variable was to disentangle people's beliefs about the predictive power of the features and their confidence in how predictive they believe the features are. We do not, however, have any hypotheses as to how confidence will change as a function of the group people were assigned to. Lastly, the action dependent variable measured people's causal beliefs about the real world in terms of their willingness to act or recommend a certain action to be done in the real world. All participants provided answers for each dependent variable.

Materials

To test the hypotheses, we used salary as a domain; it is reasonable to expect that most participants will have some familiarity regarding factors/features affecting salary and that they would already have developed certain intuitions about these factors. We chose 9 factors/features that are, to various extents, intuitively related to higher/lower salary. These were as follows: education level; the sector the employee works in (private or public); the number of hours of sleep; whether or not the employee owns a smart watch; whether or not the employee owns an office plant; whether or not the employee gets expensive haircuts; whether or not the employee wears expensive clothes; whether or not the employee goes skiing multiple times a year; and whether or not the employee rents a penthouse apartment. We aimed to have a range of factors/features whereby some are intuitively causing higher/lower salary (e.g., education level, sector), some are intuitively not relevant to salary (e.g., office plant, smart watch), and some are potential consequences or effects of higher salary rather than causing higher salary (e.g., expensive clothes, expensive haircuts, renting penthouse apartments). With these factors/features, we sought to cover possible ranges of expectation and action dependent variables. Namely, we hoped that for some factors/features, such as education level, both expectation estimates and action estimates would be high (i.e., education level is a good predictor of salary, and to increase their salary, one might consider getting a higher degree); some factors/features would have both expectation and action estimates very low (e.g., whether or not someone has an office plant does not seem to be related to salary, and buying an office plant to increase salary would seem like a futile endeavor); lastly, some factors/features such as expensive clothes and renting a penthouse apartment would have higher expectation estimates but lower action estimates (i.e., that someone is renting a penthouse apartment may be an indicator that they have a high salary, but one would not presumably rent a penthouse apartment because they believe that would increase their salary). The features/factors were not chosen from a specific dataset but were devised for the purposes of the experiment.

All collected participant data and materials as well as the analysis code are available via OSF: https://osf.io/xu7v6/?view_only=a4d11733f3a546cca4b76ad8fbc75018.

Procedure

After providing informed consent and basic demographic information (age, gender, and first language; no personally identifiable information was collected), participants were shown a welcome message. Participants then answered two preliminary questions: "How familiar are you with the factors that may affect salary?" and "How familiar are you with the AI technology, e.g. AI systems that are able to make predictions?" Answers to both questions were on a 7-point Likert scale from "1 - Not at all familiar" to "7 - Extremely familiar." The main motivation for including these questions was to check whether any differences among the three groups in the subsequent expectation or action estimates were due to differences in familiarity with the domain (salary) or familiarity with AI technology.

Following these two preliminary questions, participants saw a preamble for the specific group they were assigned to, i.e., control, AI prediction, or AI explanation (square brackets indicate which text was presented to which group):

> Your good friend Tom is looking to increase his **salary**. He's asked you for advice on how to best achieve that. [all three groups]
> There are a range of factors that are related to a higher salary. You will now consider some of these factors. [only the control group]
> In your search for ways to help your friend you have found an **AI system** that can predict whether people's yearly **salaries** are *higher than/equal to È30k (≥ È30k)* or *lower than È30k (< È30k)*. [AI prediction and AI explanation groups]
> The AI system uses a number of **factors** to make the prediction. You do not know, however, how much each factor is important for the AI system when it is making its predictions. [only the AI prediction group]
> The AI system uses a number of **factors** to make the prediction. The AI system also has an option to provide you with **explanations** regarding its predictions. [only the AI explanation group]
> You input Tom's details for all factors into the AI system and it predicts that his yearly salary is lower than **È**30k (< **È** 30k). [AI prediction and AI explanation groups]
> The AI system now provides you with explanations with respect to each factor as to why it predicts that Tom's salary is lower than *È*30k (< *È*30k). [only the AI explanation group]

The cutoff *È*30k was used as that figure was close to the median salary in the UK in 2020. After participants read the

preamble for the group they were assigned to, they proceeded to answer the three questions (expectation, confidence, action) regarding the 9 factors. The order of factors/features was randomized for each participant. Before answering the three questions, participants in the AI prediction and AI explanation groups were reminded of the AI system's prediction, and in the AI explanation group, people were additionally told the CF explanation for that factor. For example, questions and preceding text relating the education level were as follows.

> Reminder: The AI system predicts that Tom's yearly salary is ***lower than È30k (< È30k)***. [AI prediction and AI explanation groups]
> Factor: **Education level** [all three groups]
> Explanation: If Tom had **had an advanced degree** (**e.g. masters**), the AI system would have predicted that his **salary** was h**igher than/equal to È30k** (≥ **È30k**). [only the AI explanation group]
> **Q**. Would you expect that employees who **have an advanced degree (e.g. masters)** also have a **higher salary**? [expectation question, same for all three groups]
> Please rate your answer from 0 (No, not at all) to 100 (Yes, absolutely).
> **Q**. How confident are you in your response? [confidence question, same for all three groups]
> Q. Assuming Tom has the resources (time, money, etc.), would you recommend he **starts an advanced degree (e.g. masters)** with the hope of increasing his **salary**? [action question, same for all three groups]
> Please rate your answer from 0 (not at all) to 100 (totally).

