Counterfactual explanations for prediction and diagnosis in XAI *

Xinyue Dai†
School of Psychology
Trinity College Dublin
University of Dublin, Dublin, Ireland
daix@tcd.ie

Mark T. Keane
School of Computer Science
University College Dublin
Dublin, Ireland
mark.keane@ucd.ie

Laurence Shalloo
VistaMilk SFI Research Centre
Teagasc Moorepark
Cork, Ireland
laurence.shalloo@teagasc.ie

Elodie Ruelle
VistaMilk SFI Research Centre
Teagasc Moorepark
Cork, Ireland
elodie.ruelle@teagasc.ie

Ruth M.J. Byrne
School of Psychology
Trinity College Dublin
University of Dublin, Dublin, Ireland
rmbyrne@tcd.ie

ABSTRACT

We compared two sorts of explanations for decisions made by an AI system: counterfactual explanations about how an outcome could have been different in the past, and prefactual explanations about how it could be different in the future. We examined the effects of these alternative explanation strategies on the accuracy of users' judgments about the AI app's *predictions* about an outcome (inferred from information about the causes), compared to the accuracy of their judgments about the app's *diagnoses* of a cause (inferred from information about the outcome). The tasks were based on a simulated SmartAgriculture decision support system for grass growth outcomes on dairy farms in Experiment 1, and for an analogous alien planet domain in Experiment 2. The two experiments, with 243 participants, also tested users' confidence in their decisions, and their satisfaction with the explanations. Users made more accurate *diagnoses* of the presence of causes based on information about their outcome, compared to *predictions* of an outcome given information about the presence of causes. Their predictions and diagnoses were helped equally by counterfactual explanations and prefactual ones.

CCS CONCEPTS

• User studies • Human computer interaction (HCI) • HCI design and evaluation methods

AIES '22, August 1–3, 2022, Oxford, United Kingdom.

ACM ISBN 978-1-4503-9247-1/22/08...$15.00.
https://doi.org/10.1145/ 3514094.3534144

KEYWORDS

Explanation, Counterfactual, eXplainable AI (XAI), Prediction, Diagnosis, Smart Agriculture

ACM Reference format:

Xinyue Dai, Mark T. Keane, Laurence Shalloo, Elodie Ruelle, and Ruth M.J. Byrne, 2022. Counterfactual explanations for prediction and diagnosis in XAI. In *Proceedings of AIES conference (AIES'22). AIES, Oxford, UK, 12 pages.*

1 INTRODUCTION

Artificial Intelligence (AI) systems for automated decision support are increasingly prevalent in our everyday lives in diverse areas, ranging from health care to banking, and from human resources to agriculture. However, people may not trust the decisions made by an AI system or consider them to be fair, given that the reasons for decisions made by such "black box" systems are rarely obvious. EXplainable AI (XAI) aims to develop techniques to increase user understanding and trust of such systems, to improve an AI system's interpretability for users, and explain its decisions. Some XAI solutions focus on "transparency", based on attempts to explain the model directly by simplifying it or mapping the "black box" model into a more interpretable "white box" framework (e.g., using decision trees or linear regression). Other XAI solutions focus on justifying the model's decision, providing a post-hoc explanation for what it did (e.g., using feature importance for example-based explanations) [1]. Recently, counterfactual explanations have been proposed as a valuable post-hoc solution, that is, explanations that show how the model's decision outcome would have been different, under different input feature conditions. For example, an automated decision by a banking app to refuse a customer's loan application could be explained by a counterfactual such as, "*if you had asked for a lower loan, your*

application would have been approved". Counterfactual explanations have attracted significant research interest because they may meet GDPR requirements for algorithmic recourse for automated decisions [1-3] and because of their importance in human causal thinking [4, 5]. However, very little is known about how people understand the counterfactuals used in XAI and few AI studies carry out user tests of their methods. A recent review showed that fewer than 20% of studies included any form of user testing, and even fewer did so in an experimentally adequate manner [6].

Existing research suggests that counterfactual explanations are helpful to users, in that users were better able to predict what the AI system would do [7-11], and their trust and satisfaction with the system increased [9, 10, 12-17]. However, user studies have tended not to distinguish between counterfactuals about how things could have turned out differently in the past, such as "*if you had asked for a lower loan last month, your application would have been approved*", and explanations about how things could turn out differently in the future, e.g., *"if you were to ask for a lower loan next month, your application would be approved*", usually referred to as "prefactuals" [18-21]. Yet counterfactual and prefactual explanations have been found to implicate different psychological processes and to have different behavioural consequences, as we outline below [18-21, see also 22-27]. Hence, our aim in the current experiments was to compare the usefulness of counterfactuals and prefactuals as explanations for users of an AI system's decisions.

Moreover, user studies have tended to assess the accuracy of users' understanding of the decisions made by an AI system by testing their *predictive* inferences, that is, their predictions about an outcome based on information about the presence or absence of relevant causes [7-11]. Yet arguably, users' understanding of the decisions made by an AI system is best tested by their *diagnostic* inferences, that is, their diagnoses of the presence or absence of relevant causes, based on information about an outcome. After all, users of an AI system who require an explanation, for example, for the purposes of recourse, usually know the outcome, and wish to understand the causes; that is, they require a diagnostic explanation, not necessarily a predictive one. Predictions and diagnoses have also been found to implicate different psychological processes of causal reasoning, as we outline below [28-39]. Hence, the current experiments compared the usefulness of counterfactual and prefactual explanations for predictions and diagnoses.

Accordingly, we address a gap in the literature by testing counterfactuals and prefactuals as explanations, and by testing predictive and diagnostic inferences. We conducted two large-scale user studies (N = 243), using a causal model in which five causes could lead to an outcome, to examine participants' objective understanding of the AI system (measured by their decision accuracy and confidence judgments), and their subjective evaluation of explanations of its decisions (measured by their judgments of explanation helpfulness). In the next section we consider cognitive science research on counterfactual and prefactual explanations, and then we discuss predictive and diagnostic inferences.

1.1 Counterfactuals and prefactuals

We tested counterfactual and prefactual explanations, because cognitive psychological research shows that counterfactuals provide a richer set of possibilities for people to draw inferences from, compared to prefactuals. A counterfactual explanation, such as "*if you had asked for a lower loan last month, your application would have been approved*", is useful in part because it requires reasoners to mentally simulate two alternative possibilities: the conjecture, in which a lower loan was requested and it was approved, and the reality, in which a lower loan was not requested and it was not approved [22]. As a result, experimental findings show that people make many more inferences from counterfactuals about how things could have turned out differently, than from conditionals about current facts [23]. The mental simulation of such dual possibilities facilitates causal inferences [24]. In contrast, people can understand a prefactual explanation such as "*if you were to ask for a lower loan next month, your application would be approved*", by mentally simulating just the conjecture, in which a lower loan is requested and approved [22]. As a result, experimental findings show that people make just the same sorts of inferences from prefactuals about how things could turn out differently in the future, as they do from conditionals about current facts [18]. In other words, prefactuals do *not* provide the same inferential advantages as counterfactuals.

