

**How People Understand Causal and Counterfactual Explanations in the
Context of eXplainable Artificial Intelligence (XAI)**

Xinyue Dai

Thesis submitted for the degree of Doctor of Philosophy

Trinity College Dublin, the University of Dublin

2025

This research was funded by Teagasc, the Irish Agriculture and Food
Development Authority, and the VistaMilk SFI Research Centre through a
Walsh PhD Scholarship

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open-access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).



Signed: _____

Date: March 28th, 2025

Summary

In this thesis, we present the results of 13 experiments ($n = 2,756$) investigating the cognitive processes underlying causal explanation evaluation and the role of counterfactual explanations during explanation selection and risky decision-making. The experiments examined key questions in cognitive psychology relevant to understanding explanations, especially counterfactual explanations. They also explored ways to test and generate better explanations of potential relevance for the field of eXplainable Artificial Intelligence (XAI). We address three questions: What information influences people's selection of causal explanations? What are the underlying cognitive processes when people read counterfactual explanations about how things could have turned out differently? What types of tasks and explanations are helpful in people's understanding and satisfaction with automated decisions?

The first six experiments (Experiments 1, 2a, 2b, 3a, 3b and 4) investigated a novel strategy people follow when choosing between competing causal explanations. The results showed that people preferred explanations with consequent-controllable elements, such as “the symptoms are caused by a disease that can be treated with medication”, than explanations without consequent-controllable elements, such as “the symptoms are caused by a disease that has no known treatment”. These experiments also showed simplicity had a stronger impact on causal explanation selection, as we found that participants preferred simple explanations without consequent-controllable elements over complex explanations with consequent-controllable elements. Moreover, we showed that counterfactual explanations can reduce people's strong preference for the simple explanation. These findings contributed to the

debate about what guides the cognitive processes underlying causal explanation selection. The results also corroborated the mental model theory of counterfactual thinking.

The second series of two experiments (Experiments 5 and 6) showed that in a mock AI decision support app setting, counterfactual explanations (about how things could have turned out differently in the past) and prefactual explanations (about how things could turn out differently in the future) were equally effective in helping human users understand the app. They were also equally effective in terms of user satisfaction with the explanations. The results also showed that there was a difference in the objective measurement of accuracy when participants were asked to make predictive inferences (from causes to effects) compared to diagnostic inferences (from effects to causes). These findings added to the current methods of user testing and showed the importance of testing different explanation types and task types in XAI.

The third series of five experiments (Experiments 7 to 11) showed that for risky decisions (in the risky choice framing effect scenario), participants switched from the typical decision to a non-typical one proposed by an Artificial Intelligence (AI) system when they were given a counterfactual explanation. Their baseline tendency to switch was low when the non-typical decision was proposed by an AI system compared to a human expert. The results corroborated previous findings on algorithmic aversion. They also supported the mental model theory of counterfactual thinking, as participants were able to consider alternatives when given counterfactual explanations.

All 13 experiments further contribute to our understanding of the cognitive processes underlying explanation evaluation and the proposition that

counterfactual explanations prompt people to simulate alternative realities. The results have important implications for the psychological understanding of causal explanation and counterfactual thinking. The implications also extend to the current user testing methods and assumption of what makes a good explanation in XAI.

Acknowledgements

My first and deepest thanks go to my academic supervisor, Professor Ruth Byrne, for setting a role model of a rigorous and responsible researcher for me as a psychology student; for her patience, guidance, and understanding, which allowed my growth; and for reminding me to recentre and rise from falls along this very long journey. I am grateful for every meeting, conversation and rehearsal I had with her, as I always felt supported, motivated, and reassured after them. I would also like to thank my co-supervisor, Professor Mark Keane, for his ideas and insights as a multi-disciplinary scientist, which brought more depth to my research projects. Thanks also to my collaborators at VistaMilk, Dr Laurence Shalloo and Dr Elodie Ruelle, for their support and helpful discussion at different stages of my PhD.

I am grateful for all the support I have received from staff members at Trinity College Dublin, such as my thesis committee members, Professor Redmond O'Connell and Professor Sven Vanneste. I also thank everyone I have met at Teagasc, for the Walsh Scholarship funding opportunity and for giving me a platform to communicate my research with a broader audience. I thank the conference committees of AIES 2022, ESCOP 2023 and ICT 2024, for helping me with travel arrangements and giving me the chance to showcase my research.

I would not have made it this far without people in my life who have faith in me, even during moments of darkness. I am grateful to my parents, Xiaohua Dai and Yanhong Chen, for encouraging and supporting me to explore my potential and pursue a degree overseas, and for providing a safe shelter throughout my years away from home. I am grateful to my grandparents and

List of Publications

Dai, X., Keane, M. T., Shalloo, L., Ruelle, E., & Byrne, R. M. J. (2022).

Counterfactual explanations for prediction and diagnosis in XAI. In
*Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and
Society*, 215-226. <https://doi.org/10.1145/3514094.3534144>

Dai, X., Keane, M. T., Shalloo, L., Ruelle, E., & Byrne, R. M. J. (2025). *People
change their decision when an AI system contradicts it with a
counterfactual explanation* [Manuscript submitted for publication].

School of Psychology, Trinity College Dublin.

Table of Contents

Chapter 1: Introduction	14
Psychology of Explanations	18
Psychology of Counterfactual Thinking	39
Counterfactual Explanations in eXplainable AI (XAI)	57
Summary of Aims	64
Chapter 2: Cognitive Heuristics in Causal Explanation Selection.....	67
Actionability in Counterfactual XAI Explanations	68
Controllability as Heuristics.....	72
Simplicity Preference in Explanation Evaluation	74
Overview of Experiments	77
Experiment 1	77
Experiment 2a	91
Experiment 2b	103
Experiment 3a	115
Experiment 3b	123
Experiment 4	130
General Discussion.....	137
Chapter 3: Counterfactual Explanations for Prediction and Diagnosis.....	149
Counterfactuals and Prefactuals.....	150
Predictions and Diagnoses	153
User Testing of Counterfactual Explanations.....	156
Overview of Experiments	158

Experiment 5	159
Experiment 6	179
General Discussion	191
Chapter 4: AI Recommendations in the Risky Choice Framing Effect	196
Framing Effects.....	197
Counterfactual Simulation	205
AI & Human Expert Recommendations	206
Overview of Experiments	210
Experiment 7	212
Experiment 8.....	218
Experiment 9.....	224
Experiment 10.....	230
Experiment 11	235
General Discussion	240
Chapter 5: Discussion	246
Explanation Evaluation.....	248
Counterfactual Explanations.....	255
XAI Explanations	263
Limitations and Future Directions	269
Conclusion	273
References	275
Appendices	322

List of Figures

Figure 2.1. Diagrams for the Structure of Consequent-Controllability and Antecedent-Controllability.....	71
Figure 2.2. Diagrams for the Causal Structure in Experiment 1	81
Figure 2.3. The Percentages of Choices of all Explanations in Experiment 1 for the Consequent-Controllability Information Condition (top) and for the No Extra Information Condition (bottom)	89
Figure 2.4. Diagrams for the Causal Structure in Experiment 2a	93
Figure 2.5. The Percentages of Choices of all Explanations in Experiment 2a for the Antecedent-Controllability Information Condition (top) and for the No Extra Information Condition (bottom)	101
Figure 2.6. Diagrams for the Causal Structure in Experiment 2b	105
Figure 2.7. The Percentages of Choices of All Explanations in Experiment 2b for the Antecedent-Controllability Information Condition (top) and the No Extra Information Condition (bottom)	111
Figure 2.8. Diagrams of the Causal Structure in Experiment 3a.....	118
Figure 2.9. Diagrams for the Causal Structure in Experiment 3b.....	125
Figure 3.1. Examples of the Experimental Tasks in Experiment 5.....	150
Figure 3.2. Example of an Experimental Task in the Diagnosis X Prefactual Explanation Condition.....	161
Figure 3.3 A Simple Model of Grass Growth Based on Five Causes and an Outcome in Experiment 5.....	167
Figure 3.4.1. The Percentages of Correct Decisions for Predictions and Diagnoses in Experiment 5.....	171

Figure 3.4.2. The Confidence Ratings for Predictions and Diagnoses in Experiment 5.....	173
Figure 3.4.3. The Explanation Helpfulness Ratings for Predictions and Diagnoses in Experiment 5.....	176
Figure 3.4.4. The Mean Ratings for the DARPA Scale in Experiment 5.....	178
Figure 3.5. Examples of the Experimental Tasks in Experiment 6.....	182
Figure 3.6. A Simplified Farming Scenario About Five Causes and an Outcome in Experiment 6.....	185
Figure 3.7.1. The Percentages of Correct Decisions for Predictions and Diagnoses in Experiment 6.....	187
Figure 3.7.2. The Confidence Ratings for Predictions and Diagnoses in Experiment 6.....	189
Figure 3.7.3. The Ratings for the DARPA Scale in Experiment 6.....	190
Figure 4.1. The Percentages of Choices of Non-Typical Programs in Experiment 7.....	217
Figure 4.2 The Percentages of Choices of Non-Typical Programs in Experiment 8.....	223
Figure 4.3. An Example of the Experimental Task in Experiment 9.....	227
Figure 4.4. Percentages of Decision Change in Experiment 9.....	229
Figure 4.5. Percentages of Decision Change in Experiment 10.....	234
Figure 4.6. Percentages of Decision Change in Experiment 11.....	238
Figure 5.1. Examples of Experimental tasks in Chapter 3.....	266

List of Tables

Table 2.1. Frequencies of Explanation Preferences in Experiment 1.....	86
Table 2.2. Frequencies of Explanation Preferences in Experiment 2a.....	98
Table 2.3 Frequencies of Explanation Preferences in Experiment 2b.....	109
Table 2.4. Frequencies of Explanation Preferences in Experiment 3a.....	122
Table 2.5. Frequencies of Explanation Preferences in Experiment 3b.....	128
Table 2.6. Frequencies of Explanation Preferences in Experiment 4	135
Table 3.1. DARPA Explanation Satisfaction Scale.....	163

Chapter 1: Introduction

Explanations are a crucial part of people's daily cognitive processes. They play a fundamental role in people's attempts to understand the world. How explanations are evaluated and understood is a question that leads to theoretical debate and empirical investigations by philosophers, psychologists, and cognitive scientists. This thesis firstly aims to contribute to a better understanding of how explanations are evaluated, with an emphasis on counterfactual explanations: explanations based on thoughts about how things could have turned out differently.

The 13 experiments reported in this thesis address specific questions about people's evaluation of explanations. Imagine you are a farmer trying to find a satisfying explanation for the problems on your farm. Would you prefer an explanation that indicates specific consequent-controllable elements to fix the problems, such as changing the amount of fertiliser you use, compared to explanations that indicate no such elements? Our first series of six experiments in Chapter 2 addresses this question. Now imagine you are using a grass growth management app, and the app tells you that your grass growth on your farm is less than expected. Would you find a counterfactual explanation such as "If fertiliser had been present, your grass growth would have been higher in the past" more helpful in your understanding of the app's decision and your satisfaction with its explanations, compared to a prefactual explanation such as "If fertiliser were to be present, your grass growth would be high in the future"? Our second series of two experiments in Chapter 3 examines this question. And now suppose you are expecting a disease to kill 60 of your cows, and you are faced with two programs to combat this disease, one safer and one

more risky. Suppose you opt for the safe program. Would you change your decision if an AI decision support system told you to choose the risky one? Would you be convinced more if the system explained that in previous similar cases, the outcome would have been better if the risky program had been chosen? Our third series of five experiments in Chapter 4 considers this question. By answering these questions, this thesis aims to provide empirical evidence to contribute to our understanding of the cognitive processes underlying explanation evaluation. The experiments that tested counterfactual explanations aim to contribute to our understanding of the role of counterfactuals in mental simulation. The thesis also aims to test what kinds of explanations help people understand Artificial Intelligence (AI) decisions better. It demonstrates how psychological tests help develop solutions for eXplainable AI (XAI), as there is a growing interest in counterfactual explanations in XAI (e.g., Artelt et al., 2021; Verma, Dickerson & Hines, 2020; Wachter, Mittelstadt & Russell, 2018).

The first objective of this thesis concerns explanation evaluation. People are faced with competing explanations frequently, and sometimes we prefer explanations with certain qualities, such as simplicity, coherence and breadth (Thagard, 1989). Empirical studies have demonstrated that sometimes, people prefer simpler explanations to complex ones, but this preference can be reduced by providing probability information (Lombrozo, 2007). Chapter 2 examines a new heuristic in explanation evaluation as we test whether people prefer explanations that contain information on consequent-controllable elements for a problem. We also test if they prefer a complex explanation when it is described as complex but consequent-controllable more than an

explanation that is simple but consequent-uncontrollable. For example, for problems on a farm, such as when grass growth is low and weed growth is high, people tend to prefer a simpler explanation in which the problems are caused by one factor (e.g., a micro-organism *Excie* in the soil leads to grass growth being low and weed growth being high) than the combination of two separate factors (e.g., micro-organism *Pacis* leads to grass growth being low and micro-organism *Rirtan* leads to weed growth being high). We investigate whether people still prefer the simpler explanation when they are told there is no remedy for *Excie*, while *Pacis* and *Rirtan* can be remedied with fertiliser and herbicide.

The second objective of this thesis concerns the effects and underlying cognitive processes of counterfactual explanations. Counterfactual explanations entail dual possibilities and enable people to make inferences that are otherwise difficult to make (Byrne, 2016; Byrne & Tasso, 1999; Moreno-Ríos, Garcia-Madruga & Byrne, 2008; Santamaría, Espino & Byrne, 2005). As a result, we tested counterfactual explanations in situations where people usually commit to one choice. Chapter 2 investigates if counterfactuals could reduce people's preference for simpler explanations and prompt them to consider alternative complex explanations. Chapter 3 examines people's preference between different sorts of explanations, i.e., counterfactual explanations about how things could have turned out differently in the past, such as "If fertiliser had been present, your grass growth would have been higher in the past", prefactual explanations about how things could turn out differently in the future, such as "If fertiliser were to be present, your grass growth would be high in the future"? and factual causal explanations. We

examined the role of these different sorts of explanations in terms of AI decision understanding and explanation satisfaction. Chapter 4 examines the impact of explanations on people's decision change in a risky situation. We tested their effects when people were committed to one decision and were told by an AI system to change their decision, for example, to change from the safe option which will save 20 cows to the risky option which has a 1/3rd chance to save all 60 cows but a 2/3rds chance to save none of them. We found that having an explanation of the AI system's suggestion led to more decision changes compared to no explanation, and especially, AI counterfactual explanations helped people switch their decisions even more.

The third objective concerns how cognitive experiments can inform XAI research about what sorts of explanations human users find cognitively compelling and useful. We aimed to contribute to the current understanding of counterfactual explanations in XAI. Chapter 2 explores the role of consequent-controllability information on causal explanation evaluation inspired by the interest in "actionable" explanations in XAI. Through identifying a "consequent-controllability effect", we showed that XAI explanations should take into consideration consequent-controllability information but also take into account the presence of other stronger factors such as simplicity. Chapter 3 demonstrates that counterfactual and prefactual explanations lead to the same level of understanding of the AI system and that user testing should consider using not only prediction tasks (reasoning from causes to outcomes) but also diagnosis tasks (reasoning from outcomes to causes). Chapter 4 demonstrates the impact of information sources on people's decision-making in a risky situation, and we show that people are more willing to follow a human expert

than an AI decision-support system. Using counterfactual explanations for an AI system's decisions can improve user acceptance.

In the following sections of this chapter, we first discuss findings from the psychological literature about how explanations are generated and evaluated. We then consider psychological discoveries about counterfactual thinking and counterfactual explanations. We review theories of counterfactual thinking, how counterfactuals are generated, and the effects and functions of counterfactual explanations. The third section discusses the application of counterfactual explanations in the field of XAI. We review user studies on the impact of counterfactual explanations and some current issues and challenges in their application.

Psychology of Explanations

In the following section, we consider psychological findings about explanations, with a focus on cognitive processes underlying explanation evaluation. We also review psychological findings about causal explanations and how people understand causality.

Explanation Selection

In this section, we consider evidence on explanation selection, particularly the role of simplicity in explanation evaluation, given its relevance to the series of experiments in Chapter 2.

People's choice to generate an explanation is tightly related to prior beliefs, interests, and goals (Lombrozo, 2006; 2012), and people tend to seek explanations when faced with abnormal and unexpected events (Hilton &

Slugoski, 1986). In everyday life, people constantly choose between competing explanations, a process referred to as “inference to the best explanation” (Harman, 1965; Lipton, 1990). People have strong intuitions about when to seek explanations and what accounts for an adequate explanation (Lombrozo, 2012), and people are often guided by certain criteria when they select an explanation to be *the* explanation (Miller, 2019; Thagard, 1989). Research in philosophy and cognitive sciences suggests that people’s evaluation of explanations is partly guided by a series of explanatory qualities, such as simplicity, coherence, and breadth (Glymour, 2015; Liefgreen & Lagnado, 2023; Lipton, 2017; Mackonis, 2013; Thagard, 1989; Zemla et al., 2017). These qualities shape the way people make inferences and the way people learn, helping people learn about simpler rules and broader mechanisms. However, they can also lead to systematic errors, for example, people systematically over-report the prevalence of simple explanations when asked to recall the data they observed earlier (Lombrozo, 2016). The section below introduces findings about these explanatory qualities, as well as the effect of explanatory depth illusion.

Coherence. Explanation coherence is sometimes referred to as the consistency with one’s prior beliefs, background knowledge and theories of the world (Mackonis, 2013; Murphy & Medin, 1985). Other times, coherence is used to describe the consistency within an explanation, in that different elements serve as positive or negative constraints of each other (Thagard, 1989, 2000; Thagard & Verbeurgt, 1998). People usually choose an explanation that is most coherent with the target event (Read & Marcus-Newhall, 1993). Apart from scientific explanations, even in naturalistic everyday explanation settings,

internal coherence was also found to be a good predictor for explanation quality (Zemla et al., 2017). Explanatory coherence has been found to influence plausibility judgements as people assign a higher probability to facts that are more consistent with each other. To give an example, people were presented with a pair of premise and conclusion statements which were consistent (with the same underlying reason), such as “many ex-cons are hired as bodyguards” as the premise and “many war veterans are hired as bodyguards” as the conclusion. In this case, the underlying reason was that both ex-cons and war veterans tended to be tough and experienced fighters, and participants gave positive ratings of the plausibility of the conclusion given the premise. When participants were presented with a pair of premise and conclusion statements that had different reasons, such as “many ex-cons are unemployed” as the premise and “many war veterans are unemployed” as the conclusion, they rated the plausibility as negative because the underlying reasons were different for these statements: ex-cons are often considered to be not trustworthy, while war veterans sometimes suffer from trauma and are outcasts of society (Sloman, 1994). However, people might violate coherence in two ways. Sometimes, they only know fragments of a full system which could hardly qualify as coherent, and sometimes, people have contradictory sets of beliefs (Keil, 2006).

Simplicity. In philosophy, the term Occam’s razor describes the simplicity principle in problem-solving: we search for explanations with the smallest possible set of elements, and an explanation should not be overly complex beyond necessity (Baker, 2013). An explanation with a simpler inner structure and less predictive power might be preferred to an alternative with

stronger predictive power yet overwhelmingly complex and unrelated elements (Harman, 1965). Some experiments have demonstrated people's tendency to prefer simpler explanations, for example, when participants were asked to explain a woman being tired, having gained weight, and having an upset stomach, they preferred the simpler explanation of her being pregnant instead of the complex explanation that the woman had mono, had stopped exercising, and had a stomach virus (Read & Marcus-Newhall, 1993).

An important question to address is why people prefer simpler explanations. We mentioned earlier that people assign a higher probability to coherent premises and conclusions (Sloman, 1994). Similarly, some research focuses on the potential of simplicity as a cue for probability. For example, participants were tested with two competing explanations for a patient's symptoms, where the symptoms can be caused by one disease (the simple explanation) or by two diseases (the complex explanation). Probability was found to be the driving factor for participants' preferences, not simplicity, in a within-participant design, which possibly heightened participants' sensitivity to probability information as participants received situations where the simpler explanation was more probable, where the simpler explanation was equally probable as the complex one, and where the simpler explanation was less probable (Lagnado, 1994, as cited in Lombrozo, 2007). However, when simplicity was pitted against probability in a between-participant design, simplicity was shown as an important quality for causal explanation selection. It was found that when no probability information was given, 96% of participants chose the simple explanation over the other ones. When probability information was provided, participants preferred the simpler

explanation when it was more likely and equally likely compared to the complex alternative. Only when the frequency information suggested that the simpler cause was much less likely than the complex one (with a 1:10 ratio in a population of 750) did people's preference for the simpler explanation diminish (Lombrozo, 2007). Hence, people may rely on simplicity as a proxy for probability information. In a more recent study where participants were asked to draw out causal models in the criminal justice domain, it was found that when asked to draw the model, participants engaged more with probabilities, which led to a preference for the simpler explanation compared to when they were asked to give descriptions (Liefgreen & Lagnado, 2023). Bayesian Occam's razor (BOR) suggests people prefer simple explanations because Bayesian inferences penalise complex hypotheses as they contain more free parameters, thus having more "flexibility" (Sober, 2015). Experiments investigated whether people penalise more flexible hypotheses by assigning less probability to them. Participants were given some observations about two imaginary alien species and the frequency of them having red spots on their backs. One species was described to show the same frequency of having red spots and the other species was described to have different flexibility levels of having red spots; flexibility was manipulated by stating "always have two red spots on their back", "can have anywhere between 1 and 4 red spots on their back and each of these possibilities is equally likely", or "can have anywhere between 1 and 100 spots on their back". Participants were given two hypotheses and asked to choose one that was more likely, one that would be a better explanation for the observations of an alien family and their red spots. As the flexibility increased for the species, the more complex hypothesis became

increasingly penalized as participants' judgement of likelihood decreased (Blanchard, Lombrozo & Nichols, 2018a).

People's preference for simpler explanations might also arise from the tendency to simulate fewer possibilities and from an attempt to reduce cognitive effort. Studies of heuristics have proposed simple strategies as mental shortcuts, the rule of thumb, "methods for arriving at satisfactory solutions with modest amounts of computation" (Simon, 1990a, p.11), and "efficient cognitive processes, conscious or unconscious, that ignore part of the information" (Gigerenzer & Gaissmaier, 2011, p. 451). People have limited cognitive resources as their working memory capacity is restricted (Miller, 1956), and the idea of *bounded rationality* (e.g., Jones, 1999; Simon, 1990b) has been suggested to describe how people reason within the constraints imposed by their cognitive resources. The effort reduction framework was proposed to explicate the generalisation of heuristics across different domains, and some rules of this framework include integrating less information and examining fewer alternatives (Shah & Oppenheimer, 2008). From the heuristics point of view, one can argue that the preference for simpler explanations results from a tendency to reduce the amount of information in the causal links between the events, thus ignoring the alternative explanations to minimise cognitive effort. Another important element of heuristics is the idea of an "adaptive toolkit": the collection of heuristics that individuals have at their disposal and the questions of when and where to use these heuristics (Gigerenzer, Hertwig & Pachur, 2011; Raab & Gigerenzer, 2015). To account for how people negotiate the trade-off between prior probability (favouring simple explanations) and goodness-of-fit (favouring complex explanations),

opponent heuristics were proposed to explain the conflicting data. People have a stronger simplicity preference, for example, in physical systems when compared to social causal systems (Johnson, Valenti & Keil, 2019). In Chapter 2 of this thesis, we first identify a new heuristic, the consequent-controllability effect, during the process of causal explanation selection. We then compare the consequent-controllability heuristic with simplicity to investigate their interactions. Finally, we test whether we can reduce the preference for simpler explanations by helping people generate more mental simulations through counterfactual explanations.

Another important question is how to define simplicity. The studies mentioned above conveyed simplicity by changing the number of causes involved in an explanation. In contrast, other studies investigated what makes an explanation simple by differentiating node simplicity and root simplicity. Imagine a person who is extremely fatigued and has been losing weight. We know that this person is suffering from insomnia (which leads to fatigue) and decreased appetite (which leads to weight loss). Two possible explanations are: this person has depression, which leads to both insomnia and decreased appetite, or this person does not have depression, and their insomnia and decreased appetite arose independently. Node simplicity refers to the number of causes involved in an explanation, so if people's simplicity judgement is guided by node simplicity, they would prefer the explanation where insomnia and decreased appetite arose independently as it contains two elements not three. Root simplicity, on the other hand, refers to the number of unexplained causes in an explanation, and would lead to a preference towards the depression explanation, since that explanation assumes the person to be

depressed (a complete choice) instead of assuming the person to have both insomnia and decreased appetite (a proximal choice). Imaginary disease scenarios were used, and participants were asked to choose the more satisfying explanation between a complete choice, a proximal choice, and an unknown condition. Results challenged the node simplicity prediction as participants were observed to prefer the complete choice to different degrees depending on the causal structure of events, which extended to experimental conditions where probability information was involved (Pacer & Lombrozo, 2017).

Simpler explanations have been found to be more satisfying in different settings (Blanchard et al., 2018a; Bonawitz & Lombrozo, 2012; Johnson et al., 2019; Pacer & Lombrozo, 2017; Vratsidis & Lombrozo, 2022; Walker, Bonawitz & Lombrozo, 2017), and simplicity has been suggested as a guiding principle for various cognitive processes (Chater & Vitányi, 2003). However, evidence from naturalistic explanations shows that people's preference for simpler explanations is subject to the context (such as scientific versus naturalistic settings) and is not always dominant over other alternatives. In a Reddit-based survey on naturalistic explanations, it was found that people preferred more complex explanations that entailed more causal mechanisms (Zemla et al., 2017, 2023). When prompted to explain inconsistencies, people also tend to add causes rather than dispute a premise, resulting in a more complex causal explanation (Johnson-Laird, Girotto & Legrenzi, 2004; Khemlani & Johnson-Laird, 2011). Given the different findings regarding explanation simplicity, a unifying account was suggested, where the complexity of explanations should be matched with the events in question, in other words, individuals should prefer simpler explanations for simpler events

and complex explanations for complex events. Participants were asked to read scenarios and predict the complexity of satisfying explanations, generate an explanation themselves, and evaluate explanations; results indicated a match of complexity (Lim & Oppenheimer, 2020).

Breadth. Breadth, sometimes also referred to as scope, is the ability of an explanation to generalise and unify. It informs people's judgement of causal strength as people assign higher causal strength to explanations that predict more effects, thus indicating that people use explanatory power (in this case scope) as a heuristic for estimating causal strength when no statistical evidence is available (Johnson et al., 2014). Closely related to the idea of breadth is *stability*, breadth was suggested to be a virtue under the heading of stability in that a generalization which holds in a wider range of circumstances is more stable. For example, to explain Mary having gene g and bungee-jumping behaviour, an explanation that states "gene g causes risk-taking behaviour" is broader and more stable than the explanation "gene g causes bungee-jumping" (Blanchard, Vasilyeva & Lombrozo, 2018b).

Last but not least, the trade-off between an explanation's breadth and its predictive power is commonly discussed: as theories gain in explanatory breadth, their predictive power is reduced. However, it was suggested that the trade-off is not as clean-cut, and it is necessary to consider auxiliary assumptions, i.e., background knowledge which is the "gap between the non-observational terms in theories and observational terms specified by empirical hypotheses" (Trafimow & Uhalt, 2015, p. 83). For example, Halley was able to predict the exact year of "Halley's comet" using Newton's theory, with the auxiliary assumption of the comet's previous location. This example shows

how a theory can hold both breadth and predictive power. While findings indicate that people prefer explanations that account for more evidence, it remains unclear whether this preference derives from a preference for breadth as such, or from probabilistic inference (Lombrozo, 2016).

Illusion of explanatory depth. When evaluating one's own explanation, people also suffer from the "illusion of explanatory depth", where people feel they understand complex phenomena with greater accuracy and depth than they actually do. For example, when asked to rate self-confidence in how the toilet flush works, participants gave high ratings of understanding, and this rating significantly dropped after the same participants were asked to write down an explanation. This rating also further dropped after participants saw an expert explanation of the flushing system (Rozenblit & Keil, 2002). Later on, it was suggested that this illusion may be questioned as participants misconstrued how to initially rate their responses: when participants were directed to consider the mechanical process of objects rather than their functions, participants were able to provide a more accurate estimate of their understanding (Silk-Eglit & Kurtz, 2011). To account for the illusion, it was suggested that people adopt inappropriate construal styles when they give an estimate of understanding: when participants adopt a more concrete construal style from the initial abstract one, the illusion diminishes (Alter, Oppenheimer & Zemla, 2010).

To conclude, people generally show some preferences while evaluating explanations such as simplicity. However, these preferences are usually tested for scientific explanations and are subject to factors such as the influence of the context. For the three qualities we discussed above, probability judgements

have been suggested to be mediating people's preferences. The qualities have also been considered to be a heuristic that people follow. Hence, there is a debate about what makes these explanations universally more satisfying: are they linked with higher probability, or are they cognitively less effortful? These findings motivate our research question in Chapter 2, where we propose a new "consequent-controllability" heuristic according to which people choose explanations based on whether they can act on them, and we tested how people choose explanations based on consequent-controllability information compared to simplicity. We discuss research evidence that motivated our interest in this effect later in the section on counterfactual generation. The next section discusses causality and causal explanations.

Causal Explanation

In this section, we consider alternative interpretations of the cognitive processes underlying causal explanation and evidence on some of the differences between causal diagnosis and prediction, relevant to the experiments in Chapter 3. Not all explanations focus on causal relations, yet many everyday explanations involve cause and effect. Causality plays a central role in learning and reasoning (Chater & Oaksford, 2006). When causal elements co-exist with non-causal ones in an explanation, the causal ones tend to play a dominant role and influence people's judgment (Murphy & Medin, 1985). To answer the question of what causation is and how we make sense of the world causally, theorists have debated between a mechanistic and a probabilistic view: Should causal reasoning prioritise understanding the causal mechanisms or rely on statistical probabilities?

The mechanism account of causal reasoning takes a descriptive approach towards how people reason about causal relations and identify causes. Mechanisms propose the transmission of causal power, or force, as the process that forms the concept of cause and the assumption of mechanism as:

When people conceive of cause-effect relations, there is a basic assumption of mechanism; the assumption that underlying two causally linked events, there is a system of connected parts that operate or interact to make or force an outcome to occur (Ahn & Kalish, 2000, p. 4).

It has also been shown that causal mechanisms outweigh covariation information during causal reasoning. To give some examples, a series of studies with participants between three years and adulthood exhibited the tendency to interpret physical causation primarily with generative transmission rather than covariation, similarity, or temporal continuity. The wind problem in this series of experiments used a setup with a lit candle and two electric blowers positioned on either side of the candle. The candle was also shielded by a three-sided piece of glass that rotated during the experiment. One blower was activated, and the second one was activated following a brief interval. Although the second blower was activated more recently, both adults and children attributed the cause of the candle's extinguishing to the first blower, which indicated that the transmission of wind was used to attribute causality instead of temporal continuity (Shultz, 1982). Participants judged the interpretation of a constantly present factor as the cause to be more likely than the interpretation of a perfect positive covariate as the cause, and this effect persisted even when prior beliefs were controlled (White, 1995). When

participants were asked to indicate what sort of information they would need for causal attribution, they tended to seek out evidence on the underlying mechanisms; what is more, when they received evidence matching in covariation values but differing in the statement of mechanism, those containing mechanism had a stronger influence. For example, participants read about a fictitious disease and received two pieces of evidence: evidence with mechanism information described how a specific virus led to the disease by attacking certain cells; evidence without mechanism information showed a high statistical correlation between exposure to a particular environment and the incidence of the disease. Although both pieces of evidence presented similar covariation data, participants were more influenced by the evidence that included mechanism information (Ahn et al., 1995). The search for underlying mechanisms has also been found to explain cognitive effects, such as the conjunction fallacy (i.e., conjunctive explanations being rated more probable than their components). The conjunction effect occurred with mechanism-based explanations, but not with covariation-based ones. In the mechanism-based conditions, participants were given scenarios where a cause was explained using a known mechanism, such as “K had a traffic accident because she tends to not wear her glasses while driving while she is nearsighted”. In the covariation-based condition, participants read “K had a traffic accident because K is more likely to have traffic accidents than other people are”. Participants were asked to rate the probability of single causes and conjunctive explanations, and it was found that people judged the conjunctive explanation as more probable than its individual components in the mechanism-based condition, and this effect did not occur in the covariation-based condition.

(Ahn & Bailenson, 1996). People's beliefs also play an important part. These beliefs can be revised; for example, when there are contradictions, people generate explanations that revise their causal model. Participants were presented with pairs of causal statements sharing a common cause but leading to different effects, for example, jogging regularly causes an increase in fitness level, and jogging regularly causes weight loss. Initially, participants answered questions about the likelihood of one effect occurring given the cause: "What is the probability that jogging regularly increases fitness level?". Then they were provided with information contradicting the second causal statement such as "Sam jogged regularly but did not lose weight". After the contradiction, they reassessed the probability of the first effect (Walsh & Sloman, 2004). It has been suggested that when people think causally, they make errors in judgements of probability, and the process of causal thinking is beyond patterns of dependency. It is a narrative that unfolds over time, and causal reasoning includes mechanisms and narratives that go beyond probability; for example, the asymmetry of causal reasoning between predictive and diagnostic inferences indicates that people run mental simulations that depend on more information than mere probabilistic dependency (Sloman & Lagnado, 2015).

On the other hand, probabilistic accounts of causality are rooted in the nature of judgements under uncertainty (Pearl, 2009). The probabilistic account focuses on statistical dependency and the relative importance of various causal variables (Spirtes, Glymour & Scheines, 1991). In this account, people are naïve data scientists, building causal structures of the world from limited and sparse data. It was found that even young children between two and four use specialised cognitive systems to draw causal maps of the world, consistent with

Bayes nets. Children were presented with a novel device, “the blicket detector,” which lights up and plays music when certain objects are placed upon it. They observed the experimenter placing different blocks on the detector to see which ones activated the detector and then were given the opportunity to select and place blocks on the detector themselves to test their hypotheses about which blocks were “blickets” (Gopnik et al., 2004). Computational models such as the abstract causal theory suggest that domain-general statistical inference and domain-specific prior knowledge form the process of causal induction (Griffiths & Tenenbaum, 2009).

Causal models and networks contain equations used to model causes and effects using Bayesian networks, which represent interrelated conditional probabilities (Halpern & Pearl, 2005; Pearl, 2010). Causal models describe causal reasoning as a structured and mathematical process. In these models, nodes represent variables, and arrows represent a direct relation. Routes from one variable to another can be indirect, with interventions such as the removal of a node from its link (e.g., Pearl, 2009, 2011). Causal model theories have been shown to provide a good quantitative account of how causal knowledge influences categorisation. Participants were given fictitious categories including biological kinds, non-living natural kinds and artefacts. For the category of Lake Victoria Shrimp, for example, four binary features were provided to define the category, such as high amounts of ACh neurotransmitter, long-lasting flight response, accelerated sleep cycle, and high body weight. Features within each category were interconnected through causal relationships; for example, a high quantity of the ACh neurotransmitter causes the other three features, and it was found that participants’ categorisation

decisions were influenced by their causal models of category features (Rehder, 2003a). Causal models also incorporate counterfactual aspects through the form of counterfactual interventions (Pearl, 2000). Counterfactual interventions involve manipulating a variable while holding the structural relationships constant, which allows us to reason about alternative scenarios and see what would have happened if a variable had been different. These counterfactual interventions identify and evaluate the causal relations of the event. However, there is a limitation which lies in the reasoning direction: intervention is only applicable in reasoning from causes to effects. Backtracking counterfactuals, which involve reasoning from effects to causes, cannot proceed by intervention, for otherwise, they would be equivalent to their consequences (Dehghani, Iliev & Kaufmann, 2012). It was proposed to accommodate backtracking inferences through the differences between counterfactual interventions and observations. Counterfactual interventions involve actively altering a variable to observe changes in the outcome; they involve modifying the causal structure leading to changes in downstream effects. Counterfactual observations consider hypothetical scenarios where a variable is observed differently without direct intervention; they involve adjustments in upstream variables to maintain consistency with the change. This model enables backtracking by considering how changes in the outcomes can imply alternations in prior events (Lucas & Kemp, 2015).

Between the dichotomy of mechanistic and probabilistic views, other approaches incorporate both views or dispute some of their elements. For example, the dichotomy was argued to have overlooked the nature of the probabilistic implication of mechanisms (Glymour & Cheng, 1999).

Probabilistic theories have been suggested to be lacking in the relationship between causality and mechanisms, and an epistemic theory of causality is needed (Williamson, 2009).

There is also a distinction between mechanism and probability in causal explanations. Different approaches have been proposed to understand and explain the behaviour of systems, such as physical (i.e., mechanical), design (i.e., teleological), or intentional, termed “explanatory stance”. The physical stance explains behaviour using the law of physics and chemistry; for example, if a person falls, the explanation could involve muscle failure. The design stance explains behaviour in terms of purpose, function and design; for example, if a person falls, the explanation could involve the functions of human legs to stabilise, but they fail under certain conditions. The intentional stance explains behaviour with beliefs, desires and rational intentions; for example, if a person falls, the explanation involves them being distracted and not paying attention (Dennett, 1987). As the examples show, the same event can be characterised by each of the stances and yield quite different explanations (Keil, 2006). Causal explanations are diverse and can encompass both deterministic and probabilistic forms. Deterministic explanations imply a direct relationship between cause and effect, whereas probabilistic explanations acknowledge that causes influence the likelihood of an effect rather than guaranteeing it. It was argued that people across their lifespan engage with both types of explanations and apply them across various domains to support reasoning and learning (Keil, 2024).

Another prominent theory is the mental model theory, which describes how people naturally reason through simulations. The mental models theory

proposes that people represent possibilities. It postulates that people construct models that simulate the situations described by assertions. A causal assertion indicates a cause-effect relationship between two events. When individuals encounter a causal assertion, they construct a mental model that represents the causal link, enabling them to reason about potential antecedents or outcomes. For example, when people encounter the causal assertion “Pressing the switch causes the light to turn on”, people generate a mental model representation of an initial state of the switch unpressed and the lights being off. Then, the action of pressing the switch is incorporated, and the resulting state is updated to reflect the outcome of the action based on the causal assertion. With the mental models, people can make predictive and diagnostic reasoning, as well as counterfactual reasoning such as: “If I hadn’t pressed the switch, the light would still be off” (e.g., Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). According to this theory, people represent only the truth by default, but false possibilities can be prompted by counterfactuals, as the example mentioned above, people represent the truth of the switch being pressed and the light being switched on, counterfactuals such as “If I hadn’t pressed the switch, the light would still be off” suggest false possibilities. The counterfactual simulation then leads to different inferences. The mental model theory suggests three systems: system 0 contains an initial representation of the premise’s meaning, system 1 builds an initial model of the premise and draws conclusions based on heuristics, and system 2 searches for alternative models to serve as a counterexample to test whether the putative conclusion is false (Khemlani & Johnson-Laird, 2013). When people construct models about possibilities, they envisage causal relations by highlighting only the salient

cases initially, leading to a lack of differences between different sorts of causes, such as causes and enablers, until the possibilities are thought about explicitly. Causes are necessary conditions for target actions, enablers do not initiate the action but allow an action to occur. To give an example, pressing the switch causes the light to be switched on, and the presence of electricity enables the light to be switched on. According to the mental model theory, both are represented through mental simulations, but the distinction lies in the nature of the representations: causes are represented as a direct sequence, while enablers are represented as conditions that make it possible for the target event to occur. However, individuals often fail to distinguish them, as the mental models can appear similar, especially when only the primary possibilities are considered. The differences become evident when fully explicit models are constructed (Johnson-Laird & Khemlani, 2017).

Research focusing on causal reasoning is rich and burgeoning. It was found that beyond the simple cause-effect link, people make inferences about the normality of a cause as well as the causal structure of the phenomenon in question. Participants were presented with visual diagrams and descriptions illustrating a billiard table with two balls and blockers. In the conjunctive causal structure, both balls need to pass through blockers to cause an event, and in the disjunctive scenario, either ball passing through is sufficient to cause the event. It was found that participants' inferences were influenced by the interaction between causal structure and event normality, for example, participants tended to cite abnormal causes in conjunctive structure and normal causes in disjunctive structure (Kirfel, Icard & Gerstenberg, 2022). It was found that pair-wise correlations between features impact the categorising of

features, for example, when comparing a common-cause schema where one disease leads to three symptoms to a common-effect schema where three causes lead to the same disease, the three symptoms were rated to be correlated across the population. In contrast, the three causes of the disease were rated to be uncorrelated (Rehder, 2003b).

People also make different sorts of inferences about causal relations when they reason from a cause to infer its outcome (prediction), compared to when they reason from an outcome to infer its cause (diagnosis). For simple cause-effect links, they make predictions more readily than diagnoses. For example, 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 (Tversky & Kahneman, 1980). 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 (Fenker, Waldmann & Holyoak, 2005).

But in more complex cause-effect cases, where several causes lead to an effect, participants tend to make more correct diagnoses than predictions. When people make predictions from a cause to an effect, they tend to erroneously ignore alternative causes. For example, when people assess the likelihood that a person who smokes will develop lung cancer, they often focus on the link between smoking and lung cancer, neglecting other potential causes. When they make diagnoses from an effect to a cause, they tend to accurately consider the alternative causes, for example, when evaluating the likelihood that a person diagnosed with lung cancer is a smoker, people are

more likely to consider alternative causes of lung cancer (Cummins, 2014; Fernbach, Darlow & Sloman, 2011; Fernbach & Rehder, 2013). This asymmetry indicated that during prediction, there may be an overestimation of the causal strength, and that the broader consideration during diagnosis leads to a more nuanced assessment of the probability. This asymmetry also indicates that people make different sorts of inferences about the different sorts of causal relations when they make predictions compared to when they make diagnoses.

To conclude, research evidence has investigated how people build and understand causality; some suggest that causal inferences rely on causal mechanisms, while others suggest that people rely on probability and correlation to infer causality. There is also an important role for prior beliefs and the specific domain. Different models have been proposed to illustrate causal reasoning, such as Bayesian causal models and the mental model theory. There is also rich evidence on what sort of inferences people make, and the different inferences and cognitive processes between predictive and diagnostic inferences motivated our research question in Chapter 3, where we tested participants' understanding of an app with both predictive and diagnostic tasks.

When constructing explanations of “why X”, people often refer to the contrastive side of the question of “why not Y”. When we think of a causal link “A caused B”, it is implied that if A had not occurred, B would not have happened – a counterfactual conditional. This sort of “what if” thinking allows us to highlight specific factors' effects by comparing what happened to hypothetical alternatives. In the next section, we discuss psychological evidence on counterfactual thinking and its explanatory implications.

Psychology of Counterfactual Thinking

In this section we consider research on counterfactual explanations, of relevance to the comparison of counterfactual and causal explanations in Chapter 2, the comparison of counterfactual and prefactual explanations in Chapter 3, and the examination of the role of counterfactual explanations in leading people to switch to a non-typical decision proposed by an AI system in Chapter 4. The idea of contrastive explanations and counterfactuals can be traced back to early philosophical theories. In modern times, philosophers proposed the idea of “possible-world semantics” to develop the concept of comparative similarity (Lewis, 1973). Psychologists have focused on counterfactuals about how the past could have turned out differently. Intrigued by “if only” or “what if” thoughts, researchers have directed attention towards the determinants of this kind of thinking, as well as its functional and psychological consequences.

When people make judgements about an event, apart from how easily similar cases come to their mind, they also often rely on constructing a mental simulation of a temporally extended scenario to explore how an event can be “undone” (Kahneman & Tversky, 1982a). The norm theory elaborates on the normality-restoring quality, which means people tend to undo abnormal events (Kahneman & Miller, 1986). When comparing events, people set specific reference points from the alternatives that are constructed mentally on the fly. During this process, certain rules govern which aspects people mutate. The norm theory also elaborates that counterfactual thinking has consequences for people’s emotional experiences and judgements. According to the emotional amplification hypothesis, there are different affective directions of

counterfactual thinking: upward (better than reality) and downward (worse than reality) (Kahneman & Miller, 1986). More recently, studies have shown a convergent picture of how a limited set of possible actions come to people's minds, that people focus on alternatives that are valuable and likely to occur (Phillips, Morris & Cushman, 2019). People generate counterfactuals that are probable to them and can distinguish those with a higher likelihood (De Brigard, Szpunar & Schacter, 2013; Petrocelli, Seta & Seta, 2013).

Norm theory has also suggested explanations for various cognitive effects, such as the action effect, an effect in which negative outcomes elicit higher levels of regret as a result of action compared to inaction (Kahneman & Tversky, 1982b). Norm theory explains this effect by arguing that inaction is more normal (Kahneman & Tversky, 1982b; Kahneman & Miller, 1986). Replication studies found that different categories of normality, such as social normality, have distinct effects on regret (Feldman, 2020; Feldman & Albarracín, 2017). Social norms have been suggested to play an important role in incentives and prosocial behaviour (Graf et al., 2023). When it comes to the temporal changes of norms, participants were tested with the prisoner's dilemma games. Participants engaged in interactions with artificial co-players, and for each round, they were asked to choose to either cooperate or defect. Depending on the experimental conditions, the co-players predominantly chose to cooperate or defect. It was found that participants' perception of acceptable behaviour adapted quickly to the immediate behaviour of the co-players, but individual beliefs about acceptable behaviour evolved more slowly (Batzke & Ernst, 2024). Counterfactual thinking is a key principle within the norm theory, as deviation from norms influences people's mental representations of what

could have been. In the following sections, we discuss findings on how counterfactuals are generated, the underlying cognitive processes, as well as their functions.