Participants' responses to the three questions were elicited using a slider from 0 to 100 with 1-point increments. The three questions followed the same format for all other factors. The action questions sometimes had a short caveat ("Assuming Tom has the resources…") as shown above to guard against participants drifting into a cost-benefit analysis, which could deter from them providing causal estimates regarding the factor in question. The format of the CF explanations was the same for each factor, namely "If Tom had [had/worked/owned/etc. the factor/feature], the AI system would have predicted that his **salary** was **higher than/equal to È30k** (≥ **È30k**)." For example, for factor "office plant," the explanation was "If Tom had **owned an office plant**, the AI system would have predicted that his **salary** was **higher than/equal to È30k** (≥ **È30k**)," and for factor "penthouse apartment," the explanation read, "If Tom had **rented a penthouse apartment**, the AI system would have predicted that his **salary** was **higher than/equal to È30k** (≥ **È30k**)." Given that this formulation of the CF explanation implies a positive impact of the factor/feature on salary, we expected that participants' action estimates in the AI explanation group would be *higher* than the action estimates of the participants in the other two groups. At the end of the survey, participants were asked to summarize their reasoning for the estimates they provided in a free-format type textbox. This information was used to gain insight into the potential approaches participants took to answer the questions. Lastly, participants received debriefing information and were invited to provide feedback.

Results

Familiarity with the factors affecting salary and AI systems

We first analyzed the participants estimates regarding how familiar they are with factors affecting salary. We performed a one-way ANOVA for each familiarity category (i.e., salary and AI systems) with group (control, AI prediction, AI explanation) as a three-level independent variable. We found no significant effect of group on either familiarity with factors affecting salary, $F(2, 90) = 0.66$, $p = 0.52$, or familiarity with AI systems, $F(2, 90) = 1.96$, $p = 0.15$. Mean familiarity ratings indicated that participants were more familiar with factors affecting salary ($M = 3.9$) than AI systems ($M = 2.8$), which is expected. These results suggests that any potential significant differences between the groups in the further analyses cannot be accounted for by the participants familiarity with the domain (salary) or AI systems.

Main analyses

Participant estimates for each dependent variable are shown in Figure 1. To test the effect of group on each dependent variable, we initially built three linear mixed-effects models with the random intercept for each participant. However, as the distributions of participants estimates were highly skewed (especially for expectation and action dependent variables), and as residuals of the linear mixed-effects models were clearly non-normally distributed (see Appendix B), to test for the overall effect of the group, we resorted to non-parametric tests. We performed a Kruskal-Wallis rank-sum test for each dependent variable with the group as a three-level independent variable. Participants' expectation estimates were significantly affected by the group they were assigned to (control, AI prediction, AI explanation) for expectation estimates, $H(2) = 11.9$, $p = 0.003$, confidence estimates, $H(2) = 16.1$, $p < 0.001$, and action estimates, $H(2) = 27$, $p < 0.001$.

We performed post hoc pairwise comparisons between the three groups using a Wilcoxon rank sum test (the false discovery rate for multiple comparisons was controlled using the Benjamini-Hochberg procedure;[39] for more details, see Table 1 in Appendix A). We found that participants' action estimates were not significantly different between control and AI prediction groups ($p = 0.74$) but that there was a significant difference between AI prediction and AI explanation ($p < 0.001$) as well as between control and AI explanation ($p < 0.001$). From Figure 1, we can also see that participants' action estimates were higher in AI explanation group compared with the two other groups. Figure 2 suggests that this effect held across the features/factors and not just overall. These results provide support for our main hypothesis, namely that providing CF explanations would affect people's beliefs about how causal the features in the real world are.

Post hoc pairwise comparisons with respect to the participants' expectation estimates showed significant differences between control and AI prediction ($p = 0.03$) as well as control and AI explanation ($p = 0.002$); however, the AI prediction group's and the AI explanation group's estimates were not significantly different ($p = 0.36$). These results support our second hypothesis: being aware that an AI system is using certain factors/feature to make predictions and knowing what the prediction is affects people's expectation as to how well these features/factors are predicting salary. However, their expectations will not further