However, cognitive psychological research also shows that prefactuals provide a richer set of blueprints for future intentions to enable people to formulate actions, compared to counterfactuals. People tend to imagine how things could have turned out better when they reflect upon past decisions, rather than how they could have turned out worse [25], and these "better-world" counterfactuals help them to formulate intentions for the future to prevent similar bad outcomes happening again [26]. Counterfactuals are thus a key learning mechanism and they provide preparatory blueprints for behavior in the future [27]. In this regard, prefactuals about the future seem to be even more directly implicated in the formulation of future intentions. For example, when participants tried to solve a problem and failed, they created counterfactuals about how things could have turned out better for them by focusing on external aspects of the task outside their control, such as the time limit to solve the problem, or not being allowed to use pen-and-paper. In contrast, they created prefactuals about how things could turn out better for them next time by focusing on internal aspects within their control, such as their concentration, or their strategies [19-21]. In other words, prefactuals are more closely linked to the identification of actionable changes in the future than counterfactuals. Given the differences in how people think about counterfactuals and prefactuals, we compared their efficacy as explanations for AI decisions in our experiments. We wished to examine whether their relative pros and cons result in them being equally effective, or whether one is better than the other.

1.2 Predictions and diagnoses

We tested predictive and diagnostic inferences, because cognitive psychological research shows that people make different sorts of

inferences about causal relations when they reason from a cause to infer its outcome, compared to when they reason from an outcome to infer its cause. People rely on different sorts of information when they make predictions compared to when they make diagnoses. For simple cause-effect links, they make predictions more readily than diagnoses. For example, in experiments participants predicted the probability of a girl having blue eyes given that her biological mother has blue eyes, as more probable than the diagnosis of the probability of a girl's biological mother having blue eyes given the girl has blue eyes [28]. They also read causally-related words more quickly when they were in a predictive order, e.g., spark -> fire, compared to when they were in a diagnostic order, e.g., fire -> spark [29].

But in more complex cause-effect cases, in which several causes lead to an effect, participants in experiments tended to make more correct diagnoses than predictions. For example, when people made predictions from a cause to an effect, they tended to erroneously ignore alternative causes; but when they made diagnoses from an effect to a cause they tended to accurately consider the alternative causes [30-32]. There are, of course, many different sorts of causes, for example, a strong cause is necessary and sufficient for an outcome, whereas a weak cause is sufficient for an outcome but not necessary, because there are other alternative causes that can also bring about the outcome [33, 34]. In contrast, an enabling cause is necessary but not sufficient for the outcome, because other causes must occur with the enabler to bring about the outcome; nonetheless the absence of an enabler prevents the outcome, and likewise a disabling cause prevents the outcome [35]. Predictive inferences are affected by how strong a cause is, whereas diagnostic inferences are affected by how strong alternative causes are [31]. Predictive inferences about whether an effect occurred, when people were told a cause occurred, were suppressed by the absence of enablers (other causes whose absence prevented the outcome) or the presence of disablers (other causes that prevented the outcome) [31, 36]. In contrast, diagnostic inferences about whether a cause occurred, when people were told an effect occurred, were suppressed by alternative causes - that is, other causes that can bring about the effect [36-38]. In other words, people made different sorts of inferences about the different sorts of causes when they made predictions compared to when they made diagnoses.

Participants in experiments also predicted that an effect, (e.g., 'flu) is more likely, the more causal factors were present, (e.g., fever, headache, fatigue) [39]. But, in diagnosing a cause from the presence of an effect (e.g., when they know the lawn is wet), the presence of one cause (e.g., rain), led them to "discount" or doubt the presence of another cause (e.g., sprinklers) [40]. Given the differences in how people reason about predictions and diagnoses, we compared the efficacy of counterfactual and prefactual explanations for predictions and diagnoses. We made sure to include several causes for each outcome in our experimental materials, including alternative causes and disablers, and causes that had a high impact or a low impact on the outcome. For each trial in the experiments, we presented the information about causes and their outcome in a table (see one example in Figure 1), so that in our tasks all of the different sorts of causes were explicitly provided and known to be present or absent.

2 EXPERIMENTS

2.1 Experiment 1

The aim of the experiment was to test whether participants made more accurate inferences when they received counterfactual explanations or prefactual explanations, for inferences based on information about the presence or absence of five causal factors that affected grass growth on a given farm (i.e., micro-organisms in the soil, rainfall, fertilizer, clover, and cows), and information about the grass growth on the farm being high or low. The information was presented in tables, and Figure 1 provides an example of a prediction task and a diagnosis task. In the prediction tasks, we asked participants to judge what prediction the app would make, e.g., "*What do you think the app will predict about grass growth on this farm?*" and we provided them with the answer options "*High/Low*". In the diagnosis tasks, we asked them to judge what diagnosis the app would make about a specific causal factor, e.g., "*What do you think the app will diagnose about micro-organisms on this farm?*" and we provided them with the answer options "*Present/Absent*". Participants also rated how confident they were in their judgment using a 5-point scale anchored at highly unconfident (at 1) and highly confident (at 5). After they made their inference and indicated their confidence in it, we informed them of the app's decision (e.g., *grass growth was "high"*), and provided them with an explanation for the decision. We assigned participants to three groups who received either counterfactual explanations, prefactual explanations, or control descriptions. A counterfactual explanation was a past-tense subjunctive conditional (e.g., "*If rain had been absent last month, your grass growth would have been low*"); a prefactual explanation was a subjunctive conditional about the future (e.g., "*If rain were to be absent next month, your grass growth would be low*"), and a control description was merely a re-statement of the decision (e.g., "*The system's prediction is 'high'* "). The participants rated how helpful they found the explanation in assisting them to understand how the app works, using a 5-point scale anchored at very unhelpful (at 1) and very helpful (at 5).

The general information participants were given at the outset was as follows:

"Please consider the following five factors that have an impact on grass growth. For the purposes of the present study please consider only these factors and no other factors. Please also consider only the relation between a specified factor and
grass growth, please do not make assumptions about any possible relations between each of the five factors.

-Micro-organisms live in the soil that grass is planted in. The growth of grass increases if micro-organisms are present. Micro-organisms have a relatively high impact on grass growth.

-Rain is produced in various climates. The growth of grass increases if rain is present. Rain has a relatively high impact on grass growth.

Farm X - Prediction		Farm X - Diagnosis	
Micro-organisms	Present	Micro-organisms	?
Rain	Present	Rain	Present
Fertiliser	Present	Fertiliser	Present
Clover	Absent	Clover	Absent
Cows	Present	Cows	Present
Grass growth	?	Grass growth	High

Figure 1: Examples of the sorts of tables presented to participants in Experiment 1 to provide information about five causal factors (present or absent) and the outcome of grass growth (high or low). A table for a prediction task is on the left: information about the presence or absence of all five causal factors is given, and the task is to make a judgment about the outcome, as high or low. A table for a diagnostic task is on the right: information about the outcome as high or low and information about the presence or absence of four of the five causal factors is given, and the task is to make a diagnosis about the presence or absence of one of the causal factors (indicated by the question mark). Participants received one table at a time

-Fertilizer is an element that can be added to soil. The growth of grass increases if fertilizer is present. Fertilizer has a relatively low impact on grass growth.