Cognitive Processes of Counterfactuals

Counterfactual thinking is about generating possibilities. In the section below, we discuss what cognitive processes underlie the generation of counterfactuals, what sorts of inferences people make when they hear counterfactuals, and what kinds of counterfactuals people generate as default.

Dual possibilities. What do people keep in mind when they think about counterfactuals? The mechanisms underlying counterfactual reasoning maintain and update two representations (Byrne, 2016). The mental model theory proposes that counterfactual thinking calls for the simultaneous generation of an imagined world and the real world. Upon the generation of the dual possibilities, people maintain and update both of them. For example, when people hear a counterfactual statement saying, “If there had been roses in the shop, there would have been lilies”, the theory proposes that they mentally represent the counterfactual conjecture, “There were roses and there were lilies” and they also mentally represent the presupposed or known facts, “there were no roses and no lilies” (Byrne & Johnson-Laird, 2020).

Some evidence for this proposal is that when people hear the counterfactual they read “There were roses and there were lilies” as fast as “there were no roses and no lilies”, it was suggested that people keep different sorts of possibilities in mind when they read different conditionals such as counterfactual “if” and semi-factual “even if” conditionals, through measuring

the reading time of conjunctions (Santamaría et al., 2005). Later the different possibilities were extended to implicit propositions (Macbeth & Razumiejczyk, 2019).

More recently, eye-tracking experiments tested affirmative and negative counterfactuals, as well as causal assertions. It was found that participants looked at words corresponding to the conjecture and presupposed facts when they heard counterfactuals, and this effect was more pronounced in negative counterfactuals; while for causal assertions, participants looked at words corresponding to the facts only (Orenes, Espino & Byrne, 2022). The result suggests that people think about just a single possibility initially for a causal assertion. The time people spend to read the conjunction has also been used to measure if they have already envisaged the possibilities, and it was found that when people heard the conditional “if there were apples there were bananas”, they became primed to read some conjunctions very quickly such as “there were no apples and there were bananas” (Espino, Santamaría & Byrne, 2009). Such findings were in line with the mental model theories of counterfactual thinking, that people have thought about certain possibilities when they hear conditionals (Byrne, 2023b). The construction of mental representations of possible scenarios is also used when people interpret conditionals and disjunctions (Byrne & Johnson-Laird, 2020). However, a hybrid of theory-based and simulation-based models of reasoning has been suggested: participants were asked to read aloud regret-oriented counterfactual scenarios. More theorising was observed for uncontrollable scenarios and from the observer’s perspective, while more simulation was observed for controllable

scenarios, indicating a context-sensitive nature (Wilkinson, Ball & Cooper, 2010).

When “dual possibilities” are generated, inferences are made about these alternatives. When hearing the counterfactual “if there had been roses, there would have been lilies”, people believe it implies there were no roses and no lilies (Thompson & Byrne, 2002). Also, people make inferences from counterfactuals that they otherwise find difficult from ordinary conditionals. For example, they make the modus tollens inference from “There were no lilies” to “Therefore there were no roses” (Byrne & Tasso, 1999). They also make different inferences from “if only” counterfactuals compared to “even if” semi-factual, such as, “Even if there had been roses, there still would have been lilies” (e.g., Moreno-Rios et al. 2008). Functional magnetic resonance imaging (fMRI) studies showed that when people process counterfactual sentences in either visual or auditory forms, the right occipital cortex and right basal ganglia are activated, indicating that counterfactuals require a mental representation of both factual and supposed information (Kulakova et al., 2013). More recently, it was found beliefs in additional conditions have an impact on counterfactual inferences: when a counterfactual contains additional information, inferences such as modus tollens are suppressed (Espino & Byrne, 2020). However, when people read a prefactual conditional of how things could turn out differently in the future, such as “if I were to win the lottery next year I would buy a yacht”, they make different inferences than they do from a counterfactual, such as “if I had won the lottery last year, I would have bought a yacht”, which suggests they maintain a single possibility in mind for prefactuals, whereas they maintain two possibilities for counterfactuals (Byrne & Egan, 2004). We will

examine further the idea that people think explicitly about alternative possibilities when they are given a counterfactual explanation, in each of the series of experiments in Chapters 2, 3 and 4.

Counterfactual heuristics. During the generation of counterfactual thoughts, people change specific aspects of their mental simulations (Kahneman & Miller, 1986). A lot of the time, these changes are relatively minor ones that are systematically constrained (Roese, 1997). These systematic patterns of change selection are the result of counterfactual heuristics - people zoom in on pivotal joints to mutate (Kahneman & Tversky, 1982a). The section below discusses some of the heuristics people follow when generating counterfactuals.

Exceptions. Just as the norm theory indicates, counterfactuals are norm-restoring (Kahneman & Miller 1986). People choose exceptional events to be modified. For example, suppose Mr. Murphy left his office at his usual time but decided to drive home by an unusual route to enjoy the view and was killed by a truck that crashed into his car at an intersection. Most people create counterfactuals that change the route Mr. Murphy took on that day. Suppose instead that the accident happened when Mr. Murphy left the office earlier and was driving on his usual route, then most people create counterfactuals that change the time he left the office (Kahneman & Tversky, 1982a). The relation is bidirectional: when told the protagonist has generated counterfactual thoughts, it was more likely for participants to infer that the protagonist has experienced an exceptional event (Xie & Beck, 2022). However, the correspondence between the events and outcomes has been shown to impact the exceptionality heuristic: when participants were asked to undo the outcome

by changing preceding events, they altered exceptional events to undo exceptional outcomes; to undo normal outcomes, they altered events to be of greater exceptionality. For example, participants were presented with scenarios involving a student taking an exam, in the scenarios with exceptional outcomes, an outstanding student unexpectedly fails an exam, and in the unexceptional scenario, the same outstanding student passes the exam. After reading the scenarios, participants were asked to generate counterfactual thoughts. When faced with the exceptional outcome, participants' counterfactual thoughts focused more on exceptional antecedents, and with the unexceptional outcome, their counterfactual thoughts centred on normal antecedents (Gavanski & Wells, 1989). The exceptionality heuristic can also be mediated by the potential outcomes: individuals are guided by the desirability of the imagined outcome. When an exceptional action leads to a negative outcome, people often generate counterfactuals that align with unexceptional behaviour if it suggests a better result; if an alternative exceptional action could have led to a more favourable outcome, then individuals tend to imagine that alternative even if it's exceptional (Dixon & Byrne, 2011). In a more recent meta-analysis, results from 35 experimental studies that tested the exceptionality heuristic showed a medium to strong effect, and it was suggested that the effect was stronger when the routine was aligned with the status quo option and when the routine was aligned with action rather than inaction (Fillon, Kutscher & Feldman, 2021).

Controllability. People choose factors that are controllable to be modified: participants preferred to change the event corresponding to a voluntary action no matter its order in the causal chain or its normality.

Participants were given the scenario of Mr Bianchi going home after work only to find his wife dying due to a heart attack. Participants had to list the importance of counterfactuals to undo the outcome and no matter the order of events, or the exceptionality of the events, they chose to undo the event within the Mr Bianchi's control (going to a bar for a drink) rather than the uncontrollable events such as the manoeuvres of a lorry, the passage of a flock of sheep, and an impeding tree trunk (Giroto, Legrenzi & Rizzo, 1991). This controllability heuristic was later shown to be influenced by the role (actor vs reader): participants were assigned to an actor or a reader condition and were given the scenario of a game where a protagonist needed to win by completing one of two mathematical calculations in two envelopes, one easier and one harder. When people read about how the protagonist chose an envelope which contained a harder calculation and lost the game, they generated counterfactuals by modifying the protagonist's decision to select the other envelope, in other words, they chose the controllable decision. In contrast, actors who experienced the negative outcome (i.e., choosing the envelope with the harder calculation and hence losing the game) chose to modify the features of the problem, such as the time limit, in other words, they modified uncontrollable aspects (Giroto et al., 2007). In the same line of work regarding perspectives, it was later found that people imagine different alternatives when they observe versus when they read - the readers focused more on the controllable choice and the observers focused on uncontrollable factors such as alternative ways to solve the chosen task (Pighin et al., 2011). More theorising was observed during uncontrollable scenarios and more simulation was observed during controllable scenarios, and that other-perspective reasoning

involved more theorising than self-perspective reasoning (Wilkinson et al., 2010). When appropriateness comes into the picture, however, it was found that people preferred to mutate an inappropriate controllable event more than an appropriate controllable event; and that people preferred to mutate an inappropriate controllable event no matter if the outcome is normal, exceptional, positive, or negative. For example, a person decides to stop at a bar for a drink. As a result, the person arrives home late to find that a loved one has passed away. Alternatively, the person is delayed due to an unexpected traffic jam, leading to the same outcome. After reading each scenario, participants were asked to generate counterfactual thoughts, and they were more inclined to generate counterfactual thoughts about not going to the bar. Inappropriate actions (going for a drink in the bar) were more likely to be undone than appropriate ones (visiting their parents) (McCloy & Byrne, 2000). When asked to write a diary entry about an imagined negative set of events, people created more counterfactual thoughts about controllable than uncontrollable events and created an equal number of causal explanations about both, indicating that causal explanations focus on general causal relations while counterfactual explanations focus on specific antecedents (McEleney & Byrne, 2006).

In depression studies, it was found that participants with mild-to-medium depressive symptoms generated counterfactuals with more controllable than uncontrollable aspects, while those with severe symptoms generated counterfactuals that were less controllable (Markman & Miller, 2006). When people think about what could have happened in the past, they focus on uncontrollable features (such as the time limit in the game), but when

they construct prefactuals about the future, they focus on controllable features (such as a strategy to concentrate more) (Ferrante et al., 2013). It was found that unless explicitly told to do so, individuals generate relatively few counterfactuals with controllable features. Participants were asked to complete a computer-based word-search puzzle under a timer, and their goal was to find as many words as possible. They had only 15 seconds to complete the task, which guaranteed that most would fail to reach the desired number of words. It was found that participants generated few controllable counterfactuals after the task unless when they were asked to generate prefactuals or counterfactuals that could be useful for their future selves or others (Mercier et al., 2017).

Controllability has also been tested against causality. People tend to focus on enablers rather than causes in their counterfactual thoughts because they have different mental representations of causes and enablers (Byrne, 2005; Goldvarg & Johnson-Laird, 2001). According to the mental model theory, people just keep one possibility in their mind for causes, but they keep the mental simulation of both the presence of the enabler as well as the absence of the enabler. In a research example, participants were given scenarios where two antecedent events (a cause and an enabler) led to a negative outcome. For example, Richard is writing up his thesis and cannot be bothered to spend some time to back it up to the server (enabler). When he checks his emails, he opens a file containing a virus, which renders all files on his computer useless (cause). The controllability of enablers and causes were manipulated, and participants were asked whether each of the two antecedent events caused, enabled or had nothing to do with the outcome. It was found that when only the cause or the enabler is controllable, participants undid the controllable aspect

more often, but when matched in terms of controllability, participants undid the enablers more (Frosch, Egan & Hancock, 2015). Additionally, controllability has also been linked to episodic counterfactual thinking (the process of mentally simulating alternative versions of past personal experiences), and it was suggested that counterfactuals more often focus on an internally controllable action to the extent that the circumstance is self-initiated rather than others-initiated (Roese, Smallman & Epstude, 2017). In studies of trauma, it was found that people's counterfactual thinking usually focuses on their own controllable actions rather than the uncontrollable behaviour of others (Davis et al., 1995; Davis & Lehman, 1995).

Actions. People prefer to choose action rather than inaction during counterfactual generation. Imagine Lisa, who has shares in company A, considers switching to company B and decides to do so. She loses 1,000 dollars as a result. Imagine also Jennifer, who has shares in company B, considers switching to company A, but decides to stay in company B. She also loses 1,000 dollars. People judged that the former protagonist who acted was feeling worse (Kahneman & Tversky, 1982b). However, this preference can be reversed when people are prompted to think in the long term. People report that actions will lead to worse feelings in the short term, while inactions lead to worse feelings in the long term (Gilovich & Medvec, 1995). It was also found that more intense regret was obtained for action than inaction only when the information about both targets was presented, and when both targets were not related to the negative outcome, no emotional amplification was found for action, and inaction was mutated more (N'gbala & Branscombe, 1997). This effect of action was suggested to result from how actions are mentally

represented in that actions may be represented more explicitly, hence making them more mutable; inactions are only mutable from the long-term perspective when their counterfactual consequences are possibly better than the real consequences and when the consequences for the action are the same (Byrne & McEleney, 2000). More recently, the temporal patterns of action-inaction were replicated with weaker effects, and it was found that stronger responsibility was linked with action compared to inaction both in the short-term and the long-term (Yeung & Feldman, 2022).

Recency. When faced with a series of independent events in a temporal chain, people choose to mutate the most recent event. In a coin-tossing game, where two players need to toss the same side up to win, participants readily choose to change the toss of the second player (Miller & Gunasegaram, 1990). This preference can be manipulated when the context provides an alternative: when the first player had side A up, but the game restarted due to technical issues, ending in the first player tossing side B and the second player tossing side A, people choose to mutate the toss of the first player after the break— an event earlier in time (Byrne et al, 2000).

To conclude, the heuristics discussed above may reflect “fault lines” that guide mutability when people generate counterfactuals. Still, these effects have been found to be content-dependent and even reversible when other factors were added to the scenario. In Chapter 2, we consider some of the possible implications of a controllability heuristic, as found in counterfactual generation, for causal explanation. We also consider the goal of “actionability” in XAI, that AI-generated explanations should provide useful, actionable insights that help users understand and effectively respond to an AI system’s

decisions. The actionability goal enables users to take meaningful actions based on the explanation. In Chapter 2, we operationalised controllability in two ways: consequent-controllability indicated the steps one can act on to remedy the effect, for example, that a micro-organism can be treated with fertiliser; antecedent-controllability indicated the steps one can stop doing to stop the cause from occurring, for example, that a micro-organism is the result of herbicide misuse. We tested whether explanations with consequent-controllable and antecedent-controllable elements yield stronger explanation preferences.

More recently, experiments explored the perceived similarity between the imagined and the actual world and its role in judgements of counterfactual plausibility. In one set of experiments, participants were given statements that violated physical or biological laws such as “Fire that freezes when touched” and “A kitten that hatches from an egg”. Participants were asked to imagine a counterfactual alternative where these statements were true. Then, they were asked to rate how their imagined world is similar to the real world and rate how plausible the statement could have been true in the real world, and a strong correlation was found (De Brigard, Henne & Stanley, 2021). The next section discusses the relations between counterfactuals and causality. We then discuss theories of counterfactual thinking and the functions of counterfactual thinking.

Counterfactuals and Causality

Causality and counterfactuals are closely related, and causes may be identified by the differences they make, differences that must be one from what would have happened without it (Lewis, 1973). Counterfactual thoughts amplify causal judgements, we often turn to counterfactual thoughts to make

causal relations more salient (Roese, 1997). Causal explanations are also typically contrastive, as they indicate conditions which differentiate what is being explained from a counterfactual condition where the contrast holds (Lombrozo, 2012), and this contrast sometimes highlights causal relations. Counterfactual thoughts posit events with minor perturbations in our mental simulations for us to better understand causal relations: by imagining a cause without the effect, or vice versa, people amplify or reduce judgments of causality (Kahneman & Tversky, 1982a). In an example of blame assignment of a rape, participants read the description of the event and were asked to generate different types of counterfactuals. It was found that counterfactuals where changes to the victim's behaviour produced a new outcome led to the highest blame on the victim and the lowest blame on the rapist. Counterfactuals where changes in the victim's behaviour did not change the outcome, resulted in the highest rapist blame and the least victim blame (Branscombe et al., 1996). When comparing counterfactual to semi-factual "even if" thoughts, it was found that counterfactual thoughts led people to judge the event to be more causally related to the outcome, whereas semi-factual thoughts led people to judge the antecedent event to be less causally related to the outcome (McCloy & Byrne, 2002).

The counterfactual theories of causation posit that people's understanding of causality is closely related to counterfactual reasoning (Collins, Hall & Paul, 2004; Kment, 2020; Menzies & Beebee, 2001; Quillien & Lucas, 2023). In the causal case of over-determination, where multiple agents contribute to one outcome, the role of pivotality (whether an agent's action was decisive in changing the outcome) and criticality (the perceived

importance of an agent's contribution before any action is taken) were tested. Participants were assigned to groups and engaged in activities requiring cooperation, the success or failure of the team depended on combinations of individual performances. Participants assessed responsibility based on whether altering a member's action in a counterfactual scenario would change the outcome. They attributed higher responsibility to team members whose actions were pivotal to the outcome. Even before the actions were taken, those who were perceived as critical to the team's success were assigned greater responsibility (Lagnado, Gerstenberg & Zultan, 2013). In the case of causal suppression, "causal superseding" describes situations where norm violations by one agent can influence the causal attribution of other agents' actions, thus diminishing the perceived causal impact of other agents involved in the same outcome. Imagine Bill's wife, Sue, sees a left-side bookend she likes and tells Bill that she will get one tomorrow. Bill then goes to visit a friend and sees the bookend. Bill asks his friend if they are willing to sell the bookend. In a morally good condition, the friend agrees, and Bill buys the right-sided bookend from his friend. In a morally bad condition, Bill had to steal the right-sided bookend from his friend because it was a precious heirloom. The outcome of both conditions was that Sue went and bought the left-side bookend, and they had the paired set. When the varied agent (i.e., Bill) performed a morally bad action, the fixed agent (i.e., Sue) was seen as less causal. This superseding phenomenon was explained through counterfactuals, as norm violations make certain counterfactual alternatives more salient, leading observers to focus on the agent in question (Kominsky et al., 2015). The counterfactual simulation model was suggested to unify different views on

causation. The model posits that individuals assess causality by mentally simulating counterfactual scenarios, comparing what actually happened to what might have occurred under different circumstances (Gerstenberg et al., 2015).

Another close linkage between counterfactual and causal reasoning is the heuristics used during generation and inferences. We listed some heuristics during counterfactual generation such as recency and action, and counterfactuals have been proposed as explanations for the same effects in causal reasoning (Henne et al., 2019; 2021). The action effect in causal judgement describes the effect where individuals perceive actions as more causally significant than inactions. It was proposed that this effect arises because people are more inclined to consider counterfactual alternatives to actions than to inactions (Henne et al., 2019). The recency effect in causal judgements describes the effect where more recent events are judged as more causal. It was proposed that individuals are more prone to imagine counterfactuals of recent events. People's causal judgements can be altered by being prompted to consider counterfactuals for earlier events (Henne et al., 2021). These studies highlight the role of counterfactual thinking in shaping how people assign causality.

Despite the close relations, evidence has also suggested potential differences between causal and counterfactual inferences. It was argued that in people's mental simulation, mutability works as a necessity for causality (Wells & Gavanski, 1989). However, the most mutable are not necessarily the most causal. In a situation where a drunk driver crashed into Mr. Jones who was driving home on an unusual route, people identified the drunken driver as the cause. However, when they created counterfactuals, they imagined Mr. Jones

choosing his usual route (Mandel & Lehman, 1996). Causal explanations are more concerned with preventability than covariation (Spellman & Mandel, 1999). Causal thinking focuses on strong causes which covary with the outcome, while counterfactual thinking considers background enabling conditions (Byrne, 2005).

In this thesis, we explore the relations between counterfactual and causal explanations in people's explanation evaluation in Chapter 2.

Functions of Counterfactual Thinking

The functional theory of counterfactual thinking suggests that counterfactual thoughts are best explained in terms of their role in behaviour regulation and performance improvement (Epstude & Roese, 2008; Roese & Epstude, 2017), such as regulating emotional responses (e.g., Epstude & Roese, 2008; Kahneman & Miller, 1986; Roese & Olson, 1995; Sweeny & Vohs, 2012; Teigen & Jensen, 2010). Another important function is planning and justification. People rely on counterfactuals to explain events that happened in the past to excuse or defend past actions. By denying resources and control, people justify poor performance and bad outcomes accordingly. For example, for gamblers who continue gambling, it was found that they maintain both better counterfactuals ("could have been better") and worse counterfactuals ("could have been worse") to justify their behaviour (Anthony, Cowley & Blaszczynski, 2024).

Counterfactuals are focused on how things could have turned out differently in the past. Contrary to this idea is the concept of *prefactuals*, which is a conditional (if-then) that links actions with outcomes that may take place in

the future, for example, “If I take action X, it will lead to outcome Y”

(Epstude, Scholl & Roese, 2016). There are very limited empirical investigations about prefactuals. We mentioned the different sorts of possibilities people envisage when they read counterfactuals versus prefactuals (Byrne & Egan, 2004). The other example was mentioned in the controllability heuristic during counterfactual generation, in a failed situation, counterfactual thoughts mostly involve uncontrollable factors, while prefactuals involve mostly controllable elements (Ferrante et al., 2013). By generating prefactual thoughts about how things could turn out differently in the future, people generate a blueprint for the future through intention formation and decision support (Epstude & Roese, 2008; Roese & Epstude, 2017). It was also argued that task features influence prefactuals: prefactuals may show a greater variation and likelihood to occur when people have limited experience with the task (Epstude et al., 2016). When it comes to counterfactual thinking and future episodic simulation, it was suggested that from a neuroimaging perspective, both engage a common core network, but there are differences in hippocampal responses; from a behavioural perspective, both have reduced level of detail, but differ in the perceived plausibility of events (Schacter et al., 2015).

Contrary to functional theories, dysfunctional accounts of counterfactual thinking suggest that the simulation of alternatives impedes cognitive processes such as learning (Howlett & Paulus, 2013; Petrocelli et al., 2013; Sherman et al., 1995).

Given the dual possibilities account and the functions of counterfactual explanations discussed above, in this thesis we empirically tested the ability of counterfactuals to highlight causal links through simulation, and the potential

difference between counterfactuals and prefactuals as explanations. In Chapter 2, we examined counterfactual conditionals compared to causal assertions as explanations; in Chapter 3, we compared counterfactual explanations to prefactual explanations. We also investigated the effect of counterfactual explanations for risky choices in Chapter 4.

The first two sections outlined psychological findings about explanations and counterfactuals. We now turn to discuss the application of these findings, especially counterfactual explanations, in the field of AI.

Counterfactual Explanations in eXplainable AI (XAI)

The challenge of explanation in AI has not been a recent issue. In the 1970s, the discussion about the need for explanation in expert systems started (Shortliffe & Buchanan, 1975). User studies consistently showed that user satisfaction and trust can be significantly improved by explanation and justification (e.g., Bilgic & Mooney, 2005; Ye & Johnson, 1995; for a review, see Al-Ansari, Al-Thani & Al-Mansoori, 2024; Haque, Islam & Mikalef, 2023).

Following the growing number of applications of AI agents and evolving ethical and legal requirements such as the General Data Protection Regulation (GDPR) (Wachter et al., 2018), since the late 1990s, the notion of XAI has gained people's attention. XAI refers to the field and techniques in AI focused on making AI systems' decision-making processes understandable to humans. Through building systems that are more transparent, understandable and interpretable, human users will be given the opportunity to better understand and trust these systems (Arrieta et al., 2020; Biran & Cotton, 2017;

Das & Rad, 2020; Dwivedi et al., 2023; Leichtmann et al., 2023; Saeed & Omlin, 2023).

One major issue in XAI is the interpretability or explainability of AI agents, that is, the degree to which users can understand the cause of a decision made by an AI system. It was suggested that explanations can be offered without opening the “black box”, in order to help people act on the decision instead of merely understanding the AI system, and counterfactual explanations should be used as a solution (Wachter et al., 2018).

The other issue is justification, usually of why a decision is good. Creating a shared meaning is an important goal for explanation in AI: explanations should create “a shared understanding of the decision that was made between itself and a human observer” (Miller, 2019, p.8). However, the goals of the explainer and the explainee might be different (Wilkenfeld & Lombrozo, 2015), for example, the goal of the AI agent may be to gain trust and persuade, while the goal of the human user might be to understand. The goals of explanation were described in three steps: justifying why a particular decision was made, providing grounds to compare and contrast adverse decisions, and indicating what could be changed to receive a desired outcome in the future (Wachter et al., 2018).

Research on the psychology of explanation has inspired and influenced XAI methods (Byrne, 2019, 2023a; Miller, 2019; Taylor & Taylor, 2021). Amongst various methods and algorithms of XAI, we focus on counterfactual explanations because of their important role in causality and the interest in the XAI community. The next section discusses what is known and what needs to

be explored to generate better counterfactual explanations as a solution for XAI.

Counterfactual Explanations as a Solution

Recently, counterfactual explanations have been proposed as a valuable XAI solution. Counterfactual explanations are explanations that show how the AI model’s decision outcome would have been different, under different input feature conditions. For example, a banking app’s decision 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 (Karimi et al., 2020; Karimi, Schölkopf & Valera, 2021; Shang, Feng & Shah, 2022; Wachter et al., 2018) and because of their importance in human causal thinking (Byrne, 2019, 2023a; Miller, 2019). However, very little is known about how people understand the counterfactuals used in XAI, and few AI studies carry out human user tests of their methods. Few studies included any form of user testing, and even fewer did so in an experimentally adequate manner (Keane et al., 2021; Rong et al., 2023). There are also some differences between psychological research and XAI research on counterfactuals; for example, XAI research focuses on generating counterfactuals that are close neighbours to the actual decision. Suppose Alice seeks a mortgage loan and is unfortunately denied the loan. She might need to increase her income by \$10k to get approved, and even though an increase of \$50k would also fulfil the purpose, it is most pragmatic for her to make the smallest change possible.

Additionally, Alice might prefer the explanation to focus on changing a few things instead of changing many features (Verma et al., 2020). There is a lack of understanding of how findings from psychological research might be applied in XAI, as it is a new context for explanations.

Unlike traditional explanation methods, counterfactual explanations in AI are intentionally restricted and do not involve the difficulty of explaining complex rules. They provide a minimal amount of information for a decision to be changed and do not require an understanding of the internal logic of the AI system. As a result, the technical challenges of explaining algorithmic rules could be bypassed by providing a minimal amount of information through counterfactual explanations and meeting the legal requirements of the GDPR for users to claim their right to explanation (Wachter et al., 2018). A desired counterfactual explanation should provide changes that are realistic, actionable, and minimal (Verma et al., 2020). For example, it will not be beneficial if the bank loan application system advises users to “become younger”. What is more, the mutability of features needs to be carefully considered in order to generate only actionable and realistic explanations. Such systems also need to capture the user’s preferences and inform users what can be and what must not be changed (Verma et al., 2020). In this thesis, we tested the idea of “actionability”, that is, explanations should provide useful and actionable insights that help users effectively respond to an AI system’s decisions, a human user needs to know *how* to change an unfavourable decision by an AI system to achieve a desired outcome in the future.

We used controlled cognitive experiments to see whether people prefer explanations that contain actionable elements.

Human User Testing and Evaluation of Counterfactual XAI

Existing research suggests that counterfactual explanations help human users understand the system better (Bove et al., 2023; Goyal et al., 2019; Lage et al., 2019; Lee & Chew, 2023; Lim, Dey & Avrahami, 2009; Lucic, Haned & de Rijke, 2020; van der Waa et al., 2021). Following counterfactual explanations, user trust and satisfaction with the system increased (Binns et al., 2018; Celar & Byrne, 2023; Dodge et al., 2019; Förster et al., 2021; Hoffman et al., 2018; Kenny et al., 2021; Lim et al., 2009; Lucic et al., 2020; Russell, McGrath & Costabello, 2020; Warren, Byrne & Keane, 2023). Self-report scales have been suggested to test explanation satisfaction (Hoffman et al., 2018), and experiments have added objective measures to test the effect of counterfactual explanations on people's understanding of a decision support system besides subjective evaluation. For example, participants were tested with an app to predict blood alcohol content (BAC) to assist in driving decisions, where multiple features including the protagonists' gender, weight, and units of consumption are listed and a decision of whether this person's BAC is above or below the driving threshold is given by the AI system. Different sorts of explanations, such as causal explanations (e.g., Zoe's blood alcohol level was under the limit, because she drank few units of alcohol) and counterfactual explanations (e.g., Zoe's blood alcohol level would have been over the limit, if she had drunk more units of alcohol), were provided. People's understanding was measured by their accuracy when they were invited to predict what the system would say in each protagonist's case (Celar & Byrne, 2023; Warren et al., 2023). It was found that people understand explanations for categorical features such as gender (e.g., If Jenny's gender had been male,

she would have been under the limit) more readily than explanations for continuous features such as weight (e.g., If Samantha's weight had been 50 kg lighter, she would have been over the limit). A dissociation between subjective evaluation and objective performance was also found - counterfactual explanations elicited higher satisfaction as measured by self-report scales but not objective understanding as measured by the accuracy of prediction when compared to causal explanations (Warren et al., 2023). When asked to make their own decision rather than predict what the AI system will decide, it was found that counterfactual explanations were more effective than causals. Furthermore, participants' judgment of explanation helpfulness is related to whether the AI decision is correct (Celar & Byrne, 2023).

The role of domain familiarity and expertise also has been shown to have a strong impact on user acceptance of counterfactual XAI (Severes et al., 2023). For example, participants were tested with a familiar domain (blood alcohol content) and an unfamiliar one (chemical safety). It was found that when the AI decisions were correct, participants were more accurate in predicting the AI system's decisions and found the explanations of the AI system's decisions more helpful in the familiar domain, but when the AI decisions were incorrect, participants were less accurate and found the explanations less helpful in the familiar domain (Celar & Byrne, 2023). When faced with familiar and unfamiliar domains, people's understanding of explanations for correct and incorrect classifications changed on several dimensions, including reaction time and ratings of correctness (Ford & Keane, 2022). In the legal sector, a study provided US state court judges with counterfactual explanations in the context of pretrial risk assessment

instruments. It was found that the judges misinterpreted the counterfactuals as real instead of hypothetical, and once they understood their hypothetical nature, they ignored the counterfactual explanations (Yacoby et al., 2022). Visual counterfactual explanations were given to users in a bird classification task, and it was found that experienced users performed better when given access to counterfactual explanations (Goyal et al., 2019). In the healthcare sector, it was found that only non-experts benefitted from correct XAI advice during the diagnosis of X-rays (Gaube et al., 2023). Expert-level evaluations have also been incorporated into XAI methods for highly sensitive tasks such as medical diagnosis (Muddamsetty, Jahromi & Moeslund, 2021).

User studies have also tended to assess the accuracy of users' understanding of AI system decisions by testing their predictive inferences about an outcome based on information about the presence or absence of relevant causes (Goyal et al., 2019; Lage et al., 2019; Lim et al., 2009; Lucic et al., 2020; van der Waa et al., 2021).

Given the user studies above, we designed experiments reported in Chapter 3 of this thesis to test the efficacy of counterfactual explanations using a combination of objective and subjective measures. We gave participants not only predictive tasks where they had to predict grass growth on a farm based on information about causes like rainfall and fertiliser; but also, diagnostic tasks, where they had to diagnose the condition of one specific cause based on the consequent level of grass growth. We aimed to add to the current methods of testing counterfactual explanations.

Another important effect that motivated experiments in this thesis is algorithmic aversion. The term *algorithm aversion* was used to describe the

phenomenon that a human forecaster wins when compared to a statistical algorithm, even when the latter shows higher accuracy (Burton, Stein & Jensen, 2020; Dietvorst, Simmons & Massey, 2015; Kelly, Kaye & Oviedo-Trespalacios, 2023; Mahmud et al., 2022). It has been shown that people are especially averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster. Given this effect, we designed experiments reported in Chapter 4 of this thesis to test the different effects between advice from a human expert and advice from an AI system. We compared the different advice sources in a setting where people have to make a risky decision.

Summary of Aims

This chapter discussed research on explanations and in particular counterfactual explanations. Psychological research suggests that people's selection of explanations is guided by some qualities and heuristics, such as a preference for simplicity. However, such preferences have also been shown to be reduced in different settings, for example, when contradicting probability information is provided or when people evaluate explanations in a naturalistic setting. Chapter 2 reports six experiments that examine a novel heuristic when choosing causal explanations, a “consequent-controllability” heuristic, motivated by XAI proposals about “actionable” explanations, and how it interacts with the simplicity quality. We aim in Chapter 2 to provide psychological evidence as to whether people prefer explanations with actionable elements over those without actionable elements, which we operationalise in various ways by testing the consequent-controllability and the

antecedent-controllability. We test whether any preference for consequent-controllability will be reduced when tested against simplicity. We tested the treatability effect in two different domains. The results contribute to the understanding of causal explanation selection, as well as to the actionability goal in XAI.

Research in causal reasoning indicates that people make different sorts of inferences when they reason from cause to effect compared to when they reason from effect to cause. Psychological studies of simplicity in explanation have tended to focus on diagnostic tasks (from information about an effect such as a symptom to inferences about causes such as diseases). However, human User tests in XAI tend to focus on predictive tasks (from information about a cause such as an input to inferences about an effect such as a system's decision) rather than diagnostic ones. There are also limited experiments on prefactual explanations: "If I were to study harder, I will pass the exam next month" and how they might differ from counterfactual explanations "If I had studied harder, I would have passed the exam". In the two experiments reported in Chapter 3, we aim to test the efficacy of diagnostic tasks to evaluate user understanding of decision support systems. We also added prefactual explanations as a variable, to inform XAI counterfactual solutions about the difference of explanation type. The results have implications for understanding the differences between predictive and diagnostic inferences, and for the inclusion of different types of tasks when carrying out human user studies for XAI.

Psychological research on counterfactual thinking shows that counterfactual explanations elicit dual possibilities in people's mental

simulations of the world, that is, people simulate alternatives when they understand counterfactuals. In one of the experiments in Chapter 2, we tested whether counterfactual explanations reduce people's strong preference for a simpler explanation, and in the five experiments reported in Chapter 4, we tested whether a counterfactual explanation for an AI system's decision prompts people to follow a decision that is opposite to their natural tendency in risky decision-making scenarios. The results relate to the functions of counterfactuals, as we show that counterfactuals prompt people to consider alternatives, and the mental model theory of counterfactual thinking as we show that counterfactuals simulate alternative representations. The results also inform the contexts in which counterfactual explanations may be especially effective, for both psychology research and XAI application, as we show that in situations where people show a strong preference for one option, the occurrence of counterfactuals helps mitigate their dominant preferences. In the next chapter, we report six experiments in which we focused on the cognitive processes underlying how people select between competing causal explanations.

Chapter 2: Cognitive Heuristics in Causal Explanation Selection

Imagine you are a farmer experiencing a problem with low grass growth on your farm, which you know is caused by micro-organisms in the soil. Various micro-organisms can explain the problem on your farm. Suppose the micro-organisms Excie and Rirtan always cause low grass growth; Excie can be remedied with fertiliser, but there is no remedy for Rirtan. Would you prefer the explanation that your farm has Excie to the explanation that it has Rirtan? In other words, is an explanation that identifies the cause of the outcome of low grass growth more satisfying if the cause itself contains consequent-controllable elements, i.e., Excie can be remedied with fertiliser? Or suppose instead you are experiencing problems with high weed growth on your farm. The micro-organisms Cartia and Pacis always cause high weed growth. Pacis is the result of herbicide misuse, and there is no known cause for Cartia. Would you prefer the explanation that your farm has Pacis to the explanation that it has Cartia? In other words, is an explanation that identifies a cause of the outcome of high weed growth more satisfying if the cause contains antecedent-controllable elements, i.e., Pacis is the result of herbicide misuse?

In this chapter, we focus on people's selection of competing explanations and investigate what leads people to prefer one explanation over another. In six experiments, we address three research questions: Do people prefer a causal explanation that contains consequent-controllable elements compared to ones that do not contain consequent-controllable elements? And do they prefer a causal explanation that contains antecedent-controllable elements compared to ones that do not contain antecedent-controllable elements? Second, if there is a preference for these sorts of explanations, does

the preference override the extensively examined preference for simplicity (e.g., Lombrozo 2007)? The third question is, would people overcome what they usually choose (i.e., the simpler explanation) and choose a more complex explanation instead if the explanation is in the form of a counterfactual, such as “If your farm had not had micro-organism Excie, it would not have had the problems” rather than a causal assertion, “Because your farm had micro-organism Excie, it had the problems”? We discuss evidence on these research questions in the sections below, starting with eXplainable Artificial Intelligence (XAI) findings and then psychological findings.

Actionability in Counterfactual XAI Explanations

In XAI, in recent years, the concept of “actionability” has been proposed and emphasised as a criterion for explanation evaluation, and as a strong factor of AI interpretability – it is as important to tell users *how* to change a decision as to explain *why* a decision was made (Alotaibi & Singh, 2023; Bhattacharya, 2024; Degen et al., 2023; Guidotti et al., 2022; Hacker & Passoth, 2020; Kelechi & Jiao, 2023; Singh et al., 2023; Sokol & Flach, 2020). For example, a banking app that evaluates mortgage applications might explain a failed application with “not enough monthly income”, which applicants can act upon, while “being too old” as an explanation does not provide actionable solutions. One aim of AI explanation is to let people understand what needs to be changed to achieve the desired outcome in the future (Wachter et al., 2018). *Recourse* was proposed to evaluate classification models, which was defined as “the ability of a person to change the decision of the model through actionable input variables” (Ustun, Spangher & Liu, 2019, p. 10).

Some human user studies have been carried out where actionability has been suggested or applied as an explanation evaluation criterion. In a case study with a diabetes prediction system, actionable features have been proposed (features that could be acted upon by the patients or their doctors to reduce the risk of diabetes) to be an important factor of interpretability since it was observed that factors such as old age were being ranked highly by the system but simply impossible to act upon (Krause, Perer & Ng, 2016).

When it comes to XAI counterfactual explanations, it has been argued that the pathway which leads to the “closest possible world” needs to prescribe achievable and feasible goals (Karimi et al., 2020, 2021; Poyiadzi et al., 2020), such as “your mortgage application would have been approved if you had higher monthly income”. User studies have compared counterfactual explanations to other types of explanations in terms of actionability. A user study with 90 participants online was carried out, where counterfactual explanations about how the outcome would have been different (e.g., if the income were higher the mortgage would have been approved) and directive explanations about what to do to achieve the desired outcome (e.g., to rent out a room in the applicant’s house for higher income) were directly compared in a mortgage application scenario and an employee turnover scenario. Counterfactual and directive explanations were found to be almost equally actionable, and three effective metric items to assess the actionability of explanations were identified as: “breaks down any action into manageable, explicit steps”, “clearly identifies at least one action”, and “explanation addresses me directly” (Alotaibi & Singh, 2023). Another user study used a credit-refusal AI algorithm with 400 people and tested different dimensions of

interpretability, including actionability. Participants were given different types of explanations including counterfactuals and were asked to judge perceived actionability by rating “If you have the possibility to reapply, you have the means to modify your credit demand to obtain the bank loan” using a Likert scale. It was found that the presence of an explanation significantly improved actionability compared to no explanation, and counterfactual explanations significantly contributed to actionability (John-Mathews, 2021).

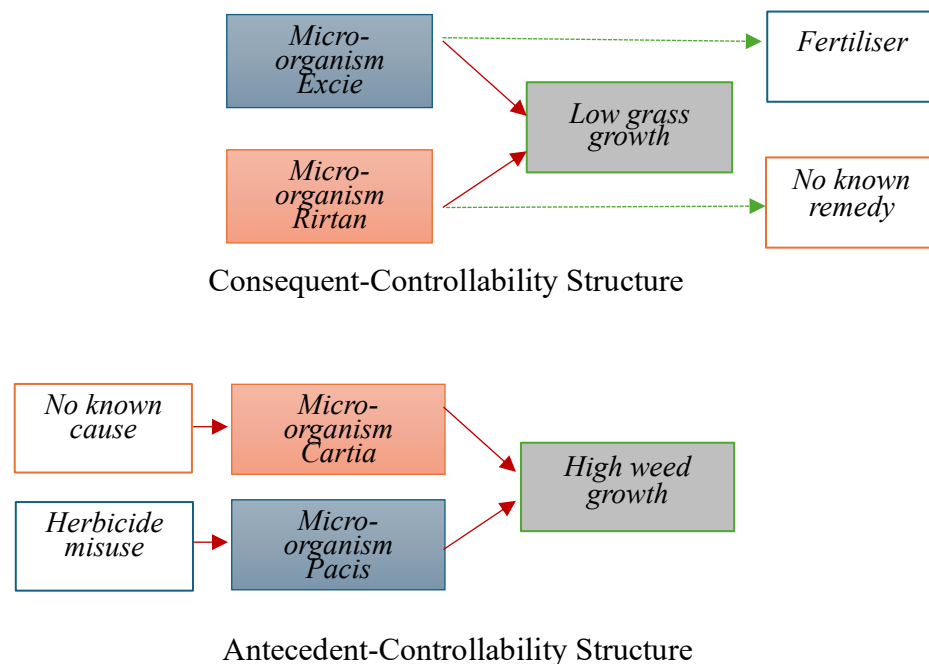
In summary, the concept of an “actionable” explanation in XAI variously references the ideas that a human user needs to know *how* to change an unfavourable decision by an AI system to achieve a desired outcome in the future, that an explanation should identify features of a decision that the human can act upon, that it should provide a human user with achievable and feasible goals, that it identifies an action broken down into explicit steps.

We wished to examine the role of actionability in the evaluation of explanations such as that the micro-organism *Excie* causes the bad outcome of low grass growth. We considered that actionability could be operationalised in at least two ways: the first one concerns what we will call the “consequent-controllability” of the outcome, e.g., whether the remedy to this micro-organism is available, e.g., *Excie* causes low grass growth, and *Excie* can be remedied by fertiliser; by consequent-controllability we mean that there is something additional the farmer can do, to intervene directly on the cause-effect link, that prevents the cause from leading to the bad outcome, something that counteracts the cause subsequently to fix it. The second one concerns what we will call the “antecedent-controllability” of causes, e.g., whether the micro-organism in the soil is caused by a factor that is within a farmer’s control, e.g.,

Pacis causes high weed growth, and Pacis results from misuse of herbicide; by antecedent-controllability we mean that there is something the farmer can stop doing, to intervene indirectly on the cause-effect link, by preventing the cause from occurring and thus from leading to the bad outcome, something that counteracts the antecedents to the cause to fix it. See **Figure 2.1** for the causal structures of these two hypotheses. We tested the effect in a causal explanation selection task to investigate whether these sorts of information affect explanation evaluation.

Figure 2.1

Diagrams for the Structure of Consequent-Controllability and Antecedent-Controllability



Note. Diagrams for the structure of the consequent-controllability and antecedent-controllability hypothesis tested in Chapter 2, using the farming content as an example in the brackets.

Controllability as Heuristics

Our interest in the role of controllability in causal explanation selection also arises because of its significance in counterfactual thinking. As we discussed in Chapter 1, controllability serves as a heuristic during the generation of counterfactuals: people choose the controllable aspects to modify, such as their own decision to choose an envelope that turned out to contain a hard sum, rather than uncontrollable aspects, such as the time constraints to calculate the sum (e.g., Girotto et al., 1991; Xie & Beck, 2022). But other factors such as the perspective people take, either as readers or actors or observers (Girotto et al., 2007; Pighin et al., 2010) and appropriateness of actions, such as being delayed by deciding to call to a bar or delayed by deciding to call to visit parents (McCloy & Byrne, 2000) have been found to interact with the controllability heuristic. In these studies, controllability is conveyed through voluntary actions, such as going for a drink or choosing one envelope between two envelopes, that reflect an intentional decision by an actor.

The idea of controllability also plays an important role in the psychology of moral responsibility and blame. For example, greater blame was assigned by participants when the agent's perceived control over actions led to negative outcomes (Alicke, 2000). Controllability was also tested with responsibility and blame, participants were asked to read a description of a male with AIDS or lung cancer, and rated controllability, responsibility and blame in relation to the illness. Findings suggest that perceived controllability influences responsibility judgements and in turn affects blame attributions, as sequential constructs (Mantler, Schellenberg & Page, 2003). The Path Model of

Blame suggests the process of canonical information processing sequence that guides blame: perceivers consider information about causality, then intentionality, then reasons or preventability. For example, participants were presented with scenarios describing negative events and were asked to seek additional information to make a blame assignment. The information corresponded to the stages of the Path Model. Causality information gave details about who or what caused the event; intentionality information provided insights into whether the action was performed intentionally or accidentally; if the action was intentional, reasons information described the agent's motives; if the action was unintentional, preventability information described whether the agent could have prevented the outcome. Participants could select which piece of information they wanted to acquire before making their judgements. The order and the speed of choices were measured (Guglielmo & Malle, 2017; Malle, Guglielmo & Monroe, 2014).