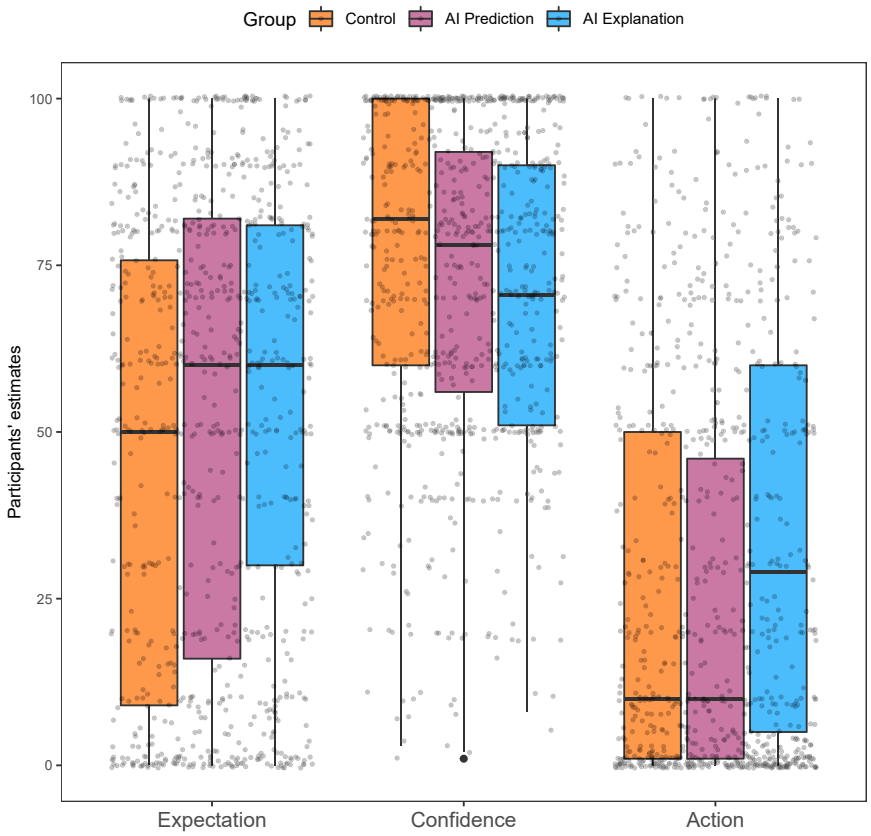

Figure 1. Experiment 1 results for each group and each dependent variable

We find that action estimates were significantly higher in the AI explanation group than in the two other groups. We also find that the AI explanation group's expectation estimates were not significantly higher than the AI prediction group's estimates, suggesting that the effect of CF explanations on action estimates is not due to participants beliefs in the AI's predictive power.

change upon learning about the CF explanations of the features/factors. This also implies that the results regarding participants' action estimates cannot be explained by the participants' expectation estimates, providing further support for the claim that CF explanations are affecting people's causal beliefs.

Post hoc pairwise comparisons on participants' confidence estimates showed significant difference between control and both AI prediction ($p = 0.02$) and AI explanation ($p < 0.001$) groups. There was not a significant difference between AI prediction and AI explanation groups ($p = 0.12$). Figure 1 shows a downward trend in estimates from the control to the AI explanation group. We speculate that this might be because some of the features/factors the AI system uses are intuitively not relevant to salary or because they are effects rather than causes of higher/lower salary. This may result in the reduction in people's confidence in the AI system's predictive accuracy. It is important to note that participants' confidence estimates were clearly different from their expectation estimates, suggesting that these two dependent variables were successfully disentangled in the experiment design.

Lastly, Figure 2 shows that the participants' estimates were dependent on the feature/factor they were asked to provide estimates for. These roughly coincided with the intuitions outlined above, namely that factors/features such as education level and sector would have both high expectation and action estimates as they seem that they can causally affect salary; factors/features such as office plant and smart watch do not seem causally relevant for salary, hence their low estimates for both expectation and action, and factors/features such as expensive clothes and renting a penthouse apartment are intuitively effects of higher salary, hence their expectation estimates would be higher, whereas their action estimates would be generally low. Critically, we found that action estimates were higher in the AI explanation group compared with the other two groups across most of the factors/features. This was not the case for the participants' action estimates for education level (degree) and sector factors. We found that the group participants were assigned to did not significantly affect their action estimates ($H(2) = 4.2$, $p = 0.12$). This is expected for two reasons. For one, as participants in the control group already have high action estimates for these two factors (following our conjecture that these two factors are intuitively causal factors), adding a CF explanation may further reinforce, but not significantly change, their causal beliefs about these two factors. On the other hand, we found that participants' action estimates were significantly different depending on the group they were assigned to for all other factors ($H(2) = 38.1$, $p < 0.001$). Secondly, as our scales were bounded at 100, it is possible that some of the non-significant finding is due to ceiling effects.

EXPERIMENT 2

Experiment 1 suggested that providing lay users with CF explanations of AI systems' predictions can (unjustifiably) affect their causal beliefs about the features/factors. The aim of Experiment 2 was to explore if we can correct the effects of CF explanations on people's causal beliefs. Inspired by the research on correcting misinformation[40] and the research on the impact of health warning messages,[41] we designed this experiment to explore if providing participants with a note communicating that AI systems are capturing correlations in data rather than causal relationships might attenuate the effect of CFs on their causal beliefs. We hypothesize that the AI explanation group presented with the note will provide lower action estimates than the AI explanation group, where the note was not present. We do not have a specific hypothesis as to how introducing the note might affect participants' expectation, confidence, or action estimates in the other groups or how the AI explanation