-Clover is a plant often found growing with grass. The growth of grass increases if clover is present. Clover has a relatively low impact on grass growth.

-Cows feed on grass. The growth of grass decreases if cows are present. Cows have a relatively low impact on grass growth.

Farmers want to achieve high growth of grass because their livelihood depends on it. To support the farmers, a smart app is available that categorizes the five factors that influence grass growth for each individual farm into either 'Present' or 'Absent'."

We constructed the materials based on a simplified model of grass growth, outlined in Figure 2. In the model we defined the causal factors as binary features, either present or absent, and the outcome as either high or low. Our decision rules assigned the numerical value of +0.8 to higher-impact positive factors if present (micro-organisms, rain), +0.4 to lower-impact positive factors if present (fertilizer, clover), and -0.4 to lower-impact negative factors if present (cows). The outcome value is determined by the sum of all factors that are present in a case. If the values of the causal factors sum to greater than 1.00, the outcome is designated as high grass growth; if their values sum to lower than 1.00, then the outcome is designated as low grass growth. There are 32 possible combinations of the presence or absence of the five causal factors, and we selected 10 cases to test (see the Appendix), for which a change to one factor would result in a change to the outcome, five with a positive outcome (high grass growth) and five with a negative outcome (low grass growth). We ensured that there were two cases related to each one of the five causal factors (to be referred to in the explanations and queried in the diagnosis tasks). The participants were given the information that a factor had a high impact on the cause (for those factors set at 0.8) or a low impact (for those set at 0.4) and they were told whether the impact was positive (for those factors set at +0.8 or +0.4) or negative (for the factor set at -0.4); they were not told the specific numeric values associated with these terms.

Participants were assigned at random first to the two explanation groups, counterfactual and prefactual, and then subsequently to the control description group. Every participant carried out two sorts of tasks, a set of prediction tasks, and a set of diagnoses tasks. To control for potential order effects, half of the participants, assigned at random, received the prediction tasks first and then the diagnoses tasks, and the other half received them in the opposite order. Within each set, the trials were presented in a different randomized order for each participant. Participants received 20 trials (10 prediction tasks and 10 diagnosis tasks); half had a positive outcome (high grass growth), and the other half had a negative outcome (low grass growth). The same 10 cases were used for prediction and diagnosis. We measured the accuracy of participants' predictions and diagnoses, how confident they were about their judgments, and how helpful they judged the explanations to be. For each trial, the participants were shown a reminder at the top of the screen about the impact of each of the causal factors on the effect.

Participants were not allowed to return to previous screens to change their responses during the experiment. They completed three comprehension questions about the five factors before they moved on to the experimental tasks to ensure they understood the factors and their impact on grass growth. Two attention checks were shown in randomized order during the tasks and a memory check was presented at the end. All of these questions are provided in the Appendix. The final task completed by the participants was the DARPA Explanation Satisfaction Scale [12] (see the Appendix).

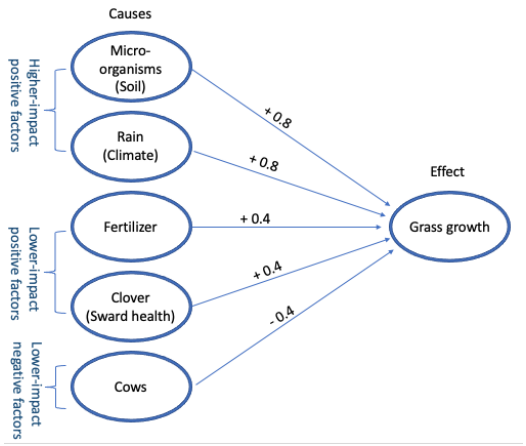

Figure 2: A simple model of grass growth based on five causal factors and an outcome. Each of the causes has an impact on the effect, and the arrows indicate the direction of the causal link, from cause to effect. The numerical values represent the strength of the causal link and in this simplified model the numerical values were set at 0.8 for a high impact and 0.4 for a low impact causal relation. The plus sign denotes a positive causal link (i.e., a cause), the minus sign indicates a negative causal link (i.e., a disabler or inhibitor).

2.1.2. Participants and procedure.

The participants were 193 members of the general public recruited from the platform Prolific. There were 69 men, 116 women, 2 non-binary individuals, and 6 preferred not to say. They were aged between 19 and 76 years with a mean age of 37 years. They were selected to be native English speakers, resident in the US, UK, Ireland, Canada, Australia, or New Zealand, and over the age of 18. To control for expertise, we recruited only participants whose self-reported occupation was not in the field of agriculture. We recruited participants to the two explanation groups first and a sample size of 130 participants was required to detect a medium effect size with 90% power at $p < .05$, according to G*power; we set our stopping rule to 142 participants to allow for failure of attention checks (and one participant had to be eliminated accordingly). We then recruited participants to a control description group, with 65 participants necessary for sufficient power based on the previous test (one participant was eliminated for failure of attention checks). A further 12 participants were eliminated due to an administrative error resulting in the absence of response-recording for one question.

The experiment took approximately 12 minutes to complete, and each participant was paid £1.75 UK sterling for their participation. Participants provided informed consent after reading an information sheet by ticking a box, they were informed that their participation was anonymous, and they could withdraw their data at any time. Ethics approval for this experiment and the next was granted on 20/02/2021 by Trinity College Dublin School of Psychology Research Ethics Committee with the reference code SPREC112020-22. Participants accessed the experiment online and the materials were presented via Alchemer, to which participants were directed by a link from Prolific.

2.1.3 Results and Discussion.

The dataset for all of the experiments is available at https://osf.io/mn7gs/. The data were analyzed using the IBM SPSS Statistics Version 25 Package. First, we established that there were no differences depending on whether participants completed the prediction trials first (N=97) or the diagnosis trials first (N=96), by independent-samples t-tests, for accuracy, t (192) = 0.06, p = .95; confidence judgments, t (192) = 0.42, p = .68, or explanation helpfulness judgments, t (192) = 0.07, p = .94. Hence, order of task had no effect.

Accuracy of judgments. Participants were very accurate in their judgments, with overall 80% correct decisions. They made more correct decisions for diagnoses than predictions, 88% vs 72%, as shown by a main effect of task type, F (1,191) = 167.63, $p < .001$, ηp2 = 0.47, in a 3 (explanation type: counterfactual vs prefactual vs control) x 2 (task type: prediction vs diagnosis) x 5 (causal factor: micro-organisms, rain, fertilizer, clover, cows) analysis of variance (ANOVA) with repeated measures on the second two factors, on the accuracy of participants' decisions, see Figure 3.

Participants also made more correct decisions for some of the causes than for others, as shown by a main effect of causal factor, F (4,188) = 51.37, $p < .001$, ηp2 = 0.21. The two variables of task type and causal factor interacted, F (4,188) = 23.20, $p < .001$, ηp2 = 0.11. The decomposition of the interaction, with paired-samples t-tests and a Bonferroni corrected alpha for 5 tests of $p < .01$, revealed that participants were more accurate for diagnoses than predictions for the low-impact causal factors of fertilizer, 85% vs 61%, t (192) = 7.44, $p < .001$, d = 0.45; clover, 86% vs 60%, t (192) = 8.54, $p < .001$, d = 0.41; and cows, 88% vs 60%, t (192) = 9.52, $p < .001$, d = 0.40; but there were no differences between predictions and diagnoses for the high-impact factors of micro-organisms, 92% vs 88%, t (192) = 1.65, p = .102, and rain, 92% vs 91%, t (192) = 0.25, p = .806.