When it comes to causal reasoning, research evidence has indicated a lack of influence from controllability. For example, causal thinking focuses more on causal links, unlike counterfactual thinking, which focuses more on background enabling conditions (e.g., Byrne, 2005; Mandel & Lehman, 1996). When asked to seek explanations for a negative event, people created an equal number of causal explanations about both controllable and uncontrollable events (McEleney & Byrne, 2006). These findings show that causal thinking does not relate to controllability. Nonetheless, we wish to explore further whether there is any role for controllability in causal explanations.

Given the research findings on the role of controllability in the psychology of counterfactual thinking, we wish to test controllability as a

potential heuristic in causal explanation. We test the idea of controllability by differentiating consequent-controllability and antecedent-controllability, as seen in **Figure 2.1**. We first test the controllability of the outcome following the explanandum (referred to as the consequent-controllability hypothesis in this thesis), influenced by the notion of actionability in the XAI literature, a criterion of interpretability that informs users on how to change the outcome. In psychological studies, controllability has been tested as a voluntary action leading to a negative outcome. Hence, we also tested the controllability of the causes antecedent to the explanandum (referred to as the antecedent-controllability hypothesis in this thesis). The consequent-controllability hypothesis was tested in Experiment 1, and the antecedent-controllability hypothesis in Experiments 2a and 2b.

Simplicity Preference in Explanation Evaluation

As mentioned in Chapter 1, research in philosophy and cognitive sciences suggests that our evaluation of explanations is partly guided by a series of explanatory qualities, such as simplicity, coherence, and breadth (Glymour, 2014; Liefgreen & Lagnado, 2023; Lipton, 2004; Mackonis, 2013; Thagard, 1989; Zemla et al., 2017). Experiments in the present chapter focused on the simplicity preference, and we compared simplicity with actionability, that is, with consequent-controllability in Experiment 3a and with antecedent-controllability in Experiment 3b.

We base our research design and question on a series of experiments conducted using materials based on a hypothetical planet (Lombrozo, 2007) and participants had to diagnose an alien's medical symptoms:

There is a population of 750 aliens that lives on planet Zorg. You are a doctor trying to understand an alien's medical problem. The alien, Treda, has two symptoms: Treda's minttels are sore and Treda has developed purple spots. Tritchet's syndrome always causes both sore minttels and purple spots. Morad's disease always causes sore minttels, but the disease never causes purple spots. When an alien has a Humel infection, that alien will always develop purple spots, but the infection will never cause sore minttels. Nothing else is known to cause an alien's minttels to be sore or the development of purple spots.

What do you think is the most satisfying explanation for the symptoms that Treda is exhibiting?

- (A) Treda the alien has Tritchet's syndrome.
- (B) Treda the alien has Morad's disease.
- (C) Treda the alien has a Humel infection.
- (D) Treda the alien has Tritchet's syndrome and Morad's disease.
- (E) Treda the alien has Tritchet's syndrome and a Humel infection.
- (F) Treda the alien has Morad's disease and a Humel infection.

It was found that when no probability information was given, 96% of participants chose the simple explanation of the alien having Tritchet's syndrome over the other ones (Lombrozo, 2007). In this chapter, we base the experimental materials on Lombrozo's (2007) experiments, and two of our

three research questions are related to simplicity preference. Firstly, we test whether the preference for a simpler explanation will be influenced by consequent-controllability and antecedent-controllability. For example, to explain an alien's symptoms, will people prefer an explanation suggesting that the symptoms are caused by two diseases that can be treated by medicine or surgery, or will people prefer an explanation suggesting that the symptoms are caused by one disease that has no known treatment?

Secondly, we discussed in Chapter 1 that people generate and update mental simulations when they think counterfactually and sometimes these simulations highlight causal links. Our research question for the last experiment in this chapter, Experiment 4, derives from these characteristics of counterfactual thinking. If counterfactuals are effective in generating mental representations of alternatives, then participants might be prompted to mentally keep track of multiple alternatives that help them to highlight the causal link between causes and effects. In other words, when the explanations are phrased as counterfactuals, participants may no longer opt for the simplest explanation because they can simulate the more complex scenario of the combination of the two diseases causing the symptoms. We investigated whether explanations constructed using counterfactual conditionals such as "If the alien Treda had not had Trichet's syndrome, Treda would not have had the symptoms" will prompt people to consider alternative possibilities, hence shifting away from a dominant preference for the simple explanation only. Counterfactual explanations have been empirically compared with causal explanations (e.g., Celar & Byrne, 2023; Warren et al., 2023). In Experiment 4, we compared

counterfactual explanations to factual causal explanations such as “Because the alien Treda had Tritchett’s syndrome, Treda had the symptoms”.

Overview of Experiments

We report six experiments to address the three research questions mentioned above. In Experiments 1, 2a and 2b, we tested the effect of consequent-controllability and antecedent-controllability when the explanations were equally simple. In Experiment 3a we investigated the interaction between consequent-controllability and simplicity, that is, we tested whether the preference for the simple explanation could be overcome when consequent-controllability information supports a complex explanation. In other words, would people prefer a complex explanation that contained consequent-controllable elements over a simple but consequent-uncontrollable explanation? The same comparison was carried out between antecedent-controllability and simplicity in Experiment 3b. Experiment 4 in this chapter addressed whether people’s preference for simplicity would be affected by counterfactual explanations. In all experiments, we tested with materials in the alien disease domain (Lombrozo, 2007). We also tested participants with scenarios in the farming domain as in the example we used above, to validate that the effects were domain general.

Experiment 1

The aim of Experiment 1 was to test whether people prefer an explanation that is consequent-controllable, for example, whether they choose, as an explanation for a symptom, treatable diseases rather than untreatable

ones. Consider a problem we developed that is somewhat similar to the one described earlier.

Suppose you are a doctor trying to understand an alien's medical problem and you are told: The alien, Treda, has two symptoms: Treda's minttels are sore and Treda has developed purple spots.

Both Trichet's syndrome and Morad's disease always cause sore minttels, but they never cause purple spots. Trichet's syndrome can be treated with medication. There is no treatment for Morad's disease.

Both Sommet's disease and a Humel infection always cause purple spots, but they never cause sore minttels. There is no treatment for Sommet's disease. A Humel infection can be treated with surgery. What do you think is the most satisfying explanation for the symptoms that Treda is exhibiting?

- (A) Treda the alien has Trichet's syndrome and a Humel infection.
- (B) Treda the alien has Trichet's syndrome and Sommet's disease.
- (C) Treda the alien has Trichet's syndrome and Morad's disease.
- (D) Treda the alien has a Humel infection and Morad's disease.
- (E) Treda the alien has a Humel infection and Sommet's disease.
- (F) Treda the alien has Morad's disease and Sommet's disease.

Option A above refers to the two causes that can be treated (i.e., consequent-controllable), and option F refers to the two causes that cannot be treated (i.e., consequent-uncontrollable). Options B and D each contain one consequent-controllable and one consequent-uncontrollable cause. Options C and E are causally insufficient because they only account for one symptom, not two.

We suggest people attempt to simulate the causal relations described in the story. They may attempt to construct models that capture the information that Trichet's and Morad's cause sore minttels, not purple spots; Trichet's can be treated with medication, and there is no treatment for Morad's disease:

Trichet: sore minttels, not purple spots treated with medication

Morad: sore minttels, not purple spots no treatment

They may also attempt to construct models that capture the information that Sommet's and Humel's cause purple spots, not sore minttels; Humel's can be treated with surgery, but there is no treatment for Sommet's:

Sommet: not sore minttels, purple spots no treatment

Humel: not sore minttels, purple spots treated with surgery

When they are asked to choose the most satisfying explanation from the set of options, there may seem to be little rational means to choose between the four causally eligible explanations. They may fall back on a short-cut heuristic, to choose the causes that are consequent-controllable. According to the idea of actionability in XAI, if you are influenced by whether an explanation suggests actionable information, then you may be inclined to choose option A above, the explanation that Treda the alien has Trichet's syndrome and a Humel infection, given that Trichet's syndrome can be treated with medication and a Humel infection can be treated with surgery. If there is such an effect, we expect that participants will tend to choose the explanation with the two consequent-controllable diseases more than any of the other explanations. The structure of this scenario is illustrated in the consequent-controllability information condition of **Figure 2.2**. We compared it to a control condition that had no consequent-controllability information:

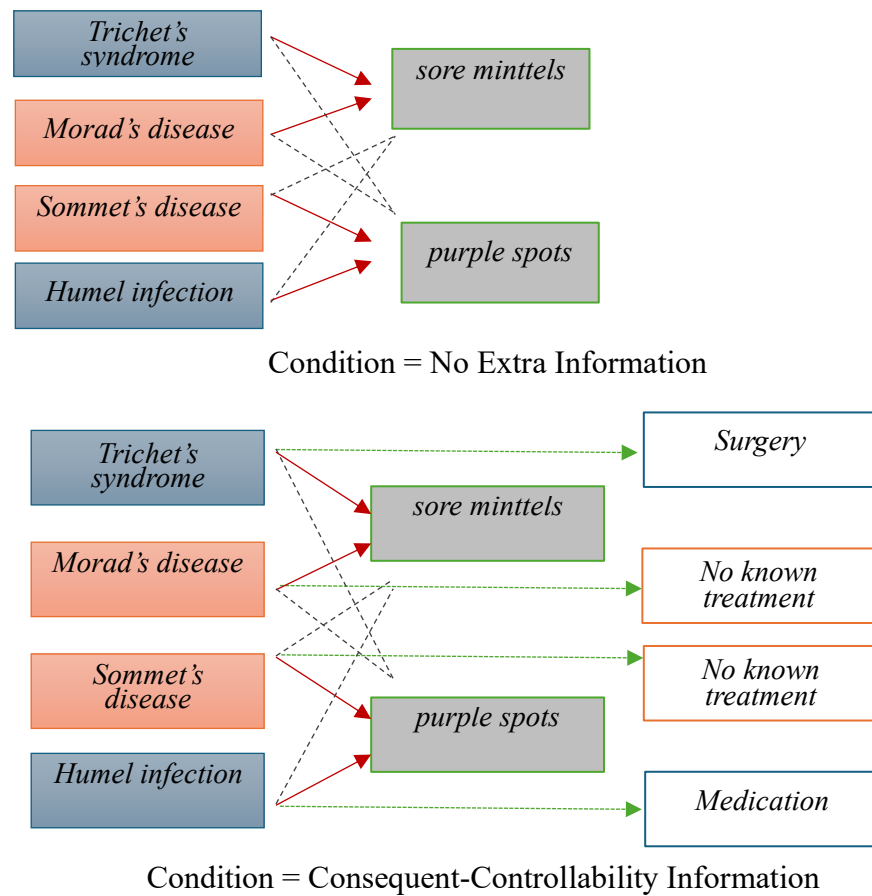
Suppose you are a doctor trying to understand an alien's medical problem. The alien, Treda, has two symptoms: Treda's minttels are sore and Treda has developed purple spots. Both Trichet's syndrome and Morad's disease always cause sore minttels, but they never cause purple spots. Both Sommet's disease and a Humel infection always cause purple spots, but they never cause sore minttels. Nothing else is known to cause an alien's minttels to be sore or the development of purple spots. What do you think is the most satisfying explanation for the symptoms that Treda is exhibiting?

- (A) Treda the alien has Trichet's syndrome and a Humel infection.
- (B) Treda the alien has Trichet's syndrome and Sommet's disease.
- (C) Treda the alien has Trichet's syndrome and Morad's disease.
- (D) Treda the alien has a Humel infection and Morad's disease.
- (E) Treda the alien has a Humel infection and Sommet's disease.
- (F) Treda the alien has Morad's disease and Sommet's disease.

Among the four causally eligible explanations, you may feel there is no real basis in the information given to enable you to choose one explanation over another. We expect that participants will tend to choose an explanation at random from these four, and so each of the four causally eligible explanations will be chosen equally, as we also randomised the order of the explanations. The structure of this problem is illustrated in the no extra information condition of **Figure 2.2**.

Figure 2.2

Diagrams for the Causal Structure in Experiment 1



Note. Diagrams of the disease domain are presented. Red arrows indicate a cause-effect relation. Dotted lines indicate that the diseases do not lead to the symptoms. These diagrams are for illustration purposes only, participants were not provided with diagrams.

Experiment 1 aimed to test the effect of consequent-controllability, i.e., whether people prefer explanations that have elements that can be subsequently fixed. We tested the consequent-controllability effect as it is close to the notion of “actionability” in XAI. We manipulated consequent-controllability by stating the specific treatment such as “the disease can be treated with

medication” versus “there is no known treatment for the disease”. We hypothesised that participants would prefer explanations that suggested only consequent-controllable factors over ones that suggested only consequent-uncontrollable factors or a mix of both, as the explanation suggests how the outcome can be acted upon. For an experimental control, we included the no extra information condition where no consequent-controllability information was provided.

Method

Participants

The participants were 184 members of the general public. In all experiments in this thesis, participants were invited to provide information on their age and gender by typing their answers into text boxes. The participants were 72 men, 107 women, three non-binary individuals, and two preferred not to say). Their ages ranged from 18 to 74 years, with a mean of 39 years. All participants from all experiments in this thesis were recruited from the online platform Prolific. No attention checks were implemented in any of the experiments in this chapter because each participant only completed two items, so no participants were eliminated in any experiment in this chapter. A sample size of 172 participants was required to detect a medium effect size with 90% power at $p < .05$, according to an a-priori G*Power test, and we collected 184 to ensure we had the minimum number of participants in each condition following the randomisation by the experiment participant recruitment platform Alchemer. The randomisation process was applied to the recruitment of all experiments in this chapter. For all experiments in this thesis, we selected

participants to be adults from Ireland, the United Kingdom, the United States, Canada, Australia and New Zealand for whom English is their first language. Each participant was paid £0.45 UK sterling for participation based on an estimated three-minute completion time for this experiment. All experiments in this thesis were approved by the Trinity College Dublin School of Psychology Research Ethics Committee (see Appendix 2.1). Approval was granted on 20/02/2021 (SPREC112020-22). Informed consent was provided by all participants by ticking a box which indicated they had read and understood consent information, which applies to all experiments in this thesis. See Appendix 2.1 for data protection and consent information for all experiments in this thesis. See Appendix 2.2 for the information leaflet and debriefing information of all experiments in this chapter.

Design

We aimed to test the effect of consequent-controllability information on people's preference for explanations, so a between-groups design was adopted to test the preference between conditions. The independent variable was the consequent-controllability information which was manipulated between groups. The participants were randomly assigned to one of the two groups: no extra information ($n = 98$), and consequent-controllability information ($n = 86$). Following Lombrozo (2007), we measured their choice of the most satisfying explanation by asking "What do you think is the most satisfying explanation for the symptoms that Treda is exhibiting?" from a set of six explanations. The dependent variable was which of the six explanations they chose; they could choose only one answer option from the six.

Materials & Procedure

We adapted experimental materials for the alien disease domain based on Lombrozo (2007). As mentioned earlier, we also tested a scenario in the farming domain to check whether the effect is domain-independent, see Appendix 2.3 for the scenario in the farming domain.

As the example given above shows, participants in the no extra information read about the diseases and their causal relations with the symptoms. Participants in the consequent-controllability information condition read that Trichet's syndrome and a Humel infection can be treated with surgery and medication and that there were no known treatments for Morad's disease and Sommet's disease. See **Figure 2.2** for a diagram illustrating the causal relations of diseases and symptoms for both experimental conditions. Participants were given six explanations based on different combinations of two causes, as shown above. The six explanations were presented in a different randomized order for each participant.

In the consequent-controllability information condition, we predicted that people would prefer the combination of Trichet's syndrome and a Humel infection, which would indicate a preference for explanations with consequent-controllable factors. Since two explanations (options C and E from our example above) were only able to explain one symptom and were causally insufficient, we treated them as fillers. In the no extra information condition, we predicted that the participants would randomly choose from the other four causally eligible combinations. We did not randomise the order of information in the scenario describing the four diseases so that we could check whether there was an effect of the order. Trichet's syndrome and Sommet's disease

were each the first mentioned causes in their pair of causes in the description, and so choices of these explanations at a greater-than-chance level would indicate a tendency to choose the first-mentioned item they read from each paragraph. Similarly, if people chose the combination of Morad's disease and a Humel infection, it would indicate a tendency to choose the last-mentioned items they read from each paragraph.

All experiments in this chapter were administered through the survey platform Alchemer. Participants accessed the experiment with their own devices in their own time.

Results and Discussion

The data were analysed using the IBM SPSS Statistics in this thesis, all data for all experiments in this chapter are shared on OSF (<https://osf.io/s6jke>). The frequencies of explanation preferences for both conditions are presented in **Table 2.1**, for each domain separately and then for both combined. Since the responses were ordinal data, we carried out non-parametric tests for experiments in this chapter.

Table 2.1*Frequencies of Explanation Preferences in Experiment 1*

Explanation options	Trichet & Humel (Both causes controllable)	Morad & Sommet (Neither cause controllable)	Trichet & Sommet (First mentioned causes)	Morad & Humel (Second mentioned causes)	Trichet & Morad (Filler)	Humel & Sommet (Filler)
Domain = Disease						
Consequent Controllability information	43 (50%)	13 (15%)	9 (10%)	10 (12%)	4 (5%)	7 (8%)
No extra information	21 (21%)	7 (7%)	31 (32%)	32 (33%)	5 (5%)	2 (2%)
Domain = Farming						
Consequent Controllability Information	37 (43%)	20 (23%)	9 (10%)	4 (5%)	7 (8%)	9 (10%)
No extra information	16 (16%)	10 (10%)	40 (41%)	17 (17%)	11 (11%)	4 (4%)
Both domains combined						
Consequent Controllability Information	80 (47%)	33 (19%)	18 (10%)	14 (8%)	11 (6%)	16 (9%)
No extra information	37 (19%)	17 (9%)	71 (36%)	49 (25%)	16 (8%)	6 (3%)

Note. All explanation options were shown to the participants in a randomised order during the experiment. No consequent-controllability information was provided in the no extra information condition, but we used here the same label to identify the explanations to match the consequent-controllability information condition. The percentages were in parentheses.

First, we checked that no differences were arising merely because of the order in which participants received the disease or farming scenario (first received scenario, second received scenario), see Appendix 2.4 for details. Thus, we combined the responses from the orders for further analysis. Two 6 X 2 Chi-square analyses compared the distribution of explanation choices

between the two conditions for each domain and showed that there was a significant relationship between consequent-controllability information and people's preferred explanations in the disease domain, $\chi^2(5) = 35.24, p < .001, V = 0.19$; and in the farming domain, $\chi^2(5) = 41.52, p < .001, V = 0.21$.

Comparison Between Conditions

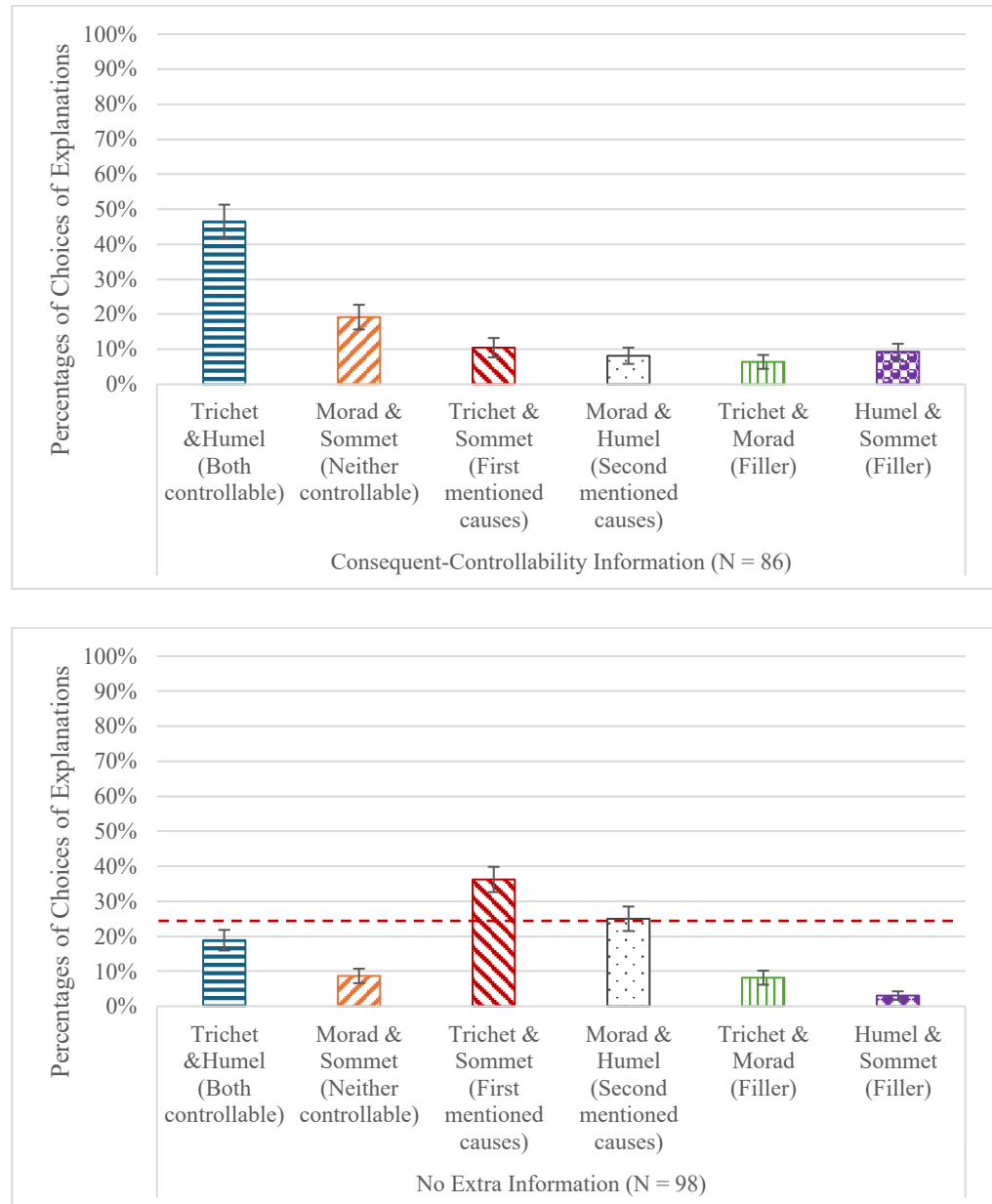
First, we checked that no differences were arising merely because of the domain (disease and farming), and there were no domain effects. See Appendix 2.4 for supplementary analysis. As a result, we combined responses from both domains. To compare the choices of the explanations between the two conditions, we carried out four Mann-Whitney U tests with a Bonferroni corrected alpha level of $p < .0125$ on participants' mean responses for the two domains. Participants chose the explanation that contained both of the causes described as consequent-controllable, more frequently in the consequent-controllability information condition than in the no extra information condition (47% vs 19%), $U = 5612.5, z = 4.331, p < .0001, r = 0.32$; they chose the explanation that contained both of the causes described as consequent-uncontrollable to the same degree in the consequent-controllability information condition as in the no extra information condition (19% vs 9%), $U = 4802.5, z = 2.253, p = .024$; they chose the explanation that contained the first-mentioned causes less in the consequent-controllability information condition (10% vs 36%), $U = 2499, z = 5.532, p < .0001, r = 0.41$; and they chose the explanation that contained the second-mentioned causes less in the consequent-controllability information condition, $U = 3140, z = 3.830, p = .0001, r = 0.28$ (8% vs 25%). Overall, participants chose the explanation that contained both of

the causes described as consequent-controllable more often in the consequent-controllability information condition than the no extra information condition. Participants did not tend to choose the explanations that contained the causes that were each mentioned first, or each mentioned second as often in the consequent-controllability information condition as in the no extra information condition, since they could choose only one answer option.

We also compared participants' choices of each of the four causally eligible explanations to a chance level of 25%, see Appendix 2.4 for supplementary analysis. In the consequent-controllability information condition, participants' choices of the explanation with consequent-controllable causes were higher than chance, the other explanations were chosen below chance level; in the no information condition, participants' choices of the explanation with the first-mentioned cause in each pair in the scenario were higher than chance, the other explanations were chosen below or at chance level, see **Figure 2.3**.

Figure 2.3

The Percentages of Choices of all Explanations in Experiment 1 for the Consequent-Controllability Information Condition (top) and for the No Extra Information Condition (bottom)



Note. The percentages were combined for both the disease and farming domains (the disease domain names are used here to illustrate). The red dotted line on the bottom graph represents a 25% chance level. Error bars are the standard error of the mean.

Results from this experiment showed that consequent-controllability information did have an impact on people's explanation selection, which was consistent with our hypothesis. Explanation options were presented in random order, so participants' choices were not determined by the order in which they received the explanations. The descriptions of causes in the scenario were presented in a fixed order, and when no consequent-controllability information is given, there is no basis to choose between the four eligible explanations. In that case, people seem to choose simply based on the order of the information provided in the scenario, they choose the explanation that contains the first mentioned disease in each pair in the scenario. Their choices in the control condition appear to be based on superficial strategies such as choosing the explanation with the first mentioned causes.

When they were given consequent-controllability information, this reliance on order no longer holds. Instead, people based their selection of an explanation on whether the causes can be treated or not, for example, they chose the two diseases that were described as being consequent-controllable. Hence, since they could choose only one response option, and they chose the explanation with the consequent-controllable causes, their tendency to choose the option with the first-mentioned causes was reduced. We did not factor out people's tendency to choose the first-mentioned or second-mentioned causes in testing, but the results suggested that the consequent-controllability effect was particularly strong because it overrode such a tendency. This tendency to choose consequent-controllable causes may reflect a heuristic people rely on when there are no other clues as to which is the best explanation and it may be based on a preference for explanations that allow you to do something, that is,

explanations that are actionable, as suggested by the idea of actionability introduced in XAI.

We followed up with Experiments 2a and 2b, where we investigated the effect of antecedent-controllability information on explanation preferences.

Experiment 2a

The aim of this experiment was to test whether people prefer an explanation that has an antecedent-controllable element, for example, whether they choose, as an explanation for a symptom, a disease with controllable antecedents rather than one without a controllable antecedent. Suppose you are trying to understand the alien's medical problem, and the information of symptoms and diseases is the same as the no information condition in Experiment 1, you may decide to choose one of the four eligible explanations simply based on the order of the information provided in the scenario. But suppose now instead you are told:

Both Trichet's syndrome and Morad's disease always cause sore minttels, but they never cause purple spots. Trichet's syndrome is the result of smoking. There is no known cause for Morad's disease. Both Sommet's disease and a Humel infection always cause purple spots, but they never cause sore minttels. There is no known cause for Sommet's disease. A Humel infection is the result of physical inactivity. What do you think is the most satisfying explanation for the symptoms that Treda is exhibiting?

- (A) Treda the alien has Trichet's syndrome and a Humel infection.
- (B) Treda the alien has Trichet's syndrome and Sommet's disease.
- (C) Treda the alien has Trichet's syndrome and Morad's disease.

(D) Treda the alien has a Humel infection and Morad's disease.

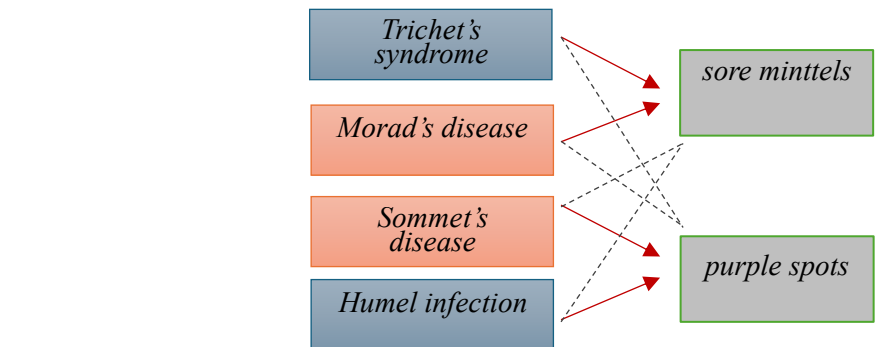
(E) Treda the alien has a Humel infection and Sommet's disease.

(F) Treda the alien has Morad's disease and Sommet's disease.

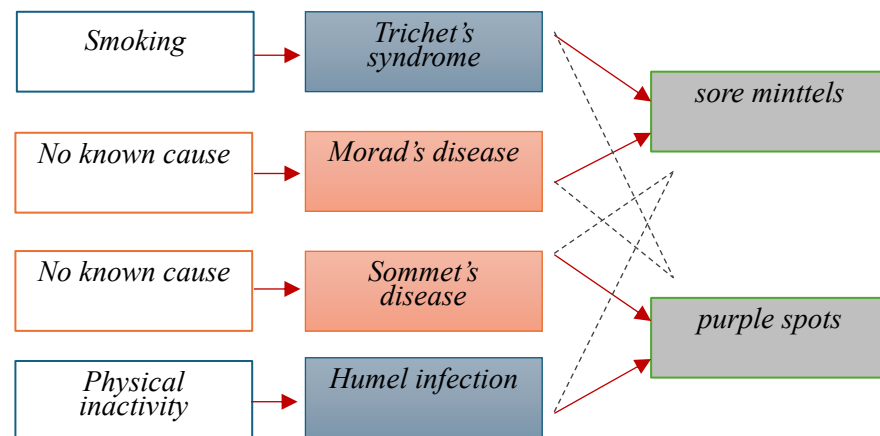
The structure of the information in the scenarios in this experiment is illustrated in **Figure 2.4**.

Figure 2.4

Diagrams for the Causal Structure in Experiment 2a



Condition = No extra information



Condition = Antecedent-controllability information

Note. Diagrams of the disease domain are presented. Red arrows indicated a cause-effect relation. Dotted lines indicated that the diseases do not lead to the symptoms. These diagrams were for illustration purposes only, participants were not provided with these diagrams during the experiment.

Similar to Experiment 1, Option A above refers to the two causes that are the results of antecedent-controllable factors, and Option F refers to the two causes that are the results of antecedent-uncontrollable factors. Options B and

D each contain one antecedent-controllable and one antecedent-uncontrollable cause. Options C and E are causally insufficient because they only account for one symptom, not two.

In this structure, we suggest people attempt to simulate models that capture the information that Trichet's and Morad's cause sore minttels, not purple spots; Trichet's is the result of smoking, and there is no known cause for Morad's disease:

Trichet: sore minttels, not purple spots result of smoking

Morad: sore minttels, not purple spots no known cause

They may also attempt to construct models that capture the information that Sommet's and Humel's cause purple spots, not sore minttels; Humel's is the result of physical inactivity, but there is no known cause for Sommet's:

Sommet: not sore minttels, purple spots no known cause

Humel: not sore minttels, purple spots result of physical inactivity

When they are asked to choose the most satisfying explanation from the set of options, there may seem to be little rational means to choose between the four causally eligible explanations. They may fall back on a short-cut heuristic, to choose the causes that are the result of actionable elements.

Suppose you are influenced by antecedent-controllable actions leading to a negative outcome that is within the protagonist's control, which in the disease scenario means whether a disease is caused by controllable actions. In that case, you may be inclined to choose the explanation that Treda the alien has Trichet's syndrome and a Humel infection, given that Trichet's syndrome is the result of smoking, and a Humel infection is the result of physical inactivity.

In this experiment, we tested the effect of antecedent-controllable causes. We gave specific causes to convey antecedent-controllable factors such as “the disease is the result of smoking”. To describe antecedent-uncontrollable factors, we told the participants that there was no known cause such as “there is no known cause for the disease”. Our aim in this experiment was to investigate the effect of explanations with antecedent-controllable factors and whether people would prefer explanations with known antecedent-controllable causes to explanations with no known causes. We predicted that participants would prefer explanations with only antecedent-controllable factors over the other ones with only antecedent-uncontrollable factors or a mix of controllable and uncontrollable ones. As in Experiment 1, we tested also the condition where no extra information was given, where people read competing explanations without antecedent-controllability information and we expected they would select one by chance, or based on superficial strategies such as the first mentioned causes.

Method

Participants

The participants were 187 members of the general public (68 men, 106 women, 4 non-binary individuals, and 9 who preferred not to say). Their ages ranged from 18 to 76, with a mean age of 40 years and no participants were eliminated. The estimated sample size was the same as in Experiment 1. Each participant was paid £0.45 UK sterling for participation based on an estimated three-minute completion time.

Design

As in Experiment 1, a between-groups design was adopted. The independent variable was the antecedent-controllability information which was manipulated between groups. The participants were randomly assigned to one of the two groups: no extra information ($n = 101$), and antecedent-controllability information ($n = 86$). We measured their choice of the most satisfying explanation by asking “What do you think is the most satisfying explanation for the symptoms that Treda is exhibiting?” from a set of six explanations, from which participants could choose one only.

Materials & Procedure

We adapted materials from Experiment 1, see **Figure 2.4**. Again, we did not randomise the order of information about diseases provided in the scenario so that we could check if there is an order effect. In the disease domain, participants in the controllability information condition read that Trichet’s syndrome and a Humel infection were the results of smoking and physical inactivity and that there was no known cause for Morad’s disease and Sommet’s disease. Participants were given six explanations based on different combinations of two causes, the six explanations were presented in a different randomized order for each participant. As in Experiment 1, two explanations were only able to explain one symptom and were causally insufficient. We treated them as fillers, and we predicted that the participants would choose from the other four causally eligible combinations. See Appendix 2.3 for the materials in the farming domain.

In the antecedent-controllability information condition, we predicted that people would prefer the combination of Trichet's syndrome and a Humel infection, which would indicate a preference for antecedent-controllable causes, rather than any of the other explanations. In the no extra information condition, we expect participants will choose the first mentioned causes, as they did in Experiment 1. The procedure of Experiment 2a was the same as in Experiment 1.

Results and Discussion

First, we checked that no differences were arising merely because of the order (first received scenario, second received scenario), see Appendix 2.4 for supplementary analysis. We combined the responses from the two orders for further analysis. Two 6 X 2 Chi-square tests compared the distribution of explanation preferences between the two experimental conditions and showed that there was a significant relationship between controllability information and people's preferred explanations in the disease domain, $\chi^2(5) = 14.89, p = .01, V = 0.13$; and in the farming domain, $\chi^2(5) = 31.47, p < .001, V = 0.18$. See **Table 2.2** for the frequencies of explanation preferences.

Table 2.2.*Frequencies of Explanation Preferences in Experiment 2a*

Explanation options	Trichet & Humel (Both causes controllable)	Morad & Sommet (Neither cause controllable)	Trichet & Sommet (First mentioned causes)	Morad & Humel (Second mentioned causes)	Trichet & Morad (Filler)	Humel & Sommet (Filler)
Domain = Disease						
Antecedent Controllability information	25 (29%)	25 (29%)	15 (17%)	11 (13%)	5 (6%)	5 (6%)
No extra information	24 (24%)	13 (13%)	33 (33%)	24 (24%)	4 (4%)	3 (3%)
Domain = Farming						
Antecedent Controllability Information	29 (34%)	25 (29%)	10 (12%)	11 (13%)	5 (6%)	6 (7%)
No extra information	12 (12%)	16 (16%)	44 (44%)	12 (12%)	5 (5%)	12 (12%)
Both domains combined						
Antecedent Controllability Information	54 (31%)	50 (29%)	25 (15%)	22 (13%)	10 (6%)	11 (6%)
No extra information	36 (18%)	29 (14%)	77 (38%)	36 (18%)	9 (4%)	15 (7%)

Note. All explanation options were shown to the participants in a randomised order during the experiment. No antecedent-controllability information was provided in the no extra information condition, but we used here the same label to identify the explanations to match the antecedent-controllability information condition. The percentages were in parentheses.

Comparison Between Conditions

We checked whether there were differences arising because of the domain content and there was no difference in the distribution of responses between the two domains when controllability information was given, however, there was a difference in responses to the disease and farming

domains in the condition in which no controllability information was given, see Appendix 2.4 for supplementary analysis, and see **Table 2.2**. We note that no differences between domains were found in any of the conditions of the other five experiments in this chapter, so there do not appear to be systematic differences in the domains. Hence, we compared participants' choices of explanations between the two conditions separately for each domain, and we also combined the responses from the two domains for further analysis for ease of comparison to the previous experiment.

To compare the choices of each of the four causally eligible explanations between the two experimental conditions, we carried out four Mann-Whitney U tests with a Bonferroni corrected alpha level of $p < .0125$ on participants' responses to the disease domain, and a similar set of comparisons on their responses to the farming domain, and then we carried out a similar set of comparisons on participants' mean responses to both domains. Participants chose the explanation with two antecedent-controllable causes more in the antecedent-controllability condition compared to the no information condition, in the farming domain, 34% vs 12%, $U = 5291.5$, $z = 3.588$, $p = .0003$, $r = 0.39$, although not in the disease domain, 29% vs 24%, $U = 4573.5$, $z = 0.820$, $p = .412$. For both domains combined, the difference was significant, 31% vs 18%, $U = 5179$, $z = 2.652$, $p = .008$, $r = 0.19$. They did not choose the explanation with two antecedent-uncontrollable causes more in the antecedent-controllability information condition than the no information condition, in the farming domain, 29% vs 16%, $U = 4917.5$, $z = 2.173$, $p = .030$, but they chose the explanation with two antecedent-uncontrollable causes more in the disease domain, 29% vs 13%, $U = 5046.5$, $z = 2.736$, $p = .006$, $r = 0.20$, and for both

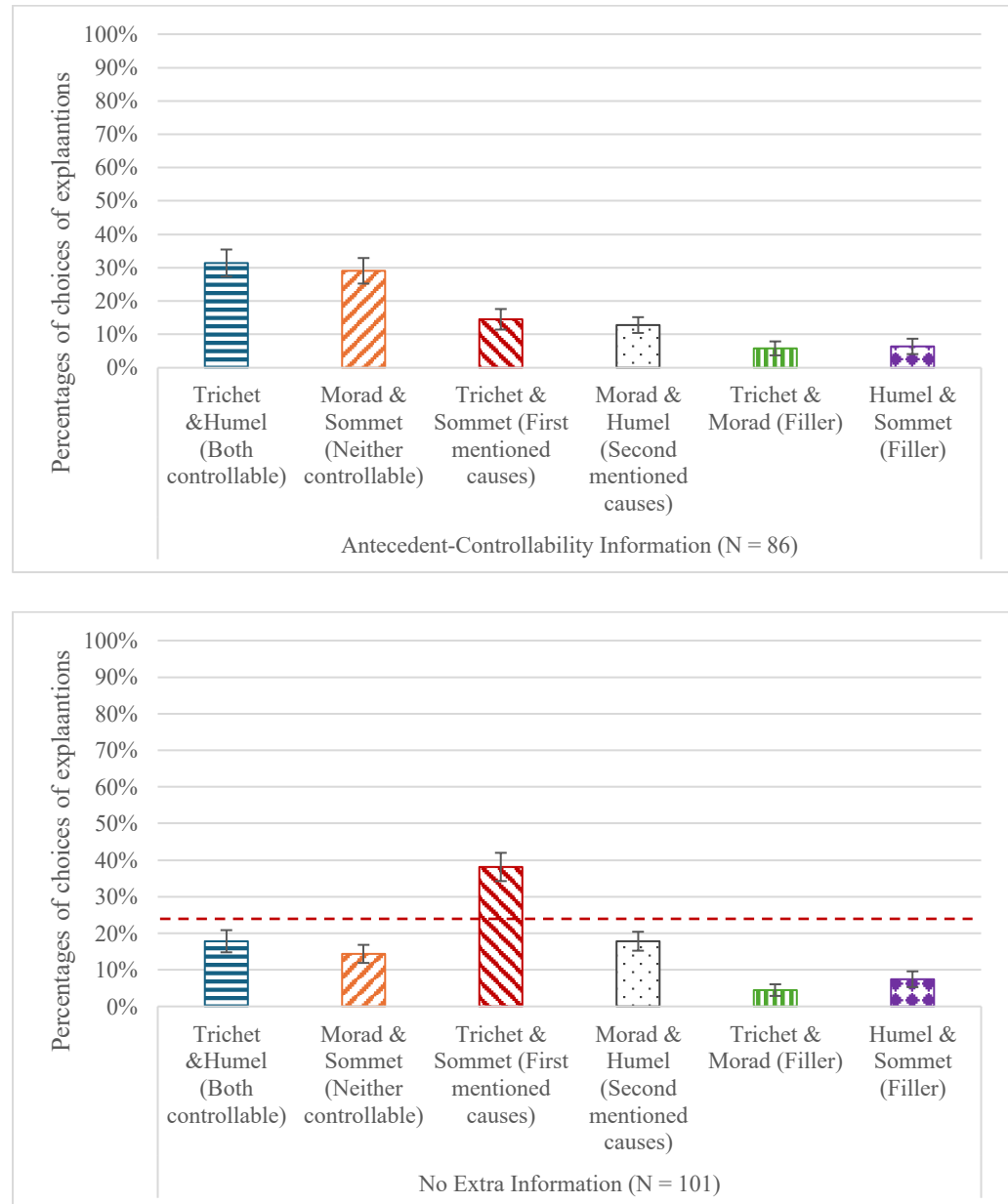
domains combined, 29% vs 14%, $U = 5261$, $z = 2.962$, $p = .003$, $r = 0.22$.

They chose the explanation with the two first mentioned causes more in the no extra information condition than the antecedent-controllability information condition, in the farming domain, 44% vs 12%, $U = 2956$, $z = 4.790$, $p < .0001$, $r = 0.35$, not in the disease domain, 33% vs 17%, $U = 3681.5$, $z = 2.370$, $p = .018$, and for both domains combined the difference was not significant on the corrected alpha, 38% vs 15%, $U = 2875$, $z = 4.538$, $p = .033$. They chose the explanation with the two second-mentioned causes to the same degree in both conditions, in the farming domain, 12% vs 13%, $U = 4382.5$, $z = 0.188$, $p = .851$, and in the disease domain, 24% vs 13%, $U = 3866.5$, $z = 1.912$, $p = .056$, and for both domains combined, 18% vs 13%, $U = 3970$, $z = 1.272$, $p = .203$. Overall, we saw an increase in selections of both antecedent-controllable and antecedent-uncontrollable explanations in the antecedent-controllability information condition compared to the no extra information condition.

We also compared participants' choices of each of the four causally eligible explanations to a chance level of 25%, see Appendix 2.4 for supplementary analysis. In the antecedent-controllability condition, participants' choices of the explanation with antecedent-controllable causes were the same as chance, and so too were their choices of the explanation with antecedent-uncontrollable causes, the other explanations were chosen below chance level; in contrast, in the no information condition, participants' choices of the explanation with the first-mentioned cause in each pair in the scenario were higher than chance, the other explanations were chosen lower chance level, see **Figure 2.5**.

Figure 2.5

The Percentages of Choices of all Explanations in Experiment 2a for the Antecedent-Controllability Information Condition (top) and for the No Extra Information Condition (bottom)



Note. The percentages were combined for both the disease and farming domains (the disease domain names are used here to illustrate). The red dotted line in the bottom graph represents a 25% chance level. Error bars are the standard error of the mean.

The results from Experiment 2a indicated that in the no extra information condition, people preferred the explanation with the first-mentioned cause in each pair in the scenario, which was consistent with Experiment 1. In the antecedent-controllability condition, we observed that participants had a tendency to select the explanation with antecedent-controllable factors compared to the control condition, overall, the difference arises in the farming domain but not in the disease domain. However, we also observed an increase in the selection of the explanation with antecedent-uncontrollable factors between the two conditions in the disease domain, and for both domains combined. Results from this experiment were somewhat inconsistent with our prediction as we hypothesised that participants would prefer the explanation with antecedent-controllable factors only. Instead, they indicated that people were satisfied with antecedent-controllable explanations but also with antecedent-uncontrollable explanations. The result is consistent with previous evidence that indicates when people spontaneously create explanations for a negative event, they generate an equal number of causal explanations for controllable and uncontrollable events (McEleney & Byrne, 2007).

The findings in this experiment suggest there may be a difference between consequent-controllability and antecedent-controllability information in their effect on causal explanation selection. A key difference is that consequent-controllability information provides people with clear actions to follow up with, e.g., to treat the symptoms with surgery. They indicate subsequent steps after the event in question that can be acted upon. Antecedent-controllability information, on the other hand, suggests

controllable antecedents and does not guarantee actions to follow to fix the problem, for example, even when people read that a disease was the result of smoking, there is no guarantee that quitting smoking could improve the symptoms. Hence, the antecedent-controllability effect we observed in this experiment was not as clear-cut compared to the consequent-controllability effect we observed in Experiment 1, with an additional presence of a tendency to select explanations in which both causes were antecedent-uncontrollable.

One potential reason for the findings is that perhaps when people heard there was no known cause for the problem, they did not interpret this information to mean the cause of the event was antecedent-uncontrollable. In the next experiment, we aim to try to ensure that the information provided to the participants can be readily interpreted to mean that the causes of the outcome were uncontrollable. We followed up with Experiment 2b, where we tested the effect of antecedent-controllability by describing the uncontrollable causes using specific factors rated by participants as uncontrollable instead of stating there is no known cause so that we could better match the information in the antecedent-controllability condition.