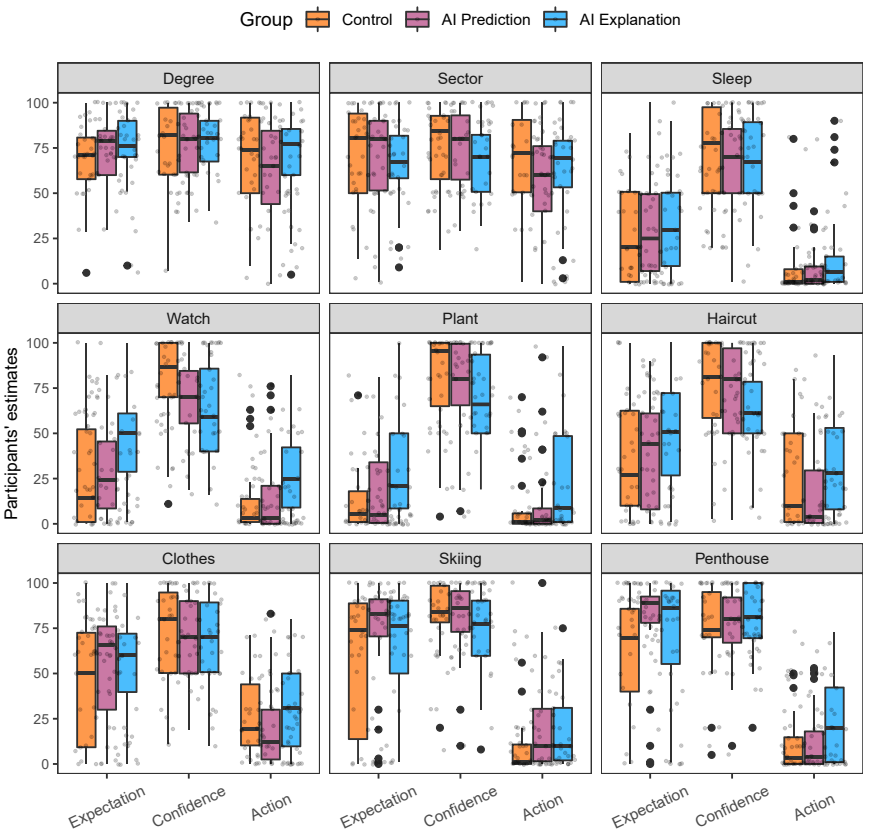

Figure 2. Experiment 1 results for each factor/feature, each group, and each dependent variable

We find a trend whereby the participants' action estimates were higher in the AI explanation group than in the other two groups across the factors. We also find that for factors that are intuitively more causal, both expectation and action estimates were high; for factors that are intuitively not causal, both expectation and action estimates were low; and for factors that are intuitively effects rather than causes of higher salary, expectation estimates were high, and action estimates were low, in agreement with the experimental predictions.

groups' expectation and confidence estimates might change due to the note.

The second aim of Experiment 2 was to provide a replication of Experiment 1 in groups that are not presented with the note. Thus, Experiment 2 will provide an additional test for the two hypotheses explored in Experiment 1.

Methods

Effect size calculations showed that the effect size of Experiment 1 results was relatively small (η^2 = 0.03), making Experiment 1 underpowered. To increase the power of Experiment 2, we increased the number of participants. We aimed to have around 45 participants in each group.

Participants, design, and materials

A total of 271 participants (n_{female} = 196, two participants identified as neither male nor female, M_{age} = 38.7, SD = 12.2) were recruited from Prolific Academic (www.prolific.ac). All participants were native English speakers residing in the UK or Ireland whose approval ratings were 95% or higher. They all gave informed consent and were paid *È*6.24 an hour for partaking in the present study, which took on average 8.6 min to complete. Participants were randomly assigned to one of 3 (control, AI prediction, or AI explanation) × 2 (correction: no note or note) = 6 between-participant groups (control and no note, n = 46; control and no note, n = 45; AI prediction and no note, n = 44; AI prediction and note, n = 46; AI explanation and no note, n = 46; AI explanation and no note, n = 44). Experiment 2 used the same three dependent variables as Experiment 1, namely expectation, confidence, and action. Materials in Experiment 2 were exactly the same as those in Experiment 1.

Procedure

The procedure for Experiment 2 was similar to the procedure for Experiment 1. The only difference is that three of the 6 groups were additionally presented with a note regarding correlation, causation, and AI systems. For groups with the note, that note was introduced in the preamble of each condition, presented on a separate page, and participants were also reminded of the note before answering the questions related to the three dependent variables. The note read slightly differently for control, AI prediction, and AI explanation groups. The note for the control group read as follows:

> **Important note**
> *Correlation does not imply causation*. Even though some factors may be highly **correlated** with higher salary that ***does not*** mean that they are **causing** higher salary.

The note does not mention the AI system, as participants in this group were not presented with any AI system. Instead, the note included general information about correlation and causation. In the AI prediction group, the note read the following:

> **Important note**
> AI systems learn ***correlations*** in data. Even though the factors the AI system uses are potentially ***correlated*** with higher salary that **does not** mean that they are ***causing*** higher salary.