Participants made the same number of correct decisions when they were given counterfactual explanations (80%), prefactual explanations (80%), and descriptions (81%), as indicated by the lack of a main effect of explanation type, F (2,190) = 0.19, p = .83, ηp2=0.002; and explanation type did not interact with task type, F (2,190) = 0.65, p = .52, ηp2 = 0.007, or causal factor, F (8,184)=0.56, p = .81, ηp2=0.006; and the three variables did not interact, F (8,184) = 0.55, p = .82, ηp2 = 0.006.

Confidence judgments. Consistent with the accuracy results, participants were more confident about diagnoses than predictions for some of the low-impact causal factors, and there were no differences between predictions and diagnoses for some of the high-impact causal factors, as shown by the interaction of task type and causal factor, F (4,188) =8.31, $p < .001$, ηp2 =0.04; there was no main effect of task type, F (1,191) =0.11, p = .74, ηp2 =0.001, and there was a main effect of causal factor, F (4,188) =25.94, p <.001, ηp2 =0.12. None of the other effects were significant; once again, there was no main effect of explanation

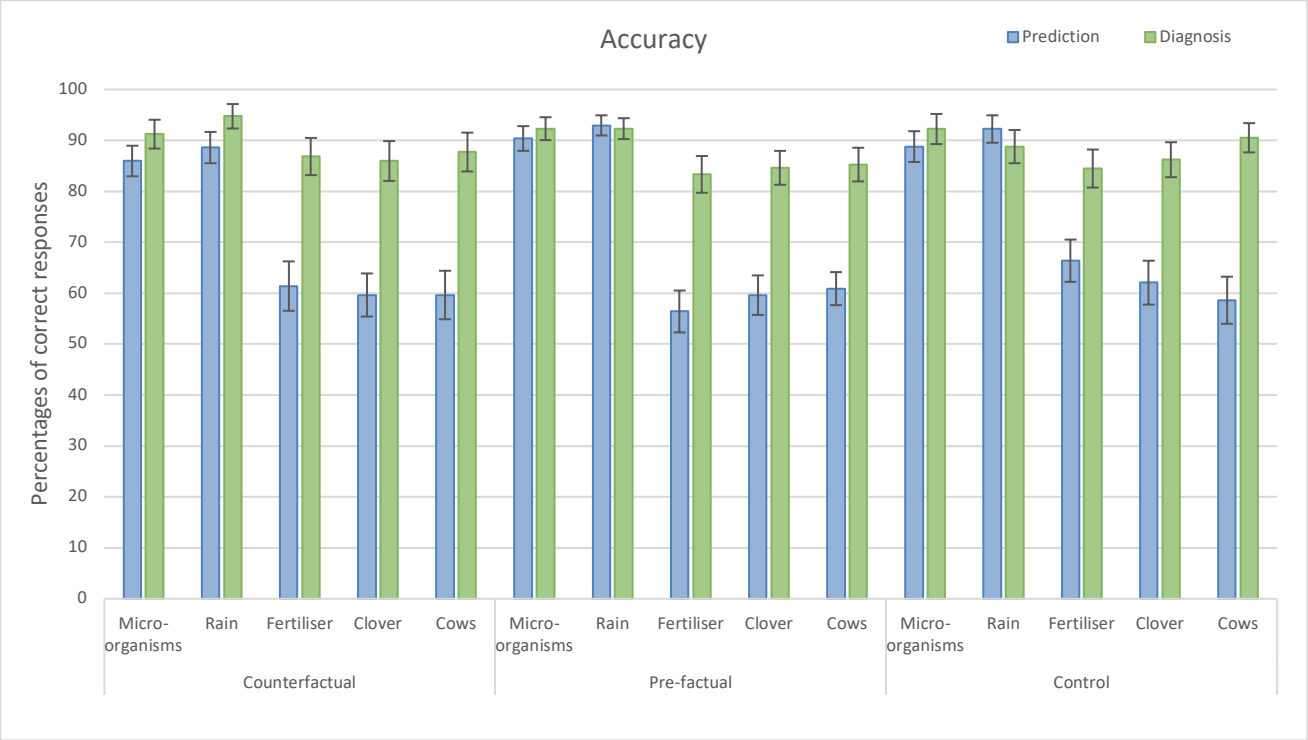

Figure 3: The percentages of correct decisions for predictions and diagnoses in Experiment 1. Participants were given counterfactual explanations, prefactual explanations, or control descriptions, for each of five causal factors in a farming domain. Error bars are standard error of the mean.

type, $F(2,190) = 0.80$, $p = .45$, $\eta p2 = 0.008$, and explanation type did not interact with task type, $F(2,190) = 0.37$, $p = .70$, $\eta p2 = 0.004$; or causal factor, $F(8,184) = 0.90$, $p = .52$, $\eta p2 = 0.009$; and the three variables did not interact, $F(8,184) = 1.22$, $p = .28$, $\eta p2 = 0.01$.

Explanation helpfulness judgments. Consistent with the accuracy and confidence results, participants found the explanations more helpful for diagnoses than predictions for low-impact causal factors, and there was no difference between diagnoses and predictions for high-impact causal factors, as shown by the interaction of task type and causal factor, $F(4,188) = 4.3$, $p = .002$, $\eta p2 = 0.022$; there was a main effect of task type, $F(1,191) = 10.05$, $p = .002$, $\eta p2 = 0.05$, and a main effect of causal factor, $F(4,188) = 14.65$, $p < .001$, $\eta p2 = 0.072$. None of the other effects were significant: once again, there was no main effect of explanation type, $F(2,190) = 2.59$, $p = .08$, $\eta p2 = 0.027$, and explanation type did not interact with task type, $F(2,190) = 0.37$, $p = .69$, $\eta p2 = 0.004$, or causal factor, $F(8,184) = 0.81$, $p = .45$, $\eta p2 = 0.008$; and the three variables did not interact, $F(8,184) = 0.80$, $p = .61$, $\eta p2 = 0.008$.

A one-way ANOVA on the mean scores for the DARPA Explanation Satisfaction Scale comparing counterfactual, prefactual and control descriptions showed no differences between them, $F(2,190) = 0.99$, $p = .37$, $\eta p2 = 0.010$.

The experiment shows that participants made fewer correct decisions for prediction tasks than diagnosis ones, for decisions about low-impact causal factors. Similarly, they were less confident for prediction tasks than diagnosis ones for low-impact causal factors, and they judged explanations to be more helpful for diagnosis tasks than prediction ones for low-impact causal factors.

The counterfactual and prefactual explanations did not have any effect on accuracy, confidence, or explanation helpfulness judgments, compared to control descriptions. One possible reason why the explanations had no effect is that participants found the task relatively easy - their overall accuracy was above 80%, and so they may have had no need for explanations. We addressed this issue in the next experiment.