Experiment 2b

The aim of this experiment was to test whether providing specific antecedent-uncontrollable information, instead of stating there is no known cause, will decrease the choices of uncontrollable explanations. In Experiment 2a, we stated “there is no known cause for the disease” to convey antecedent-uncontrollability. In this experiment, we provided instead specific information such as “the disease is the result of genetic predisposition” to attempt to

indicate the lack of antecedent-controllability in a clearer way. If we observe a tendency to choose the explanation with antecedent-controllable causes, and no tendency to choose the explanation with antecedent-uncontrollable causes, then such results would provide a clear indication of an antecedent-controllability effect in explanation selection.

Suppose you are trying to understand the alien's medical problem, and the information of symptoms and diseases is the same as the no information condition in Experiment 1 and Experiment 2a, you may decide to choose the first mentioned diseases, as the previous experiments indicate.

But suppose now instead you are told:

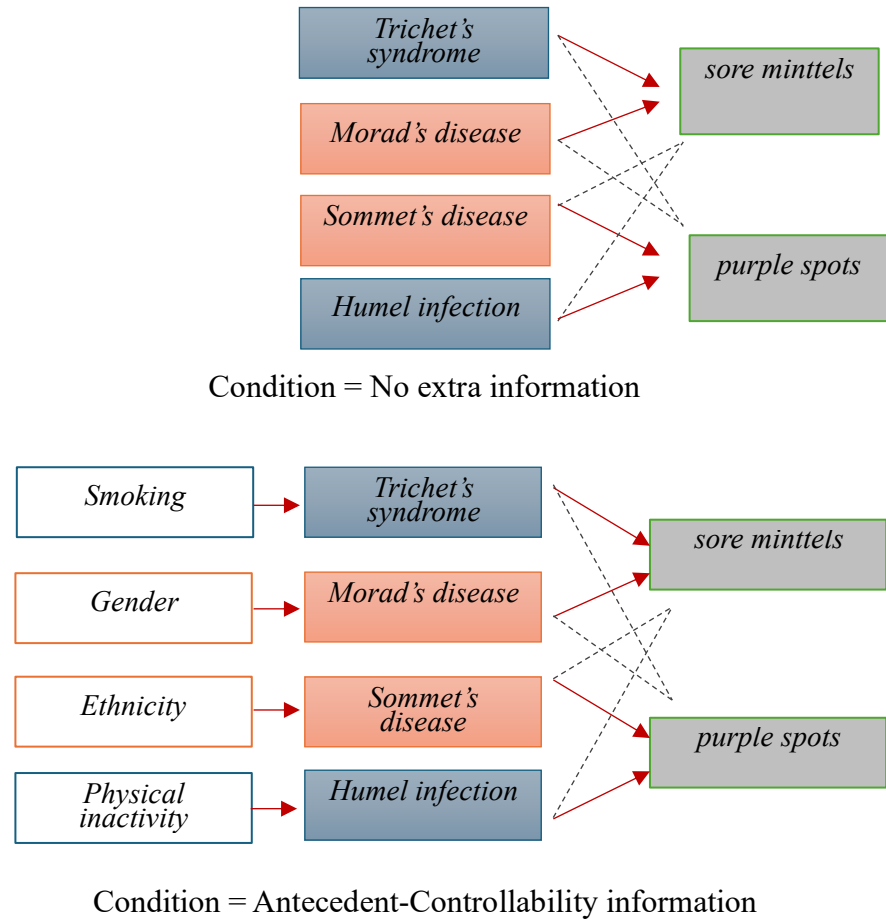
Both Trichet's syndrome and Morad's disease always cause sore minttels, but they never cause purple spots. Trichet's syndrome is the result of smoking. Morad's disease is the result of gender. Both Sommet's disease and a Humel infection always cause purple spots, but they never cause sore minttels. Sommet's disease is the result of ethnicity. A Humel infection is the result of physical inactivity. What do you think is the most satisfying explanation for the symptoms that Treda is exhibiting?

- (A) Treda the alien has Trichet's syndrome and a Humel infection.
- (B) Treda the alien has Trichet's syndrome and Sommet's disease.
- (C) Treda the alien has Trichet's syndrome and Morad's disease.
- (D) Treda the alien has a Humel infection and Morad's disease.
- (E) Treda the alien has a Humel infection and Sommet's disease.
- (F) Treda the alien has Morad's disease and Sommet's disease.

See **Figure 2.6** for the structure of the scenarios in Experiment 2b.

Figure 2.6

Diagrams for the Causal Structure in Experiment 2b



Note. The diagrams for the disease domain are presented. Red arrows indicated a causal relation. Dotted lines indicated that the diseases do not lead to the symptoms. Participants did not see these diagrams during the experiment. The controllable and uncontrollable antecedents were rated by participants in a material check, see Appendix 2.5.

In this structure, we suggest people attempt to simulate the causal model as:

Trichet:	sore minttels, not purple spots	result of smoking
Morad:	sore minttels, not purple spots	result of gender

Sommet: not sore minttels, purple spots result of ethnicity

Humel: not sore minttels, purple spots result of physical inactivity

If you are influenced by antecedent-controllability during explanation selection, then you may be inclined to choose the explanation that Treda the alien has Trichet's syndrome and a Humel infection, given that Trichet's syndrome is the result of smoking, and a Humel infection is the result of physical inactivity. If the specific information about uncontrollable causes, such as gender and ethnicity, ensures that the causes are interpreted as uncontrollable, then you will no longer choose the antecedent-uncontrollable explanation that Treda the alien has Morad's disease and Sommet's disease, given they are the result of gender and ethnicity. We expect that participants will tend to choose the explanation with the two antecedent-controllable diseases and will not choose the explanation with the two antecedent-uncontrollable diseases.

Experiment 2a has shown that participants did not choose explanations with antecedent-controllable factors more than antecedent-uncontrollable ones for both domains and that antecedent-uncontrollable factors, as well as antecedent-controllable factors, increased explanation satisfaction. Experiment 2b aimed to test if this preference still existed when the competing explanations contained specific factors as antecedent-uncontrollable, such as gender and ethnicity. We hypothesised that participants would choose the antecedent-uncontrollable explanations less and choose the antecedent-controllable explanations instead. As in the previous experiments, participants were given tasks in two domains: disease and farming.

Method

Participants

The participants were 174 volunteers (61 men, 111 women, 1 non-binary individual, and 1 preferred not to say). Their ages ranged from 18 to 76 years, with a mean of 39 years, and no participants were eliminated. The estimated sample size was the same as in previous experiments. Each participant was paid £0.45 UK sterling for participation based on an estimated three-minute completion time.

Design

A between-groups design was adopted. The independent variable was the antecedent-controllability information which was manipulated between groups. The participants were randomly assigned to one of the two groups: no extra information ($n = 88$), and antecedent-controllability information ($n = 86$). We measured participants' choice of the most satisfying explanation by asking, "What do you think is the most satisfying explanation for the symptoms that Treda is exhibiting?" from a set of six explanations.

Materials & Procedure

We first carried out a materials check, which tested a small group of 10 participants from the general public on their judgments of the controllability of different factors in the disease and farming domain, to support the construction of experimental materials for Experiment 2b. See Appendix 2.5 for the materials and results of the materials check. In the disease domain, the following items were rated as most controllable: *Smoking*, *Physical inactivity*,

and *Unhealthy diet*; and the following items were rated as most uncontrollable: *Gender*, *Ethnicity*, and *Genetic predisposition* (The information for the farming domain is provided in Appendix 2.5). Participants in the antecedent-controllability information condition read that Trichet's syndrome was the result of smoking, a Humel infection was the result of physical inactivity, Morad's disease was the result of gender, and Sommet's disease was the result of ethnicity.

Again, participants were given six explanations based on different combinations of two causes in a different randomized order for each participant. Once again, we did not randomise the order of information about the diseases provided in the scenario so that we could check for an order effect. The explanation choices were the same as in Experiment 2a. In the antecedent-controllability information condition, we predicted that people would prefer the combination of Trichet's syndrome and a Humel infection, which would indicate a preference for antecedent-controllable causes, and there would be no preference for antecedent-uncontrollable causes. In the no extra information condition, we expect participants will choose the first mentioned causes, as they did in previous experiments. The procedure was the same as in previous experiments, and the experiment took about 3 minutes to complete.

Results and Discussion

First, we checked that no differences were arising merely because of the order (first received scenario, second received scenario), and there were no order effects. See Appendix 2.4 for supplementary analysis. As a result, we combined responses from both orders. Two 6 X 2 Chi-square analyses showed

that there was a significant relationship between antecedent-controllability information and participants' preferred explanations in the disease domain, $\chi^2(5) = 16.27, p = .006, V = 0.14$; and in the farming domain, $\chi^2(5) = 30.18, p < .001, V = 0.19$. See **Table 2.3** for the frequencies of explanation preferences.

Table 2.3

Frequencies of Explanation Preferences in Experiment 2b

Explanation options	Trichet & Humel (Both causes controllable)	Morad & Sommet (Neither cause controllable)	Trichet & Sommet (First mentioned causes)	Morad & Humel (Second mentioned causes)	Trichet & Morad (Filler)	Humel & Sommet (Filler)
Domain = Disease						
Antecedent Controllability information	22 (26%)	21 (24%)	10 (12%)	20 (23%)	7 (8%)	6 (7%)
No extra information	21 (24%)	8 (9%)	30 (34%)	16 (18%)	7 (8%)	6 (7%)
Domain = Farming						
Antecedent Controllability Information	21 (24%)	19 (22%)	13 (15%)	16 (19%)	13 (15%)	4 (5%)
No extra information	13 (15%)	9 (10%)	43 (49%)	7 (8%)	6 (7%)	10 (11%)
Both domains combined						
Antecedent Controllability Information	43 (25%)	40 (23%)	23 (13%)	36 (21%)	20 (12%)	10 (6%)
No extra information	34 (19%)	17 (10%)	73 (41%)	23 (13%)	13 (7%)	16 (9%)

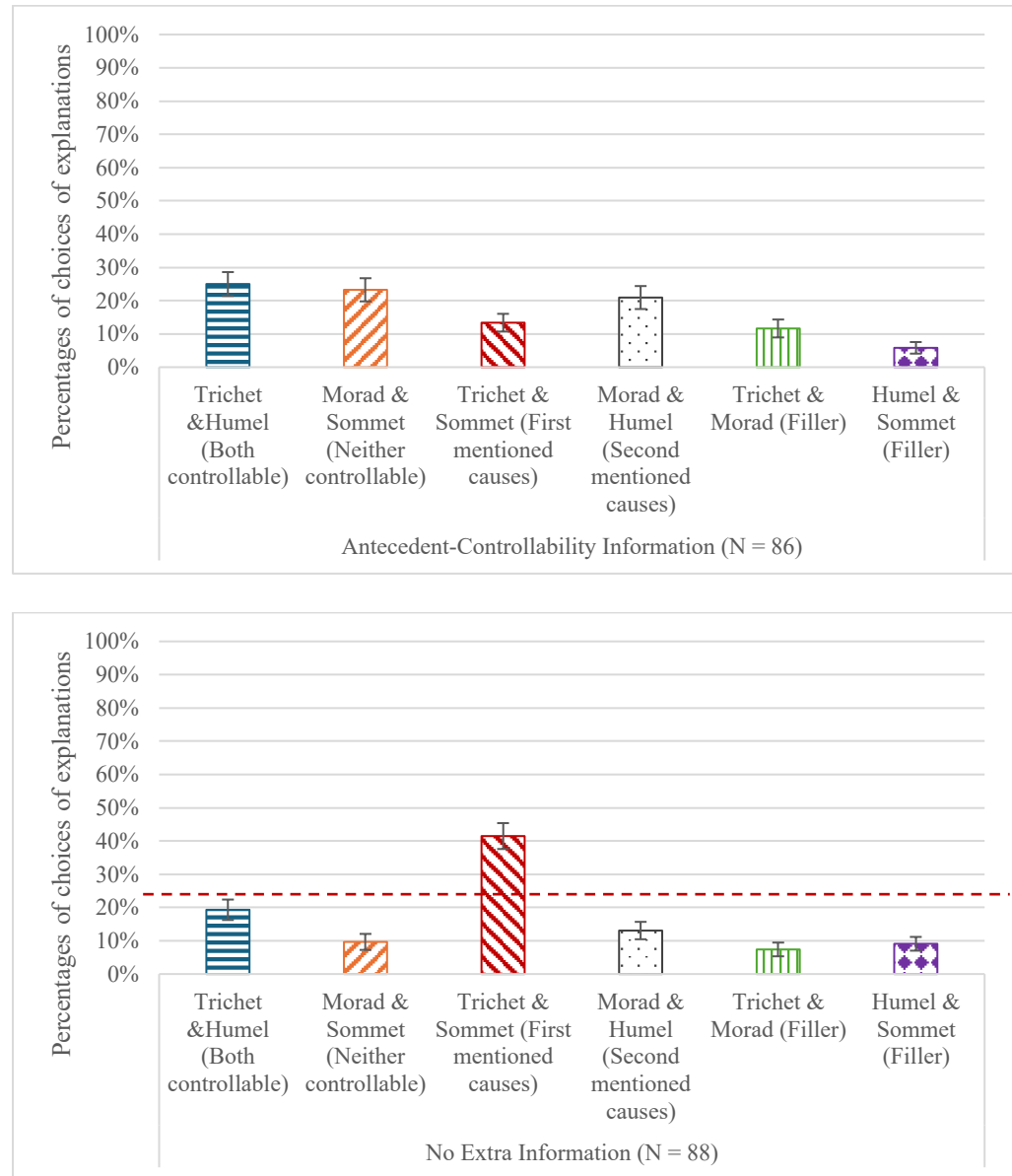
Note. All explanation options were shown to the participants in a randomised order during the experiment. No antecedent-controllability information was provided in the no extra information condition, but we used here the same label to identify the explanations to match the antecedent-controllability information condition. The percentages were in parentheses.

Comparison between Conditions

First, we checked that no differences were arising merely because of the domain (disease and farming), and there were no domain effects. See Appendix 2.4 for supplementary analysis. As a result, we combined responses from both domains. To compare the choices of each of the four causally eligible explanations between the two experimental conditions, we carried out four Mann-Whitney U tests on participants' mean responses to both domains with a Bonferroni corrected alpha level of $p < .0125$. Participants did not choose the antecedent-controllable explanation more in the antecedent-controllability condition than in the no extra information condition, 25% v 19%, $U = 4084$, $z = 1.06$, $p = .289$; they chose the antecedent-uncontrollable explanation more in the antecedent-controllability condition than in the no extra information condition, 23% v 10%, $U = 4610.5$, $z = 3.188$, $p = .001$, $r = 0.24$; they chose the first mentioned causes less in the antecedent-controllability information condition than the no extra information condition, 13% v 41%, $U = 2177.5$, $z = 5.446$, $p < .0001$, $r = 0.41$; and they chose the second mentioned causes to the same degree in both conditions, 21% v 13%, $U = 4201.5$, $z = 1.59$, $p = .112$. See **Figure 2.7** for the distribution of percentages of choices for both conditions.

Figure 2.7

The Percentages of Choices of All Explanations in Experiment 2b for the Antecedent-Controllability Information Condition (top) and the No Extra Information Condition (bottom)



Note. The percentages were combined for both the disease and farming domains (the disease names are used here to illustrate). The red dotted line on the bottom graph represents a 25% chance level. Error bars are the standard error of the mean.

Overall, the increase in choices of the antecedent-controllable explanation given antecedent-controllable information versus no information observed in Experiment 2a was not replicated in this experiment, even though the same controllable factors were provided. As **Table 2.3** shows, there is again no effect in the disease domain and a reduced difference in the farming domain. In contrast, the increase in choices of the antecedent-uncontrollable explanation given antecedent-controllability information versus no information observed in Experiment 2a was replicated in this experiment, this time even when the antecedent-uncontrollable factor was described in specific terms, e.g., “gender causes Trichet’s disease”, rather than general terms, “there is no known cause for Trichet’s disease”.

The results show that, overall, there was no increase in choices of the antecedent-controllable explanation given antecedent-controllability information compared to no information unlike Experiment 2a, even though the same controllable factors were provided (e.g., smoking, physical inactivity). We can conclude from the results of both Experiment 2a and 2b that there is no preference for antecedent-controllable explanations. Contrary to our predictions, we have observed a preference for antecedent-uncontrollable explanations when antecedent-controllability is provided compared to no information, whether the uncontrollable factor is described in specific terms, e.g., “gender causes Trichet’s disease” as in Experiment 2b or in general terms, “there is no known cause for Trichet’s disease” in Experiment 2a.

We note that in all three experiments, Experiment 1, 2a, and 2b, when participants are given no extra information, they consistently exhibit a

tendency to select the explanation that contains the first named disease in each pair (Trichet and Sommet), which we interpret as a superficial strategy induced by there being no reason to choose any one of the four causally viable explanations above the other. This artefact of order is eliminated when participants are provided with consequent-controllability or antecedent-controllability information.

Overall, the results from Experiments 1, 2a, and 2b indicated that there is a consequent-controllability effect, no antecedent-controllability effect as we originally conceived it but instead an antecedent-uncontrollability effect when it comes to causal explanation selection. To attempt to explain the results, we return to the definition and evidence that led us to examine antecedent-controllability and consequent-controllability, and we highlight again that there are key differences between consequent-controllability and antecedent-controllability. Consequent-controllability occurs after the negative event and is perhaps most similar to the idea of actionability in XAI to propose recourse that enables “the ability of a person to change the decision of the model through actionable input variables” (Ustun et al., 2019, p. 10). Hence, consequent-controllability focuses on changing the negative outcome after it has happened, regardless of what led to the negative event. Antecedent-controllability, on the contrary, is antecedent to the negative event, and evidence indicates that in their counterfactual thinking people often focus on events leading up to the outcome that contain voluntary actions within the protagonist’s control. The key difference lies in the temporal relations they have with the negative outcome in question. Why might antecedent-uncontrollability affect explanation selection? Comparisons of counterfactual

thinking about the past and prefactual thinking about the future suggest that when participants were asked to generate counterfactuals about the past, they focused on uncontrollable elements, and when they were asked to generate prefactuals about the future, they focused on controllable elements (Ferrante et al., 2013). Given that we provided information about controllable and uncontrollable antecedents that occurred prior to the cause, participants may have engaged in counterfactual thinking, which may have led them to focus more on the uncontrollable factors. We suggest that there is an important role for actionability as operationalised in terms of consequent-controllability-in explanations.

In Experiments 1, 2a, and 2b, we tested consequent-controllability and antecedent-controllability hypotheses when all the explanations were equally simple, i.e., each of the explanations suggests two diseases to explain the symptoms. In the next two experiments, we tested consequent-controllability and antecedent-controllability for explanations that differed in their simplicity. There are many qualities that guide explanatory preferences, such as simplicity and coherence. In the real world, satisfying explanations need to incorporate various qualities. Hence, we wish to examine whether people prefer an explanation that has consequent-controllable causes even when it is not the simplest explanation. In Experiment 3a, we test the effect of consequent-controllability information versus simplicity, to investigate if the preference for consequent-controllable explanations still holds when the competing explanations include explanations that are consequent-uncontrollable but simpler. In Experiment 3b, for completeness, we compared the effect of antecedent-controllability versus simplicity.

Experiment 3a

This experiment aimed to compare consequent-controllability and simplicity information in causal explanation selections. We wished to establish whether participants prefer simpler explanations even when they contain factors that are consequent-uncontrollable, i.e., whether participants prefer explanations that are consequent-controllable but more complex or explanations that are consequent-uncontrollable but simpler. We used a different scenario where the explanations differ in terms of simplicity, based on the original study of Lombrozo (2007), as we described in the introduction section earlier. The scenario below resembles the examples earlier such as the names of diseases and symptoms, but there is a difference in simplicity, as the symptoms can be explained by either one disease or two diseases.

Suppose you are a doctor trying to understand an alien's medical problem. The alien, Treda, has two symptoms: Treda's minttels are sore and Treda has developed purple spots.

Tritchett's syndrome always causes both sore minttels and purple spots. Morad's disease always causes sore minttels, but the disease never causes purple spots. When an alien has a Humel infection, that alien will always develop purple spots, but the infection will never cause sore minttels. Nothing else is known to cause an alien's minttels to be sore or the development of purple spots. What do you think is the most satisfying explanation for the symptoms that Treda is exhibiting?

- (A) Treda the alien has Trichet's syndrome.
- (B) Treda the alien has Morad's disease.
- (C) Treda the alien has a Humel infection.

(D) Treda the alien has Trichet's syndrome and Morad's disease.

(E) Treda the alien has Trichet's syndrome and a Humel infection.

(F) Treda the alien has Morad's disease and a Humel infection.

Recall that as described earlier, the symptoms can be explained solely with one single disease (option A above, Trichet's syndrome) or two diseases (option F above, Morad's disease and a Humel infection). Hence Trichet's syndrome has been considered the simple explanation whereas Morad's disease and a Humel infection have been considered the complex explanation (e.g., Lombrozo 2007). Note that the explanations in option D and option E also account for both symptoms but they each subsume A, which is itself sufficient, and so they are over-determined. In contrast, options B and C are causally insufficient explanations because they do not account for both symptoms. As described earlier previous findings have shown that participants decide that option A is the more satisfying explanation, as it is simpler compared to other explanations – it accounts for both symptoms with only one disease (Bonawitz & Lombrozo, 2012; Blanchard et al., 2018a; Johnson et al., 2019; Lombrozo, 2007; Pacer & Lombrozo, 2017; Vrantsidis & Lombrozo, 2022; Walker et al., 2017). In this control condition, We expect that we will replicate the findings and that participants will tend to choose the simpler option A.

But suppose now instead you are told:

Trichet's syndrome always causes both sore minttels and purple spots. There is no treatment for Trichet's Syndrome. Morad's disease always causes sore minttels, but the disease never causes purple spots. Morad's disease can be treated with medicine. When an alien has a Humel infection, that alien will always develop purple spots, but the infection will never cause sore minttels. A

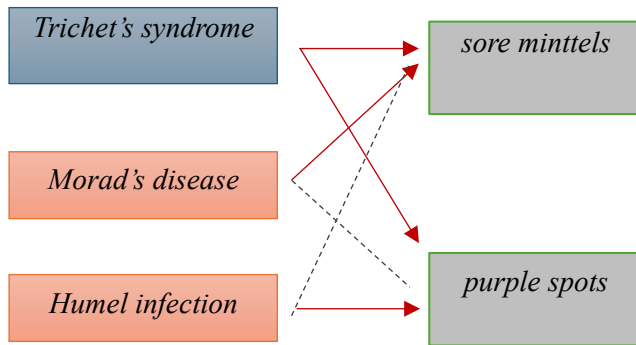
Humel infection can be treated with surgery. Nothing else is known to cause an alien's minttels to be sore or the development of purple spots. what do you think is the most satisfying explanation for the symptoms that Treda is exhibiting?

- (A) Treda the alien has Trichet's syndrome.
- (B) Treda the alien has Morad's disease.
- (C) Treda the alien has a Humel infection.
- (D) Treda the alien has Trichet's syndrome and Morad's disease.
- (E) Treda the alien has Trichet's syndrome and a Humel infection.
- (F) Treda the alien has Morad's disease and a Humel infection.

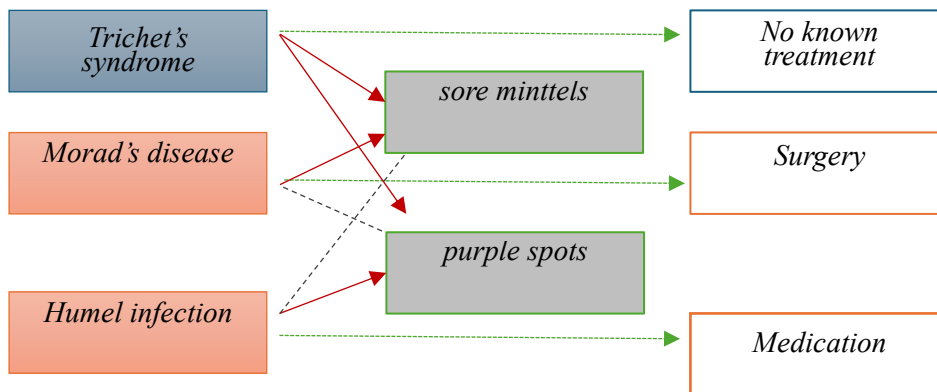
In this scenario, the simple explanation, Trichet's syndrome, contains a cause that is consequent-uncontrollable; whereas the complex explanation, Morad's disease and a Humel infection, contains causes both of which are consequent-controllable. **Figure 2.8** illustrates the causal structure of the scenarios in this experiment.

Figure 2.8

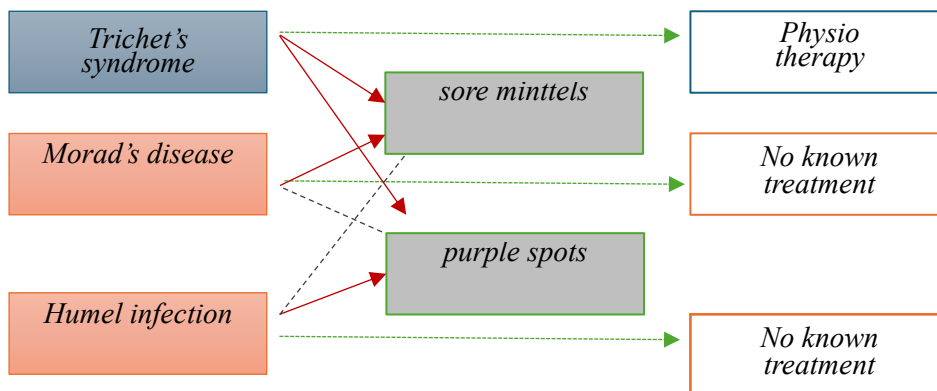
Diagrams of the Causal Structure in Experiment 3a



Condition = No extra information (Lombrozo, 2007)



Condition = Simple cause uncontrollable



Condition = Simple cause controllable

Note. The diagrams of the disease domain are presented. Red arrows indicated a causal relation. Dotted lines indicated that the diseases do not lead to the symptoms. The participants did not see the diagram during the experiment.

We found in Experiment 1 that people prefer consequent-controllable explanations, and in this experiment, we aim to pit consequent-controllability against simplicity. If the consequent-controllability effect still holds, then the participants may be inclined to choose the explanation that Treda has Morad's disease and a Humel infection, given that Morad's disease can be treated with medication and a Humel infection can be treated with surgery. If the consequent-controllability effect is strong enough to over-ride the simplicity preference, we expect that participants will tend to choose the explanation with consequent-controllable factors even though it is the more complex explanation (that Treda has Morad's and Humel), more often compared to the simple but consequent-uncontrollable explanation (that Treda has Trichet's). If instead consequent-controllability and simplicity are equally strong and operating against each other, we expect there to be a similar frequency of choices of both. Hence, we investigated whether people's preference for the simple explanation would be decreased when the simple explanation contains consequent-uncontrollable factors.

For completeness, we also included a third condition in which the simple explanation was also the most consequent-controllable, and the complex explanation contained causes that were consequent-uncontrollable. We investigated whether this preference for the simple explanation would be increased when it contained consequent-controllable causes, albeit the preference for simplicity was already very high.

In this experiment, consequent-controllability information was given in the same way as in Experiment 1. Given past research on scenarios where the simplicity preference was challenged (e.g., Johnson-Laird et al., 2004;

Khemlani & Johnson-Laird, 2011; Lombrozo, 2007; Zemla et al., 2017; 2023), and our findings from Experiment 1, we hypothesised that the preference for the simple explanation would decrease when it was described to be consequent-uncontrollable.

Method

Participants

The participants were 217 volunteers (87 men, 123 women, 1 non-binary individual, and 6 preferred not to say). Their ages ranged from 18 to 72 years, with a mean of 38 years, and no participants were eliminated. An a-priori G*Power test showed that a sample size of 207 participants was required to detect a medium effect size. Each participant was paid £0.60 UK sterling for participation.

Design

A between-groups design was adopted. In this experiment, consequent-controllability information was manipulated between groups. The participants were randomly assigned to one of the three groups: no extra information ($n = 78$); simple cause uncontrollable and the other two causes were described as controllable ($n = 69$); and simple cause controllable and the other two causes were described as uncontrollable ($n = 70$). We measured their choice of the most satisfying explanation from a set of six explanations.

Materials & Procedure

Participants received scenarios for two domains, disease and farming. The materials for the no extra information condition are the same as Lombrozo (2007). We developed materials for other conditions (see **Figure 2.8**). See Appendix 2.3 for the materials in the farming domain. In the disease domain, participants in the simple cause uncontrollable condition read that there was no treatment for Trichet's syndrome and that Morad's disease and a Humel infection could be treated by surgery and medication respectively. For those in the simple cause controllable condition, they read that Trichet's syndrome could be treated with physiotherapy, while there was no treatment for Morad's disease and a Humel infection. The procedure of the current experiment was the same as in the previous experiment. The experiment took about four minutes to complete.

Results and Discussion

First, we checked whether differences were arising merely because of the order (first received scenario, second received scenario), and there were no differences. See Appendix 2.4 for supplementary analysis. As a result, we combined responses from both orders. Two 6 X 3 Chi-square tests showed that there was a significant relationship between the consequent-controllability information in the three conditions and participants' preferred explanations in the disease domain, $\chi^2(10) = 14.728, p = .0001, V = 0.08$; and in the farming domain, $\chi^2(10) = 11.091, p = .0009, V = 0.07$. See **Table 2.4** for the frequencies of explanation preferences.

Table 2.4.*Frequencies of Explanation Preferences in Experiment 3a*

Explanation options	Trichet (Simple)	Morad (Insufficient)	Humel (Insufficient)	Trichet & Morad (Over-determined)	Trichet & Humel (Over-determined)	Morad & Humel (Complex)
Domain = Disease						
No extra information	70 (90%)	0 (0%)	1 (1%)	3 (4%)	4 (5%)	0 (0%)
Simple cause uncontrollable	53 (77%)	1 (1%)	3 (4%)	4 (6%)	1 (1%)	7 (10%)
Simple cause controllable	60 (86%)	1 (1%)	1 (1%)	4 (6%)	1 (1%)	3 (4%)
Domain = Farming						
No extra Information	65 (83%)	2 (3%)	4 (5%)	3 (4%)	2 (3%)	2 (3%)
Simple cause uncontrollable	56 (81%)	0 (0%)	1 (1%)	5 (7%)	1 (1%)	6 (9%)
Simple cause controllable	55 (79%)	2 (3%)	0 (0%)	6 (9%)	1 (1%)	6 (9%)
Both domains combined						
No extra information	135 (87%)	2 (1%)	5 (3%)	6 (4%)	6 (4%)	2 (1%)
Simple cause uncontrollable	109 (79%)	1 (1%)	4 (3%)	9 (7%)	2 (1%)	13 (9%)
Simple cause controllable	115 (82%)	3 (2%)	1 (1%)	10 (7%)	2 (1%)	9 (6%)

Note. All explanation options were shown to the participants in a randomised order during the experiment. The percentages were in parentheses.

We compared the choices of the simple explanation to the other explanations. In the no extra information condition, 87% of participants chose the simple explanation of Trichet's syndrome, in the simple cause uncontrollable condition, 79% of participants chose the simple explanation, and in the simple cause controllable condition, 82% of participants chose the simple explanation. A Kruskal-Wallis test showed there was no difference in the choices of choices of the simple explanation across three conditions, $H(2) = 1.453, p = .483$. In other words, participants chose the simple explanation as

often in all experimental conditions. See Appendix 2.4 for a supplementary analysis of the choices of the complex explanation.

We hypothesised that people's preference for the simple explanation would be decreased when this simple explanation was described as an uncontrollable factor and the complex option controllable. However, the results showed that the manipulation of consequent-controllability did not affect people's simplicity preference, as the majority (around 80%) of participants still preferred the simple explanation. The results indicate that the preference for simplicity is stronger than the preference for consequent-controllability. We followed up with Experiment 3b, to pit antecedent-controllability against simplicity.

Experiment 3b

The aim of this experiment was to compare antecedent-controllability and simplicity information in causal explanation selections. Results from Experiment 2a indicated that there was a preference for antecedent-controllable explanations when participants were given antecedent-controllability information vs no information, but it occurred in one domain only, and this preference was not replicated in Experiment 2b. Hence, we concluded there is no clear antecedent-controllability effect. Instead, they unexpectedly showed an effect of antecedent-uncontrollability. Hence, we examine whether participants prefer explanations that are antecedent-uncontrollable but more complex. Recall in Experiment 3a, when no extra information is provided, people prefer the simple explanation and tend to choose option (A), Trichet's syndrome.

But suppose you are told:

Tritchett's syndrome, which is the result of gender, always causes both sore minttels and purple spots. Morad's disease, which is the result of physical inactivity, always causes sore minttels, but the disease never causes purple spots. When an alien has a Humel infection, which is the result of an unhealthy diet, that alien will always develop purple spots, but the infection will never cause sore minttels. Nothing else is known to cause an alien's minttels to be sore or the development of purple spots.

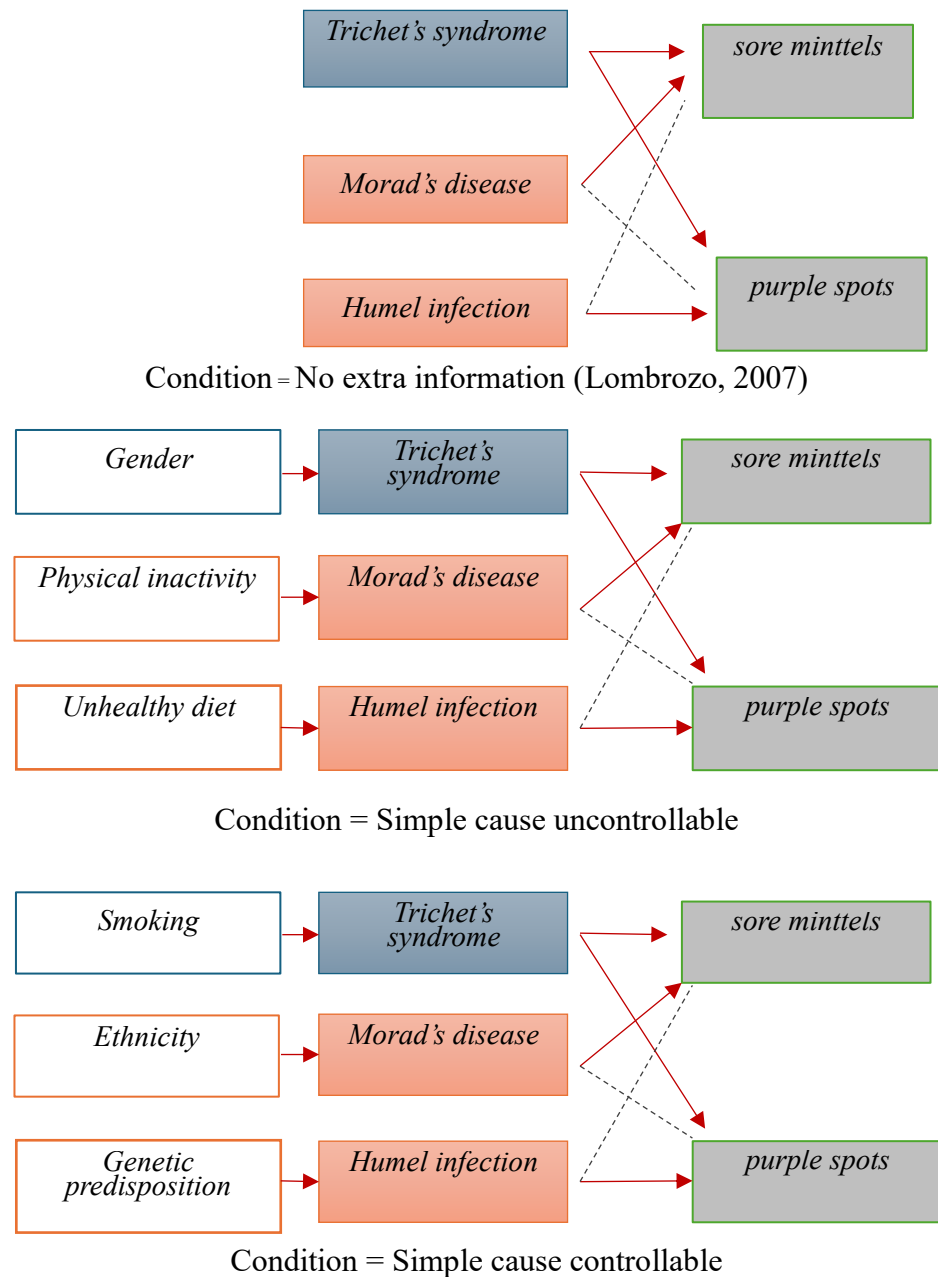
What do you think is the most satisfying explanation for the symptoms that Treda is exhibiting?

- (A) Treda the alien has Tritchett's syndrome.
- (B) Treda the alien has Morad's disease.
- (C) Treda the alien has a Humel infection.
- (D) Treda the alien has Tritchett's syndrome and Morad's disease.
- (E) Treda the alien has Tritchett's syndrome and a Humel infection.
- (F) Treda the alien has Morad's disease and a Humel infection.

In this version, the simple explanation contains an antecedent-uncontrollable cause, whereas the complex explanation contains antecedent-controllable causes. In the third version, the simple explanation contained an antecedent-controllable cause, whereas the complex explanation contained antecedent-uncontrollable causes. See **Figure 2.9** for the causal structure of the materials in this experiment.

Figure 2.9

Diagrams for the Causal Structure in Experiment 3b



Note. Diagrams for the disease domain are presented. Red arrows indicated a causal relation. Dotted lines indicated that the diseases do not lead to the symptoms. Participants were not provided with these diagrams during the experiment.

Different descriptions were used in Experiments 2a and 2b to convey antecedent-uncontrollability: we stated “there is no known cause for ...” in Experiment 2a, and we gave specific information rated by participants as uncontrollable, such as “the disease is the result of gender” in Experiment 2b. In this experiment, we used the same way to describe antecedent-controllability as in Experiment 2b, where we told people specific causes that were uncontrollable (e.g., the disease is the result of gender).

Method

Participants

The participants were 224 volunteers (74 men, 145 women, 1 non-binary individual, and 4 preferred not to say). Their ages ranged from 18 to 82 years, with a mean of 40 years and no participants were eliminated. The estimated sample size was the same as in Experiment 3a. Each participant was paid £0.50 UK sterling for participation.

Design

A between-groups design was adopted. The independent variable was the antecedent-controllability information which was manipulated between groups. The participants were randomly assigned to one of the three experimental conditions: No extra information ($n = 70$), simple cause uncontrollable ($n = 72$), i.e., the simple cause is uncontrollable and the other two causes are controllable, and simple cause controllable ($n = 82$), i.e., the simple cause is controllable and the other two causes are uncontrollable. We

measured participants' choice of the most satisfying explanation from a set of six explanations.

Materials & Procedure

Participants received scenarios for two domains, disease and farming. For the disease content, participants in the simple cause uncontrollable condition read that Trichet's syndrome was the result of gender, that Morad's disease was the result of physical inactivity and that a Humel infection was the result of an unhealthy diet. In the simple cause controllable condition, participants read that Trichet's syndrome was the result of smoking, while Morad's disease was the result of ethnicity, and a Humel infection was the result of genetic predisposition. See Appendix 2.3 for the experimental materials for the farming domain. See **Figure 2.9** for a diagram illustrating the causal relations of the scenarios in the experimental conditions.

As in Experiment 3a, target explanations (A vs. F) varied in simplicity. D and E are over-determined, and B and C are causally insufficient. The procedure of the current experiment was the same as in previous experiments. The experiment took about three minutes to complete.

Results and Discussion

First, we checked that no differences were arising merely because of the order (first received scenario, second received scenario), and there were no differences. See Appendix 2.4 for supplementary analysis. As a result, we combined responses from both orders. Two 6 X 3 Chi-square tests showed that there was a significant relationship between the antecedent-controllability

information in the three conditions and participants' preferred explanations, in the disease domain, $\chi^2(10) = 7.65, p = .0057, V = 0.06$, and in the farming domain, $\chi^2(10) = 7.76, p = .0053, V = 0.06$. See **Table 2.5** for the frequencies of explanation preferences.

Table 2.5.

Frequencies of Explanation Preferences in Experiment 3b

Explanation options	Trichet (Simple)	Morad (Insufficient)	Humel (Insufficient)	Trichet & Morad (Over-determined)	Trichet & Humel (Over-determined)	Morad & Humel (Complex)
Domain = Disease						
No extra information	60 (86%)	0 (0%)	2 (3%)	4 (6%)	2 (3%)	2 (3%)
Simple cause uncontrollable	61 (85%)	1 (1%)	1 (1%)	1 (1%)	2 (3%)	6 (8%)
Simple cause controllable	62 (76%)	1 (1%)	4 (5%)	5 (6%)	3 (4%)	7 (9%)
Domain = Farming						
No extra Information	53 (76%)	2 (3%)	2 (3%)	8 (11%)	2 (3%)	3 (4%)
Simple cause uncontrollable	54 (75%)	3 (4%)	0 (0%)	9 (13%)	4 (6%)	2 (3%)
Simple cause controllable	65 (79%)	3 (4%)	1 (1%)	5 (6%)	7 (9%)	1 (1%)
Both domains combined						
No extra information	113 (81%)	2 (1%)	4 (3%)	12 (9%)	4 (3%)	5 (4%)
Simple cause uncontrollable	115 (80%)	4 (3%)	1 (1%)	10 (7%)	6 (4%)	8 (6%)
Simple cause controllable	127 (77%)	4 (2%)	5 (3%)	10 (6%)	10 (6%)	8 (5%)

Note. All explanation options were shown to the participants in a randomised order during the experiment.

We compared the choices of the simple explanation to the other explanations, as in the previous experiment. In the no extra information condition, 81% of participants chose the simple explanation, in the simple cause uncontrollable condition, 80% of participants chose the simple

explanation, and in the simple cause controllable condition, 77% of participants chose the simple explanation. A Kruskal-Wallis test showed there was no difference in the choices of choices of the simple explanation across three conditions, $H(2) = 0.412, p = .814$. In other words, participants chose the simple explanation as often in all experimental conditions. See Appendix 2.4 for a supplementary analysis of the choices of the complex explanation.

If antecedent-controllability had an effect, participants' preference for the simple explanation would be decreased when the simple explanation was described as uncontrollable and the complex option controllable, and if antecedent-uncontrollability had an effect, their preference for the simple explanation would be decreased when the simple explanation was described as controllable and the complex option uncontrollable. However, the results showed that the manipulation of antecedent-controllability and antecedent-uncontrollability did not affect people's simplicity preference, as the majority (over 70%) of participants still preferred the simple explanation. These findings were consistent with Experiment 3a in showing that simplicity serves as a strong explanatory preference when people reason about explanations and overrides actionability as operationalised by consequent-controllability and antecedent-controllability. It is consistent with previous findings which have shown the preference for simplicity over-rides probability, at least until the simple explanation becomes highly improbable (Lombrozo 2007). In Experiments 1, 2a and 2b, when simplicity was not varied, consequent-controllability had a clear effect on people's explanation selection, whereas antecedent-controllability had little systematic effect. In Experiments 3a and 3b, when the explanations varied in simplicity, neither consequent-

controllability information nor antecedent-controllability information over-rode people's preference for the simple explanation. The results indicated that the simplicity preference has a stronger influence on people's explanation evaluation. The next experiment examines one final potential influence on the preference for simplicity in explanations, namely the phrasing of the explanation as a causal or counterfactual assertion.

Experiment 4

The aim of this experiment was to test whether stating the explanation options as counterfactuals will decrease the preference for the simpler explanation. Consider the scenario given in Experiments 3a and 3b in the no extra information conditions:

Suppose you are a doctor trying to understand an alien's medical problem. The alien, Treda, has two symptoms: Treda's minttels are sore and Treda has developed purple spots. Trichet's syndrome always causes both sore minttels and purple spots. Morad's disease always causes sore minttels, but the disease never causes purple spots. When an alien has a Humel infection, that alien will always develop purple spots, but the infection will never cause sore minttels. Nothing else is known to cause an alien's minttels to be sore or the development of purple spots. What do you think is the most satisfying explanation for the symptoms that Treda is exhibiting?

Now suppose you were given the explanations below, phrased as counterfactuals instead:

- A) If Treda had not had Trichet's syndrome, Treda would not have had the symptoms

- B) If Treda had not had Morad's disease, Treda would not have had the symptoms.
- C) If Treda had not had a Humel infection, Treda would not have had the symptoms.
- D) If Treda had not had Trichet's syndrome and Morad's disease, Treda would not have had the symptoms.
- E) If Treda had not had Trichet's syndrome and a Humel infection, Treda would not have had the symptoms.
- F) If Treda had not had Morad's disease and a Humel infection, Treda would not have had the symptoms.

Which one would you choose as the most satisfying explanation now?

In Experiment 4, we aimed to examine the effect of explanation types on the preference for simpler explanations, i.e., to compare causal explanations and counterfactual explanations. We gave one group of participants counterfactual explanations as above, in the past tense and subjunctive mood, using the connective "if", and the other group matched causal explanations, in the past tense and indicative mood, using the connective "because",

- A) Because Treda had Trichet's syndrome, Treda had the symptoms.
- B) Because Treda had Morad's disease, Treda had the symptoms.
- C) Because Treda had a Humel infection, Treda had the symptoms.
- D) Because Treda had Trichet's syndrome and Morad's disease, Treda had the symptoms.
- E) Because Treda had Trichet's syndrome and a Humel infection, Treda had the symptoms.