Here, participants are told information regarding correlation and causation that is more relevant to the AI systems. Specifically, they are told that AI systems capture relationships that are correlational and should not be interpreted as causal. In the AI explanation condition, the note read the following:

> **Important note**
> AI systems learn ***correlations*** in data. Even though the factors the AI system uses are potentially ***correlated*** with higher

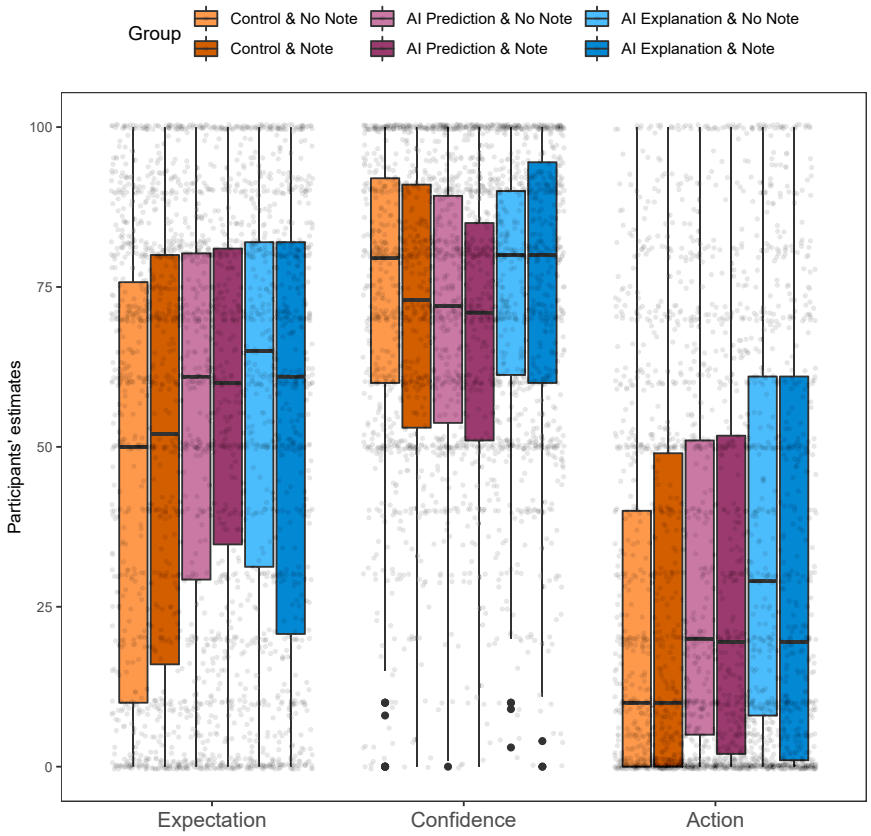

Figure 3. Experiment 2 results for each dependent variable

Like in Experiment 1, we find that action estimates are higher in the AI explanation group and cannot be explained by expectation estimates when the note is not communicated to the participants. When the note is presented to the participants, action estimates in the AI explanation group are at the level of the AI prediction group and not significantly higher.

salary that **does not** mean that they are ***causing*** higher salary. Similarly, the **explanations** of the AI systems' predictions are about the *correlations* the AI system has identified and not about which factors are *actually causing* higher salary.

In addition to being told that AI systems capture correlations, participants in this group were also told that the explanations of the AI system's predictions are explanations of these correlations and are not necessarily of causal relations.

Results

Familiarity with the factors affecting salary and AI systems

Like in Experiment 1, we first analyzed the participants' estimates regarding how familiar they are with factors affecting salary. We performed a two-way ANOVA for each familiarity category (i.e., salary and AI systems) with group and correction as two factors. We found no significant effect of group (control, AI prediction, AI explanation) on either familiarity with factors affecting salary, $F(2, 265) = 0.99$, $p = 0.37$, or familiarity with AI systems, $F(2, 265) = 0.13$, $p = 0.88$. We found no significant effect of correction (no note, note) on either familiarity with factors affecting salary, $F(1, 265) = 0.55$, $p = 0.46$, or familiarity with AI systems, $F(1, 265) = 0.06$, $p = 0.8$. Finally, we found no significant interaction effect between the two independent variables on either familiarity with factors affecting salary, $F(2, 265) = 1.68$, $p = 0.19$, or familiarity with AI systems, $F(2, 265) = 1.19$, $p = 0.31$. Mean familiarity ratings indicated that participants were more familiar with factors affecting salary ($M = 3.9$) than AI systems ($M = 2.9$). These results are very similar to those in Experiment 1.

Main analyses

Participants' estimates for each dependent variable as shown in Figure 3. Similar to Experiment 1, the distributions of participants' estimates were skewed (especially for expectation and action dependent variables), and residuals of the linear mixed-effects models were clearly non-normally distributed (see Appendix B). So, to test for the overall effect of group, we performed a Kruskal-Wallis rank-sum test for each dependent variable. Participants' expectation estimates were significantly affected by group (control and no note, control and note, AI prediction and no note, AI prediction and note, AI explanation and no note, AI explanation and note) for expectation estimates, $H(5) = 38.9$, $p < 0.001$, confidence estimates, $H(5) = 33.8$, $p < 0.001$, and action estimates, $H(5) = 67.7$, $p < 0.001$.