2.2 Experiment 2

In the second experiment our aim was to compare predictions to diagnoses, and the effects of counterfactual explanations to prefactual ones, in scenarios based not on farming but on an unfamiliar analogous domain about an alien planet. The rationale for this modification was to remove any background knowledge or assumptions participants may have had about the domain, and thus to increase the difficulty of the tasks. Our aim was to ensure that we could test whether participants find counterfactual or prefactual explanations helpful for predictions or diagnoses in unfamiliar situations. Hence, we created scenarios analogous to the farming scenarios of the first experiment, based on information about an alien planet. Such content has been used

previously in experiments to control for the effects of background knowledge on human reasoning [41,42]. We also modified the design of the first experiment to require participants in this experiment to choose between counterfactual and prefactual explanations, to ensure that they compared them explicitly.

2.2.1 Materials and design.

Participants were asked to imagine that they were testing a SmartAgriculture AI decision-support app built for an alien planet called Kronus. This app was proposed to provide information about the presence or absence of five causal factors that affected the growth of an alien plant called Elos on a given alien farmland called a Delc. The five factors mentioned in this domain were: (i) Besloor as a property of the growing medium, (ii) Thardon as a climate substance, (iii) Fropa as a chemical element, (iv) Corrick as a plant species, and (v) Umbrat as an animal. The information was presented in tables akin to those used in Experiment 1. In the prediction tasks, we asked participants to judge what prediction the app would make, e.g., "*What do you think the app will predict about Elos growth on this Delc*?" and we provided them with the answer options "*High/Low*". In the diagnosis tasks, we asked them to judge what diagnosis the app would make, e.g., "*What do you think the app will diagnose about Fropa on this Delc?*" and we provided them with the answer options "*Present/Absent*". Participants also judged how much they agreed with the statement "*I am confident in my judgment of the app's prediction/diagnosis*" using a 5-point scale anchored at strongly disagree (at 1) and strongly agree (at 5).

We then informed them of the app's decision, e.g., *Elos growth was "high"*, and provided two explanations for the decision, a counterfactual explanation based on a past tense subjunctive conditional, e.g., "*If Thardon had been absent last month, your Elos growth would have been low*", and a prefactual explanation based on a subjunctive conditional about the future, e.g., "*If Thardon were to be absent next month, your Elos growth would be low*". The participants were asked to choose which one of the two explanations provided was more helpful in assisting them to understand how the app works. For each task, the participants were shown a reminder at the top of the screen about the impact of each of the factors on the outcome.

We constructed the materials based on an alien scenario adapted from the farming example in Experiment 1 (see Figure 4). The alien model is entirely analogous to the farming one with the five causal factors as binary features (present or absent), the outcome as either high or low, and similar numerical values set for two higher-impact positive factors (Besloor, Thardon), two lower-impact positive factors (Fropa, Corrick), and one lower-impact negative factor (Umbrat).

The information participants were given about scientists working on Planet Kronus was as follows:

"*The scientists observed five factors that have an impact on Elos growth. There are no other observed factors that influence Elos growth. The relations between these five factors are still unknown to the scientific team.*

-Besloor is a property of the medium that Elos is planted in. Agricultural scientists confirmed that the growth of Elos increases if Besloor is present. Besloor has a high impact on Elos growth.

-Thardon is a substance that is produced by the climate of planet Kronus. Meteorologists confirmed that the growth of Elos increases if Thardon is present. Thardon has a high impact on Elos growth.

-Fropa is an element that can be added to Besloor. Chemical scientists confirmed that the growth of Elos increases if Fropa is present. Fropa has a low impact on Elos growth.

-Corrick is often found growing with Elos. Botanists confirmed that the growth of Elos increases if Corrick is present. Corrick has a low impact on Elos growth.

-Umbrat is an animal species that feeds on Elos. Farming experts confirmed that the growth of Elos decreases if Umbrat is present. Umbrat has a low impact on Elos growth.

Aliens on Planet Kronus want to achieve high growth of Elos. To support the aliens, the scientists developed a smart app that categorizes the five factors that influence Elos growth for each individual Delc into either 'Present' or 'Absent'."

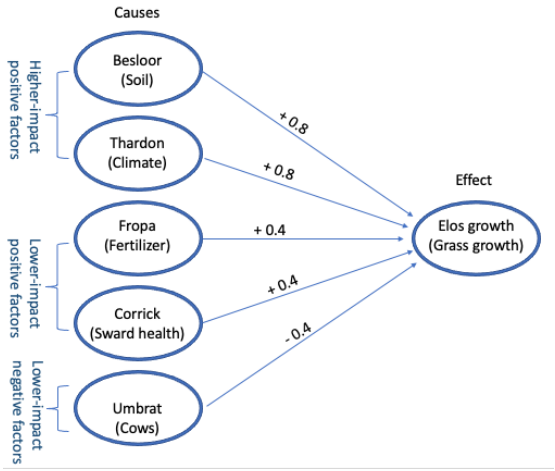

Figure 4: A simplified alien scenario about five causal factors and an outcome. In the diagram, the properties of the causal factors and their relations to the outcome are matched with the grass growth content in Figure 1. The alien names were non-English words with two syllables.

Every participant carried out a set of prediction tasks and a set of diagnoses tasks, as they did in the previous experiment. For each task they were provided with a counterfactual explanation and a prefactual explanation, and they indicated which one they found most helpful. We measured the accuracy of their predictions and diagnoses, their confidence in their judgments, and which of the two explanations they chose as more helpful. Hence the design was a within-participants one.

Participants received 20 trials (10 prediction tasks and 10 diagnosis tasks) based on the same cases used in Experiment 1 (see the Appendix). Once again, to control for potential order effects, half of the participants, assigned at random, received the prediction tasks first and then the diagnoses tasks, and the other

half received them in the opposite order. Within each set, the trials were presented in a different randomized order for each participant. The materials were presented in the same way as in Experiment 1, with similar comprehension questions, and attention and memory check questions (see the Appendix). The final task completed by the participants was the DARPA Explanation Satisfaction Scale [12].

2.2.2 Participants and procedure.

The participants were 50 members of the general public recruited from the platform Prolific. There were 26 men and 24 women. They were aged between 18 and 54 years with a mean age of 31 years. They were selected to be native English speakers, resident in the US, UK, Ireland, Canada, Australia, or New Zealand, over the age of 18. We did not control for participants' self-reported occupation in this experiment. A sample size of 44 participants was required to detect a medium effect size for 90% power at $p < .05$, according to G*power and we set our stopping rule to 50 to allow for failure of attention checks effect size for 90% power at $p < .05$, according to G*power and we set our stopping rule to 50 to allow for failure of attention checks (although no one had to be eliminated accordingly). The experiment took approximately 12 minutes to complete, and each participant was

2.2.3 Results and Discussion.

First, we established that there were no differences depending on whether participants completed the prediction trials first (N=28) or the diagnosis trials first (N=22), by independent-samples t-tests for accuracy, $t(49) = 0.34$, $p = .74$, and confidence judgments, $t(49) = 0.16$, $p = .88$. Hence, order of task had no effect.