- F) Because Treda had Morad's disease and a Humel infection, Treda had the symptoms.

We know from previous research that counterfactual explanations help people to simulate alternatives (e.g., Byrne & Egan, 2004; Santamaría et al., 2005). When thinking about counterfactual and causal explanations, people have different mental representations. When they understand a counterfactual, such as "If Treda had not had Trichet's syndrome, Treda would not have had the symptoms", they simulate the events in the conjecture, Treda did not have Trichet's syndrome and she did not have the symptoms, and they also think about the presupposed facts, Treda had Trichet's syndrome, and she had the symptoms. But when they understand a causal assertion, such as "Because Treda had Trichet's syndrome, Treda had the symptoms", they simulate only the events in the assertion, at least at the outset, Treda had Trichet's syndrome, and she had the symptoms. Eye-tracking experiments tested affirmative and negative counterfactuals, as well as causal assertions. It was found that participants looked at words corresponding to the conjecture and presupposed facts when they heard counterfactuals, and this effect was more pronounced in negative counterfactuals, while for causal assertions, participants looked at words corresponding to fact only (Orenes et al., 2022). Moreover, when participants spontaneously provided explanations for a negative event, they created twice as many causal explanations as counterfactual explanations (McEleney & Byrne, 2007). Given these findings, we suggest that if counterfactuals lead to richer mental representations, then participants would be prompted to consider alternative possibilities following reading them. Since they may consider alternative possibilities, that may lead them to make choices of explanations other than the

simple one. As a result, we predicted that people would choose the simpler explanation less frequently when the explanations were phrased as counterfactuals than as causals.

Method

Participants

The participants were 185 volunteers (72 men, 112 women, and 1 non-binary individual). Their ages ranged from 19 to 81 years, with a mean of 38 years, and no participants were eliminated. The estimated sample size was the same as in Experiment 1, 2a and 2b. Each participant was paid £0.45 UK sterling for participation based on an estimated three-minute completion time.

Design

A between-groups design was adopted. Explanation type was manipulated between groups. The participants were randomly assigned to one of the two groups: causal explanation ($n = 99$), and counterfactual explanation ($n = 86$). We measured their choice of the most satisfying explanation from a set of six explanations.

Materials & Procedure

Participants received scenarios for two domains, disease and farming. See Appendix 2.3 for the experimental materials in the farming domain. The testing materials were the same as the no extra information conditions in experiments 3a and 3b in this chapter. We manipulated the phrasing of explanation options. In the causal explanation condition, explanations were

stated as causals such as “Because Treda had Trichet’s syndrome, Treda had the symptoms”, in contrast, in the counterfactual explanation conditions, explanations were stated as counterfactuals such as “If Treda had not had Trichet’s syndrome, Treda would not have had the symptoms.” The procedure of this experiment was the same as previous experiments in this chapter.

Results and Discussion

First, we checked that no differences were arising merely because of the domain (disease and farming), or the order (first received scenario, second received scenario), and there were no differences. See Appendix 2.4 for supplementary analysis. As a result, we combined responses from both orders. Two Chi-square tests showed that the distribution of choices was different between the experimental conditions in the disease domain, $\chi^2(5) = 7.62, p = .0058, V = 0.09$; and in the farming domain, $\chi^2(5) = 14.38, p = .0001, V = 0.12$. See **Table 2.6** for the frequencies of explanation preferences.

Table 2.6*Frequencies of Explanation Preferences in Experiment 4*

Explanation options	Trichet (Simple)	Morad (Insufficient)	Humel (Insufficient)	Trichet & Morad (Over-determined)	Trichet& Humel (Over-determined)	Morad& Humel (Complex)
Domain = Disease						
Causal explanation	80 (81%)	2 (2%)	2 (2%)	6 (6%)	5 (5%)	4 (4%)
Counterfactual explanation	55 (64%)	3 (3%)	2 (2%)	10 (12%)	6 (7%)	10 (12%)
Domain = Farming						
Causal explanation	80 (81%)	2 (2%)	1 (1%)	4 (4%)	9 (9%)	3 (3%)
Counterfactual explanation	48 (56%)	3 (3%)	2 (2%)	12 (14%)	14 (16%)	7 (8%)
Both domains combined						
Causal explanation	160 (81%)	4 (2%)	3 (2%)	10 (5%)	14 (7%)	7 (4%)
Counterfactual explanation	103 (60%)	6 (3%)	4 (2%)	22 (13%)	20 (12%)	17 (10%)

Note. All explanation options were shown to the participants in a randomised order during the experiment.

We compared the choice of the simple explanation to the rest of the explanations. In the causal explanation condition, 81% of participants chose the simple explanation, and in the counterfactual explanation condition, 60% of participants chose the simple explanation. A Mann-Whitney U test indicated that there was a significant difference as participants chose the simple explanation less when the explanations were stated as counterfactual explanations, $U = 3053$, $z = 3.783$, $p < .001$, $r = 0.278$. See Appendix 2.4 for a supplementary analysis of the choices of the complex explanation.

The results were consistent with our prediction, that when people read the explanations as counterfactuals, their choices of the simple explanation decreased, compared to when people read the same explanations options as

causal explanations. In other words, when they read a counterfactual explanation such as “ If Treda had not had Trichet’s syndrome and a Humel infection, Treda would not have had the symptoms.”, they were able to construct a mental simulation where these two diseases caused the symptoms the alien is exhibiting and also to think about the situation where these two diseases were not contracted and no symptoms showed, as explicitly mentioned in the negative counterfactual. This tendency for people to be able to simulate more possibilities for counterfactuals enabled them to think more explicitly about the causal link between these two diseases and the symptoms, leading them to choose these explanations more than they would without counterfactual explanations. The results were also consistent with what the dual possibility theory of counterfactuals would predict: people maintain imagined alternatives in counterfactual conditionals and with this, they can make otherwise difficult inferences. In our experiment, they were able to shift the strong preference for a simple explanation and give more consideration to the complex alternative. However, it is important to note that they still showed a strong preference for the simple explanation (60%), even in the counterfactual format. The counterfactual phrasing does not override the strong preference for simple explanations. We suggest that counterfactuals helped participants simulate alternatives and think more about the causal links between the causes and outcomes. Hence, we observed a modest increase in their choices of the more complex options. The results showed that participants’ choices were still guided by the simplicity preference but influenced by the linguistic phrasing in the form of counterfactuals.

General Discussion

Six experiments in this chapter examined how people evaluate and choose explanations from competing causal explanations. We aimed to answer three research questions: Do people prefer actionable explanations, which we operationalised as a causal explanation that contains consequent-controllable factors compared to ones that do not contain consequent-controllable factors; and a causal explanation that contains antecedent-controllable factors compared to ones that do not contain antecedent-controllable factors? Second, if there is a preference for these sorts of explanations, does the preference override the preference for other factors, such as simplicity? The third question is, would people overcome what they usually choose (i.e., the simpler explanation) and choose a more complex explanation instead if they receive explanations in the form of a counterfactual, such as “If your farm had not had micro-organism Excie, you would not have had the problems”?

Experiment 1 showed that when explanations were equally simple, participants preferred the explanation with consequent-controllable information such as “*The symptoms are caused by diseases that can be treated with surgery and medication*”. Experiment 2a and Experiment 2b showed that when explanations were equally simple, antecedent-controllability information did not cause a preference for the explanation with antecedent-controllable factors such as “*The symptoms are caused by diseases that are the result of smoking and physical inactivity*”. There was however a surprising increase in the choices of antecedent-uncontrollable explanations such as “*The symptoms are caused by diseases that are the result of gender and genetic ethnicity*”. Experiment 3a and Experiment 3b showed that neither consequent-

controllability information nor antecedent-controllability information overrides an overwhelming preference for the simple explanation. Experiment 4 showed that stating the explanations as counterfactuals decreased people's strong preference for simple explanations compared to causal explanations, even though people still chose the simple explanation very highly. We discuss the results and their implications for the psychology of causal explanation and XAI below. We then consider some limitations of the experiments of this chapter and future experimental directions.

Consequent-Controllability and Antecedent-Controllability

Our first research question in this chapter was inspired by the findings about controllability in counterfactual thinking, and the idea of actionability in XAI. We operationalised actionability in two ways, as consequent-controllability, and as antecedent-controllability.

Given a scenario in which you are told that someone has two symptoms, and you are told about two diseases that lead to the first symptom, one of which can be treated e.g., by surgery, and the other cannot, and two other diseases which lead to the second symptom, one of which can be treated e.g., by medication, and the other cannot, the explanation you tend to prefer is that the person's symptoms are caused by the two diseases that can be treated. In Experiment 1 participants chose the explanation that contained both of the causes that were described as consequent-controllable, more frequently than they chose that explanation when they were given no information about consequent-controllability (46% vs 19%). Similar results were found for a scenario about diseases as for one about a farm that has problems with low

grass growth and high weed growth, and various micro-organisms that lead to these outcomes, some of which can be remedied, e.g., by fertiliser, and some of which cannot.

This tendency to choose consequent-controllable causes may reflect a heuristic people rely on when there are no other clues as to which is the best explanation (e.g., Gigerenzer & Gaissmaier, 2011; Gigerenzer et al., 2011; Girotto et al., 1991; Roese, 1997; Raab & Gigerenzer, 2015; Shah & Oppenheimer, 2008; Simon, 1990a; Xie & Beck, 2022). It is unlikely that it could be considered an explanatory “virtue” such as simplicity, coherence, or breadth because it has little rational basis explanation (e.g., Jones, 1999; Simon, 1990b). Instead, it may be based on a preference for explanations that allow you to do something, as suggested by the idea of actionability introduced in XAI explanation (Alotaibi & Singh, 2023; Bhattacharya, 2024; Degen et al., 2023; Guidotti et al., 2022; Hacker & Passoth, 2020; Kelechi & Jiao, 2023; Singh et al., 2023; Sokol & Flach, 2020; Verma et al., 2020). However, it could lead people astray so that they make errors in their explanation selection since the most consequent-controllable explanation is not necessarily the accurate explanation.

Given instead a scenario in which you are told that someone has two symptoms, and you are told about two diseases that lead to the first symptom, one of which has itself been caused by something controllable, e.g., smoking, and the other has not, and two other diseases which lead to the second symptom, one of which has itself been caused by something controllable, e.g., physical inactivity, and the other had not, the explanation you tend to prefer is, contrary to our expectations, not that the person’s symptoms are caused by the

two diseases that can be controlled. In Experiment 2a participants chose the explanation that contained both of the causes that were described as antecedent-controllable, more than they chose that explanation when they were given no information about controllability (31% vs 18%) but the difference occurred for just one of the scenarios (the farming one); and in Experiment 2b, they chose the explanation that contained both of the causes that were described as antecedent-controllable, just as frequently as they chose that explanation when they were given no information about antecedent-controllability (25% vs 19%). Hence there does not seem to be a robust effect of antecedent-controllability. Hence antecedent-controllability may not be a prominent feature in causal explanation, unlike its prominent role in counterfactual thinking (e.g., Ferrante et al., 2013; Girotto et al., 1991, 2007; Markman & Miller, 2006; McCloy & Byrne, 2000; McEleney & Byrne, 2007; Mercier et al., 2017; Pighin et al., 2010).

A key difference between consequent-controllability and antecedent-controllability is that consequent-controllability information provides people with clear actions to follow up, e.g., to treat the symptoms with surgery. Antecedent-controllability information, on the other hand, suggests antecedents to the event and does not guarantee actions to follow to fix the problem, for example, even when people read that a disease was the result of smoking, there is no guarantee that quitting smoking could improve the symptoms. An important implication of these results is that in any attempts in XAI to ensure actionability in automated explanations, it is crucial that actionability is instantiated as consequent-controllability rather than antecedent-controllability. For example, an actionable explanation for a failed mortgage application

should state “Your failed application can be changed to approval through an increase of your monthly income”, rather than “Your failed application is the result of a monthly income that is too low”. Satisfying actionable explanations need to be constructed with clear indications of actions for a desired outcome, and the fact that it comes after the negative outcome is important. What is more, these findings also may suggest that on a temporal scale, causal explanations may be less concerned with past causes but more concerned with future actions, at least when there is no other information on which to base a choice of explanation.

However, unexpectedly, there was a preference for explanations that contained antecedent-uncontrollable causes, whether described generally as in Experiment 2a, e.g., “there is no known cause” or described specifically as in Experiment 2b, e.g., gender or ethnicity. Participants chose the explanation that contained both of the causes that were described as antecedent-uncontrollable, more than they chose that explanation when they were given no information about antecedent-controllability, in Experiment 2a (29% vs 14%), and Experiment 2b (23% vs 10%). The finding is in line with findings about the difference in focus between counterfactual and prefactual thinking, and we note that it is a smaller effect than the finding that consequent-controllability affects explanation choices.

For the investigation of consequent-controllability and antecedent-controllability, one limitation in our experiments 1, 2a and 2b, was that we did not combine them in one experiment, for example, one could say that disease A is the result of smoking and can be treated by quitting smoking, disease B is the result of physical activity but cannot be treated by higher activity level. Future

studies could consider combining and comparing them for further consolidation of the effects. Another limitation of our experiments is the nature of the outcome. We use a negative event or an unwanted outcome, but some evidence shows that when we compare a positive outcome to a negative one, we can see differences in the direction of counterfactual thoughts people generate: people generate downward counterfactuals about how things could have been worse following a positive outcome to serve an emotional regulatory function, and upward counterfactuals about how things could have been better following a negative event to serve a preparative function (Roese & Olson, 1997). There is also evidence that outcome controllability affects counterfactual direction: people generate more upward counterfactuals following controllable outcomes and more downward counterfactuals following uncontrollable outcomes (Roese & Olson, 1995). Future studies can also explore how the nature of the event impacts what sorts of explanations people prefer. Furthermore, participants in these experiments were asked to choose the most satisfying explanation. We used this question directly from Lombrozo (2007), and it is a self-report measure focusing on a general evaluation of explanations. It may also be somewhat ambiguous to participants what a “satisfying” explanation is, and in some senses, it could be said that an explanation that can be acted on would be more satisfying than one that cannot be acted on, even though it may not be the accurate explanation. In the next chapter, we include more measurements where we test explanation efficacy, such as an objective performance accuracy measurement.

We now turn to the second research question, and we discuss how the findings may further indicate what cognitive processes underly causal explanation evaluation.

Controllability and Simplicity

In Experiment 1, we found that when all explanations were equally simple, participants showed a preference for consequent-controllable explanations. From a psychological perspective, this consequent-controllability effect adds to our understanding of how people select between competing explanations as a heuristic that people follow: in the disease example, all four causally qualified explanations can explain the symptoms and the above-chance choices of the explanation with consequent-controllable diseases indicated that in such circumstances, the cognitive processes people may rely on to select an explanation may include some superficial heuristics such as consequent-controllability.

Given a scenario in which you are told that someone has two symptoms, and you are told about one disease that leads to the two symptoms, and two other diseases which each lead to one of the two symptoms, people tend to choose the simple explanation based on one disease rather than the complex explanation based on two diseases. Experiment 3a tested whether the consequent-controllability effect will override simplicity preference. Participants were told that the single disease that led to the two symptoms was consequent-controllable, e.g., by surgery, and the two other diseases had no known treatment, or that the single disease that led to the two symptoms had no known treatment, and the two other diseases each had treatments e.g.,

medication, and physiotherapy. In this chapter, we replicated the findings of a simplicity preference in Experiment 3a and Experiment 3b, as participants showed an overwhelming preference (around 80%) for the simple explanation, even when the simple explanation is consequent-uncontrollable (Experiment 3a) and when the simple explanation is antecedent-controllable (Experiment 3b). The results showed that the simplicity preference has a robust effect in this experimental setting (in both disease and farming domains) and serves as a much stronger quality for explanation preference than controllability. In Experiment 4, even when we saw the preference for the simple explanation significantly decreased by stating the explanations as counterfactuals, more than half of the participants (60%) still preferred the simple explanation.

We have discussed some theoretical claims for the simplicity preference and the debate about what guides causal explanation selection. Some argue that people rely on cognitive processes that detect whether an explanation is most likely, and people use its simplicity as a clue, for example, it is more likely that one cause results in the two symptoms than that two causes separately cause the two symptoms (Lombrozo, 2007; Liefgreen & Lagnado, 2023). One possible suggestion is that participants used consequent-controllability also as a proxy for probability (for example, having a disease that can be treated with medication is more likely than having a disease with no known treatment). However, if that was the case then they should not overwhelmingly prefer the simple but consequent-uncontrollable explanation, as they would cancel each other out. Hence our findings suggest that consequent-controllability is not understood as a proxy for likelihood. Instead, it is possible that when participants considered consequent-controllability information, they did not

relate it to any probability assumption. As a result, when there was a conflict between simplicity and consequent-controllability, they became influenced by the probability and chose the simpler one as more satisfying. This shows an important limitation of Experiments 1 and 3a, as we did not incorporate probability with consequent-controllability, such as providing an explanation that is consequent-controllable but has low probability. This will be an important issue to address in future research, as it will better inform the processes underlying the consequent-controllability effect.

We did not find a clear effect of antecedent-controllability in Experiments 2a and 2b. However, we observed a preference for antecedent-uncontrollability. In Experiment 3b, we pit antecedent-uncontrollability with simplicity. As in Experiment 3a, people still dominantly preferred the simple explanation, even when it was described as antecedent-controllable, and they did not prefer the complex but antecedent-uncontrollable explanation more. The findings indicated that both consequent-controllability and antecedent-controllability information are not understood as a proxy for likelihood.

From the heuristics point of view, the results contribute to the idea that people use heuristics as shortcuts sometimes during causal explanation selection. In our case, it was clear that the participants gave priority to simplicity information given the context. Participants did not sort through the manipulation between controllability information and simplicity potentially as a way to arrive at a decision with less cognitive effort. This point of view is further consolidated in Experiment 4, where we saw that simplicity preference could be mitigated when the participants were prompted by counterfactuals to put in more cognitive effort.

For XAI methods, results from Experiments 3a and 3b gave a clear indication that there is a difference between various explanatory qualities in their influence on explanation satisfaction. Again, a good explanation should incorporate different qualities. The emphasis in XAI on actionable explanations may need to be balanced with an emphasis on explanations that also have other qualities such as simplicity.

Counterfactual Explanation Mitigates Simplicity Preference

The last research question concerns whether counterfactual explanations can mitigate the simplicity preference. This question derives from findings that counterfactuals prompt people to mentally simulate alternative probabilities and can even help people make inferences that are otherwise difficult to make (Byrne & Tasso, 1999; Santamaría et al., 2005; Thompson & Byrne, 2002). Given these characteristics, it is especially interesting to test counterfactuals in scenarios where people usually commit to one decision and ignore the other alternatives. In this chapter, we tested counterfactuals in the alien disease scenario (Lombrozo, 2007) where the preference for the simple explanation is robust. We found that by comparing counterfactual explanations (If Treda had not had Morad's disease and a Humel infection, Treda would not have had the symptoms) to causal ones (Because Treda had Morad's disease and a Humel infection, Treda had the symptoms), there was a significant effect of counterfactual explanations in mitigating people's tendency to choose the simple explanation (80% vs 60%). The differences between these explanations were that the counterfactual one prompts people to consider the scenario where Treda had Morad's disease and a Humel infection leading to the symptoms, and

also the scenario where Treda did not have these diseases and exhibited no symptoms. The mental simulation allows people to consider the causal links between diseases and symptoms. This richer representation of possible scenarios also prompts people to process information with more cognitive effort, allowing them to overcome shortcut reliance and consider alternatives. One implication is that XAI automated explanations should consider using counterfactual explanations especially when there is a need to overcome a tendency to stick to a simple decision.

Experiment 4 tested the effect of counterfactual explanations on the simplicity preference. One potential limitation is that we did not consider the potential interactions between counterfactual explanations and controllability (both consequent and antecedent). As we discussed before, controllability serves as a heuristic in counterfactual thinking, for example, when people choose what aspects to mutate when they generate counterfactuals, they tend to mutate the controllable aspects in some circumstances (Giroto et al., 1991). Future experiments could investigate if counterfactual explanations could elicit a stronger effect of consequent-controllability information.

To conclude, six experiments in this chapter identified a consequent-controllability effect in that consequent-controllable explanations were preferred to consequent-uncontrollable ones. We did not observe a preference for antecedent-controllable explanations. When we tested consequent-controllability information against simplicity, we found that simplicity is a much stronger preference for explanation selection. This simplicity preference was mitigated when we provided explanations in the form of counterfactuals. The results replicated findings such as the simplicity preference in causal

explanation evaluation. The effect of consequent-controllability and the lack of effect of antecedent-controllability indicate the need to consider how actionability of explanations is operationalised. Counterfactual explanations were effective in decreasing the simplicity preference. These results corroborated with heuristic accounts of explanation selection and contributed to our understanding of how people rely on a variety of strategies when they attempt to explain.

We discussed that consequent-controllability and antecedent-controllability differ in their temporal relations with the outcome in question. In the next chapter, we explore this temporal comparison further and in a different direction by examining two sorts of simulations: counterfactual explanations about what could have been different in the past and prefactual explanations about what could be different in the future. Addressing these questions will further inform us of the role of temporal direction in explanation evaluation. In the next chapter, we also endeavour to test explanations in a mock AI decision support system setting.

Chapter 3: Counterfactual Explanations for Prediction and Diagnosis

Imagine a farmer using a smart agriculture grass growth management app. The app collects data on factors that affect grass growth on a farm, including the number of micro-organisms in the soil, rain level, fertiliser usage, beneficial weed growth and the number of livestock on this farm, and it suggests whether the grass growth for the next period is higher or lower than expected. The farmer may want to know how the app works and may require an explanation of why a certain suggestion was made on the grass growth levels. In this chapter, we aim to explore if counterfactual explanations and prefactual explanations are equally effective in helping farmers understand the app. Counterfactual explanations are framed in terms of how things could have turned out differently in the past, such as “if your fertiliser had been present, your grass growth would have been high”. Prefactual explanations are framed in terms of how things could turn out differently in the future, such as “if fertiliser were to be present, your grass growth would be high”. We also aimed to test participants’ understanding of such apps by asking them to make two types of inferences about the app’s decisions. We asked them to make a prediction from cause to effect, i.e., predict the level of grass growth the app would suggest given all the causes, as well as a diagnosis from effect to cause, i.e., diagnose the condition of one of the causes given the grass growth level and condition of other causes (see **Figure 3.1**). We discuss evidence on these research questions in the sections below, again starting with psychological evidence, and then XAI findings.

Figure 3.1

Examples of the Experimental Tasks in Experiment 5

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

Note. Examples of the instances presented to participants in Experiment 5 to provide information about five causes (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 causes 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 causes is given, and the task is to make a diagnosis about the presence or absence of one of the causes (indicated by the question mark).

Counterfactuals and Prefactuals

Thinking about the past and thinking about the future elicit different cognitive processes (Huang, Xie & Chen, 2021). We test 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 (Byrne, 2016). 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 facts (Byrne & Tasso, 1999). Participants were given counterfactual conditionals such as “If Linda had been in Dublin, then Cathy would have been in Galway”, or factual conditionals such as “If Linda was in Dublin, then Cathy was in Galway”. Participants made more Modus Tollens inferences of “Linda was not in Dublin” from “Cathy was not in Galway” from counterfactuals compared to factual. They also made more Denial of Antecedent inferences of “Cathy was not in Galway” from “Linda was not in Dublin” from counterfactuals than factual. The difference in the frequency of inferences indicates that people keep two possible situations in mind for counterfactual conditionals. The mental simulation of such dual possibilities facilitates causal inferences (Gerstenberg et al., 2021). In contrast, when people understand prefactual conditionals, for example when they read a prefactual conditional of “if Linda were in Cork tomorrow then Cathy would be in Galway”, they envisioned a scenario where the antecedent and consequent both occur, they mentally simulate just the conjecture for prefactual (Byrne & Egan, 2004). To conclude, prefactuals do not provide the same inferential advantages as counterfactuals. On the other hand, understanding prefactual conditionals may require less cognitive effort since the initial simulation only requires one possibility, unlike 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 (Rim & Summerville, 2014), and these “better-world” counterfactuals help them to formulate intentions for the future to prevent similar bad outcomes happening again (Markman, McMullen & Elizaga, 2008). Counterfactuals are thus a key learning mechanism, and they provide preparatory blueprints for behaviour in the future (Roese & Epstude, 2017). 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 level, or their strategies (Epstude et al., 2016; Fernandez, Gan & Chan, 2022; Ferrante et al., 2013; Mercier et al., 2017; Park & Seo, 2023). In other words, prefactuals are more closely linked to the identification of actionable changes in the future than counterfactuals. Recall that in Chapter 2, we explored the actionability idea in XAI and operationalised it through controllability in causal explanations. The intention formation function of prefactuals is in line with the controllability idea, as experimental evidence shows that prefactual thoughts focus on actions that are within the protagonist’s

control. Given the differences in how people think about counterfactuals and prefactuals, in this chapter, we compare their efficacy as explanations for AI decisions in our experiments. Some experiments asked participants to generate counterfactual or prefactual thoughts following a prompt about their performance (Fernandez et al., 2022; Ferrante et al., 2013), while other experiments have provided participants with counterfactual or prefactual conditionals and ask for their inferences (Byrne & Egan, 2004; Beck et al., 2006). We provided participants with counterfactual and prefactual explanations in our experiments and asked them to provide their ratings of helpfulness (in Experiment 5) or choose between these two types of explanations (in Experiment 6). We aim to examine whether counterfactual and prefactual explanations are equally effective and satisfying, or whether one is better than the other.

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 cause(s) to outcome(s), compared to when they reason from outcome(s) to infer cause(s). As we outlined in Chapter 1, 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, people 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 (Fenker et al., 2005). People also assign stronger causal links between

the cause and effect for prediction than diagnosis. For example, 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 (Tversky & Kahneman, 2015).

But in more complex cause-effect cases, such as a common-effect model in which several causes lead to an effect, participants in experiments tended to make more correct diagnoses than predictions. For example, during predictive inferences, people tend to make errors and ignore alternative causes that may contribute to the outcome, but during diagnostic inferences, people tend to take into account the alternative causes and make more accurate judgements (Cummins, 2014; Fernbach, Darlow & Sloman, 2010; Fernbach et al., 2011; Fernbach & Rehder, 2013). Research on predictive and diagnostic inferences contributes to our understanding of the underlying cognitive processes of causal inferences (Fernbach & Rehder, 2013). For example, neglect of alternative causes in prediction (Fernbach et al., 2011) is an example of error-prone causal inference, and these errors “seem to share the hallmarks of heuristics, shortcuts that are applied liberally and unconsciously” (Fernbach & Redher, 2013, p.83).

There is also a discussion about the differences between various kinds 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 (Goldvarg & Johnson-Laird, 2001; Johnson-Laird & Khemlani, 2017). 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 (Frosch & Byrne, 2012). Predictive inferences are affected by how strong a cause is, whereas diagnostic inferences are affected by how strong alternative causes are (Fernbach & Rehder, 2013). 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) (Byrne, 1989; Cummins, 2014; Fernbach & Rehder, 2013). 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 (Byrne, 1989; De Neys, Schaeken & D’Ydewalle, 2003; Espino & Byrne, 2020). In other words, people make different inferences about the different sorts of causes when they make predictions compared to when they make diagnoses. Furthermore, people predicted that an effect, (e.g., ‘flu) is more likely, the more causal factors were present, (e.g., fever, headache, fatigue) (Rehder & Waldmann; 2017). 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) (Khemlani & Oppenheimer, 2011).

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 (for an example see **Figure 3.1**), so that in our tasks all of the different sorts of causes were explicitly provided and known to be present or absent.

The tasks we used were based on a simulated SmartAgriculture decision support system for grass growth outcomes on realistic farming scenarios in Experiment 5 and an analogous alien planet domain in Experiment 6. We compared two sorts of explanations for decisions made by this 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 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 two experiments also tested users' self-reported confidence in their decisions and their /preference satisfaction with the explanations provided.

User Testing of Counterfactual Explanations

As discussed in Chapter 1, existing research suggests that counterfactual explanations help users improve their understanding of systems, and their trust and satisfaction (Binns et al., 2018; Celar & Byrne, 2023; Dodge et al., 2019; Förster et al., 2021; Goyal et al., 2019; Hoffman et al., 2018; Kenny et al., 2021; Lage et al., 2019; Lim et al., 2009; Lucic et al., 2020;

Russell et al., 2020; van der Waa et al., 2021). However, user studies have tended not to distinguish between counterfactuals about how things could have turned out differently in the past, such as “your grass growth would have been higher if you had used more fertiliser”, and explanations about how things could turn out differently in the future, e.g., “your grass growth would be higher if you were to use more fertiliser”, usually referred to as “prefactuals” (Byrne & Egan, 2004; Delaney et al., 2022; Epstude et al., 2016; Ferrante et al., 2013; Mercier et al., 2017). Yet counterfactual and prefactual explanations have been found to implicate different psychological processes and to have different behavioural consequences, as we discussed earlier. Hence, our aim in this experimental chapter was to compare the usefulness for users of counterfactuals and prefactuals as explanations of an AI system’s decisions, to see whether prefactual explanations elicit higher levels of user satisfaction and understanding. We did this by providing prefactual explanations together with counterfactual ones and testing user satisfaction. Our experiments contribute to the current methods of using past counterfactuals as XAI explanations and to our knowledge of the differences between prefactual counterfactual explanations.

Moreover, user studies have tended to assess the accuracy of users’ understanding of AI system decisions by testing their predictive inferences about an outcome based on information about the presence or absence of relevant causes (Goyal et al., 2019; Lage et al., 2019; Lim et al., 2009; Lucic et al., 2020; van der Waa et al., 2021). Predictive inferences require making an inference from causes to effects. In the farming app example mentioned earlier, a predictive inference would require users to decide whether the outcome

variable grass growth is higher or lower than expected given information about all the causally related factors. We aimed to investigate the approach of testing user understanding with not only predictive but also diagnostic inferences, where users were tested on their diagnoses of the presence or absence of relevant causes, based on information about an outcome. In the farming app example, this type of inference would require users to decide the condition of one of the causally related factors to be either present or absent, based on the outcome condition and other factors, as shown in **Figure 3.1**. After all, users of an AI system who require an explanation, for example, for recourse, usually know the outcome, and wish to understand the causes; that is, they require a diagnostic explanation, not necessarily a predictive one.

Overview of Experiments

Through two experiments, we addressed our research questions by testing counterfactuals and prefactuals as explanations, and by testing predictive and diagnostic inferences as an indicator of user understanding. We conducted two large-scale user studies ¹ (N = 243), using a common-effect causal model in which multiple causes could lead to an outcome. A pilot study tested the model where four causes lead to an outcome. Experiment 5 tested a realistic farming model with five causes, and Experiment 6 tested an alien

¹ Experiments of this chapter have been published in the Proceedings of the 2022 AAI/ACM Conference on AI, Ethics, and Society (Dai et al., 2022). farming model with five causes. We examined the participants' objective

understanding of the AI system (measured by their decision accuracy), their subjective evaluation of their performance (measured by their ratings of confidence) and explanations of its decisions (measured by their ratings of explanation helpfulness, or their choice between two explanations, as well as by an established explanation satisfaction scale).

Experiment 5

In this experiment, we tested two hypotheses. First, we predicted that participants would show different levels of accuracy between prediction tasks and diagnosis tasks given previous findings that people make more accurate diagnoses than predictions in a common effect causal structure as they are able to consider alternative factors when making diagnoses. Second, since prefactual thinking has an emphasis on future planning actions and requires fewer possibilities to be envisaged in their initial mental representation, we test whether prefactual explanations will result in a higher level of user understanding and satisfaction compared to counterfactual explanations. We tested by constructing a grass growth app used to predict the grass growth on a farm to be either high or low or to diagnose factors on the farm that are causally linked to grass growth to be either present or absent. In a pilot study, we tested participants with four causes (see Appendix 3.2) with only positive causes and a strong ceiling effect was observed as participants' accuracy levels were very high. In Experiment 5 we gave participants text to read that included five causes including a disabler, with a realistic farming setting:

“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.

- 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'."

Participants then saw an example of prediction tasks and diagnosis tasks before the corresponding testing block. For each task, participants made their decision first and were asked to report their confidence level in the decision they just made, then they received the app's actual decision together

with an explanation, after which they were asked to rate their judgement of the helpfulness of the explanations (see **Figure 3.2**).

Figure 3.2

Example of an Experimental Task in the Prediction X Counterfactual

Explanation Condition

Farm NK - Prediction	
Micro-organisms	Present
Rain	Present
Fertiliser	Absent
Clover	Present
Cows	Present
Grass growth	?

What prediction will the app make about grass growth on this farm? *

☐ High ☐ Low

How confident are you in your judgement of the app's prediction? *

Highly unconfident Unconfident Neutral Confident Highly confident

☐ ☐ ☐ ☐ ☐

The app's prediction is: High.

Explanation: If micro-organisms had been absent last month, your grass growth would have been low.

Please rate how helpful the explanation above is in helping you to understand how the app works.

Very unhelpful Unhelpful Neutral Helpful Very helpful

☐ ☐ ☐ ☐ ☐

Note. Participants saw the prediction task in a table and were asked to provide their judgement on the app's decision and their self-rated confidence level in their judgement (top). After that participants were provided with the app's actual decision, and an explanation of the decision, followed by a rating on the helpfulness of the explanation(bottom).

This experiment compared prediction tasks and diagnosis tasks, as well as the effect of counterfactual explanations and prefactual explanations for the app's decisions to no-explanation controls. Participants were assigned to one of three explanation groups (Counterfactual, Prefactual, or No-explanation control) and each completed a block of prediction tasks and another block of diagnosis tasks. The control description was merely a re-statement of the decision (e.g., "The system's prediction is 'high' "). We used this kind of statement to ensure we had consistent task layout and material length for participants, while not providing any sort of explanation. Participants were presented with 10 items in each block, which were systematically varied in terms of the five factors and balanced across the materials (i.e., two items for each cause). The full set of 20 materials are presented in Appendix 3.3. Explanation helpfulness judgements were measured following each item, and the DARPA scales for explanation satisfaction (Hoffman et al., 2018) were measured at the end of the experiment (see **Table 3.1**).

Table 3.1*DARPA Explanation Satisfaction Scale*

Scale Items
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.

Note. The scale (Hoffman et al., 2018) 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.

The objective measurement was the accuracy of participants' decisions when compared to the app's decisions, which indicated participants' understanding of the system. The subjective measurement included self-reported confidence level and self-reported evaluation of the explanations for each task, as well as the DARPA explanation satisfaction scale at the end. 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 cause, 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. 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). For each trial, the participants were shown a reminder in a written paragraph at the top of the screen about the impact of each of the causes 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 3.3. See Appendix 3.1 for the information and debrief sheets for all experiments in this chapter.

Method

Participants

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. We selected participants to be adults from Ireland, the United Kingdom, the United States, Canada, Australia and New Zealand for whom English was their first language. To control for expertise, we recruited only participants whose self-reported occupation was not in the field of agriculture. The initial experimental design contained only two types of explanation conditions (counterfactual and prefactual) and we recruited participants to the two explanation groups first, 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 recruited 142 participants to allow for failure of attention checks, and one participant had to be eliminated accordingly. After data collection and initial analysis, we decided to add a control condition. 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). Across three conditions, a further 12 participants were eliminated due to an administrative error resulting in the absence of a 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.

Design

A mixed design was adopted since we aimed to test the difference in accuracy between prediction and diagnosis tasks with each participant as a within-participants factor to avoid variance in performance levels in a group,

and we chose to only expose each participant to one type of explanation to avoid the influence of other explanation types as a between-participants factor. Hence the explanation type (counterfactual vs. prefactual vs control) was manipulated between groups and the task (prediction vs diagnosis) was tested within participants with counterbalancing. There were three explanation conditions: counterfactual ($n = 57$), prefactual ($n = 78$), and control ($n = 58$).

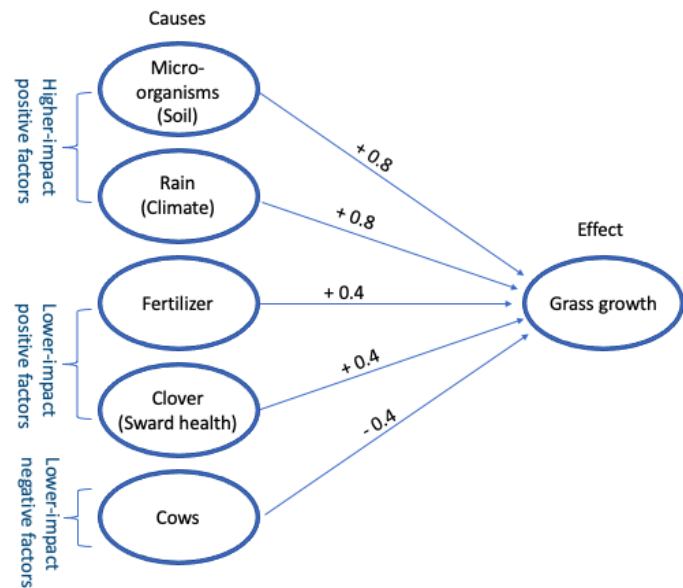
Materials & Procedure

We constructed the materials based on a simplified model of grass growth, outlined in **Figure 3.3**. We 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 (fertiliser, 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 causes 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 causes, and we selected 10 cases to test (see Appendix 3.3), 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 causes (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.

Figure 3.3

A Simple Model of Grass Growth Based on Five Causes and an Outcome in Experiment 5



Note. 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), and the minus sign indicates a negative causal link (i.e., a disabler or inhibitor). Participants were not provided with this diagram.

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.

Results and Discussion

The dataset for all of the experiments is available at <https://osf.io/pvn7y/>. There was no effect of the order of tasks, see Appendix 3.4 for supplementary analysis. We calculated percentages of correct judgments and the mean ratings for confidence and explanation evaluation for both task types. Hence, we used ANOVA to analyse all the data in this chapter.

Accuracy. Participants were very accurate in their judgments, with overall 80% correct judgements. We included not only the explanation type and task type in the analysis but also the different causes. We ran a 3 (explanation type: counterfactual vs prefactual vs control) x 2 (task type: prediction vs diagnosis) x 5 (causes: micro-organisms, rain, fertilizer, clover, cows) analysis of variance (ANOVA) on the accuracy of participants' decisions, see **Figure 3.4.1**. Recall that in the experimental materials, micro-organisms and rain were described as high-impact positive causes, fertiliser and clover were described as low-impact positive causes, and cows were the low-impact negative cause. See Appendix 3.4 for supplementary analyses where we categorised the causes into three groups (high-impact positive causes, low-impact positive causes, and low-impact negative causes). Supplementary analysis showed the same results.

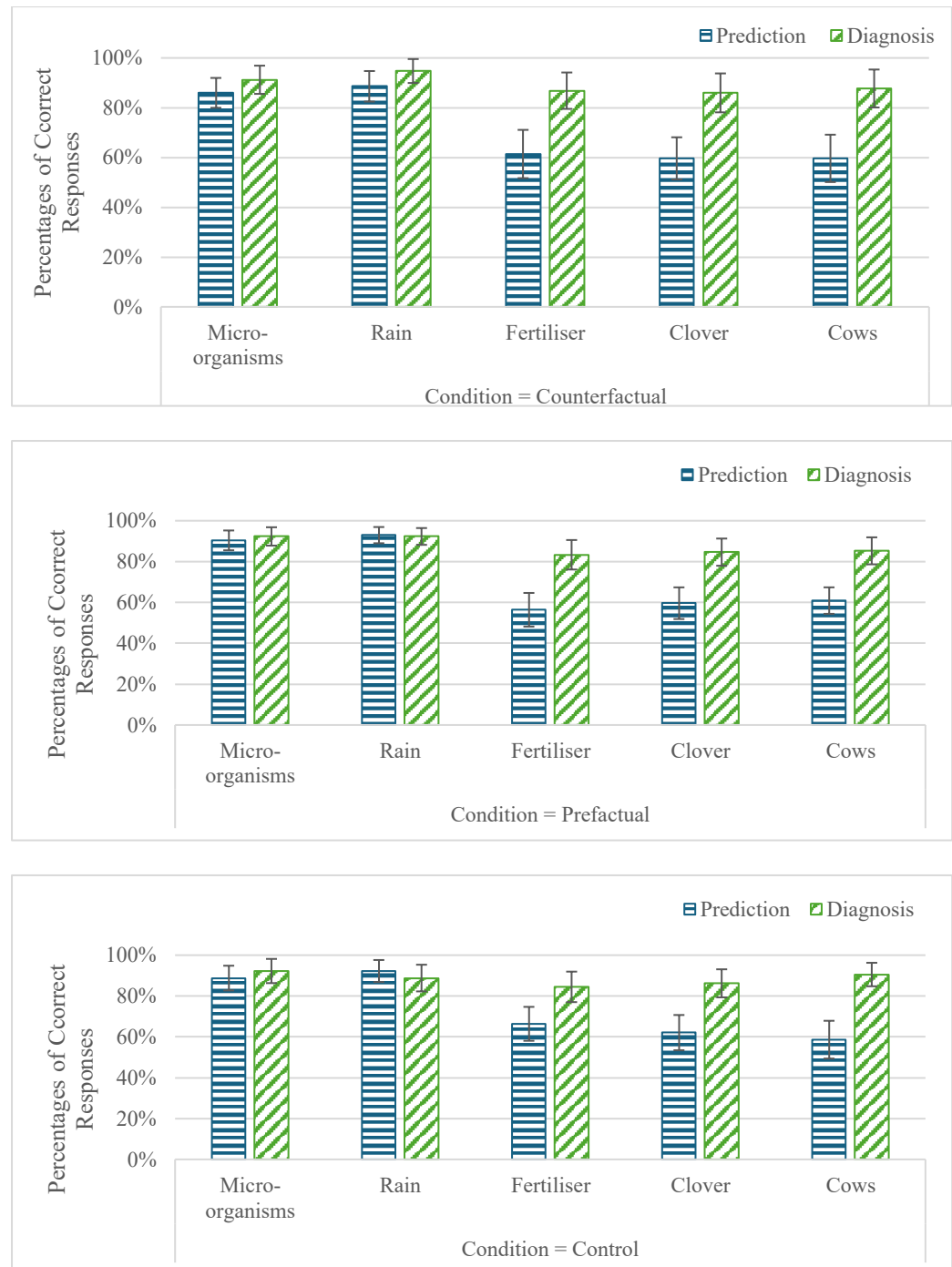
Participants made the same number of correct decisions when they were given counterfactual explanations (80%), prefactual explanations (80%), and control descriptions (81%), as indicated by the lack of a main effect of explanation type, $F(2,190) = 0.19, p = .83$; and explanation type did not interact with task type, $F(2,190) = 0.65, p = .52$, or causes, $F(8,184) = 0.56, p = .81$; and the three variables did not interact, $F(8,184) = 0.55, p = .82$. The results were not consistent with our hypothesis of a difference in how counterfactual and prefactual explanations help participants understand how the system works better. The result goes against previous findings, for example, in a user study with an imagined blood alcohol prediction app, it was found that providing explanations improved user accuracy (Warren et al., 2023). We suggest that the ability to capture a difference in user accuracy depending on explanations may be weakened by the simplicity of tasks. Participants were very accurate in their judgments, over 80%, which suggests they found the task simple to do. Even though we added in a fifth cause as a disabler (cows) based on our pilot experiment (Appendix 3.2), the overall accuracy level was still very high. In other words, participants might not have needed explanations at all as they found the tasks very simple.

Participants made more correct decisions for diagnoses than predictions, 88% vs 72%, as shown by a main effect of task type, $F(1,191) = 167.63, p < .001, \eta^2 = 0.47$. They made more correct decisions for some of the causes than for others, as shown by a main effect of causes, $F(4,188) = 51.37, p < .001, \eta^2 = 0.21$. The two variables of task type and causes interacted, $F(4,188) = 23.20, p < .001, \eta^2 = 0.11$. 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 causes which were 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 causes which were micro-organisms, 92% vs 88%, $t(192) = 1.65$, $p = .102$, and rain, 92% vs 91%, $t(192) = 0.25$, $p = .806$. Low-impact factors were closer to the decision threshold and were more difficult to judge. The results suggested that firstly, accuracy level variance between prediction and diagnosis tasks only occurred for instances that were more difficult to judge (i.e., low-impact causes), as indicated by the absence of difference for easier instances (i.e., high-impact causes). Secondly, there was a difference in performance level between prediction and diagnosis tasks for low-impact instances, since participants found prediction tasks more difficult.