From Figure 3, we can also see that participants' action estimates were significantly higher in the AI explanation and no note group compared with the two other no note groups. This mirrors the findings from Experiment 1 and further supports the first hypothesis from Experiment 1. Unlike in Experiment 1, the difference between control and AI prediction condition was also significant, $p = 0.02$ (for more details, see Table 4 in Appendix A). Pairwise comparisons for dependent variable expectation show significant difference only between control and both AI prediction ($p = 0.001$) and AI explanation ($p < 0.001$) groups. No significant difference was found between AI prediction and AI explanation ($p = 0.31$). This result provides support to our second hypothesis from Experiment 1 and suggests that even though there were significant difference between all three no note groups in the action dependent variable, the significant difference between AI prediction and AI explanation cannot be accounted for by differences in expectation estimates.

Pairwise comparisons across all three dependent variables show that the only significant difference between no note and note conditions was between AI explanation and no note and AI explanation and note for the action dependent variable ($p = 0.01$) (see Tables 2, 3, and 4 in Appendix A for more details).

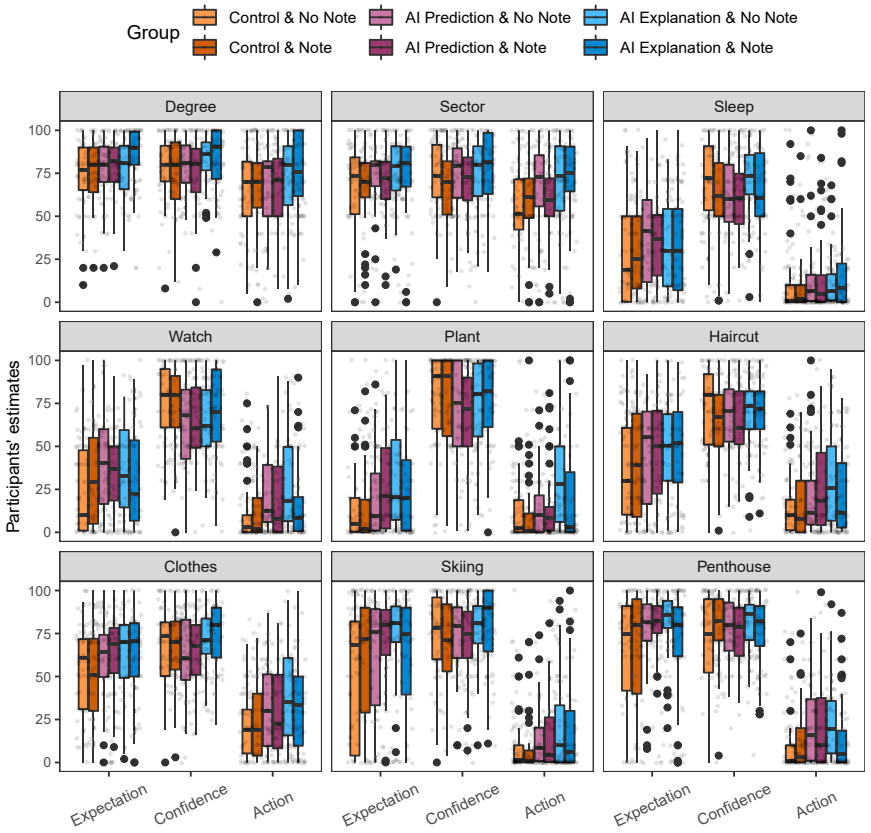

Figure 4. Experiment 2 results for each factor/feature and each dependent variable

We find that the similar distributions of estimates for different factors as in Experiment 1. Furthermore, we also find the trend of reduction in action estimates when the note is present compared with when it is not.

Participants' action estimates in group AI explanation and note were lower than those in group AI explanation and no note and were not significantly different than those from AI prediction and no note or AI prediction and note. This implies that the effect of CF explanations on participants' causal beliefs was attenuated and no different from that in groups where CF explanations of AI systems' predictions were not shown. Moreover, Figure 3 suggests that participants' action estimates in AI explanation and note were lower than those in explanation and no note for almost all features/factors. These results directly support our hypothesis.

Finally, participants' confidence estimates were again different from their expectation estimates. But, unlike in Experiment 1, where there was a downward trend in participants' confidence estimates, Experiment 2 found that AI prediction groups' estimates were lower than both control groups' and AI explanation groups' estimates and that there was no significant difference between control groups' and AI explanation groups' estimates. In Experiment 1, we speculated that confidence estimates might be driven by some factors/features not being relevant to salary or in an anti-causal relationship to (i.e., effects of) salary. However, the data form Experiment 2 does not seem to support this supposition.

DISCUSSION

If one of the aims of XAI is to provide human users with information that will help them better understand how an AI system came to a prediction and how the system will behave in the future, then we need to communicate to that user as clearly as possible the predictive and associative (rather than the causal) nature of these systems so that the mental models humans create based on that information are more genuine and representative of the AI system's nature.

Two experiments showed that participants' causal estimates were significantly higher when they were presented with a CF explanation compared with both the baseline and when only the prediction was communicated. We further found that this was not the case for people's beliefs regarding how good the feature/factors are at predicting salary and that there was no significant difference in expectation estimates difference between the group where only predictions were presented and the group where both the prediction and a CF explanation was included. This result suggests people's expectation estimates cannot account for the differences in their causal beliefs and that these differences were in fact due to CF explanations alone. This implies that CF explanation of AI systems' predictions can (unjustifiably) skew people's causal beliefs about the world.