Accuracy of judgments. Participants were accurate in their judgments, with overall 73% correct decisions in this unfamiliar domain, somewhat less than the 80% accuracy of decisions found for the familiar farming scenarios in the first experiment. Once again, as in Experiment 1, participants made more correct decisions for diagnoses, 79%, than for predictions, 67%, as shown by a main effect of task type, $F(1,49) = 16.10$, $p < .001$, $\eta p^2 = 0.25$, in a 2 (task type: prediction vs diagnosis) x 5 (causal factor: Besloor, Thardon, Fropa, Corrick, Umbrat) ANOVA with repeated measures on both factors, on accuracy of participants' decisions (see Figure 5).

Participants made more correct decisions for some of the causes than for others, as shown by a main effect of causal factor, $F(4,46) = 15.89$, $p < .001$, $\eta p^2 = 0.25$. The two variables, task type and causal factor, interacted, $F(4,46) = 4.55$, $p = .002$, $\eta p^2 = 0.09$. The decomposition of the interaction, by paired-samples t-tests with a Bonferroni corrected alpha for 5 tests of $p < .01$, revealed that participants tended to be more accurate in diagnoses than predictions for the low-impact causal factors, for Umbrat, 81% vs 53%, $t(49) = 4.88$, $p < .001$, $d = 0.41$; Fropa, 74% vs 57%, $t(49) = 2.45$, $p = .018$, $d = 0.50$; and Corrick, 70% vs 56%, $t(49) = 2.04$, $p = .047$, $d = 0.48$ (although these latter two are not significant on the corrected alpha of $p < .01$); and there were no differences between diagnoses and predictions for the high-impact factors of Besloor, 86% vs 88%, $t(49) = 0.42$, $p = .67$, and Thardon, 82% vs 81%, $t(49) = 0.21$, $p = .84$.

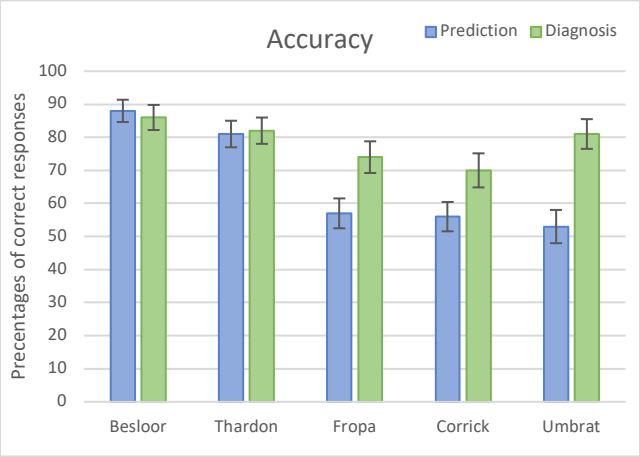

Figure 5: The percentages of correct responses for predictions and diagnoses in Experiment 2. Participants were given five causal factors in an alien planet domain. Error bars are standard error of the mean.

Confidence judgments. Participants were more confident about diagnoses than predictions for some of the low-impact causal factors, but there was no difference between predictions and diagnoses for the high-impact causal factors, as shown by the interaction of task type and causal factor, $F(4,46) = 2.53$, $p = .04$, $\eta p^2 = 0.05$; there was no main effect of task type, $F(1,49) = 2.01$, $p = .16$, $\eta p^2 = 0.04$, but there was a main effect of causal factor, $F(4,46) = 5.74$, $p < .001$, $\eta p^2 = 0.11$.

Explanation choices. Participants chose counterfactual explanations, 56%, as often as prefactual ones, 45%, and there was no difference in their tendency to prefer counterfactuals to prefactuals for prediction or diagnosis tasks, 49% vs 61%, $t(49) = 0.78$, $p = .44$, $d = 0.10$.

The experiment shows that participants made fewer correct decisions for prediction tasks than diagnosis ones, for decisions about low-impact causal factors. Similarly, they were less confident about prediction tasks than diagnosis ones, for some low-impact causal factors. They chose counterfactual explanations as often as prefactual explanations, for predictions and diagnoses. Experiment 2 replicates the findings of Experiment 1 for an unfamiliar domain.

2.3 General Discussion

Participants made more correct inferences for diagnoses than predictions, in the farming domain of Experiment 1, and in the alien planet domain of Experiment 2. The results are consistent with psychological findings that inferences about complex causal relations differ for predictions and diagnoses [31-40]. The discovery has important implications for the use of counterfactuals in XAI. It indicates that the assessment of the impact of explanations on users' accuracy in understanding an AI system's decisions needs to examine diagnosis accuracy, as well as prediction accuracy. In the past, prediction has been the only

task considered in user studies and our results suggest that it would be useful to extend the range of tasks given to users to assess their understanding.

Importantly, the greater accuracy of participants' inferences for diagnoses than predictions occurred only for low-impact causal factors in both domains, and there were no differences for high-impact causal factors. It may be more difficult to work out the effect of a low-impact causal factor for an outcome, compared to a high-impact causal factor, perhaps because the effect of a low-impact causal factor on the outcome may be more dependent on the occurrence of other causal factors as well. This difficulty is greater for prediction where the outcome is unknown, than for diagnosis where the outcome is known. There were no differences in the effects of alternative causes and disablers for predictive and diagnostic tasks. Hence, another important implication of the results for the use of counterfactuals in XAI is that counterfactual explanations that focus on more than one cause merit much closer further examination.

Participants made the same number of correct inferences whether they were given counterfactual explanations or prefactual ones in the first experiment, and they choose counterfactual explanations as often as prefactual explanations as most helpful in the second experiment. The results indicate that counterfactual and prefactual explanations have similar effects on participants' judgments. However, we note that accuracy in both experiments was high, even in the unfamiliar domain, and so it may be the case that participants did not require explanatory assistance to make their judgments.

It is also worth noting that in our experimental tasks, a certain set of inputs led to one specific outcome, which set an upper limit of simplicity on our tasks. In contrast, real-life farming situations, as well as existing machine-learning techniques, contain a higher level of uncertainty. In our experiments, participants may have relied on superficial strategies to make their judgments, due to the simplicity of the tasks. Moreover, the need for an explanation often arises from a gap in one's knowledge and facilitates learning to fill that gap [43]. People also tend to seek explanations when faced with unexpected events [44, 45]. Explanations in our experiments were provided after each task, independent of decision accuracy or user requirement. Future research to examine the relative merits of counterfactual and prefactual explanations for assisting prediction and diagnostic inferences in more difficult and uncertain decision-making contexts in XAI may be helpful. A final observation is that one significant advantage of counterfactual and prefactual explanations is that they provide a blueprint for how to bring about a desired outcome. Accordingly, future studies may benefit from asking participants to evaluate the helpfulness of such explanations, not only for understanding the decisions made by an app, but also for their own decision-making as an app user.

3 CONCLUSIONS

The experiments show that the accuracy of users' judgments about an AI system's decisions is greater when they make diagnostic inferences rather than prediction ones. The current AI literature on explanation methods has neglected the importance of such reasoning task differences. The present results show that the psychological context in which an AI system's decision-making is presented may be as important, if not more so, than the algorithmic details of particular explanatory strategies. Recently, it has been noted that very few of the key algorithmic features of over a hundred different AI counterfactual methods have been corroborated in controlled user studies [6]. Our work shows that experimentally adequate user testing is needed, not merely to corroborate algorithmic proposals, but because it can fundamentally change our conception of how these algorithms might be used in real-life decision scenarios.