Figure 3.4.1

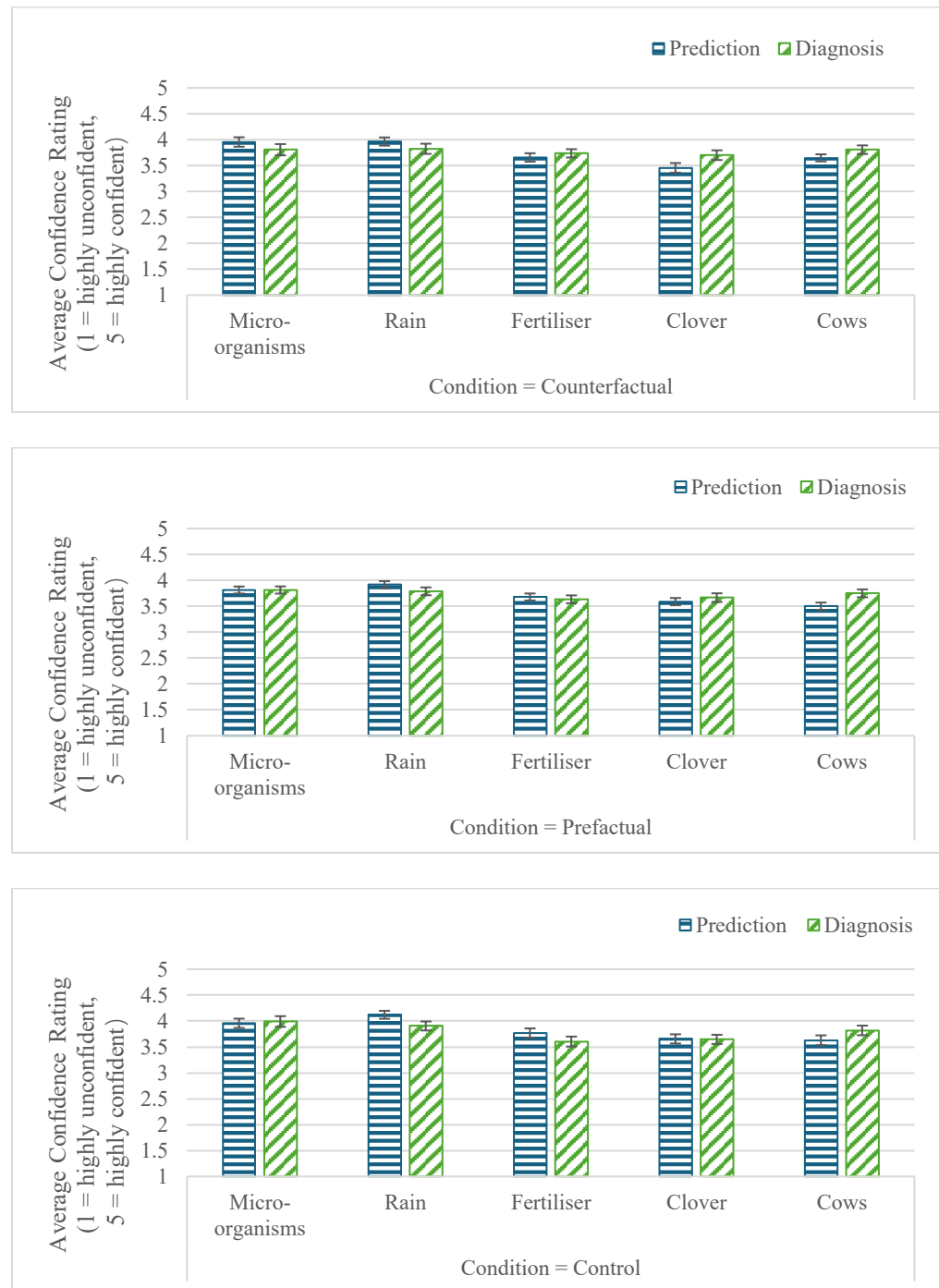
The Percentages of Correct Decisions for Predictions and Diagnoses in Experiment 5



Note. Participants were given counterfactual explanations (top), prefactual explanations (middle), or control descriptions (bottom), for each of the five causes in a farming domain. Error bars are standard errors of the mean.

Confidence ratings. Consistent with the accuracy results, there was no main effect of explanation type, $F(2,190) = 0.80, p = .45$, and explanation type did not interact with task type, $F(2,190) = 0.37, p = .70$; or causes, $F(8,184) = 0.90, p = .52$; and the three variables did not interact, $F(8,184) = 1.22, p = .28$. Participants were more confident about diagnoses than predictions for some of the low-impact causes, and there were no differences between predictions and diagnoses for some of the high-impact causes, as shown by the interaction of task type and causes, $F(4,188) = 8.31, p < .001, \eta^2 = 0.04$; there was no main effect of task type, $F(1,191) = 0.11, p = .74$, and there was a main effect of causes, $F(4,188) = 25.94, p < .001, \eta^2 = 0.12$, see **Figure 3.4.2**.

The decomposition of the interaction between causes and task type, with paired-samples t-tests and a Bonferroni corrected alpha for 5 tests of $p < .01$, revealed that participants were more confident for diagnoses than predictions for the negative low-impact cause cows, $t(192) = 3.737, p < .001, d = 0.76$; and for the high-impact positive cause rain, $t(192) = 2.738, p = .007, d = 0.80$. They were equally confident in prediction and diagnosis tasks for the rest of the causes, for micro-organisms, $t(192) = 0.586, p = .559$; for fertiliser, $t(192) = 0.826, p = .41$; and for clover, $t(192) = 1.777, p = .077$. We suggest the differences found for cows and rain factors may have occurred merely by chance since as we will see, we did not replicate the differences in the next experiment. This also indicated a reliability issue in self-reported confidence measures, as it does not always align with their objective performance levels.

Figure 3.4.2*The Confidence Ratings for Predictions and Diagnoses in Experiment 5*

Note. Participants were given counterfactual explanations (top), prefactual explanations (middle), or control descriptions (bottom), for each of the five causes in a farming domain. Error bars are standard errors of the mean.

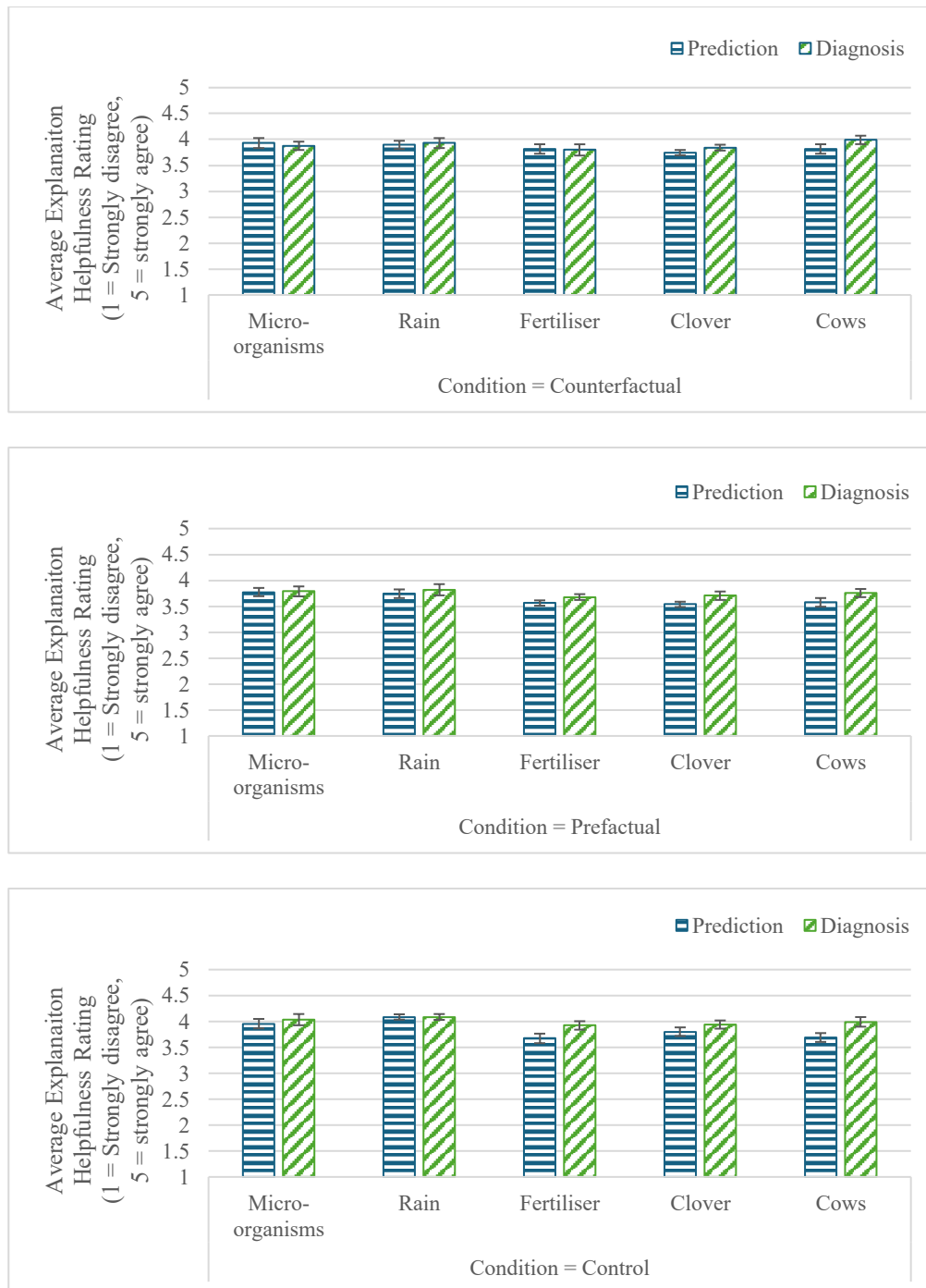
Explanation helpfulness ratings. Consistent with the accuracy results, there was no main effect of explanation type, $F(2,190) = 2.59, p = .08$, and explanation type did not interact with task type, $F(2,190) = 0.37, p = .69$, or causes, $F(8,184) = 0.81, p = .45$; and the three variables did not interact, $F(8,184) = 0.80, p = .61$. Once again, there was a main effect of task type, $F(1,191) = 10.05, p = .002, \eta^2 = 0.05$, and a main effect of causes, $F(4,188) = 14.65, p < .001, \eta^2 = 0.072$. Participants found the explanations more helpful for diagnoses than predictions for low-impact causes, and there was no difference between diagnoses and predictions for high-impact causes, as shown by the interaction of task type and causes, $F(4,188) = 4.3, p = .002, \eta^2 = 0.022$. The decomposition of the interaction between causes and task type, with paired-samples t-tests and a Bonferroni corrected alpha for 5 tests of $p < .01$, revealed that participants found explanations to be more helpful for diagnoses than predictions for the low-impact factors, clover, $t(192) = 2.848, p = .005, d = 0.67$; and for cows, $t(192) = 2.848, p = .005, d = 0.68$. For the other factors, they found explanations to be equally useful in predictions and diagnoses, for micro-organisms, $t(192) = 0.270, p = .787$, for rain, $t(192) = 1.033, p = .303$, and for fertiliser, $t(192) = 2.226, p = .027$, see **Figure 3.4.3**.

The results for judgments about explanation helpfulness were opposite our prediction, that because prefactual explanations elicit more future-planning thinking they would lead to a higher level of helpfulness rating. Furthermore, counterfactual and prefactual explanations were not rated to be more helpful than the control description. This result is consistent with some previous findings, for example, Warren et al. (2023) also found that in a hypothetical blood alcohol content related driving decision support app, participants did not

subjectively rate counterfactual explanations as more satisfying or trustworthy than no-explanations. Similarly to the accuracy measurement, we suggest that given people's tendency to subjectively report the same level of evaluation of helpfulness of explanations between these conditions, we need to take into account the high accuracy and simplicity of tasks for participants in Experiment 5. Participants considered explanations more helpful for diagnoses than predictions, and they were more accurate on diagnosis tasks than prediction tasks. It may be that the explanations helped them to be more accurate on the diagnosis tasks than the prediction tasks. They considered explanations more helpful for the diagnosis tasks than the prediction tasks for most of the low-impact causes which are more difficult tasks, rather than for the high-impact causes which are easier tasks; and they were also more accurate on diagnosis tasks than prediction tasks for the low impact causes; again it may be that their judgment of helpfulness reflects that the explanations were more helpful to them in the diagnosis tasks.

Figure 3.4.3

The Explanation Helpfulness Ratings for Predictions and Diagnoses in Experiment 5

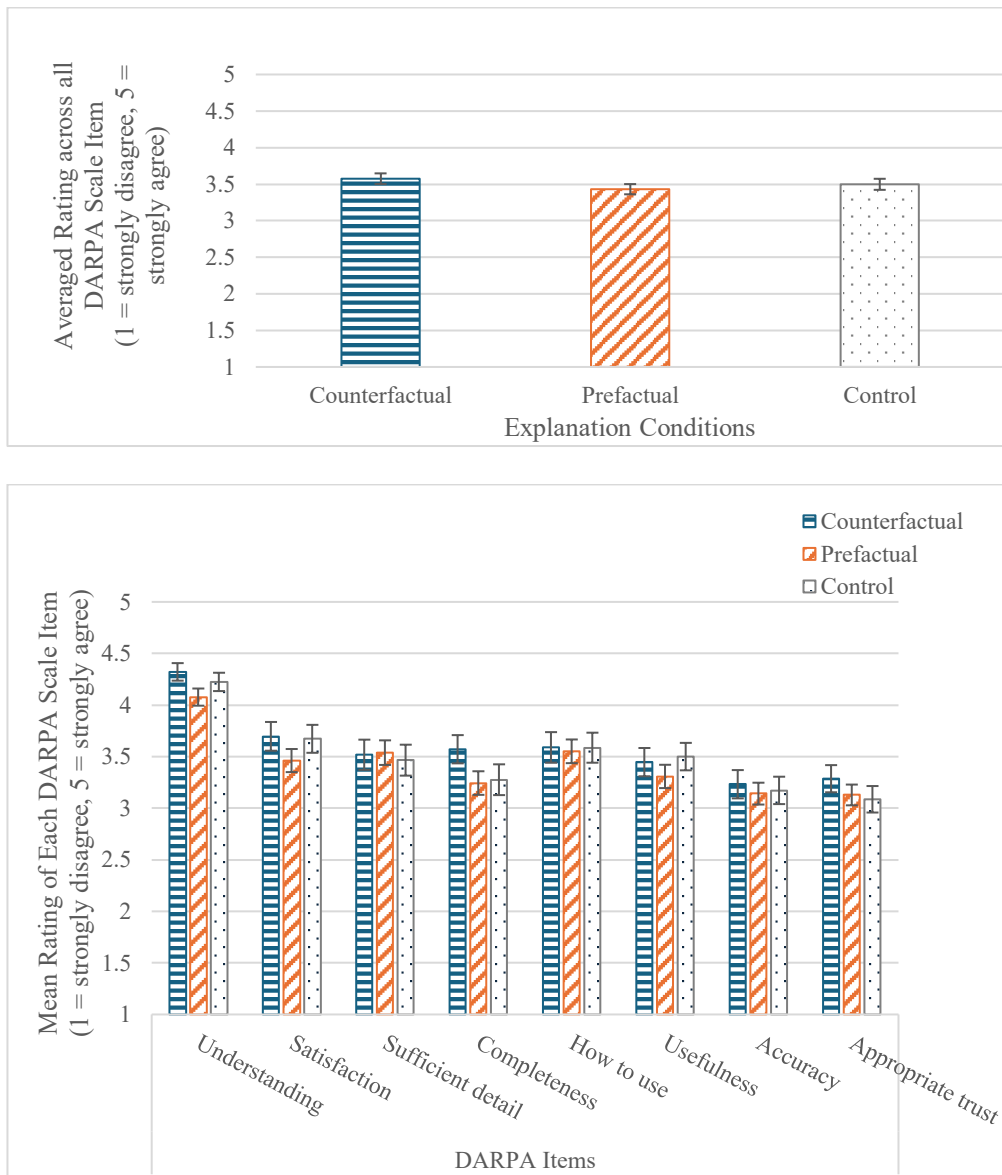


Note. Participants were given counterfactual explanations (top), prefactual explanations (middle), or control descriptions (bottom), for each of the five causes in a farming domain. Error bars are standard errors of the mean.

Finally, 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, $F(2,190) = 0.99, p = .37$. A 3 (explanation: counterfactual vs prefactual vs control) $\times$ 8 (each DARPA scale item) ANOVA showed there was no significant effect of explanation type, $F(2,189) = 1.091, p = .338$, and no interaction between explanation type and scale item, $F(14,189) = 0.534, p = .914$. DARPA scale items (see **Table 3.1**) had a main effect, $F(7,189) = 27.807, p < .001, \eta^2 = 0.13$, as **Figure 3.4.4** shows.

Figure 3.4.4

The Mean Ratings for the DARPA Scale in Experiment 5



Note. The mean ratings for the DARPA scale for each of the three experimental conditions in Experiment 5 (top) and of each DARPA scale item (bottom).

Error bars are standard errors of the mean.

To summarise, there was no difference between the explanation types, which was against our hypothesis. Participants showed an equal level of accuracy following counterfactual explanations, prefactual explanations, and

control descriptions. We suggest that this lack of difference may be due to the simplicity of the tasks.

Participants were more accurate in diagnosis tasks than in prediction tasks, for the more difficult tasks (judgments about low-impact causes rather than high-impact ones). Their subjective judgments of explanation helpfulness were consistent with their objective accuracy, in that they judged explanations to be more helpful for diagnosis tasks than for prediction tasks, for the more difficult tasks (judgments about low-impact causes rather than high-impact ones). The self-reported confidence level was inconsistent with participants' accuracy as there was only a difference in confidence between diagnosis and prediction tasks for two causes.

In the following experiment, we aimed to first test in an alien planet setting to see if the results depended on general knowledge of farming activities. We also aimed to modify our measurements to see whether counterfactual and prefactual explanations are processed differently by participants.

Experiment 6

The second experiment aimed to test whether participants' accuracy level for prediction and diagnosis tasks also occurs in an alien domain, where less general knowledge of the causal structure is involved. We constructed similar tasks based on plant growth scenarios on a hypothetical alien planet. Such content has been used previously in experiments to control for the effects of background knowledge on human reasoning (Dias & Harris, 1990; Kuhl, Artelt & Hammer, 2023; Markovitz, Forgues & Brunet, 2010). 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 causes 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, (v) Umbrat as an animal feeding on the plant. The general information participants were given at the outset 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 four 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’.”

Once again, participants were given prediction and diagnosis tasks, as shown in **Figure 3.5**.

Figure 3.5*Examples of the Experimental Tasks in Experiment 6*

Delc PW - Prediction		Delc PP - Diagnosis	
Besloor	Present	Besloor	?
Thardon	Present	Thardon	Present
Fropa	Present	Fropa	Present
Corrick	Absent	Corrick	Absent
Umbrat	Present	Umbrat	Present
Elos growth	?	Elos growth	High

Note. These tables provide information about five causes (present or absent) and the outcome of elos growth (high or low). A table for a prediction task is on the left: information about the presence or absence of all five causes 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 causes is given, and the task is to make a diagnosis about the presence or absence of one of the causes (indicated by the question mark). Participants received one table at a time.

The measurements of accuracy and confidence used in Experiment 6 were the same as in Experiment 5. The measurement of explanation helpfulness was different. In Experiment 6 participants were asked to choose between a counterfactual and a prefactual explanation, as we describe below. In the prediction tasks, we first 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: “Which of the following explanations helps you to understand the app’s decision?”. For each task, the participants were shown a reminder in the form of written paragraphs at the top of the screen about the impact of each of the factors on the outcome.

Method

Participants

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. The selection criteria were the same as in Experiment 5. Since we constructed materials based on an alien planet, in Experiment 6 we did not control for participants’ self-reported occupation. 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 recruited 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 paid £1.75 for their participation.

Design

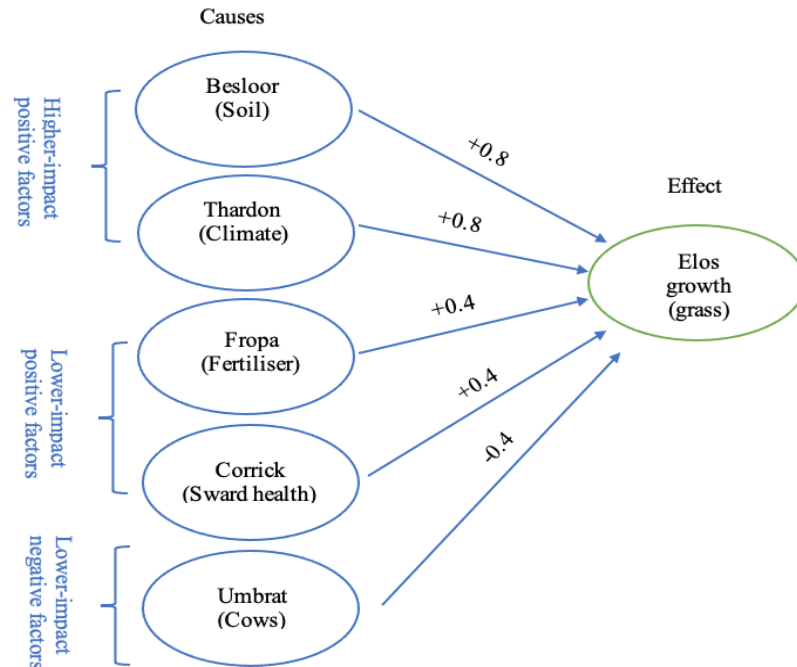
A fully within-participant design was adopted. Each participant was tested on two task blocks consisting of different task types (prediction vs. diagnosis) in a randomised order. Twenty-eight participants received the prediction tasks block first and twenty-two received the diagnosis tasks block first. We measured the accuracy of people's judgement of the app's decision, how confident they were about their judgements, and their preference between a counterfactual explanation and a prefactual explanation. To maintain consistency across experiments, we also administered the DARPA scales for explanation satisfaction (see **Table 3.1**) at the end of this experiment.

Materials & Procedure

We constructed the materials based on an alien scenario adapted from the farming example in Experiment 5 (see **Figure 3.6**). The alien model is entirely analogous to the farming one with the five causes 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 decision rules were also the same as in Experiment 5, see Appendix 3.3. The procedure of Experiment 6 was the same as Experiment 5.

Figure 3.6

A Simplified Farming Scenario About Five Causes and an Outcome in Experiment 6



Note. The diagram is adapted from the diagram of Experiment 5 and the names were matched with non-English words with 2 syllables and used in Experiment 6. The plus sign indicated the factors positively attributed to the effect, while the minus sign denoted the factor negatively attributed to the effect.

Results and Discussion

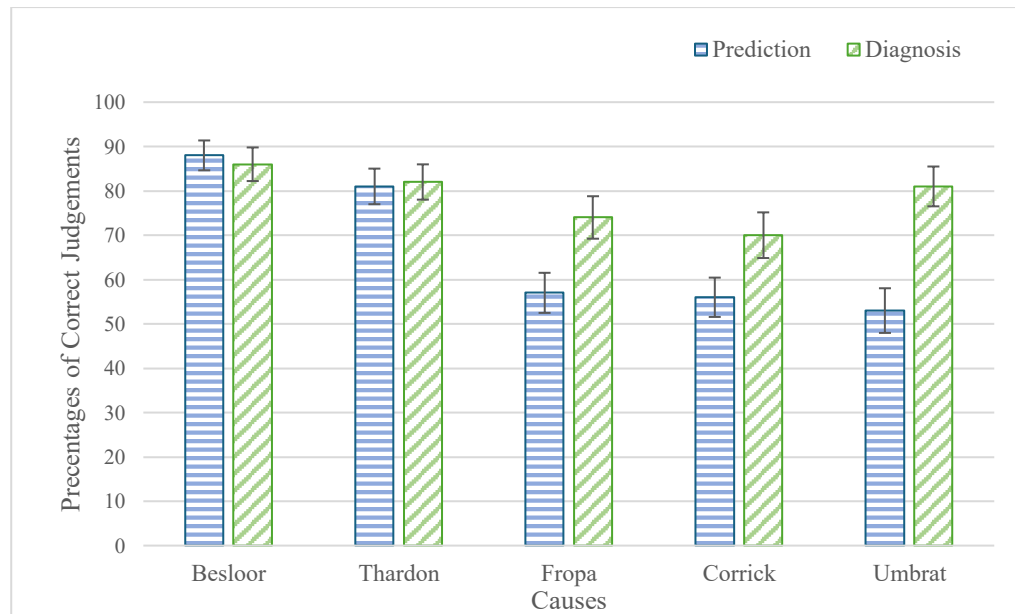
First, we established that there were no differences depending on the order of tasks, see Appendix 3.4.

Accuracy. Participants were accurate in their judgments, with overall 73% correct decisions in this alien domain. Participants made more correct decisions for diagnoses, 79%, than for predictions, 67%, as shown by a main effect of task type, $F(1,49) = 16.10, p < .001, \eta^2 = 0.25$, in a 2 (task type:

prediction vs diagnosis) $\times$ 5 (causes: Besloor, Thardon, Fropa, Corrick, Umbrat) ANOVA on the accuracy of participants' judgements (see **Figure 3.7.1**). There was no main effect of causes, $F(4,46) = 15.89, p < .001, \eta^2 = 0.2$, but there was a significant interaction of the two variables, $F(4,46) = 4.55, p = .002, \eta^2 = 0.09$. 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 causes, 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 $p < .01$). There were no differences between diagnoses and predictions for the high-impact causes of Besloor, 86% vs 88%, $t(49) = 0.42, p = .67$, and Thardon, 82% vs 81%, $t(49) = 0.21, p = .84$. The results were consistent with Experiment 5, indicating that we replicated the higher objective performance level for diagnostic tasks in Experiment 5, and that the effect extends to this alien domain where less general knowledge is assumed. The decomposition was also consistent with Experiment 5, for which participants showed different levels of performance between task types in low-impact causes only, suggesting that their accuracy is sensitive to the difficulty of tasks. See Appendix 3.4 for supplementary analyses where we categorised the causes into three groups (high-impact positive causes, low-impact positive causes, and low-impact negative causes). Supplementary analysis showed the same results.

Figure 3.7.1

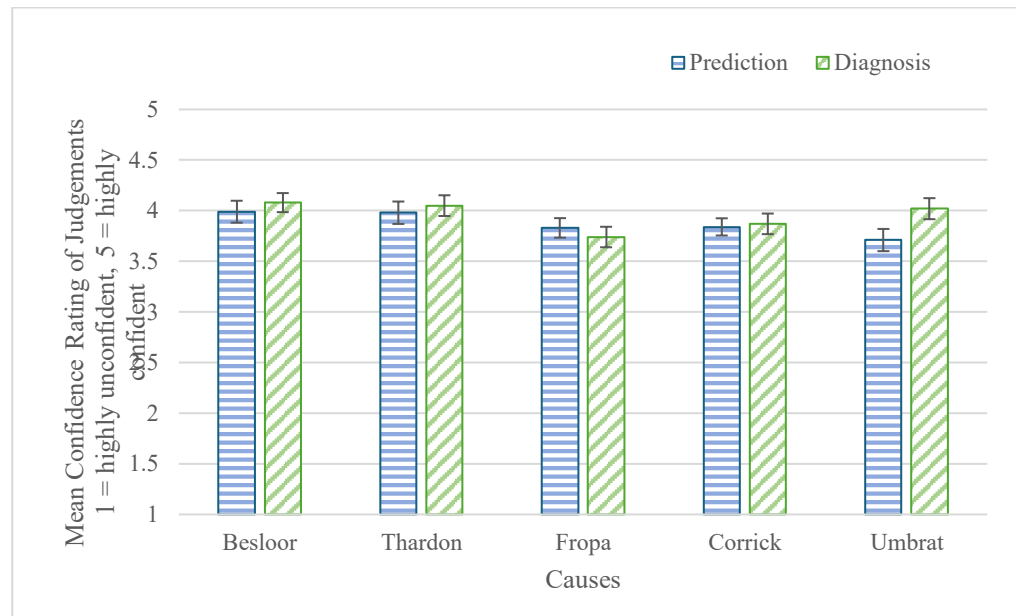
The Percentages of Correct Decisions for Predictions and Diagnoses in Experiment 6



Note. Participants were given five causes in an alien planet domain. Error bars are standard errors of the mean.

Confidence ratings. A 2 (task) X 5 (causes) ANOVA on average confidence ratings showed that there was a main effect of causes, $F(4,46) = 5.74, p < .001, \eta^2 = 0.11$; and no main effect of task type, $F(1,49) = 2.01, p = .16$. There was also a significant interaction between the two, $F(4,46) = 2.53, p = .04, \eta^2 = 0.05$. 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 confident in diagnoses ($M = 4.02, SD = 0.74$) than predictions ($M = 3.71, SD = 0.78$) for the negative cause Umbrat $t(49) = 3.275, p = .002, d = 0.67$; and their confidence level was the same for prediction and diagnosis tasks for the other four positive causes, Besloor, $t(49)$

= 0.829, $p = .411$; Thardon, $t(49) = 0.638$, $p = .527$; Fropa, $t(49) = 0.922$, $p = .361$; and Corrick, $t(49) = 0.339$, $p = .736$. See **Figure 3.7.2** for the distribution of confidence rating. The observed difference in confidence between diagnosis and prediction tasks for the low-impact negative cause is consistent with what we saw in accuracy measurement, in that people found low-impact causes objectively more difficult to judge. What is more, this difference in subjective confidence provided additional information as to what participants subjectively perceive as more difficult - having a low-impact cause that has a different sort of impact on the outcome.

Figure 3.7.2*The Confidence Ratings for Predictions and Diagnoses in Experiment 6*

Note. Participants were given five causes in an alien planet domain. Error bars are standard errors of the mean.

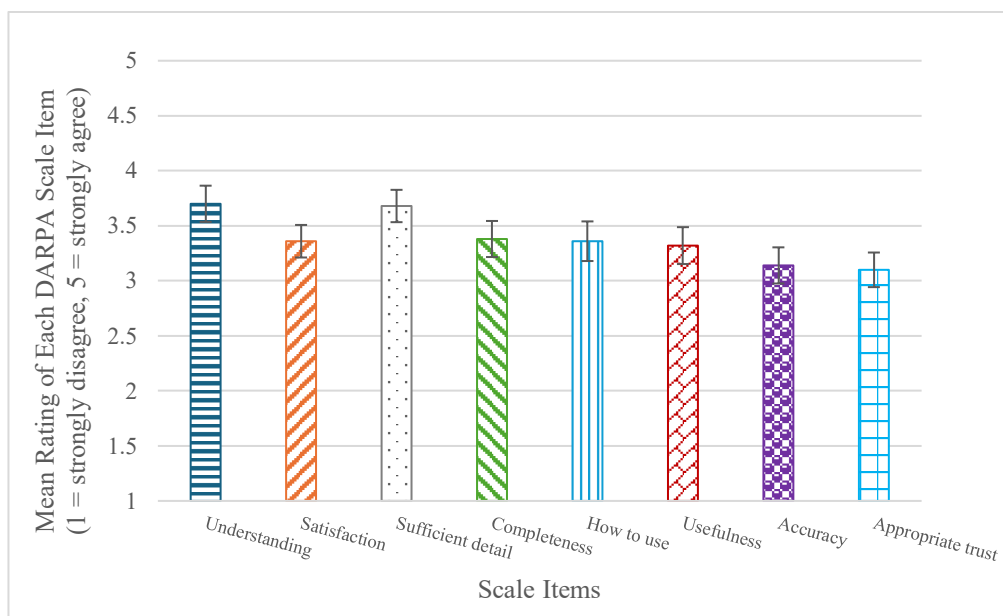
Explanation choices. Participants chose counterfactual explanations, 56%, as often as prefactual ones, 45%, and there was no difference in their tendency to select counterfactuals as their preferred explanation in prediction or diagnosis tasks, 49% vs 61%, $t(49) = 0.78$, $p = .44$. The results were consistent with Experiment 5 as we did not observe a preference for counterfactual or prefactual explanations, even with a forced choice task. The results indicate that there was no difference in people's subjective rating of explanation helpfulness, providing support that either form of explanation can be useful as an explanation for AI system decisions.

For the DARPA explanation satisfaction scale, since we only had one experimental group and the scale was administered at the end of the

experiment, we could not run a comparison analysis on scale responses, participants gave an average rating of 3.38, with a standard deviation of 0.664. See **Figure 3.7.3** for the mean ratings of each scale item.

Figure 3.7.3

The Ratings for the DARPA Scale in Experiment 6



Note. Ratings of DARPA explanation satisfaction scale by item. Error bars are standard errors of the mean.

To summarise, the accuracy level of participants was relatively high in the alien farming domain. We found a difference in accuracy between diagnosis and prediction tasks, but only for low-impact causes such as Umbrat. The confidence rating results were consistent with the accuracy measure, as participants reported feeling less confident when they made diagnostic judgements compared to predictive judgements for the low-impact cause of Umbrat. These findings indicate that people's performance and their

confidence levels were sensitive to the difficulty of tasks. In this experiment, even though we changed our measurement of explanation satisfaction by asking participants to choose between a counterfactual and a prefactual explanation for each task, we still did not see a difference in their preference.

General Discussion

The present studies report findings on the differences in prediction versus diagnosis tasks, as well as the efficacy of prefactual explanations compared to counterfactual ones. Two experiments showed that in a common effect structure where multiple causes affect one outcome, participants showed a higher level of accuracy for diagnosis when compared to prediction, especially for the tasks that were objectively more difficult (reflected in their closeness to decision boundary) and subjectively more challenging (reflected in participants' accuracy and ratings in confidence levels). We did not see a subjective preference between prefactual and counterfactual explanations, as both were rated as helpful and satisfying as mere control descriptions in Experiment 5, and there was no preference when participants were required to choose between a counterfactual and a prefactual explanation in Experiment 6. The subjective measurements of confidence did not always align with task accuracy, and explanation satisfaction mostly aligned with participants' task accuracy. In the following section, we discuss the implications of these findings and the limitations of these experiments.

The predictive and diagnostic inferences required by the experimental tasks were different from some tasks from previous research. We did not give descriptions of scenarios where the alternative causes were either present or

completely absent (e.g., Farnbach et al., 2010). Instead, in the experimental tasks we used, all alternative causes were listed, and their conditions were provided. In other words, the experimental tasks encouraged participants to consider alternative causes for prediction inferences. However, our results showed that participants were more accurate when they made diagnostic inferences. We now turn to explain the findings in light of different theories of the reasoning process.

Previous research used likelihood to measure and model inference accuracy and suggested that when reasoning from cause to effect, individuals often neglect alternative causes, leading to an overestimation of the likelihood (Cummins, 2014; Farnbach et al., 2011, 2013). Although we did not ask participants to judge likelihood in these experiments, and we gave information as a binary state instead of probability information, the results were consistent with the asymmetry (in our cases for the low-impact causes).

Another prominent theory is the effort reduction one: participants could be relying on mental shortcuts to reduce cognitive efforts when they make these inferences. It was suggested by previous research that people find it easier to make causal inferences when the relevant knowledge can be represented simply. When causal reasoning becomes computationally demanding, reasoners tend to reduce effort by invoking shortcuts that can lead to errors in their judgements. For example, a reasoner may assume the alternative cause is always present, always absent, or always effective. It was empirically found that when alternative causes were made to be explicitly present, participants were sensitive to alternatives for both prediction and diagnosis. Hence, the neglect of alternative causes was suggested to be due to

errors in representational and computational steps during the reasoning (Fernbach & Redher, 2013). The tasks in this chapter, however, ruled out some of the shortcuts, for example, we clearly stated the identity and the condition of all causes and effects in one moment of time. Participants also did not have the need to retrieve variables from memory as they were provided with all information in a table for each task. Hence, the asymmetry we observed in this chapter was relatively unlikely to be based on these sorts of shortcuts during the reasoning process.

The results added to the current body of empirical evidence that shows an advantage in diagnosis tasks in common effect models. The limitations of the present experiments lie in the asymmetry of prediction and diagnosis tasks, in that for prediction tasks participants had to consider the conditions of all causes before they make an inference about the outcome, while for diagnosis tasks participants could potentially just consider the cause in question, as the example illustrated above. In other words, the focal cause in prediction tasks was not made clear to participants before they saw the explanation, while the focal cause was made clear in the tasks for diagnostic inferences. This limitation could be mitigated in future studies, for example, to test prediction tasks while highlighting one of the causes as the focal one and the rest as alternatives. By doing this, one could rule out the possibility that diagnostic inferences were more accurate due to it being a much easier task. Another important direction is to test not only the common-effect structure but also the common-cause structure, where one cause leads to multiple outcomes.

Importantly, the greater accuracy of participants' inferences for diagnoses than predictions occurred only for low-impact causes in both domains and

there were no differences for high-impact causes. Participants also showed almost the same pattern in their confidence ratings. This indicated that participants were sensitive to the difficulty of tasks, even though they were generally very accurate in both experiments.

Participants made the same number of correct inferences whether they were given counterfactual explanations or prefactual ones in the first experiment, and they chose 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 all experiments was high, even in the unfamiliar domain, so it may be the case that participants did not require explanatory assistance to make their judgments.

Our results also have 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 mainstream task considered in user studies and our results suggested that it would be useful to extend the range of tasks given to users to assess their understanding. 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. Moreover, the need for an explanation often arises from a gap in one's knowledge and facilitates learning to fill that gap (Keil, Rozenblit & Mills, 2004). People also tend to seek explanations when faced with unexpected events (Hilton &

Slugoski, 1986; Quinn & Keane, 2022). Explanations in our experiments were provided after each task, independent of decision accuracy or user requirements. 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.

Finally, the experiments in this chapter examined the use of counterfactuals as an explanation strategy for people's understanding of AI decisions, and we have tested prefactual explanations together with counterfactual explanations. In the next chapter, we examine the effects of counterfactual explanations on people's acceptance of an AI's recommendation in a risky situation, to extend our understanding of how counterfactuals help people to simulate alternative possibilities.

Chapter 4: AI Recommendations in the Risky Choice Framing Effect

Suppose you are a farmer with a dairy farm. You are told that a new disease is expected to kill 60 of your cows. You hear there are two possible programs of treatment to combat the disease that you can take: program A will result in 40 of your cows dying, program B has a $1/3^{\text{rd}}$ probability that none of your cows will die and a $2/3^{\text{rd}}$ probability that all of them will die. Which program would you choose? Perhaps you may decide to choose program B. But suppose now you check with an Artificial Intelligence (AI) decision support app you use to see what it recommends. It tells you to choose program A, and it also says that if program A had been chosen in previous cases the outcome would have been better. Would you change your mind and choose program A? In this chapter, the goal is to test whether participants will follow the recommendation of an AI decision support system to switch their choice from one that most people typically judge is the best choice in this situation, to a different non-typical choice, and whether a counterfactual explanation offered by the AI system affects their tendency to do so. The secondary goal is to test the effect of recommendation sources and whether participants will follow recommendations from a human expert more than those from a decision support system. The research focus of this chapter is also motivated by the goal of eXplainable AI (XAI) such as providing transparent and reliable explanations for humans to understand the AI system's decision-making process (e.g., Fiok et al., 2022). In the sections below, we discuss different topics that motivated our research questions in this chapter, starting with psychological findings about the framing effect and counterfactual explanations, then followed by AI versus human recommendations.

Framing Effects

People's decisions are influenced by the way information is presented or framed. It was found that individuals tend to be risk-averse when facing gains and risk-seeking when facing losses (Tversky and Kahneman, 1981). In the original experiment, an imaginary disease was described to potentially kill 600 people, and participants were given two programs to combat this disease. When the programs were framed as gains, i.e., how many lives could be saved, most people (72%) chose the safe option (saving 200 people), not the risky one (a $1/3^{\text{rd}}$ probability of saving everyone and a $2/3^{\text{rd}}$ probability of saving no one). When the same programs were framed as losses, i.e., how many people would die, most (78%) did not choose the safe option (400 people dying), but the risky one ($1/3^{\text{rd}}$ probability that no one dies and a $2/3^{\text{rd}}$ probability that 600 people die) (Tversky and Kahneman, 1981). This flip of choices in a risky decision-making situation was termed the "risky choice framing effect" (Levin et al., 1998), and we use this term to describe it throughout this chapter. There has been extensive research on the risky choice framing effect and other framing effects since then; we discuss the cognitive processes that underlie this effect and the robustness of framing effects below.

Cognitive Theories

In this section, we outline prominent theories that have been proposed to explain the framing effects.

Loss Aversion and Prospect Theory. Loss aversion is the phenomenon where a loss is felt greater than a gain in an equivalent value (Kahneman & Tversky, 1984). For example, imagine walking on the street and finding a 10

euro note on the floor, you pick it up with a boost of happiness. Later that day, you accidentally dropped 10 euros and lost it. The negative feeling from losing money will likely be stronger than the positive feeling of gaining money, even if the amount of the money is the same. Along the same line, the prospect theory (Kahneman & Tversky, 1979) suggests that people exhibit asymmetrical preferences when they assess gain and loss, which provides a cognitive explanation for the risky choice framing effect: this cognitive bias to avoid losses leads participants to be more risk-seeking when facing potential losses. When asked to directly compare whether the two options in the framing effect problem should be treated the same, participants gave sensible justifications for why the two situations (saving 200 people vs 400 people dying) should be treated differently, indicating a lack of recognition of equivalence (Frisch, 1993). It was found that a broad set of brain areas, including midbrain dopaminergic regions, are sensitive to potential gains and the activities in these areas decrease for potential losses. Individual differences in behavioural loss aversion are predicted by activities in several regions including the ventral striatum and prefrontal cortex (Tom et al., 2007). These findings provide a biological explanation for loss aversion.

Effort Reduction. Other theories have considered the cognitive effort it requires to process gains and losses. The cost-benefit trade-off model explains framing effects by suggesting that individuals try to determine a good decision while expending minimum cognitive effort when examining alternatives. It was found through reaction time and functional Magnetic Resonance Imaging (fMRI) data that people take longer to select a risky gain than a sure gain, while it took a similar amount of time for people to select a sure loss or a risky

loss. Studies with fMRI revealed that the activities were localised in the prefrontal and parietal cortices, indicating the involvement of working memory and imagery (Gonzalez et al., 2005). Using eye-tracking technologies, it was also found that cognitive effort asymmetry plays a crucial role in framing effects, as the cognitive effort in terms of fixations during eye movement was higher in the negative framing than in the positive (Kuo, Hsu & Day, 2009).

The fuzzy-trace cognitive theory (Reyna & Brainerd, 1991; Reyna et al., 2021) suggests that the framing effect is the result of radically simplifying decision information, as “the key principle of the theory is that reasoning prefers to operate on simple gist, as opposed to exact detail” (Reyna & Brainerd, 1991, p. 249). Experiments showed that the framing bias is not caused by numerical processing, or the hypothesis that participants select the positive option over the negative one. Instead, participants “transformed the problems into uniformly positive representations to avoid the complexity of negation” (Reyna & Brainerd, 1991, p. 249). According to the fuzzy-trace theory, the gain frame could be represented as:

A: Some people will be saved.

B: Some people will be saved, or no one will be saved.

In this gain frame, “some people will be saved” is the same across alternatives and the decision focuses on saving some versus saving none, and presumably the majority prefers saving some.

In the loss frame, the representations might look like:

C: Some people will die.

D: Nobody will die, or some people will die.

In the loss frame, “some people will die” is the same across alternatives and the decision focuses on the situation where nobody will die, and presumably the majority prefers this situation. In line with the fuzzy-trace theory is “the zero effect”, where participants expressed strong preferences for options offering a possibility of zero death (“no one will die”) only in the loss frame (Zhang & Slovic, 2019). When tested with different working memory loads (with random letters), it was also found that participants showed the framing effect in both load and no-load conditions, indicating that people employ a heuristic to make satisfactory decisions with minimal effort (Whitney, Rinehart & Hinson, 2008).

Linguistic Ambiguity. The linguistic ambiguity hypothesis (Chick, Reyna & Corbin, 2016) proposes that people don’t perceive partially specified options the same way as they perceive fully specified options. In the gain framing, for example, the partially specified safe option of “200 people will be saved” poses ambiguity for the fate of the remaining 400 people. If the incomplete option is perceived as lower-bounded, e.g., when interpreted as “at least 200 are saved”, it becomes preferable. Similarly, in the loss framing, the partially specified safe option of “400 people will die” is ambiguous about the fate of the remaining 200, and if perceived lower-bounded could be interpreted as “at least 400 will die”, then the other option becomes preferable. This ambiguity hypothesis posits the framing effect in line with rationality since the option with a higher value is chosen (Mandel, 2014).

Schema Theory. The schema theory argues that people don’t equate saving 200 people with 400 people dying, indicating they didn’t have a complete mental representation of this disease and the programs, leading to a

lack of understanding of the relationship between choices (Jou, Shanteau & Harris, 1996; Okder, 2012). It was found that people usually consider the relationship to be independent, in other words, it does not occur to them that saving 200 lives meant causing 400 deaths, leading to the framing effect. However, when a description of limitations (what the researchers termed as a causal schema) was introduced into the scenario, the framing effect disappeared. For example, in the disease problem, participants were told why it was impossible to save all lives and why some could be saved only at the expense of others. They were told that there was a limited supply of the medication, enough to save only 200. If the medication is distributed to all 600 people, there would be a $1/3^{\text{rd}}$ probability that all 600 would be saved and a $2/3^{\text{rd}}$ probability that no one would be saved (Jou et al., 1996). Linguistic ambiguity (i.e., lack of complementary relationship between options) has been shown to be sufficient to cause the risky choice framing effect. For example, when the quantifiers were explicitly defined as exact such as “it is certain that exactly 200 people will be saved”, the framing effect disappeared (Mandel, 2014). However, it was later found that this type of ambiguity is not necessary for the risky choice framing effect, as participants who rejected the forms of ambiguity hypothesised (e.g., the word “exactly” added to the sure option) showed a robust framing effect nonetheless (Chick et al., 2016).