We also found that one might be able to guard against the unwanted effect of CF explanations on causal beliefs. Inspired by the work on misinformation and health warning messaging, we designed a note communicating to the participants the correlational character of AI systems rather than causal. Adding the note reduced the effect of CF explanation on the participants' causal beliefs.

Future work

In this study, we have used salary as a domain. This is because we expected that participants are largely familiar with this domain, which would allow us to test whether their expectation and action estimates were in line with what we expected. As their estimates were in line with our expectations, this provided evidence for the appropriateness of the metrics we used. Further research should explore other domains; in particular, the domains that people are not as familiar with. We expect, however, that the effect of adding a CF explanation on action estimates will be at least as great as in this study. From Figures 2 and 4, we can see that when participants' expectation estimates were relatively low (e.g., owning a smart watch, owning an office plant, and getting expensive haircuts), implying that they believed the

association between the features/factors and salary was weak, their action estimates were invariably higher in the AI explanation condition compared with the other two conditions. If, as seems plausible, when reasoning in a relatively unfamiliar domain, participants believe there is a weak association between the features/factors and the label, then we would expect to find a greater impact of adding CF explanations on their action estimates compared with this study, as this study also included features that are intuitively causally efficacious with respect to the label (e.g., having an advance degree or sector the employee works in).

A wealth of research on explanation and explanatory goodness suggests that simpler explanations have a bigger impact on our (causal) beliefs.[42–45] In our studies, only one feature/factor was included in a CF explanation at a time, so our CF explanations were on the simpler side of the spectrum. This could imply that the CF explanations that include multiple factors and are a combination of these factors have less impact on our causal beliefs about the world than the simpler CF explanation that uses one or two factors. Further research should explore how more complex CF explanations of AI systems' predictions affect people's causal beliefs about the world. It should be noted that although increasing the complexity of CF explanations may reduce their undesired impact on our causal beliefs, it may at the same time increase the time needed to process these explanations and reduce satisfaction.[46]

Finally, we have only briefly discussed the role of the participants' confidence in their expectation estimates. We found that confidence estimates are clearly different from the expectation ones. However, we have not explored in further detail how confidence estimates may depend on whether people are just told about the AI system's prediction or they are also told the CF explanation. It may be interesting to explore how confidence estimates interact with people's estimates of how accurate they believe the AI system is in predicting the label.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100635.

ACKNOWLEDGMENTS

Marko Tešić holds a research fellowship from the Royal Academy of Engineering (RAEng) and we would like to thank the RAEng for supporting this research. We would also like to thank Kirsty Phillips for insightful discussions and comments on the experimental materials.

REFERENCES

1. DARPA (2016). Explainable Artificial Intelligence (XAI) Program. Retrieved from. https://www.darpa.mil/program/xplainable-artificial-intelligence.

2. Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.1702.08608.

3. Gunning, D., and Aha, D. (2019). Darpa's explainable artificial intelligence program. AI Mag. *40*, 44–58.

4. Montavon, G., Samek, W., and Müller, K.R. (2018). Methods for interpreting and understanding deep neural networks. Digit. Signal Process. *73*, 1–15.

5. Rieger, L., Chormai, P., Montavon, G., Hansen, L.K., and Müller, K.R. (2018). Structuring neural networks for more explainable predictions. In Explainable and Interpretable Models in Computer Vision and Machine Learning, H.J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, and U. Güçlü, et al., eds. (Springer), pp. 115–131.

6. Samek, W., Wiegand, T., and Müller, K.R. (2017). Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. Preprint at arXiv. https://doi.org/10.48550/arXiv.1708.08296.

7. Bansal, A., Ali, F., and Parikh, D. (2014). Towards transparent systems: semantic characterization of failure modes. In European Conference on Computer Vision (Springer), pp. 366–381.

8. Chen, J.Y., Procci, K., Boyce, M., Wright, J., Garcia, A., and Barnes, M. (2014). Situation Awareness-Based Agent Transparency (Army research lab Aberdeen proving ground MD human research and engineering). Technical report.

9. Fallon, C.K., and Blaha, L.M. (2018). Improving automation transparency: addressing some of machine learning's unique challenges. In International Conference on Augmented Cognition (Springer), pp. 245–254.

10. Hayes, B., and Shah, J.A. (2017). Improving robot controller transparency through autonomous policy explanation. In 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (IEEE), pp. 303–312.

11. Mercado, J.E., Rupp, M.A., Chen, J.Y.C., Barnes, M.J., Barber, D., and Procci, K. (2016). Intelligent agent transparency in human–agent teaming for multi-uxv management. Hum. Factors *58*, 401–415.

12. Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv. JL & Tech. *31*, 841.

13. Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. *30*.

14. Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (Association for Computing Machinery), pp. 1135–1144.

15. Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. Preprint at arXiv.

16. Koh, P.W., and Liang, P. (2017). Understanding black-box predictions via influence functions. In International conference on machine learning (PMLR), pp. 1885–1894.