ACKNOWLEDGMENTS

The research reported in this paper was funded by Teagasc, the Irish Agriculture and Food Development Authority, and the VistaMilk SFI Research Centre through a Walsh PhD Scholarship.

REFERENCES

[1] Verma, S., Dickerson, J. and Hines, K. 2020. Counterfactual explanations for machine learning: A review. arXiv:2010.10596. Retrieved from https://arxiv.org/abs/2010.10596

[2] Wachter, S., Mittelstadt, B. and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*,31 (2017), 841.

[3] Karimi, A.H., Barthe, G., Schölkopf, B. and Valera, I. 2020. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. arXiv:2010.04050. Retrieved from https://arxiv.org/abs/2010.04050

[4] Byrne, R. M. J. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *IJCAI*, 6276 – 6282.

[5] Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267 (2019), 1-38.

[6] Keane, M. T., Kenny, E. M., Delaney, E. and Smyth, B. 2021. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. In *IJCAI-21*.

[7] Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D. and Lee, S. 2019. Counterfactual visual explanations. In *PMLR*, 2376 - 2384.

[8] Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J. and Doshi-Velez, F. 2019. Human evaluation of models built for interpretability. In *Proceedings of the AAAI conference on Human Computation and Crowdsourcing*, 7, 59-67.

[9] Lim, B. Y., Dey, A. K. and Avrahami, D. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 2119 - 2128.

[10] 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*, 90-98.

[11] van der Waa, J., Nieuwburg, E., Cremers, A. and Neerincx, M. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291 (2021), 103404.

[12] Hoffman, R. R., Mueller, S. T., Klein, G. and Litman, J. 2018. Metrics for explainable AI: Challenges and prospects. arXiv:1812.04608. Retrieved from http://arXiv:1812.04608

[13] Kenny, E. M., Ford, C., Quinn, M. and Keane, M. T. 2021. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, 294 (2021), 103459.

[14] Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J. and Shadbolt, N. 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*, 1-14.

[15] Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. and Dugan, C. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*, 275 - 285.

[16] 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*, 1274.

[17] Russell, C., Mc Grath, R. and Costabello, L. 2020. Learning Relevant Explanations.
Retrieved from http://whi2020.online/static/pdfs/paper_54.pfd

[18] Byrne, R. M. J. and Egan, S. M. 2004. Counterfactual and prefactual conditionals. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 58, 2 (2004), 113-120.

[19] Ferrante, D., Girotto, V., Stragà, M. and Walsh, C. 2013. Improving the past and the future: A temporal asymmetry in hypothetical thinking. *Journal of Experimental Psychology: General*, 142, 1 (2013), 23-27.

[20] Mercier, H., Rolison, J. J., Stragà, M., Ferrante, D., Walsh, C. R. and Girotto, V. 2017. Questioning the preparatory function of counterfactual thinking. *Memory & cognition*, 45, 2 (2017), 261-269.

[21] Epstude, K., Scholl, A. and Roese, N. J. 2016. Prefactual Thoughts: Mental Simulations about What Might Happen. *Review of General Psychology*, 20, 1 (2016), 48-56.

[22] Byrne, R. M. J. 2016. Counterfactual thought. *Annual review of psychology*, 67 (2016), 135-157.

[23] Byrne, R. M. J. and Tasso, A. 1999. Deductive reasoning with factual, possible, and counterfactual conditionals. *Memory & Cognition*, 27, 4 (1999), 726-740.

[24] Gerstenberg, T., Goodman, N. D., Lagnado, D. A. and Tenenbaum, J. B. 2021. A counterfactual simulation model of causal judgments for physical events. *Psychological review*, 128, 5 (2021), 936.

[25] Rim, S. and Summerville, A. 2014. How far to the road not taken? The effect of psychological distance on counterfactual direction. *Personality and Social Psychology Bulletin*, 40, 3 (2014), 391-401.

[26] Markman, K. D., McMullen, M. N. and Elizaga, R. A. 2008. Counterfactual thinking, persistence, and performance: A test of the reflection and evaluation model. *Journal of Experimental Social Psychology*, 44, 2 (2008), 421-428.

[27] Roese, N. J. and Epstude, K. 2017. The Functional Theory of Counterfactual Thinking: New Evidence, New Challenges, New Insights. Academic Press, 1 - 79.

[28] Tversky, A. and Kahneman, D. 2015. Causal schemas in judgments under uncertainty. *Progress in social psychology*, 1 (2015), 49-72.

[29] Fenker, D. B., Waldmann, M. R. and Holyoak, K. J. 2005. Accessing causal relations in semantic memory. *Memory & cognition*, 33, 6 (2005), 1036-1046.

[30] Fernbach, P. M., Darlow, A. and Sloman, S. A. 2011. Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, 140, 2 (2011), 168-185.

[31] Fernbach, P. M. and Rehder, B. 2013. Cognitive shortcuts in causal inference. *Argument & Computation*, 4 (2013), 64-88.

[32] Cummins, D. D. 2014. The impact of disablers on predictive inference. *Journal of experimental psychology: learning, memory, and cognition*, 40, 6 (2014), 1638.

[33] Goldvarg, E., and Johnson-Laird, P. N. 2001. Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive science*, 25 (2001), 4, 565-610.

[34] Johnson-Laird, P. N., and Khemlani, S. 2017. Mental models and causation. In Waldman, M. (ed). Oxford handbook of causal reasoning, Oxford: Oxford University Press.

[35] Frosch, C. A., and Byrne, R. M. J. 2012. Causal conditionals and counterfactuals. *Acta psychologica*, 141, 1 (2012), 54-66.

[36] Byrne, R. M. J. 1989. Suppressing valid inferences with conditionals. *Cognition*, 31, 1 (1989), 61-83.

[37] De Neys, W., Schaeken, W. and D'Ydewalle, G. 2003. Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory & cognition*, *31*, 4 (2003), 581-595.

[38] Espino, O. and Byrne, R. M. J. 2020. The suppression of inferences from counterfactual conditionals. *Cognitive science*, 44, 4 (2020), e12827.

[39] Rehder, B. and Waldmann, M. R. 2017. Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & cognition*, 45, 2 (2017), 245-260.

[40] Khemlani, S. S. and Oppenheimer, D. M. 2011. When one model casts doubt on another: A levels-of-analysis approach to causal discounting. *Psychological bulletin*, 137, 2 (2011), 195.

[41] Markovits, H., Forgues, H. L. and Brunet, M.-L. 2010. Conditional reasoning, frequency of counterexamples, and the effect of response modality. *Memory & Cognition*, 38, 4 (2010), 485-492.

[42] Dias, M. and Harris, P. L. 1990. The influence of the imagination on reasoning by young children. *British Journal of Developmental Psychology*, 8, 4 (1990), 305-318.

[43] Keil, F., Rozenblit, L. and Mills, C. 2004. What lies beneath? Understanding the limits of understanding. Thinking and seeing: Visual metacognition in adults and children. Cambridge, MA: MIT Press, 227-249.