Role of Emotions. Apart from theories that centre around the cognitive processes of the framing effect, there is also a rich body of research on the importance of emotions in decision-making (Nabi et al., 2020). It was found that emotions have a significant influence on participants’ tendencies to take risks and on the impact of a frame on choices (Druckman & McDermott,

2008). The “risk-as-feelings” hypothesis highlights the role of affect experienced at the moment of decision-making, as research shows that emotional reactions to risky situations diverge from cognitive assessments and when the divergence occurs, emotional reactions often override cognitive assessment. For example, feelings about risks are insensitive to the changes in probability, but cognitive evaluations take probability into account (Loewenstein et al., 2001). Studies have shown that the framing effect is specifically associated with amygdala activity, indicating a key role of emotions (De Martino et al., 2006; Kahneman & Frederick, 2007).

Pervasiveness and Robustness

Framing effects have been established in a broad range of content domains, including legal decisions such as accepting and rejecting decisions of child custody (Shafir, 1993), medical decisions such as choosing treatment plans for specific diseases (Levin et al., 2002; McNeil et al., 1982), social dilemmas such as distribution of public goods (Aquino, Steisel & Kay, 1992), and political judgements (Amsalem & Zoizner, 2022). These studies showed how pervasive framing effects are in different aspects of our daily lives. Framing effects were tested to be robust in a large-scale replication project “Many Labs” (Klein et al., 2014). Multiple meta-analyses have established framing effects with a small to medium effect size (Kühberger, 1998; Maule & Villejoubert, 2007; McDonald et al., 2021; Piñon & Gambara, 2005; Steiger & Kühberger, 2018). When participants were encouraged to give more thought to the problem through justification of choices, they still showed substantial framing effects (LeBoeuf & Shafir, 2002). Levels of expertise have also been

shown to have little effect on this bias: a group of philosophers showed substantial framing effects compared to non-philosopher academic participants regardless of familiarity with the framing effects (Schwartzgebel & Cushman, 2015). The framing effect also extends to national security intelligence agents who make risky decisions routinely, as they exhibit larger framing biases and higher confidence levels when compared to college students and post-college adults (Reyna et al., 2013). More recently, the framing effect has been tested during the COVID-19 pandemic and was found to be larger than under typical circumstances (Rachev et al., 2021). The framing effect has also been applied to guide people in various contexts towards better decisions, e.g., climate crisis and public health messaging during COVID-19 (e.g., Hameleers, 2021; Homar, & Cvelbar, 2021; Rachev et al., 2021; Spence & Pidgeon, 2010).

While a lot of empirical evidence has shown framing effects as robust and pervasive effects, many factors have also been shown to challenge them. In a meta-analysis with 136 empirical papers, factors that influence the robustness of framing effects were coded. One important factor is whether framing information is manipulated through a different reference point. For example, whether the disease problem is referenced as currently existing conditions or a future risky event. The other important factor is outcome salience, for example, describing one's chance to suffer from a disease as 10% or the chance of being healthy as 90%. The task participants receive also matters - whether participants were asked to respond by making a choice or giving a rating (Kühberger, 1998).

Earlier studies focused on the *need for cognition* (NC), an indicator of people's "proclivities to give decisions greater thought" (LeBoeuf & Shafir,

2003, p. 79), and “their tendency to engage in and enjoy thinking” (Cacioppo & Petty, 1982, p. 116). It was found that participants in the low-NC group showed framing effects while those in the high-NC group did not (Smith & Levin, 1996). Later on, it was found that NC scores did not reduce the incidence of framing effects, however, when participants were presented with both the gain frame and the loss frame, participants with high NC responded more consistently across different frames of a problem (LeBoeuf & Shafir, 2003), and that effortful processing did not increase the framing effects (Igou & Bless, 2007). Other factors such as wording have been shown to weaken the framing effect, by substituting “not saved” for “die” and “not die” for “saved” (Kühberger, 1995). When participants were given formal training in decision theory, the preference for the risky option in the loss framing condition was no longer observed (Fagley & Miller, 1987).

More recently, research evidence shows that perspective-taking acts like a strong moderator for the framing effect, as a focus on feelings is linked to an increase and a focus on thoughts is linked to a decrease in the effect (Fagley, Coleman & Simon, 2010). The amount of risk (indicated by different probabilities, the probabilities of people dying being 0.3, 0.4, 0.6, and 0.7), the scale of the disease (number of people affected), and time constraints have been shown to impact framing effects in a within-participants design. The larger the probabilities are, the more risk-taking participants become; when there is a larger number of people’s lives being affected, participants are more risk-averse; the effect of framing is more pronounced under severe time constraints (Diederich, Wyszynski & Ritov, 2018; Wyszynski & Diederich, 2022). When the probability of gain was manipulated (30% chance gain vs

70% chance gain), it was found that a low probability of gain suppresses the framing effect (Gosling & Moutier, 2017). When the standard choice among two options measurement was tested together with other response modes such as rating and ranking, it was shown that there were changes in the evaluation of only the riskless option and there was little or no effect on the risky option (Kühberger & Gradl, 2013). Working memory capacity has been shown to be correlated with the scale of framing effects exhibited by individuals (Corbin, McElroy & Black, 2010). However, in a replication study, it was found that working memory capacity did not interact with framing manipulations (Bogdanov et al., 2023).

To conclude, framing effects have been tested in various settings and have been observed as reliable and robust effects. Some factors have been shown to mitigate this effect. In this thesis, we use the standard disease problem with no variations in scale, risk level or response mode. We adapted the materials to be in a farming context and we tested the farming scenarios together with the disease one. We gave participants recommendations to choose the non-typical choices, which were the risky option in the gain frames and the risk-averse option in the loss frames, and we aimed to see the effect of non-typical recommendations on the framing effects.

Counterfactual Simulation

Echoing the research question we raised in Chapter 2, we take into account that counterfactuals call for the simultaneous generation of dual possibilities (Byrne, 2016), and that people can also make inferences from counterfactuals which are otherwise difficult to make (Santamaría et al., 2005;

Thompson & Byrne, 2002). Motivated by the cognitive effects of counterfactual thinking, we tested participants with counterfactual explanations in some of the experiments in this chapter, to see whether a counterfactual explanation from an AI system and a human expert can prompt people to follow recommendations even more. Recall that some theories we listed earlier explain framing effects by suggesting that people have an incomplete representation of the problem of the options, such as the fuzzy-trace theory (Reyna & Brainerd, 1991; Reyna et al., 2021) and the schema theory (Jou et al., 1996; Okder, 2012). We hypothesised that having a counterfactual explanation would help participants to maintain a mental representation of the typical choice, but also an alternative one which involves the non-typical choice, and results in a bigger proportion of switches in decisions. We also extended the investigation from XAI counterfactual explanations to human expert counterfactual explanations.

AI & Human Expert Recommendations

One of the research questions we address in this chapter is how people react to AI system recommendations, and we tested that by comparing the acceptance of the same recommendations from a human expert to those from an AI system. We investigated if people would switch their decisions when an AI system recommends a decision that is opposite to their preferred choice. We chose a very robust cognitive phenomenon as our testing material, the risky choice framing effect, which was discussed in the previous section. We based our hypothesis on the effect of recommendations from research evidence in the field of expert recommendations and AI recommendations.

Acceptance of Expert Recommendations

People show a willingness to trust and follow expertise: when participants were asked to judge the certainty of a premise in a deductive argument, they perceived the likelihood of a conclusion to be greater when the premises were uttered by an expert (a professor of medicine) rather than a novice (a medical student) (Stevenson & Over, 2001). Expertise also has an impact on communication and people's estimations of what other people know, for example, when computer experts and laypersons estimate the distribution of Internet concepts and general knowledge distribution among students, laypersons showed a stronger bias than experts (Bromme, Rambow & Nückles, 2001).

The impact of others, especially of experts' opinions exists throughout the cognitive process of taking advice, for example, advice-taking has been conceptualised as three cognitive processes, and participants were tested using functional magnetic resonance imaging (Meshi et al., 2012). The first process is the valuation of advice, during which people use expert advice to a greater degree than novices (Harvey & Fischer, 1997; Jungermann & Fischer, 2005; Sniezek, Schrah & Dalal, 2004). The second process is the assessment of the "opinion difference" (the comparison between the advisor's opinion and one's own opinion). This difference is used to the degree to which the person is influenced by the advice since the weight of advice decreases when its distance from the initial opinion increases (Yaniv, 2004). The third process is combining valuation and the opinion difference to being an actual utilisation. Our research question was based on people's tendency to follow expert recommendations, and we were interested in testing the effect of the recommendation of an expert

that is opposite to what people normally choose. However, this trust towards expert recommendations can be challenged, for example, when expert testimony was disclosed to be false (Moreno-Ríos & Byrne, 2018).

AI Appreciation and AI Aversion

The term *algorithmic appreciation* was used to describe the phenomenon that users chose algorithmic advice more than human advice in a head-to-head comparison (Logg, Minson & Moore, 2019), for example, it was shown that AI decisions, or automated decision-making, tended to be perceived as more rational and objective than a human counterpart (Dijkstra, Liebrand & Timminga, 1998). However, extensive research has also shown a contradictory phenomenon when it comes to AI advice: the term *algorithm aversion* was used to describe the phenomenon that a human forecaster wins when compared to a statistical algorithm even when the latter shows higher accuracy. It has been shown that people are especially averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster. People more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake (Burton et al., 2020; Dietvorst et al., 2015; Mahmud et al., 2022). In some scenarios, the role of humans cannot be replaced by AI, no matter the perceived usefulness and ease of use (Kelly et al., 2023). For example, AI involvement was rejected in Christian education by Christian respondents (Tran & Nguyen, 2021). In another example, Burgundy wine producers were reluctant to use AI technologies in the wine industry as they have a preference for tradition and do not want their culture to be disrupted (Atwal, Bryson & Williams, 2021). The advice between a human

expert and a machine expert, as well as advice between a human novice and a machine novice, were directly compared and it was found that automated novices were trusted more than human novices and that human experts were trusted more than machine experts (Madhavan & Wiegmann, 2007). This aversion has been tested in different domains, and we consider a few examples here. When receiving stock price forecasting advice from either a human expert or a statistical forecasting method, participants were able to incorporate both and improve their accuracy but to a lesser degree for the statistical method (Önköl et al., 2009). When it comes to human resource and employee selection, human evaluation was reported to be more useful, fair and flexible compared to algorithm-based evaluation (Diab et al., 2011), and people were less comfortable outsourcing emotion-based jobs (compared to cognition-based jobs) to robots when they rated different tasks on MTurk to be more “thinking” or “feeling” and rated their perception of bot-sourcing these two types of tasks (Waytz & Norton, 2014). In health care, it was found that people prefer to receive treatments from humans, and they think algorithms neglect the uniqueness of their cases (Longoni, Bonezzi & Morewedge, 2019). In marketing and consumer behaviour studies, it has been shown that even though AI recommendations can mimic a human expert, user acceptance and perception might differ (Li, Rau & Huang, 2020). When it comes to hedonic products (e.g., buying a new laptop for entertainment purposes), humanised AI are more effective than non-humanised AI, while these two AI approaches were found to be equally effective when it comes to utilitarian products (e.g., buying a new laptop for work purposes) (Wien & Peluso, 2021). Consumer’s responses towards recommendations from an AI system were compared to

recommendations from humans when it comes to vice vs virtue frame products, i.e., products that provide short-term pleasure such as tastiness vs long-term benefits such as healthiness. A more negative response towards AI recommendations was found (Yang et al., 2024). There have been some studies that looked into the role of AI on decision change. For example, participants were asked to complete a questionnaire to assess their attitude towards risk. Then, they were asked to choose a level of risk in a lottery game that aligns with their preferences. If their choice deviates from their optimal choice, then the decision support system intervenes by recommending a more suitable option (Lammert et al., 2024). In a situation where virtual advisors were utilised for health-related behaviour change, participants engaged in dialogues with the virtual advisor. The advisor aimed to help manage study stress for students and provided recommendations for behaviours such as studying in a group, even if some students consider it a waste of time. The advisor provided different versions of recommendations, some based on the individual's goal to get better grades and some focused on the individual's belief that studying in a group is a waste of time (Abdulrahman, Richards & Bilgin, 2022). In this chapter we designed the experiments to test AI system recommendations and human expert recommendations, using a robust psychological effect – the risky choice framing effect, to test differences between these recommendation sources on decision change.

Overview of Experiments

As we discussed above, we have three primary hypotheses that were motivated by the literature: firstly, we aimed to test if non-typical

recommendations (to be risky in the gain frame and be risk-averse in the loss frame) mitigate framing effects, and we hypothesised that recommendations would mitigate the framing effects; secondly, we aimed to test the differences in the effects between AI agent recommendations and human expert recommendations on people's decision-making, we hypothesised that participants would follow human expert recommendations more; thirdly, we aimed to test if providing a counterfactual explanation could prompt people to consider alternative options in risky decision-making, and we hypothesised that having counterfactual explanations together with non-typical recommendations would mitigate the framing effects even more.

Experiment 7 tested AI system recommendations (e.g., An Artificial Intelligence decision support system recommends that you choose Program B) and Experiment 8 tested human expert recommendations (e.g., A medical expert recommends that you choose Program B). We gave participants the framing effect scenarios in the disease domain and the farming domain too, so we can gain insight into what sorts of explanations benefit AI decision support systems for agriculture. We gave participants a non-typical recommendation together with the scenario and asked for their decisions. Experiment 9 directly compared information sources as a between-groups factor, e.g., if people had chosen the typical options (Program A) in the gain framing, would they change their decisions after reading an AI recommendation or a human expert recommendation? In Experiment 10 and Experiment 11, we added counterfactual explanations such as “The medical expert recommends that you choose Program A. The expert says that in previous cases the outcome would have been better if Program A had been chosen” to the recommendations, and

we tested AI recommendations with counterfactual explanations and human expert recommendations with counterfactual explanations accordingly.

Experiment 7

In this experiment, we tested whether participants' tendency to choose the typical decisions (the safe program A for gains and the risky program B for losses) would decrease when an AI system recommends the non-typical choice, and we compared that with the control condition where no recommendations were provided. Given that the risky choice framing effect is a robust phenomenon, we hypothesised that firstly AI system non-typical recommendations would change participants' choice in the frame they were represented with, but not completely. As an active control, we also compared participants' choices when an AI recommended the opposite choice to the typical one, to participants' choices when an AI recommended the same choice as the typical one. We hypothesised that the framing effect would still occur when participants were recommended to make the typical choice, and maybe even as a stronger effect. We tested a new farming content as well as the disease content.

Method

Participants

The participants were 297 members of the general public recruited from the platform Prolific. There were 112 men, 174 women, 1 non-binary individual, and 10 preferred not to say. They were aged between 18 and 76 years with a mean age of 39 years. As in previous chapters, we selected

participants in this chapter to be adults from Ireland, the United Kingdom, the United States, Canada, Australia and New Zealand for whom English was their first language. A sample size of 270 participants was required to detect a medium effect size with 90% power at $p < .05$, according to an a-priori G*Power test, and we recruited 297 to ensure we had enough participants in each condition after randomisation done by the survey platform Alchemer. No attention checks were implemented since participants completed only two items and no participants were eliminated. Each participant was paid £0.50 UK sterling for participation, which took three minutes to complete on average.

Design

We aimed to test the framing effect in the standard version, where participants were only exposed to one frame, hence a fully between-groups design was adopted. Both framing (gain vs. loss) and information condition (no recommendation vs. typical recommendation vs. non-typical recommendation) were manipulated between groups. The participants were randomly assigned to one of the six groups: gain & no recommendation ($n = 54$), loss & no recommendation ($n = 58$), gain & typical recommendation ($n = 50$), loss & typical recommendation ($n = 45$), gain & non-typical recommendation ($n = 45$), and loss & non-typical recommendation ($n = 45$). We measured one dependent variable: their choice between the safe program A and the risky program B.

Materials & Procedure

Scenarios in the disease domain were taken from Tversky and Kahneman (1981). We adapted the materials to a farming content, where a disease was expected to impact a farm and kill 60 cows. See Appendix 4.1 for information and debrief sheets of this chapter and see Appendix 4.2 for the full set of materials for all experimental conditions. After reading the scenario, participants in the typical recommendation condition and the non-typical recommendation condition read a line at the end of the scenario which said, “An Artificial Intelligence decision support system recommends that you choose Program A/B”. Those who were assigned to the typical recommendation conditions received a recommendation from an Artificial Intelligence decision support system of choosing program A in the Gain condition, and program B in the Loss condition. Those who were assigned to the non-typical recommendation conditions received a recommendation from an Artificial Intelligence decision support system of choosing program B in the Gain condition, and program A in the Loss condition. Participants accessed the experiment using their own devices in their own time. The materials were presented via Alchemer. Participants were directed to the survey through a link from Prolific. After reading background information and providing consent, the participants were asked to voluntarily provide their age and gender. Then each participant read the scenarios in the disease content and farming content in a randomised order. The participants were asked “Which of the two programs would you choose” and they had to choose between the two programs, the order of which was randomised. In the end, they were given debriefing

information and provided with an open text box to leave their comments if they wished to do so.

Results and Discussion

All data for all experiments in this chapter are shared on OSF (<https://osf.io/ywd7h>). There were no differences between the disease and farming domains and no effects of the order in which participants received the two domains (See Appendix 4.3.1 for supplementary analysis). We labelled the choices of non-typical options as 1, and the choices of typical options as 0. Each participant's score of the two domains was summed up and indicated as 0, 1, or 2.

When they received no recommendation, in the gain frame, 77% of participants made the typical choice of safe Program A and 23% of participants made the non-typical choice of the risky Program B, and in the loss frame, 66% of participants made the typical choice of risky Program B and 34% of participants made the non-typical choice of the safe Program A. We note this pattern of results corresponds to the standard framing effect observed in previous research. When they received recommendations that were the same as the typical choices, in the gain frame, 17% of participants made the non-typical choice of the risky Program B, and in the loss frame, 23% of participants made the non-typical choice of the safe Program A. When they received recommendations that were opposite typical choices, in the gain frame, 23% of participants made the non-typical choice of the risky Program B, and in the loss frame, 51% of participants made the non-typical choice of the

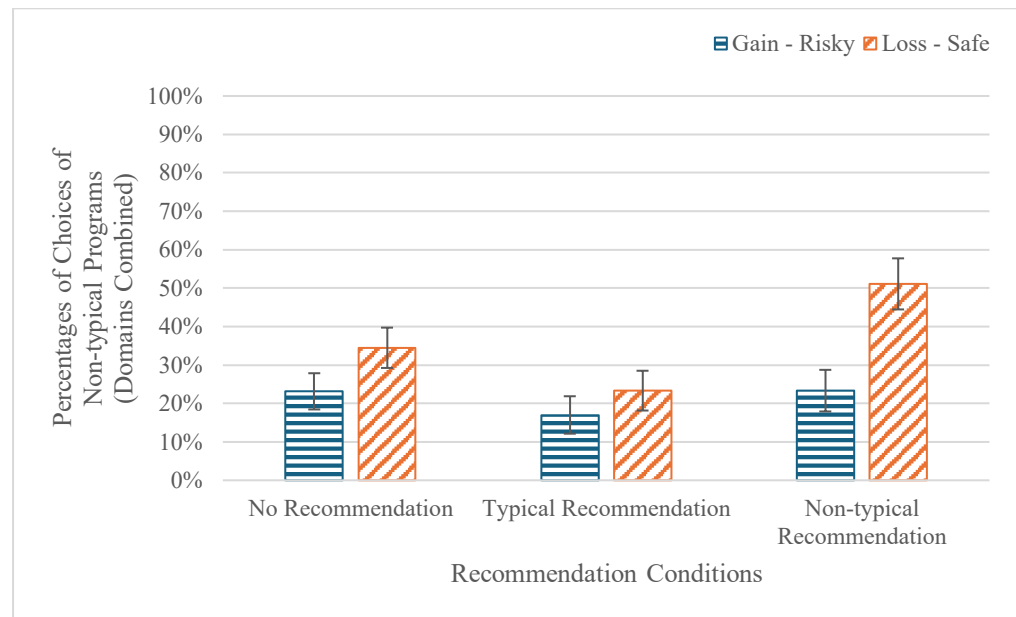
safe Program A. **Figure 4.1** shows the responses from all participants, combined between the two domains.

Since the responses were ordinal data and the normality assumption was violated, we carried out non-parametric tests for experiments in this chapter. A Mann-Whitney U test compared each participant's summed score from two domains and indicated there was a significant difference between the gain frame and the loss frame overall, $U = 13199.5$, $z = 3.365$, $p = .0008$, $r = 0.20$. A Kruskal -Wallis test across the three recommendation conditions overall was significant, $H(2) = 8.799$, $p = .012$, $\eta^2 = 0.02$. To examine this result further, We carried out 4 comparisons, with a Bonferroni-adjusted alpha level of $p < .0125$. We compared the non-typical choices between the non-typical recommendation condition and the no recommendation condition, as well as between the non-typical recommendation condition and the typical recommendation condition. In the gain frame, the risky program B was chosen as often by participants in the non-typical recommendation condition, as in the no recommendation condition, $U = 1204.5$, $z = -0.088$, $p = .930$, and the same as the typical recommendation condition, $U = 1009.5$, $z = 1.101$, $p = .271$, indicating that the AI recommendation had no effect in the gain frame. In the loss condition, the safe program A was chosen somewhat more often by participants in the non-typical recommendation condition than the no recommendation condition, $U = 1039$, $z = 1.902$, $p = .057$, and more often compared to the typical recommendation condition, $U = 668.5$, $z = 3.034$, $p = .002$, $r = 0.32$, indicating that the AI recommendation increased the tendency to choose the safe program A in the non-typical recommendation condition for losses. Hence, the differences in recommendations were due to participants

choosing the non-typical program more often in the non-typical recommendation condition than in the typical recommendation condition.

Figure 4.1

The Percentages of Choices of Non-Typical Programs in Experiment 7



Note. The percentages of choices of non-typical programs when participants were given no recommendation or a typical recommendation by an AI decision support system, that is, program A in the Gain frame, or program B in the Loss frame, or the non-typical recommendation, that is, program B in the Gain frame, or program A in the Loss frame in Experiment 8. Error bars are standard errors of the mean.

The results showed firstly that the framing effect still existed for almost all conditions since participants still chose the typical program more than chance (>50%), no matter the recommendations. This showed that the risky

choice framing effect is robust and implied that prompting participants to consider different options only mitigates this effect to a relatively small degree.

Secondly, participants made more non-typical choices when an AI system recommended the non-typical choice for losses, whereas AI recommendations did not affect their choices of the typical decision of the safe program A for gains. The results indicated that when an AI system provides a recommendation that is non-typical, the recommendation was effective in restraining people from making risky choices in losses but not effective in encouraging people to be more risk-seeking in gains. This difference we observed was consistent with the prospect theory and loss aversion accounts for the framing effect: since people exhibit cognitive asymmetry towards gains versus losses, and since a loss feels greater than an equivalent gain, people would be more sensitive towards potential losses, resulting in their reluctance to be riskier.

Given that we observed an effect of AI system recommendations only in some experimental conditions, we followed up with Experiment 8, where we looked at the effect of human expert recommendations in the same experimental design.

Experiment 8

The overall objective of Experiment 8 was to test the effect of human expert recommendations on the risky choice framing effect. Participants were given a recommendation in the form of “a medical/farming expert recommends that you choose...”. Given the research evidence on algorithmic aversion, we aimed to investigate the influence of human expert recommendations. Given

the algorithmic aversion phenomena, we hypothesised that recommendations from a human expert would have a stronger influence on participants' decisions and that the framing effects would be mitigated when a human expert recommends the non-typical program.

Method

Participants

The participants were 312 members of the general public recruited from the platform Prolific. There were 133 men, 162 women, 5 binary individuals and 12 preferred not to say. They were aged between 18 and 73 years with a mean age of 36 years. We selected participants based on the same criteria as Experiment 7, and the estimated sample size was also the same as Experiment 7 due to the same design.

Due to an administrative error in randomisation, we recruited more participants in two of the groups, loss & no recommendation ($n = 108$) and gain & typical recommendation ($n = 96$). In total, 412 participants were recruited. To control for sample sizes between groups, we calculated the average sample size of the other four conditions ($n = 52$) and used the random selection function in SPSS to randomly select 52 participants from these two groups. After selection 312 data were retained for analysis. Results from the analysis of 412 participants were consistent with the results of the selected sample of 312 and can be found in Appendix 4.3.2. No attention checks were implemented because each participant only completed two items, and no participants were eliminated. Each participant was paid £0.50 UK sterling for participation, which took three minutes to complete on average.

Design

The design was the same as the previous experiment. The participants were randomly assigned to one of the six groups: gain & no recommendation ($n = 45$), loss & no recommendation ($n = 52$), gain & typical recommendation ($n = 52$), loss & typical recommendation ($n = 59$), gain & non-typical recommendation ($n = 54$), and loss & non-typical recommendation ($n = 50$).

Materials & Procedure

The materials were the same as in Experiment 7; we only changed “the Artificial Intelligence decision support system” to “a medical/farming expert”. In the disease content, participants were told “a medical expert recommends that...”, while in the farming content, participants were told, “a farming expert recommends that...”. See Appendix 4.2 for the full set of materials for all experimental conditions. The procedure was the same as the previous experiment.

Results and Discussion

There were no differences between the disease and farming domains, and no effects of the order in which participants received the two domains, see Appendix 4.3.1 for supplementary analysis. We combined responses from both domains and orders.

When they received no recommendation, in the gain frame, 79% of participants made the typical choice of safe Program A, and 21% of participants made the non-typical choice of risky Program B. In the loss frame, 49% of participants made the typical choice of risky Program B and 51% of

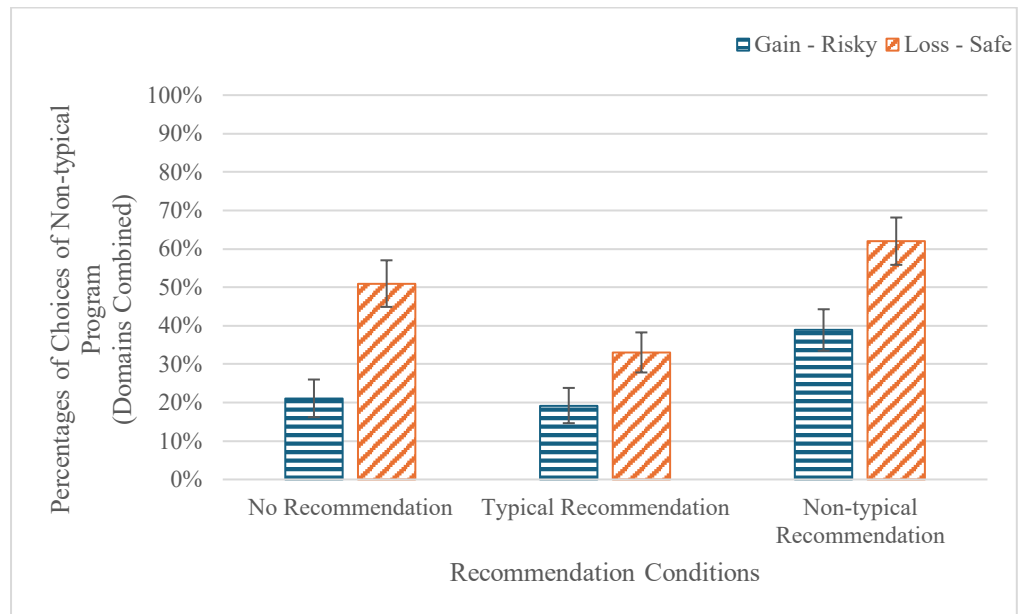
participants made the non-typical choice of the safe Program A. We note the pattern of results for losses departs from the standard framing effect observed in previous research. When they received recommendations that were the same as typical choices, in the gain frame, 19% of participants made the non-typical choice of the risky Program B, and in the loss frame, 33% of participants made the non-typical choice of the safe Program A. When they received recommendations that were opposite typical choices, in the gain frame, 39% of participants made the non-typical choice of the risky Program B, and in the loss frame, 63% of participants made the non-typical choice of the safe Program A. **Figure 4.2** shows the responses from 312 participants.

A Mann Whitney U test indicated there was a significant difference in the choices of the non-typical programs between gain and loss frames overall, $U = 15305.5$, $z = 4.306$, $p < .0001$, $r = 0.24$; for information conditions, Kruskal-Wallis test by ranks was significant, $H(2) = 16.862$, $p = .0002$, $\eta^2 = 0.042$. We carried out 4 comparisons, with a Bonferroni-adjusted alpha level of $p < 0.0125$. We compared the non-typical choices between the non-typical recommendation condition and the no recommendation condition, as well as between the non-typical recommendation condition and the typical recommendation condition. In the gain condition, the risky program B was chosen somewhat less often by participants in the non-typical recommendation condition, compared to the no recommendation condition, $U = 915$, $z = 2.345$, $p = .019$, and it was chosen less when compared to the typical recommendation condition, $U = 1014$, $z = 2.778$, $p = .005$, $r = 0.27$, indicating that the human expert recommendation reduced the tendency to choose the safe option A in the gain frame. In the loss condition, the safe program A was chosen as often by

participants in the non-typical recommendation condition as in the no recommendation condition, $U = 1120$, $z = 1.297$, $p = .195$ (because of the unusually high choices of the non-typical program A in the no recommendation condition), and more often than in the typical recommendation condition, $U = 953$, $z = 3.402$, $p < .001$, $r = 0.33$, indicating that the human expert recommendation increased the tendency to choose the safe program A in the non-typical choice condition. Again, participants chose the non-typical program more when provided with the non-typical recommendation compared to when provided with the typical recommendation.

Figure 4.2

The Percentages of Choices of Non-Typical Programs in Experiment 8



Note. The percentages of choices of non-typical programs when participants were given no recommendation or the typical recommendation, that is, program A in the Gain frame, or program B in the Loss frame, or the non-typical recommendation, that is, program B in the Gain frame, or program A in the Loss frame in Experiment 8. Error bars are the standard errors of the mean.

The results showed that participants chose the non-typical program more when a human expert recommended the non-typical choice for both gain and loss frames. Unlike the effect of an AI system's recommendations in Experiment 7, the results from this experiment showed an effect of human recommendations not only for losses but also for gains. This result indicated that firstly, human expert recommendations were more effective than AI recommendations in changing decisions since the recommendations worked for both frames; and that participants chose the non-typical option more when

compared to Experiment 8 (50% vs 37%). Secondly, the differences in the tendency to change decisions for gains versions and losses versions showed a different pattern when compared to Experiment 7.

In Experiments 7 and 8, we gave participants the recommendations immediately with the scenarios and asked for their choice between the two programs. One shortcoming of this approach was that there will be a certain percentage of participants who would choose the non-typical option even without a recommendation, hence the observation of the effect of a recommendation is somewhat masked. We followed up with Experiment 9, where we gave participants the scenario and asked for their choices before being given a recommendation, and then asked for their choices after they were given a non-typical recommendation, to examine more directly any decision change. We also factored in the recommendation source as an independent variable, so we could compare directly the difference between AI recommendations and human recommendations.

Experiment 9

This experiment aimed to test whether participants' choices of the typical decisions (the safe program A for gains and the risky program B for losses) decrease when an AI system or a human expert recommends the non-typical choice when participants have committed to their own choice before being given a recommendation. As a result, in this experiment, we had to screen participants and retain data only from those who made typical choices before the recommendation, in other words, we expected to have to eliminate about one-fifth of participants. We compared participants' choices after they

received a non-typical recommendation to their choices before they received a recommendation, to test the effect of a non-typical recommendation. We hypothesised that we would replicate our findings from the previous experiments in that we would observe a decrease in people's tendency to choose the typical choice given either an AI system's recommendations or a human expert's recommendations, and we also hypothesised that human expert recommendations would have a stronger impact than AI system recommendations.

Method

Participants

The participants were 242 members of the general public, recruited from the platform Prolific. There were 105 men, 134 women, and 3 preferred not to say. They were aged between 19 and 74, with a mean age of 37 years. We selected participants with the same demographics as in previous experiments. According to a g-power test with an estimated medium effect size and power level of 0.9, an estimated 232 participants were required.

We screened participants by whether they made the typical choice in their pre-recommendation response in the first domain they received, i.e., whether they chose the safe program A if given a gain version and whether they chose the risky program B if given a loss version. Each participant was given two versions of the scenario, one based on the disease content, and one based on farming content, which were presented in a different random order to each participant. However, in the end, we analysed only participants' responses to the first content they were given because of the necessary requirement to

eliminate participants who made the non-typical decision initially, to avoid eliminating more participants than necessary.

The recruitment included two waves: in the first wave we recruited 262 participants, and we recruited 94 more participants in our second wave. Two hundred forty-two participants (68% of the overall number of participants tested) made the typical decision, consistent with earlier research, for the first content they received. Of these, 167 participants also made the typical decision for the second content they received (50% of the overall number of participants tested). No attention checks were implemented since each participant only completed two items. Each participant was paid £ 0.45 UK sterling for participation, which on average lasted three minutes.

Design

A fully between-groups design was adopted. Both framing (gain vs. loss) and information source (AI system vs human expert) were manipulated between groups. The participants were randomly assigned to one of the four groups: gain & AI system ($n = 64$), loss & AI system ($n = 59$), gain & human expert ($n = 61$), and loss & human expert ($n = 58$). We measured participants' choices before they were given a recommendation and their choices after the non-typical recommendation.

Materials & Procedure

Scenarios were the same as in previous experiments, firstly, participants were asked to make a choice in the pre-recommendation phase before recommendations were given. After participants had made their initial choice,

their choice remained on the screen and they received a pop-up recommendation with the non-typical recommendation (risky program B for the gain condition and safe program A for the loss condition), after which they were asked to make a choice again (see **Figure 4.3**). The procedure was the same as in previous experiments.

Figure 4.3

An Example of the Experimental Task in Experiment 9

Imagine that your country is preparing for an outbreak of an unusual disease which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows: If program A is adopted, 200 people will be saved. If program B is adopted, there is a 1/3rd probability that 600 people will be saved, and 2/3rd probability that no-one will be saved. Which of the two programs would you choose? ☒ Program A ☐ Program B	Imagine that your country is preparing for an outbreak of an unusual disease which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows: If program A is adopted, 200 people will be saved. If program B is adopted, there is a 1/3rd probability that 600 people will be saved, and 2/3rd probability that no-one will be saved. Which of the two programs would you choose? ☒ Program A ☐ Program B An AI decision support system recommends that you choose program B. Which of the two programs would you choose now? ☐ Program A ☐ Program B

Note. An example of the presentation of the experimental task. Participants could choose program A or program B by ticking a box (illustrated on the left, for the choice of program A). They were then presented with the recommendation of an expert and could choose program A or program B by ticking a box again (illustrated on the right).

Results and Discussion

Again, we labelled the typical choices as 0, and the non-typical choices as 1. We analysed the responses of participants who made the typical choice

on the first presented domain (242 participants). There were no differences between the disease and farming domains, and no effects of the order in which participants received the two domains, see Appendix 4.3.1. We carried out a supplementary analysis with responses from participants who made the typical choice on both the first and second presented domains (167 participants), which showed similar results, see Appendix 4.3.3.

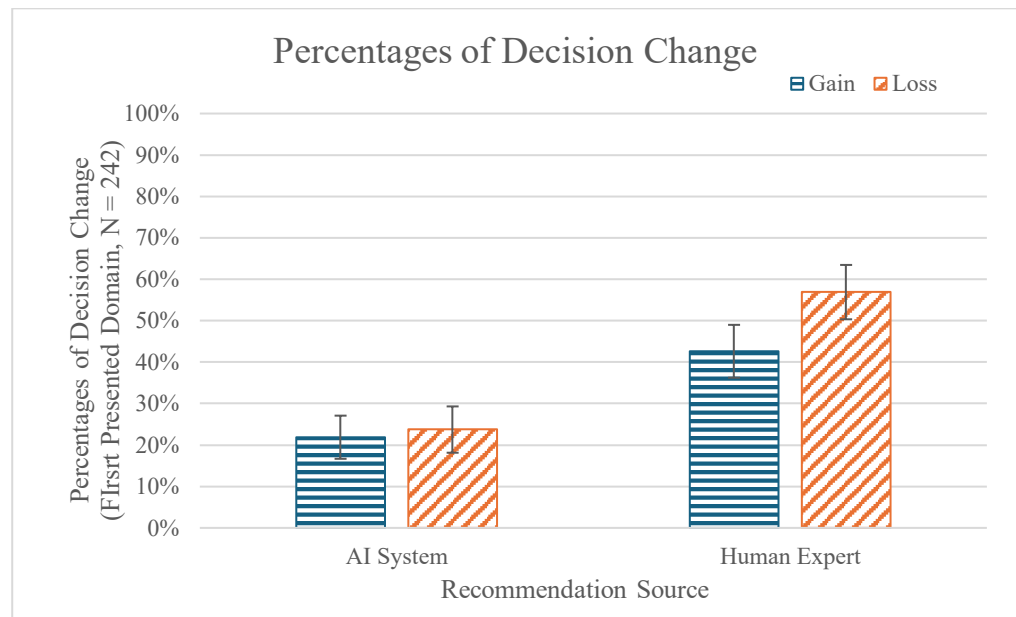
Figure 4.4 shows the responses from the 242 participants to the first scenario they received. When they were recommended to make the non-typical choice to the typical one that they had made for this first scenario, by an AI system, in the gain frame, 78% of them continued to choose the safe option and 22% changed to the recommended risky one; in the loss frame, 76% continued to choose the risky option and 24% changed to the recommended safe one. Given a human recommendation, in the gain frame, 57% of participants continued to choose the safe option, and 43% changed to the recommended risky one; in the loss frame, 43% continued to choose the risky option, and 57% changed to the recommended safe one.

We analysed responses from the first presented domain from 242 participants using non-parametric tests as in previous experiments. There was a significant effect of recommendation source, $U = 9281$, $z = 4.337$, $p < .0001$, $r = 0.279$, and no effect of framing, $U = 6715$, $z = 1.321$, $p = .187$. Four comparisons with a Bonferroni adjusted p-value of $p < .0125$ showed that for gains, participants switched their choices from the safe option to a risky one following a human expert's recommendation more than an AI system's recommendations, $U = 1547$, $z = 2.476$, $p = .013$, $r = 0.221$; and they also switched more for losses, $U = 1143.5$, $z = 3.643$, $p = .0003$, $r = 0.337$. There

were no differences between gains and losses for AI recommendations, $U = 1853$, $z = 0.244$, $p = .807$; or human ones, $U = 1516.5$, $z = 1.550$, $p = .121$.

Figure 4.4

Percentages of Decision Change in Experiment 9



Note. Percentages of switches from a safe choice to a risky one (for gains) and from a risky choice to a safe one (for losses) switches following a recommendation from an AI system or a human expert, for the first scenario received in Experiment 9. Error bars are standard errors of the mean.

The results showed that a recommendation to make the non-typical choice changed participants' otherwise very robust natural tendencies: they made more risky choices (for gains), and more risk-averse choices (for losses). In Experiment 7, we observed the effect of AI system recommendations only for losses, not for gains, whereas in this experiment we observed the effect of AI recommendations for both frames. One explanation for the inconsistency is

the change in the tasks – in Experiment 7, we did not ask for participants' decisions first before we examined the influence of a recommendation. As a result, the responses included participants who would have made a non-typical recommendation anyway, hence their responses cannot provide any insight in terms of the effect of an opposite recommendation. We suggest that responses from Experiment 9 paint a clearer picture of the influence of non-typical AI system recommendations. What is more, we did not see a difference in the percentages of switches between the gain frame and the loss frame following either AI system recommendations or human expert recommendations, consistent with Experiments 7 and 8.

The results also showed that participants changed their choices more often when a human expert made the non-typical recommendation opposite to their original choice, rather than when an AI decision support system did so, consistent with previous experiments. The results were consistent with our prediction that people are more inclined to follow human expert recommendations than AI ones in making risky or safe choices. In the following experiments, we continued to test participants in the pre and post-recommendation setup, and we investigated the effect of counterfactual explanations in decision change.

Experiment 10

In this experiment, we tested the hypothesis that counterfactual explanations about how things could have turned out differently in the past prompt people to mentally simulate alternatives to reality and, hence, prompt participants to change their decision following a non-typical recommendation

to a greater degree. We based our hypothesis on assumptions of theories such as the fuzzy trace theory that people exhibit framing effects because they have not fully flashed out the disease problem and the two options. Hence, when they are prompted by counterfactual explanations to mentally keep track of more possibilities, they could have a more complete mental representation of the problem and be less biased by the framing effect. Given the growing interest in the application of counterfactual explanations in eXplainable AI, we tested counterfactual explanations such as “An Artificial Intelligence decision support system recommends that you choose program B. It says that in previous cases the outcome would have been better if Program B had been chosen”. We continued to test participants in a pre-recommendation and post-recommendation setup, to test the decision changes of each individual.

Method

Participants

The participants were 248 members of the general public, recruited from Prolific, there were 89 men, 151 women, 2 non-binary individuals and 6 preferred not to say. They were aged between 19 and 80 with a mean age of 39 years. We selected participants with the same demographics and the same sample size as in Experiment 9. We screened participants in the same way as in Experiment 9.

The participant recruitment included 2 waves: in the first wave we recruited 278 participants. After screening, 203 participants were retained for analysis and 75 participants were eliminated. We recruited 70 more

participants, of whom 25 were eliminated and 45 were retained for analysis after screening.

Design

A fully between-groups design was adopted. Both framing (gain vs. loss) and explanation condition (no explanation vs counterfactual explanation) were manipulated between groups and we tested only an AI system's recommendations and explanations in this experiment. The participants were randomly assigned to one of the four groups: gain & no explanation ($n = 65$), loss & no explanation ($n = 61$), gain & counterfactual explanation ($n = 64$), and loss & counterfactual explanation ($n = 58$). We measured participants' choices before they were given a recommendation and their choices after the recommendation.

Materials & Procedure

The scenarios and experiment set-up were the same as in Experiment 10. Participants in the no explanation condition received the same AI system recommendation as previous experiments: "An Artificial Intelligence decision support system recommends that you choose program A/B", whereas participants in the Counterfactual Explanation condition received an additional counterfactual explanation: "An Artificial Intelligence decision support system recommends that you choose program A/B. It says that in previous cases the outcome would have been better if Program A/B had been chosen". The procedure was the same as in Experiment 9.

Results and Discussion

We labelled our data in the same way as in Experiment 9. We analysed the responses of participants who made the typical choice on the first presented domain (248 participants). There were no differences between the disease and farming domains, and no effects of the order in which participants received the two domains, see Appendix 4.3.1. As a supplementary analysis, we also analysed the responses of participants who made the typical choice on both the first and second presented domains (174 participants), see Appendix 4.3.3.

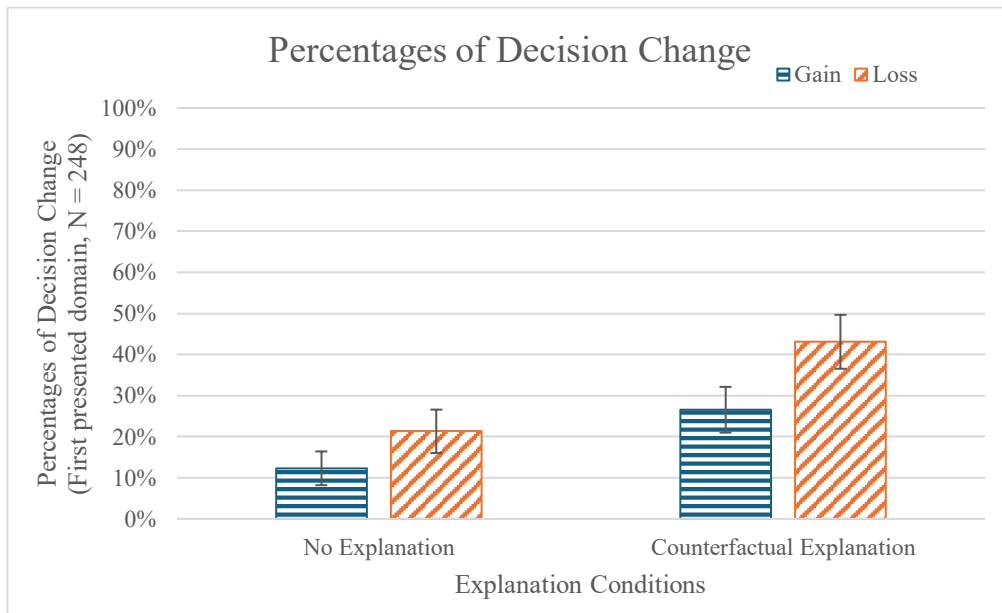
Figure 4.5 shows the responses from the 248 participants to the first scenario they received. When they were recommended to make the opposite choice to the typical one that they had made with no explanation, in the gain frame, 88% of them continued to choose the safe option and 12% changed to the recommended risky one; in the loss frame, 79% continued to choose the risky option and 21% changed to the recommended safe one. Given a recommendation and a counterfactual explanation, in the gain frame, 73% of participants continued to choose the safe option, and 27% changed to the recommended risky one; in the loss frame, 57% continued to choose the risky option, and 43% changed to the recommended safe one.