17. Förster, M., Hühn, P., Klier, M., and Kluge, K. (2021). Capturing users' reality: a novel approach to generate coherent counterfactual explanations. In Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS community), p. 1274.

18. Galhotra, S., Pradhan, R., and Salimi, B. (2021). Explaining black-box algorithms using probabilistic contrastive counterfactuals. In Proceedings of the 2021 International Conference on Management of Data (Association for Computing Machinery), pp. 577–590.

19. Karimi, A.-H., Barthe, G., Schölkopf, B., and Valera, I. (2020b). A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. Preprint at arXiv. https://doi.org/10.48550/arXiv.2010.04050.

20. Kommiya Mothilal, R., Mahajan, D., Tan, C., and Sharma, A. (2021). Towards unifying feature attribution and counterfactual explanations: different means to the same end. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (Association for Computing Machinery), pp. 652–663.

21. Lucic, A., Haned, H., and de Rijke, M. (2020). Why does my model fail? contrastive local explanations for retail forecasting. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Association for Computing Machinery), pp. 90–98.

22. Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. (2020). Face: feasible and actionable counterfactual explanations. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (Association for Computing Machinery), pp. 344–350.

23. van der Waa, J., Nieuwburg, E., Cremers, A., and Neerincx, M. (2021). Evaluating xai: a comparison of rule-based and example-based explanations. Artif. Intell. *291*, 103404.

24. Verma, S., Dickerson, J., and Hines, K. (2020). Counterfactual explanations for machine learning: A review. Preprint at arXiv. https://doi.org/10.48550/arXiv.2010.10596.

25. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. *267*, 1–38.

26. Byrne, R.M.J. (2007). Precis of the rational imagination: how people create alternatives to reality. Behav. Brain Sci. *30*, 439–453.

27. Byrne, R.M.J. (2016). Counterfactual thought. Annu. Rev. Psychol. *67*, 135–157.

28. Byrne, R.M.J. (2019). Counterfactuals in explainable artificial intelligence (xai): evidence from human reasoning. In IJCAI, S. Kraus, ed. (International Joint Conferences on Artificial Intelligence Organization), pp. 6276–6282.

29. Thompson, V.A., and Byrne, R.M.J. (2002). Reasoning counterfactually: making inferences about things that didn't happen. J. Exp. Psychol. Learn. Mem. Cogn. *28*, 1154–1170.

30. Gerstenberg, T., Goodman, N.D., Lagnado, D.A., and Tenenbaum, J.B. (2021). A counterfactual simulation model of causal judgments for physical events. Psychol. Rev. *128*, 936–975.

31. Wells, G.L., and Gavanski, I. (1989). Mental simulation of causality. J. Pers. Soc. Psychol. *56*, 161–169.

32. McCloy, R., and Byrne, R.M.J. (2002). Semifactual "even if" thinking. Think. Reas. *8*, 41–67.

33. Del Giudice, M. (2022). The Prediction-Explanation Fallacy: A Pervasive Problem in Scientific Applications of Machine Learning. https://doi.org/10.31234/osf.io/4vq8f.

34. Dillon, E., Jacob, L.R., Scott, L., Roth, J., and Syrgkanis, V. (2021). Be Careful when Interpreting Predictive Models in Search of Causal Insights (Medium). Retrieved from. https://medium.com/towards-data-science/be-careful-when-interpreting-predictive-models-in-search-of-causal-insights-e68626e664b6.

35. Molnar, C., Gunnar, K., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M., and Bischl, B. (2020). Pitfalls to Avoid when Interpreting Machine Learning Models.

36. Barocas, S., Selbst, A.D., and Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Association for Computing Machinery), pp. 80–89.

37. Karimi, A.-H., Barthe, G., Balle, B., and Valera, I. (2020a). Model-agnostic counterfactual explanations for consequential decisions. In International Conference on Artificial Intelligence and Statistics (PMLR), pp. 895–905.

38. Warren, G., Keane, M.T., and Byrne, R.M.J. (2022). Features of explainability: how users understand counterfactual and causal explanations for categorical and continuous features in xai. Preprint at arXiv. https://doi.org/10.48550/arXiv.2204.10152.

39. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B *57*, 289–300.

40. Irving, D., Clark, R.W.A., Lewandowsky, S., and Allen, P.J. (2022). Correcting statistical misinformation about scientific findings in the media: causation versus correlation. J. Exp. Psychol. Appl. *28*, 1–9.

41. Hammond, D. (2011). Health warning messages on tobacco products: a review. Tob. Control *20*, 327–337.

42. Lombrozo, T. (2007). Simplicity and probability in causal explanation. Cogn. Psychol. *55*, 232–257.

43. Lagnado, D. (1994). The psychology of Explanation: A Bayesian approach (Schools of Psychology and Computer Science, University of Birmingham). Unpublished Masters Thesis.

44. Read, S.J., and Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: a parallel distributed processing account. J. Pers. Soc. Psychol. *65*, 429–447.

45. Thagard, P.R. (1978). The best explanation: criteria for theory choice. J. Philos. *75*, 76–92.

46. Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., and Doshi-Velez, F. (2018). How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.00682.