[44] Hilton, D.J., and Slugoski, B.R. 1986. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, 93 (1986), 75–88.

[45] Quinn, M.S. and Keane, M.T. 2022. Factors affecting "expectations of the unexpected": the impact of controllability and valence on unexpected outcomes. *Cognition*, 225(2022), 105142.

A Appendix

A.1 Test cases.

The 32 cases generated from 5 binary factors are shown, with the 10 cases selected to test in the Experiments highlighted in bold. The cases were instantiated with farming content in Experiment 1 and with alien content in Experiment 2. The factor to be diagnosed, and used in explanations, is underlined. The counterfactual and prefactual explanations are provided for the test cases. The information was presented to participants in tabular form, see Figure 1.

Case	Micro-organism/ Besloor (+0.8)	Rain/ Thardon (+0.8)	Fertiliser/ Fropa (+0.4)	Clover/ Corrick (+0.4)	Cows/ Umbrat (-0.4)	Grass growth/ Elos growth	Explanation (Counterfactual first, prefactual second)
1	Present	Present	Present	Present	Present	High (2.0)	-
2	**<u>Present</u>**	**Present**	**Present**	**Absent**	**Present**	**High (1.6)**	If micro-organisms had been absent last month, your grass growth would have been low If micro-organisms were to be absent next month, your grass growth would be low
3	Present	Present	Absent	Present	Present	High (1.6)	-
4	Present	Present	Absent	Absent	Present	High (1.2)	-
5	**Present**	**Absent**	**<u>Present</u>**	**Present**	**Present**	**High (1.2)**	If fertilizer had been absent last month, your grass growth would have been low If fertilizer were to be absent next month, your grass growth would be low
6	**Present**	**Absent**	**Present**	**Absent**	**<u>Present</u>**	**Low (0.8)**	If cows had been absent last month, your grass growth would have been high If cows were to be absent next month, your grass growth would be high
7	Present	Absent	Absent	Present	Present	Low (0.8)	-
8	Present	Absent	Absent	Absent	Present	Low (0.4)	-
9	Absent	Present	Present	Present	Present	High (1.2)	-
10	**Absent**	**Present**	**Present**	**<u>Absent</u>**	**Present**	**Low (0.8)**	If clover had been present last month, your grass growth would have been high If clover were to be present next month, your grass growth would be high
11	Absent	Present	Absent	Present	Present	Low (0.8)	-
12	**<u>Absent</u>**	**Present**	**Absent**	**Absent**	**Present**	**Low (0.4)**	If micro-organisms had been present last month, your grass growth would have been high If micro-organisms were to be present next month, your grass growth would be high
13	Absent	Absent	Present	Present	Present	Low (0.4)	-
14	Absent	Absent	Present	Absent	Present	Low (0)	-
15	Absent	Absent	Absent	Present	Present	Low (0)	-
16	Absent	Absent	Absent	Absent	Present	Low (-0.4)	-
17	Present	Present	Present	Present	Absent	High (2.4)	-
18	Present	Present	Present	Absent	Absent	High (2.0)	-
19	Present	Present	Absent	Present	Absent	High (2.0)	-
20	Present	Present	Absent	Absent	Absent	High (1.6)	-
21	Present	Absent	Present	Present	Absent	High (1.6)	-
22	Present	Absent	Present	Absent	Absent	High (1.2)	-
23	**Present**	**Absent**	**Absent**	**Present**	**<u>Absent</u>**	**High (1.2)**	If cows had been present last month, your grass growth would have been low If cows were to be present next month, your grass growth would be low
24	**Present**	**Absent**	**<u>Absent</u>**	**Absent**	**Absent**	**Low (0.8)**	If fertilizer had been present last month, your grass growth would have been high If fertilizer were to be present next month, your grass growth would be high
25	**Absent**	**<u>Present</u>**	**Present**	**Present**	**Absent**	**High (1.6)**	If rain had been absent last month, your grass growth would have been low If rain were to be absent next month, your grass growth would be low
26	Absent	Present	Present	Absent	Absent	High (1.2)	-
27	**Absent**	**Present**	**Absent**	**<u>Present</u>**	**Absent**	**High (1.2)**	If clover had been absent last month, your grass growth would have been low If clover were to be absent next month, your grass growth would be low
28	Absent	Present	Absent	Absent	Absent	Low (0.8)	-
29	Absent	Absent	Present	Present	Absent	Low (0.8)	-
30	**Absent**	**<u>Absent</u>**	**Present**	**Absent**	**Absent**	**Low (0.4)**	If rain had been present last month, your grass growth would have been high If rain were to be present next month, your grass growth would be high
31	Absent	Absent	Absent	Present	Absent	Low (0.4)	-
32	Absent	Absent	Absent	Absent	Absent	Low (0)	-

A.2 Comprehension questions.

Before beginning the experimental trials, participants were asked 3 comprehension questions. The questions are instantiated with farm content as an example here, and with the correct answers astrixed. If participants selected a wrong answer, they were instructed to please try again (for one more try), when they selected a correct answer, they were told so. Participants were allowed to move on to the experimental trials after completing the comprehension questions regardless of the accuracy of their performance.

1. Which two of the following statements about the grass growth are true?

- Micro-organisms, rain, fertilizer, and clover have a negative effect on grass growth, which means the presence of these factors will lead to a decrease in grass growth.
- Micro-organisms, rain, fertilizer, and clover have a positive effect on grass growth, which means that the presence of these factors will lead to an increase in grass growth. *
- Cows have a positive impact on grass growth, which means that the presence of cows will lead to an increase in grass growth.
- Cows have a negative impact on grass growth, which means that the presence of cows will lead to a decrease in grass growth. *

2. Which three of the following factors have a relatively low impact on grass growth?

Micro-organisms. Rain. Fertiliser* Clover* Cows *

3.Which two of the following factors have a relatively low positive impact on grass growth?

Micro-organisms. Rain. Fertiliser *. Clover * Cows.

A.3 Attention check questions.

There were two attention checks for both experiments, one randomly interspersed with trials in the prediction block, and one randomly interspersed in the diagnoses block. The attention checks were presented in a Table similar to the experimental trials but instead of asking for a prediction or diagnosis, participants were asked about information in the table, such as,

What is the condition of micro-organisms on this farm?

A4. Memory check questions.

The memory check question occurred at the end of all tasks; the correct answers are astrixed here.

Please choose ALL FIVE factors that have an impact on grass growth from the options below:

Fertiliser*. Rain*. Sunlight. Clover*. Sand. Micro-organisms*. Cows*.

A5. DARPA explanation satisfaction questions.

The scale was administered after the memory check question. All responses were recorded with a 5-point scale anchored at (1) Strongly disagree and (5) strongly agree.

1. From the explanation, I understand how the app works.
2. The explanation of how the app works is satisfying.
3. The explanation of how the app works has sufficient detail.
4. This explanation of how the app works seems complete.
5. This explanation of how the app tells me how to use it.
6. This explanation of how the app works is useful to my goals.
7. This explanation of the app shows me how accurate the app is.
8. This explanation lets me judge when I should trust and not trust the app.