Mann Whitney U non-parametric tests indicated that there was a significant effect between the explanation conditions, $U = 9051$, $z = 3.205$, $p = .001$, $r = 0.204$, and a significant effect of framing, $U = 8639$, $z = 2.264$, $p = .024$, $r = 0.144$. The results showed that people are influenced more by the AI system's recommendations with a counterfactual explanation than without an explanation. Subsequent comparisons with a Bonferroni adjusted p-value of $p < .0125$ showed that for gains, there was no difference in participants'

tendency to switch their choices following a counterfactual explanation versus no explanation, $U = 1783.5$, $z = 2.040$, $p = .041$ (on the Bonferroni corrected alpha); but for losses they switched more following a counterfactual explanation versus no explanation, $U = 1383.5$, $z = 2.583$, $p = .011$, $r = 0.233$. There were no differences between gains and losses following no explanation, $U = 1804$, $z = 1.350$, $p = .177$; or a counterfactual explanation, $U = 1549$, $z = 1.912$, $p = .056$.

Figure 4.5

Percentages of Decision Change in Experiment 10



Note. Percentages of switches from a safe choice to a risky one (for gains) and from a risky choice to a safe one (for losses) switches following a recommendation with no explanation or with a counterfactual explanation for Experiment 10, for the first scenario received. Error bars are standard errors of the mean.

The results showed that participants changed their choices more often when a counterfactual explanation was provided with an AI recommendation rather than when no explanation was provided, as indicated by the significant main effect of explanation conditions. This was consistent with our prediction that counterfactual explanations would be effective in further changing people's decisions because counterfactual explanations require a richer mental presentation of alternatives to reality. Various theories of framing effects such as the effort reduction theory (Reyna & Brainerd, 1991; Reyna et al., 2021) and the schema theory (Jou et al., 1996; Okder, 2012) point out that people don't have a clear representation of the risky disease problem as they struggle to realise the equivalence of 200 people living and 400 people dying. Counterfactual simulations, on the other hand, prompt participants to mentally process more than one reality and are therefore useful in changing people's decisions for framing effects.

The results also indicated that participants switched their decisions to the same degree in gains and losses, consistent with Experiments 8 and 9. In the next experiment, we test the efficacy of counterfactual explanations when they come from human experts.

Experiment 11

Counterfactual explanations have been shown to improve user satisfaction and understanding of AI decisions, but it is still unknown how counterfactual explanations would impact decisions when we compare AI counterfactual explanations and human counterfactual explanations. Humans offer counterfactual explanations all the time in various social contexts (e.g.,

Catellani, 2012; Catellani & Bertolotti, 2014; Catellani & Covelli, 2013).

Experiment 11 aimed to explore the efficacy of counterfactual explanations from human experts in the context of risky decision-making. We compared decision changes when participants were told “A medical/farming expert recommends that you choose program A/B . The expert says that in previous cases the outcome would have been better if Program A/B had been chosen”, compared to when they were given no explanations at all. Given the effect of counterfactual explanations in Experiment 10, we hypothesised that those who received a counterfactual explanation would change their decision to a greater degree.

Method

Participants

The participants were 243 members of the general public recruited from Prolific, there were 120 men, 115 women, 3 non-binary individuals and 5 preferred not to say. They were aged between 19 and 78 with a mean age of 42 years. We selected the participants with the same criteria and sample size as in Experiments 9 and 10.

In the first wave, we recruited 240 participants. After screening, 171 participants were retained for analysis and 69 participants were eliminated. We recruited 111 more participants, of whom 39 were eliminated and 72 were retained for analysis after screening.

Design

The design was the same as Experiment 10. The participants were randomly assigned to one of the four groups: gain & no explanation ($n = 60$),

loss & no explanation ($n = 58$), gain & counterfactual explanation ($n = 66$), and loss & counterfactual explanation ($n = 59$).

Materials & Procedure

The scenarios and experiment set-up were the same as in Experiments 9 and 10. Participants in the no explanation condition received the same human expert recommendation as previous experiments: “A medical/farming expert recommends that you choose program A/B”, whereas participants in the counterfactual explanation condition received a counterfactual explanation: “A medical/farming expert recommends that you choose program A/B . The expert says that in previous cases the outcome would have been better if Program A/B had been chosen”. The procedure was the same as in Experiments 9 and 10.

Results and discussion

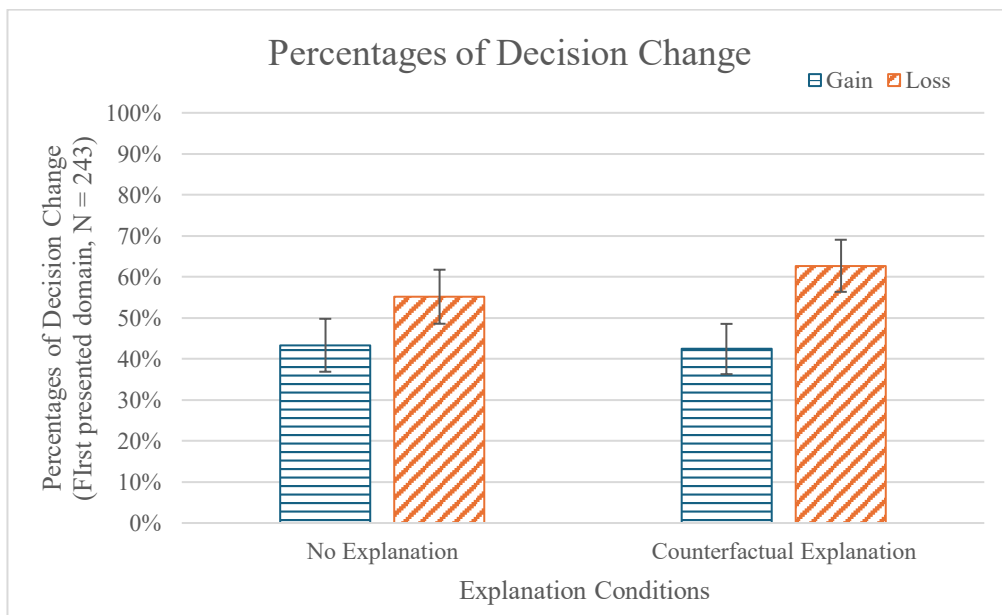
We labelled the typical choices as 0, and the non-typical choices as 1. There was a difference between domains, see Appendix 4.3.1 for supplementary analyses on the difference between the two domains, as well as an effect of the order of domains received. See Appendix 4.3.3 for a supplementary analysis of participants who made the typical decision in both domains. However, domain and order effects have not been observed in the previous four experiments in this chapter, and we suggest they are systematic. Hence, we analysed each participant’s response to the first received scenario here, consistent with previous experiments.

Figure 4.6 shows the responses from the 243 participants who made the typical response in the pre-recommendation phase to the first scenario they

received. When they were recommended to make the opposite choice to the typical one that they had made for this first scenario, with no explanation, in the gain frame, 58% of them continued to choose the safe option and 43% changed to the recommended risky one; in the loss frame, 45% continued to choose the risky option and 55% changed to the recommended safe one. Given a counterfactual explanation, in the gain frame, 58% of participants continued to choose the safe option, and 42% changed to the recommended risky one; in the loss frame, 37% continued to choose the risky option, and 63% changed to the recommended safe one.

Figure 4.6

Percentages of Decision Change in Experiment 11



Note. Percentages of switches from a safe choice to a risky one (for gains) and from a risky choice to a safe one (for losses) following a recommendation by a human expert with no explanation or with a counterfactual explanation, for the

first scenario received in Experiment 11. Error bars are standard errors of the mean.

Mann Whitney U non-parametric tests indicated that there was no significant effect between the explanation condition, $U = 7585$, $z = 0.443$, $p = .658$, and a significant effect of framing, $U = 8559$, $z = 2.506$, $p = .012$, $r = 0.161$. Four comparisons (with a Bonferroni adjusted p -value of $p < .0125$) indicated that there was no difference in participants' tendency to switch their choices following no explanation versus a counterfactual explanation in gains, $U = 1962$, $z = 0.103$, $p = .918$, or in losses, $U = 1840$, $z = 0.825$, $p = .409$. There were no differences between gains and losses in the no explanation condition, $U = 1534$, $z = 1.281$, $p = .200$; or the counterfactual explanation condition, $U = 1552$, $z = 2.257$, $p = .024$.

The results showed that participants did not change their choices more when a counterfactual explanation was provided with human expert recommendations, rather than when no explanation was provided. This was not as we predicted since we had observed a significant effect of counterfactual explanations in Experiment 10 for an AI system's recommendations. A potential explanation is that the percentages of decision switches following human expert recommendations were higher than AI system recommendations. In Experiment 10, we observed a decision switch of about 15% given an AI recommendation and no explanation and about 35% given a counterfactual explanation. In Experiment 11, we observed a decision switch of about 49% given a human recommendation and no explanation and about 53% given a counterfactual explanation. In other words, we observed a lack of further

effect following counterfactual explanations for human recommendations, which may indicate that it is difficult to switch people's decisions to a higher degree than 50%. The differences in percentages of switches in these two experiments again indicated algorithmic aversion: participants were more likely to switch their decisions following a human expert without explanation compared to when they received an AI system recommendation with or without an explanation. Again, there was no difference in the percentages of decision changes when we compared the gain frame and the loss frame.

General Discussion

The experiments in this chapter explored the effect of AI system recommendations and human expert recommendations on the risky choice framing effect – a very robust cognitive bias. We also investigated the effect of counterfactual explanations, from both AI systems and human experts. We had three hypotheses: participants would follow a human expert's recommendations more than they would for recommendations from an AI system; receiving a non-typical recommendation would mitigate the framing effect; and having a counterfactual explanation with the recommendation would prompt people to change their decisions even more.

Overall, the results from all experiments showed that a recommendation to make a non-typical choice changed participants' robust tendency in risky framing scenarios: they made more risky choices for gains and more safe choices for losses, for both disease and farming domains. In the following sections, the results and their implications are discussed.

Framing Effects

The risky choice framing effect, alongside other framing effects, has been established as a robust effect. Through the results from the five experiments, it was observed that the framing effect is robust, as people generally chose the safe option more in the gain frame and the risky option more in the loss frame. With the non-typical recommendations and counterfactual explanations, we observed the framing effect being mitigated and even reversed (e.g., when people receive non-typical recommendations from human experts in the loss frame). These findings contribute to the literature on the framing effect, as we showed that non-typical recommendations were shown to be effective in changing people's natural decision tendencies. The results from all five experiments also indicated that generally, participants switched their decisions to the same degree in the gain frame compared to the loss frame. We saw a lack of change in the gain frame in Experiment 7, which we suggest could be due to chance variation. We did not design our tasks to test specific theories of why framing effects occur, however, the fact that we could use non-typical recommendations and counterfactual explanations to mitigate this effect is consistent with proposals that people lack a complete mental representation of the risky choice task when they were given just the scenarios without recommendations and/or explanations, in line with some theories of framing effects such as schema theory (Jou et al., 1996; Okder, 2012), or effort reduction theories (Gonzalez et al., 2005; Reyna & Brainerd, 1991; Reyna et al., 2021; Whitney et al., 2008; Zhang & Slovic, 2019) we discussed above.

Counterfactual Explanations

Experiment 10 and Experiment 11 implemented counterfactual explanations. In Experiment 10, where explanations were provided with AI system recommendations, we saw a significant effect of counterfactual explanations in decision changes. In Experiment 11 where recommendations were from human experts, we did not see a further improvement in decision change. However, even without explanations, a non-typical recommendation from a human expert leads to decision switches of about 45% – 55%, which is greater than decision switches following an AI recommendation with counterfactual explanations, which averaged about 15% - 35%. A possible explanation for the lack of effect of counterfactual explanations in Experiment 11 would be that a non-typical human expert recommendation is very effective, and an additional counterfactual explanation did not amplify this effect further.

Our findings showed that providing a counterfactual explanation with an AI system recommendation in the framing effect changed people's decisions more compared to when no explanation was given. We tested counterfactual explanations, especially because of their ability to prompt people to simulate alternatives since people may not compare the possible alternatives and their outcomes fully in the risky framing effect task. The results were consistent with our hypothesis of AI counterfactual explanations and consistent with the mental models theory of thinking: people may not be able to mentally fully represent all options of the risky choice framing problem by default, and with counterfactuals, they were able to simulate the possibility of the alternative program (e.g., Byrne, 2002; Byrne & Tasso, 1999; Johnson-Laird & Byrne, 1991). By reading counterfactuals, participants may also build a stronger causal

link (e.g., Kahneman & Tversky, 1982a; Roese, 1997) between the non-typical program and a positive outcome, leading them to change their decisions.

Counterfactual explanations can come in opposite directions, i.e., upward (positive) counterfactuals about how things could have turned out better, and downward (negative) counterfactuals about how things could have turned out worse (Byrne, 2016). These two types of counterfactuals can be triggered by different events and serve different functions, for example, upward counterfactuals explore causal relations and prepare us for future events, while downward counterfactuals regulate emotions and justify decisions (Roese & Olson, 2014). In our experiments, we explored the effectiveness of upward counterfactuals for AI recommendations. Future studies should also the effectiveness of downward counterfactuals for AI recommendation, e.g., “the AI system says that in previous cases the outcome would have been worse if the program you chose was implemented”. Given the loss aversion effect where losses are amplified, it would be interesting to see if the counterfactual explanation could be even more effective if described in the downward direction (Dai et al., 2025).

Recommendation Source and Algorithmic Aversion

Experiment 7 and Experiment 10 tested the effect of AI system recommendations, Experiment 8 and Experiment 11 tested the effect of human expert recommendations, while Experiment 9 directly compared AI system recommendations and human expert recommendations. Results from Experiment 9 indicated that participants followed human expert recommendations more when compared to AI system recommendations, which

was consistent with our hypothesis and what we know from the literature on algorithmic aversion, where a preference for human recommendations was found over algorithmic ones (e.g., Dietvorst et al., 2015; Madhavan et al., 2007). When we looked at results from Experiments 10 and 11, people switched their decisions following AI system recommendations for around 15% - 35% depending on the explanation condition, and around 45% - 60% following human expert recommendations depending on the explanation condition, indicating that the effect of human expert recommendations in the framing effect was stronger.

Our results from Experiments 10 and 11 suggested that counterfactual explanations improved user acceptance of the system's opposite recommendation. We showed that even with a very generic counterfactual statement such as "the AI system says that in previous cases the outcome would have been better if Program A/B had been chosen", we could reduce the robust framing effect, which can be considered as a useful strategy to combat algorithmic aversion. The counterfactual explanation we chose in these experiments, however, could indicate that the AI system has evidence for these two programs after reviewing all past cases. The evidence-based and causal nature of counterfactual explanations should be addressed in follow-up studies (Dai et al., 2025).

As mentioned above, the experimental manipulation we used for the recommendation source was simply stating "An AI system recommends..." versus "An expert recommends...", we did not alter the content or the wording of the recommendations to be more machine-like or human-like. This indicates that the algorithmic aversion we observed in these experiments is a very robust

effect given that we saw a consistent difference with this manipulation. Our results contribute to the list of tasks where algorithmic aversion is observed, and this clear difference between recommendation sources is insightful for AI-based explanation strategies in that a more human-like approach may yield more acceptable explanations.

Conclusion

In this chapter, we aimed to test whether participants will follow the AI recommendation to switch their choice to a different non-typical choice and whether a counterfactual explanation offered by the AI system affects their tendency to do so. We also aimed to test the differences between recommendation sources (AI system vs human expert). The results from five experiments in this chapter addressed our goals and were consistent with our hypotheses. We showed that by providing a non-typical recommendation from an AI system or a human expert, we could mitigate, or even reverse the framing effect. We also established the difference in acceptance of AI recommendations versus human recommendations, i.e., algorithmic aversion, in the very robust cognitive bias. Counterfactual explanations from AI systems were able to increase decision switch, and we did not observe a further decision change following human expert counterfactuals. Our results imply that XAI strategies should take into account the strong effect of algorithmic aversion and counterfactual explanations.

Chapter 5: Discussion

This thesis had three main objectives, as discussed in Chapter 1. The first objective concerned how people evaluate causal explanations.

Experiments 1 to 3 reported in Chapter 2 examined how people choose between competing causal explanations, and we tested the role of consequent-controllability and antecedent-controllability on explanation selection. The results suggest that people may rely on heuristics when they choose between competing causal explanations.

The second objective concerned the effects of counterfactual explanations. Experiment 4 in Chapter 2 examined whether participants' tendency to choose the simpler explanation could be mitigated when they received counterfactual explanations compared to just causal explanations. Experiments 5 and 6 reported in Chapter 3 compared counterfactual and prefactual explanations in terms of their effectiveness in helping participants make accurate judgements of a mock AI system and in terms of user satisfaction. Experiments 10 and 11 in Chapter 4 examined whether counterfactual explanations from AI systems and human experts could increase participants' decision change in a risky situation. The results support the idea that counterfactual explanations prompt people to simulate alternatives. As a result, counterfactual explanations can sometimes help people override their natural tendencies and biases during reasoning.

The third objective was to contribute empirical evidence on the sorts of explanations that people find helpful to inform XAI research. In Chapter 2, we examined the effects of controllability based on its role in counterfactual thinking and its closeness to the XAI goal of actionability. We investigated

ways to operationalise controllability to generate more satisfying causal explanations. In Chapter 3, we tested different sorts of explanations (counterfactual and prefactual) as well as different sorts of causal tasks (predictive inferences and diagnostic inferences) with a mock AI system. In Chapter 4, we pitted recommendations from an AI system against recommendations from human experts in the context of risky decision-making. We also examined the role of counterfactual explanations in people's decision change. The results suggested that it is important to design cognitive experiments and gather empirical evidence with different types of tasks and explanations, to help construct more satisfying and effective XAI explanations.

In this chapter, the key results from the reported experiments are summarised. Their implications for both cognitive psychological theory and XAI methods are outlined, and some directions for future research are suggested. The first section addresses the first objective and discusses the results for explanation evaluation and their implications for understanding cognitive processes in explanation. The second section addresses the second objective and discusses the results for counterfactual explanations and their implications for understanding cognitive processes in counterfactual thinking. The third section addresses the third objective and discusses the implications of all of the experimental results for XAI. We then conclude this chapter by considering some of the limitations in the experiments and directions for future research.

Explanation Evaluation

In Chapter 2, we investigated the impact of controllability on people's preference between competing explanations. We tested controllability because of its role in counterfactual thinking: people choose the controllable factors in a scenario to be modified when they imagine how things could have been different (Giroto et al., 1991; 2007; McEleney & Byrne, 2007; Pighin et al., 2010). Although no clear connections between controllability and causal thinking have been established in previous research (McEleney & Byrne, 2007), we aimed to test the potential role of controllability in causal explanations. We gave participants scenarios of various sorts, for example, where their farm is experiencing a problem with low grass growth, a problem caused by micro-organisms in the soil. They were also tested with similar materials in the disease domain. They were provided with different combinations of micro-organisms as explanations for the farm's problem and were asked to choose the most satisfying explanation among competing explanations. The rationale was that participants' selection of the most satisfying explanation would be influenced by heuristic cues, i.e., the manipulation of controllability. We operationalised controllability in two ways: the first one being the "consequent-controllability" of the outcome. Suppose the micro-organisms Excie and Rirtan always cause low grass growth; Excie can be remedied with fertiliser, but there is no remedy for Rirtan. We examined whether people preferred the explanation that the farm has Excie instead of Rirtan. In other words, consequent-controllability identifies the additional actions one can take to intervene directly on the causal link to prevent the cause from leading to the outcome. Consequent-controllability counteracts the cause

subsequently to fix it. The second way to operationalise was the “antecedent-controllability” of the outcome. Suppose the micro-organisms *Cartia* and *Pacis* always cause high weed growth. *Pacis* is the result of herbicide misuse, and there is no known cause for *Cartia*. We examined whether people preferred the explanation that the farm has *Pacis* instead of *Cartia*. In other words, antecedent-controllability identifies the actions that one can stop doing to intervene indirectly on the causal link by preventing the cause from occurring. Antecedent-controllability counteracts the antecedents to the cause to fix it. The results reported in Chapter 2 demonstrated a consequent-controllability effect in that participants preferred the explanation which referred to controllable-consequent causes (Experiment 1). The results also showed a lack of effect from antecedent-controllability information in that participants did not prefer the explanation which referred to controllable-antecedent causes (Experiments 2a and 2b). However, the results showed that simplicity served as a much stronger criterion for explanation selection when compared to consequent-controllability (Experiment 3) and antecedent-controllability (Experiment 4), in that participants preferred the simpler explanation rather than the explanation with controllable-consequent or controllable-antecedent causes.

In Experiment 1, participants were told that some micro-organisms could be subsequently fixed (e.g., *Excie* can be remedied with fertiliser, *Pacis* can be remedied with herbicide) and that some micro-organisms cannot be subsequently fixed (e.g., there is no known remedy for *Rirtan* or *Cartia*). Participants chose an explanation from a set that included such controllable causes and uncontrollable ones, and they preferred the explanation with both

controllable causes, that “Alien farm X has micro-organisms Excie and Pacis” to other explanations with other combinations of micro-organisms. What is more, in a control condition where participants were not told about the consequent-controllability information of micro-organisms, their selection of this particular explanation was lower (46% vs 19%). The results showed for the first time a consequent-controllability effect on explanation selection and corroborated that people follow cognitive shortcuts during explanation selection: there are no other cues (such as probability information) supporting one explanation over the other ones, and consequent-controllability served as a heuristic, leading participants to strongly prefer the explanation with causes that can be subsequently fixed. The results were consistent with inferential heuristics accounts of explanations, where people rely on shortcuts to reduce the complexity of tasks such as probability judgments to simpler judgemental operations (Gigerenzer & Gaissmaier, 2011; Gigerenzer et al., 2011; Raab & Gigerenzer, 2015; Shah & Oppenheimer, 2008; Simon, 1990a; Tversky & Kahneman, 1974). This effect of consequent-controllability is also in line with the idea of actionability in XAI (Alotaibi & Singh, 2023; Bhattacharya, 2024; Degen et al., 2023; Guidotti et al., 2022; Hacker & Passoth, 2020; Kelechi & Jiao, 2023; Singh et al., 2023; Sokol & Flach, 2020; Verma et al., 2020), where explanations for AI decisions have been proposed to be most effective if they identify manageable actions explicitly.

The next two experiments in Chapter 2 tested the role of antecedent-controllability. In Experiment 2a, participants were told that some micro-organisms were caused by something controllable (e.g., Excie is the result of fertiliser misuse, Pacis is the result of herbicide misuse) and that the cause of

some micro-organisms is unknown (e.g., there is no known cause for Rirtan or Cartia). Participants chose an explanation from a set that included such controllable causes and uncontrollable ones, and they preferred the explanation with both controllable causes, that “Alien farm X has micro-organisms Excie and Pacis” to other explanations (31% vs 18%), although only for the farming domain. We also observed an unexpected preference for explanations without antecedent-controllability (i.e., with both uncontrollable causes, “Alien farm X has micro-organisms Rirtan and Cartia”) compared to when they were not given any controllability information (29% vs 14%). To rule out the possibility that people failed to interpret “no known cause” as an indication that the cause of the factor is uncontrollable, we followed up with Experiment 2b.

Participants were told instead, “Rirtan is the result of day length. Cartia is the result of mountainous terrain”. These causes were previously rated by a small group of participants as uncontrollable in a material check. However, in this experiment, participants chose the explanation with antecedent-controllability just as frequently as they chose that explanation when no controllability information was given (25% vs 19%). Hence, we did not observe a robust effect of antecedent-controllability. We again observed the unexpected preference for the explanation without antecedent-controllability when people were given controllability information compared to when no controllability information was given (23% vs 10%). Together, these results of Experiments 2a and 2b suggest that antecedent-controllability, notwithstanding its close link with controllability in counterfactual thinking, may not have a prominent role in causal explanation. However, the preference for the explanation without antecedent-controllability could be considered to be in line with previous

research findings about past and future counterfactual thinking: when people generate counterfactuals about the past, they focus on uncontrollable factors (such as “if I had been allowed more time I would have solved more puzzles”), and when they generate prefactuals about the future, they focus on controllable factors (such as “if I switch strategy earlier next time I will solve more puzzles”) (Ferrante et al., 2013). It was also found that unless explicitly told to do so, people usually generate few counterfactuals with controllable elements in such circumstances (Mercier et al., 2017). In our experiments, antecedent-controllability identifies factors that were antecedent to the cause-effect link in the past. Our results showed that for antecedent-controllability, people showed a preference for uncontrollable factors, perhaps similar to their tendency to generate counterfactuals that focus on uncontrollable events when they think about a past game. Hence, our results may add to this asymmetry in focus between past and future thinking.

In Experiments 3a and 3b, consequent-controllability and antecedent-controllability were compared to simplicity: participants were given explanations that differed in their level of complexity, the simpler explanation contained one cause (e.g., the micro-organism *Excie* leads to both low grass growth and high weed growth on the farm), and the complex explanation contained two causes (e.g., *Pacis* leads to low grass growth but it does not cause high weed growth; *Rirtan* leads to the high weed growth but it does not cause low grass growth). Participants were given different options based on conditions: In a baseline condition, there was no information on controllability, and the explanations only varied in simplicity (following Lombrozo 2007). In the second experimental condition in Experiment 3a, participants were told that

Excie could be remedied consequently, e.g., by a specialised seed choice, and the two other micro-organisms Pacis and Rirtan had no known remedy, and in the third condition, that Excie had no known remedy, and Pacis and Rirtan can be treated with fertiliser and herbicide accordingly. Experiment 3b had three similar conditions except that participants were told about antecedent-controllable causes, e.g., Excie was caused by herbicide misuse. The results of both Experiment 3a and 3b showed that people still overwhelmingly chose the simpler explanation (around 80%) even though the simpler explanation had no consequent-controllability or antecedent-controllability. Moreover, unlike Experiments 2a and 2b, where we observed a preference for the cause without antecedent-controllability (e.g., the micro-organisms that have no known cause), here we did not observe a preference for the complex but uncontrollable cause. The results of Experiments 3a and 3b replicated findings from previous research that people preferred the simpler explanation (Lombrozo, 2007; Liefgreen & Lagnado, 2023). Simplicity was proposed as one factor that “boosts the explanation’s perceived quality” (Lombrozo, 2012, p.15): moreover, an interaction between simplicity and probability showed that when facing probabilistic uncertainty, people use simplicity as a probabilistic cue (Johnson et al., 2019; Lombrozo, 2007; Vrantsidis & Lombrozo, 2022). Could the effects of controllability also arise because participants interpret a controllable cause as a more probable one, and so controllability is a cue to likelihood? For example, participants may make the assumption that a disease that has a remedy must occur often and hence a treatment has been developed for it, whereas they may assume that a disease that has no known treatment is rare. However, we suggest that in Experiments 3a and 3b, consequent-

controllability and antecedent-controllability were not processed by participants as probability indications. Otherwise, people should select the simple but uncontrollable cause as frequently as the complex but controllable cause, as the probability indicated by simplicity and controllability should compete. Under the theory that simplicity serves as a cue to probability in explanation selection (Lombrozo 2007; 2012), we suggest that when there are no probability cues, such as simplicity, people follow consequent-controllability as a heuristic and a shortcut strategy, maybe because it allows them to do something to remedy a bad outcome in the future. However, when there were probability cues, such as simplicity, they followed the probability cues instead. In contrast, heuristics such as controllability serve as shortcuts and sometimes lead to errors and biases.

The finding of a consequent-controllability effect has potentially important implications for our understanding of causal explanation. It adds to many heuristics that people may rely on. It also brings further consideration to what is typically tested as “controllability” in cognitive experiments, such as counterfactual thinking. This effect is also in line with the differences in terms of controllability between counterfactual and prefactual thinking, in that antecedent-controllability is about the past whereas consequent-controllability is about the future. This asymmetry between backward past and forward future thinking suggests a systematic tendency in people’s reasoning. We will explore this tendency further by examining the asymmetry between future predictive and past diagnostic inferences tested in Chapter 3 in the next section. The results of Experiments 1-3 show that participants were guided by heuristics such as consequent-controllability when there were no other cues such as

simplicity. The reliance on heuristics could lead them to make errors during causal explanations. When there were such cues, they based their evaluation on them instead. Hence, controllability may be relied upon only in a constrained set of circumstances when no other information is available.

Counterfactual Explanations

The second research question we addressed was about the effects specifically of counterfactual explanations. We embedded counterfactual explanations in all three experimental chapters in this thesis to help us better understand counterfactual explanations.

In Chapter 2, Experiment 4 tested whether stating the competing explanations as counterfactuals (e.g., “If Treda had not had Trichet’s syndrome, Treda would not have had the symptoms”) would decrease the overwhelming preference for the simpler explanation, compared to when the explanations were stated as causal statements (e.g., “Because Treda had Trichet’s syndrome, Treda had the symptoms”). Research evidence has shown that counterfactual reasoning prompts people to maintain and update two representations (Byrne, 2016; Byrne & Johnson-Laird, 2009; Espino et al., 2009; Macbeth & Razumiejczyk, 2019; Orenes et al., 2022; Santamaría et al., 2005). It has also been shown that counterfactual explanations elicit higher user satisfaction than causal explanations (Warren et al., 2023) and that counterfactual explanations were more effective than causal ones when participants were asked to make their own decision instead of predicting an AI system’s decision (Celar & Byrne, 2023). Our rationale for this experiment was based on these research findings, and we aimed to see whether counterfactual explanations, compared

to causal ones, would lead to a decrease in people's preference for the simple explanation. We suggest that counterfactuals encourage people to think more about the explanations and simulate alternatives. The results suggested that counterfactual explanations mitigated the preference for simpler explanations compared to causal ones (60% vs 81%). When participants read counterfactual explanations that were based on the more complex situation of two causes for two symptoms such as "If Treda had not had Morad's disease and a Humel infection, Treda would not have had the symptoms", they were better able to mentally flesh out the situation where these two diseases occurred, and the alien showed the symptoms; as well as the situation where these two diseases did not occur, and the alien did not show the symptoms. This type of simulation helps participants to think more about alternatives and may also establish a causal link between the diseases and symptoms, prompting participants to surpass their initial preference for the simple explanation in at least some cases.

In Chapter 3, we investigated the efficacy of counterfactual explanations referring to the past, such as "if fertiliser had been present last month, your grass growth would have been high" and prefactual explanations referring to the future, such as "if fertiliser were to be present next month, your grass growth would be high" in a mock AI grass growth management system. The rationale was based on research on prefactuals, for example, prefactuals help people generate a blueprint for the future through intention formation (Epstude & Roese, 2008; Roese & Epstude, 2017). Prefactual thoughts also focus on controllable factors, while counterfactual thoughts focus on uncontrollable factors (Ferrante et al., 2013). People keep in mind a single possibility given prefactuals (Byrne & Egan, 2004). In other words,

counterfactual and prefactual thinking differ in terms of the focus on mutability, goal, and mental representations. We pitted prefactual explanations against counterfactual explanations in Experiments 5 and 6. Both experiments, one based on real-life farming content, and one based on an alien farming domain, showed similar results. Firstly, we asked participants to judge what kind of prediction or diagnosis the system would make in each instance. Secondly, we asked participants to rate their confidence level in each judgement they made. Lastly, we measured their satisfaction with the explanations. In Experiment 5, we provided participants with counterfactual explanations, prefactual explanations or just a control description in a between-groups design, and participants were asked to “please rate how helpful the explanation is in helping you to understand how the app works”. In Experiment 6, they were given the counterfactual and the prefactual explanations and asked to choose “Which of the explanations helps you to understand the app’s decision?”. In Experiment 5, all three explanation types were equally helpful in participants’ understanding of the system’s decision, as participants showed the same level of accuracy in their judgements when they were given counterfactual explanations, prefactual explanations or just the control description. We note that, however, the accuracy was generally very high for participants (around 80%) for both experiments. There were no differences between self-rated confidence levels between the explanation conditions in Experiment 5. Despite the different measurements we adopted, in both experiments, participants showed an equal level of satisfaction with both types of explanation. The results suggested that counterfactual and prefactual explanations had similar effects in the current experimental setting in both

objective measurement of understanding and subjective ratings of confidence and satisfaction. The result was unexpected given previous evidence that people think about more possibilities when they understand counterfactuals than prefactuals, and that people focus on different aspects of an event when they generate counterfactuals and prefactuals (e.g., Byrne & Egan, 2004; Ferrante et al., 2013; Mercier et al., 2017). It was also surprising that these explanations did not show an advantage over just the control description, as previous research suggests that counterfactual explanations improve user trust and satisfaction (Binns et al., 2018; Celar & Byrne, 2023; Dodge et al., 2019; Förster et al., 2021; Hoffman et al., 2018; Kenny et al., 2021; Lim et al., 2009; Lucic et al., 2020; Russell et al., 2020; Warren et al., 2023). It could be related to the simplicity of the tasks and that participants did not experience the need for explanations. When participants can readily understand and make accurate predictions about an AI system's decisions, explanations of different sorts were no more effective than re-descriptions in improving their prediction accuracy, or in affecting their evaluations of the explanations, as Experiments 5 and 6 showed. People tend to seek explanations when an abnormality or unexpected event occurs (Hilton & Slugoski, 1986). Hence, it could be the case that participants were very accurate in their judgements and did not need explanations. We also note that differences between prefactuals and counterfactuals were empirically tested and found when people were asked to generate counterfactual or prefactual thoughts (e.g., Ferrante et al., 2013) or make inferences after they read counterfactual or prefactual conditionals (e.g., Byrne & Egan, 2004). Our findings could be restricted to the comparison between them as explanations, as discussed above. One putative conclusion is

that differences between counterfactual and prefactual explanations may be observed only in situations in which people require an explanation, for example, when they find it difficult to understand an AI system or predict its decisions.

In Chapter 4, we tested participants with the risky framing problem, and we examined their tendency to switch their decision following a recommendation that was opposite to what they tended to choose. Recall that in Experiment 4, we compared counterfactual explanations to causal explanations and found that counterfactual explanations mitigated people's dominant preference for the simple explanation. Following the mental models theory of counterfactuals (e.g., Byrne & Tasso, 1999; Byrne & Johnson-Laird 2020; Khemlani, Byrne & Johnson-Laird, 2018) that people simulate alternative possibilities when they understand counterfactuals, we tested counterfactual explanations here again in a setting where participants usually commit to one decision and do not consider the alternative(s). Our rationale was based on the effects of counterfactual explanations, as well as theories of framing effects that suggest framing effects occur because people simplify the decision (Reyna & Brainerd, 1991), they adopt a heuristic to reduce the effort of thinking about all the aspects of the options (Whitney et al., 2008), and they do not have a complete mental representation of the disease and the options (Jou et al., 1996; Okder, 2012). In Experiment 10, we provided recommendations from an AI system such as "The AI decision support system recommends Program B", with an explanation-for example, "It says that in previous cases the outcome would have been better if Program B had been chosen". In Experiment 11, we provided recommendations with explanations

from a human expert, such as “The expert says that in previous cases the outcome would have been better if Program B had been chosen”. We compared participants’ willingness to follow the recommendations and switch their choices when they saw counterfactual explanations to the condition where no explanation was given. In Experiment 10, we found that having an explanation together with the AI recommendation led to more decision changes compared to no explanation (35% vs 15%). In Experiment 11, we did not see an increase following an explanation from a human expert (55% vs 50%), but we noted even without an explanation the percentage of decision change was very high, indicating that the lack of difference given human expert counterfactuals could be due to a ceiling effect. In contrast, participants did not tend to follow the recommendations of an AI system without an explanation. The results were in line with the findings from counterfactual user testing that counterfactual explanations increase user trust (Binns et al., 2018; Celar & Byrne, 2023; Dodge et al., 2019; Förster et al., 2021; Hoffman et al., 2018; Kenny et al., 2021; Lim et al., 2009; Lucic et al., 2020; Russell et al., 2020; Warren et al., 2023). The results are also consistent with the proposal earlier that explanations seem to be effective only in circumstances in which people experience some need for an explanation.

Moreover, when participants received a recommendation to make a different choice to one they had made, an explanation was somewhat effective in changing people’s decisions when it accompanied an AI system’s recommendation, but not when it accompanied a human expert’s recommendation, for which people may not have felt as much need for an explanation, as Experiments 10 and 11 showed.

Together the findings from the experiments mentioned above were consistent with previous findings that counterfactuals help people to mentally simulate alternatives, and the results have implications for theoretical accounts of counterfactual thinking. The mental models theory suggests that counterfactual thinking is a process where people mentally represent alternative possibilities (Byrne, 2002; Johnson-Laird & Byrne, 1991). The results of Experiment 4 in Chapter 2 and Experiments 10 and 11 in Chapter 4, were consistent with the mental model theory, as we observed an increase in participants' tendency to consider options that were opposite to what they would choose by default, such as a simple explanation based on one disease leading to two symptoms, or the typical framing choice of a safe option for gains, or a risky option for losses. We suggest that the processes of how they override their default choice could be explained by the mental models theory in terms of ensuring participants can think through the alternative possibilities more explicitly when prompted by counterfactuals. The results of Experiments 5 and 6 did not corroborate the model theory's predictions however, given its proposals that people think about more possibilities when they understand a counterfactual compared to when they understand a prefactual, yet we did not find any difference in participants' accuracy or preferences given counterfactual or prefactual explanations, or re-descriptions. We tentatively suggest the task may have been so easy that people were readily able to think about the possibilities without needing counterfactuals to help them further.

The functional theory of counterfactual thinking explains counterfactuals through their role in emotional regulation and intention formation (Epstude & Roese, 2008; Roese & Epstude, 2017). For example, a

student who just failed the exam might generate upward counterfactual thoughts such as “If only I had studied more, I would have passed the exam”. This type of thought elicits the feeling of regret and thus motivates the individual to better prepare for future exams. It also helps the individual to identify steps to improve future outcomes. Now, think of another example where someone narrowly avoids a car accident who might have downward counterfactual thoughts such as “If I had left home 30 seconds later, I would have crashed into the other car”. This sort of thinking elicits feelings of gratitude and relief. The functional theory also incorporates prefactual thinking, pointing out its role in proactive planning and goal-directed behaviour (Epstude et al., 2016). Results from Chapter 3, however, failed to show a difference between counterfactual conditionals and prefactual conditionals as explanations, and we suggest that the lack of difference could be due to the simplicity of the tasks.

We note that our results about counterfactual explanations also have important implications for XAI, especially in situations where AI advice aims to change a user’s decision. We suggest that, for example, when an AI system is designed to assist end users in choosing an appropriate risk level in a lottery game (Lammert et al., 2024), a counterfactual explanation that states, “If you had chosen this level of risk, the choice would have been more aligned with your attitude towards risks” might be helpful in changing people’s mind. When an AI advisor is designed to help students reduce stress by convincing them to take actions such as studying in a group (Abdulrahman et al., 2022), the advisor may recommend, “If you had not joined a study group, you would not have familiarised yourself with the materials for the discussion”. In the next

section, we discuss further how our findings from all chapters inform XAI explanations.

XAI Explanations

The third objective of this thesis was to contribute further cognitive empirical evidence to help inform XAI research about the sorts of explanations that people find helpful in understanding AI decisions.

In Chapter 2, we incorporated the “actionability” goal of XAI, that explanations should provide useful insights that people can carry out actions about and inform users of how to change an unfavourable decision to achieve a desired outcome. We operationalised the XAI “actionability” idea in two ways, by examining causes that had consequents that could be remedied, e.g., a disease that could be treated by surgery, which we referred to as consequent-controllability, and causes that had antecedents that could be remedied, e.g., a disease that was caused by smoking, which we referred to as antecedent-controllability. We tested whether participants preferred an explanation with consequent-controllability compared to other explanations with no controllability, and whether they preferred an explanation with antecedent-controllability information compared to other explanations with no controllability. Results from Experiment 1 showed that consequent-controllability information affected people’s explanation preferences, and the results from Experiment 2 indicated that antecedent-controllability had less consistent effects. The experiments provide some support for the importance ascribed to the idea of “actionability” in XAI research, at least insofar as it relates to consequent-controllability. However, Experiment 3a showed that the

consequent-controllability effect was over-ridden by simplicity, as participants preferred simple explanations (a single cause leads to two effects) to complex explanations (two independent causes lead separately to the two effects), regardless of whether the causes were consequent-controllable or not. The results implied that careful experimentation is required to establish how relevant factors may interact to determine how effective an explanation is. For example, when trying to come up with better explanations with actionable information, it is important to consider how complex this explanation is to achieve the maximum benefit from actionability. In a banking app, for example, a system might provide actionable explanations such as “Your mortgage application failed because your annual income was below 25k, and you reside outside of Dublin”. The system might also provide unactionable explanations such as “Your mortgage application failed because you are over 55 years old, and your gender is female”. In this case, as indicated by our results, people should prefer the former explanation as it contains consequent-controllability information and provides actionable steps. However, imagine that a system provides an explanation that says, “Your mortgage application failed because you are over 55 years old”. In this case, according to our results, people may prefer the new explanation of “being over 55 years old” to the explanation of “income below 25k and residing outside of Dublin”, simply because the former explanation is simpler. We note, however, that experiments in Chapter 2 were tested in scenarios about aliens with diseases and farms with grass growth problems, and the consequent-controllability effect needs to be empirically examined further in a user-testing setting such as an AI decision support app.

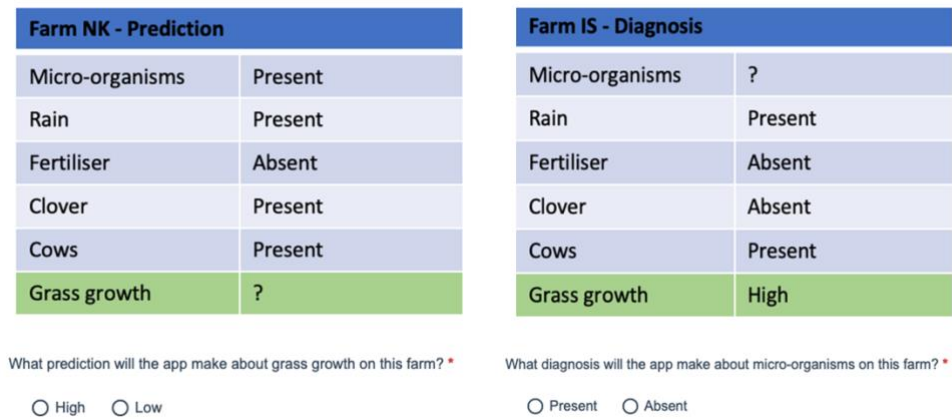
In Chapter 3, Experiments 5 and 6 compared counterfactual and prefactual explanations in a mock AI decision support app. Prefactual explanations have been suggested in AI case-based reasoning due to their function to realise actions (Delaney et al., 2022; Park & Seo, 2023). As outlined earlier, The results indicated that participants considered counterfactuals and prefactuals are equally satisfying and helpful in our experimental setting, albeit a setting in which participants appeared to be able to understand the AI system's decisions very easily. The experiments in Chapter 3 also investigated different sorts of inferences as the measurement of user understanding. As described in Chapter 2, cognitive research has shown an asymmetry between predictive inferences (reason from cause to effect) and diagnostic inferences (reason from effect to cause). People assign a higher probability of the effect given the cause when they make predictions, for example, they assign a higher probability to a baby having blue eyes given the mother has blue eyes compared to the probability of a mother having blue eyes given the baby has blue eyes (Tversky & Kahneman, 1980). People process causal pairs more readily when they are in the predictive order (e.g., spark -> fire) than when they are in the diagnostic order (e.g., fire -> spark) (Fenker et al., 2005). In causal structures where multiple causes lead to one outcome, people ignore alternative causes during prediction, and they can take into account alternative causes during diagnosis, resulting in a more accurate probability judgement during diagnosis (Cummins, 2014; Fernbach et al., 2011; Fernbach & Rehder, 2013). Many XAI user studies have used prediction to assess users' understanding of AI systems (Celar & Byrne, 2023; Goyal et al., 2019; Lage et al., 2019; Lim et al., 2009; Lucic et al., 2020; van der Waa et

al., 2021; Warren et al., 2023). We aimed to test diagnostic inferences as an additional measurement.

Participants were introduced to a causal structure where five causes lead to one outcome. They were given instances in the form of tables, where they could see whether each of the causes was either present or absent and whether the effect was either high or low, as illustrated in **Figure 5.1** below:

Figure 5.1

Examples of Experimental tasks in Chapter 3



Note. Example of the prediction task and the diagnosis task in Experiment 5.

For prediction, they were given the condition of the causes and asked to judge whether the app would predict the outcome to be high or low. For diagnosis, they were given the condition of four of the causes and were given the condition of the outcome. They were asked to judge whether the app would diagnose the one cause left to be present or absent. Both experiments differed in the content (one imagined alien farm and one real-life farm) but showed similar results: participants were more accurate in their diagnostic judgements.

The findings imply that XAI could usefully expand its measurements of user accuracy to include diagnosis rather than only prediction. They suggest that user testing should take into account various sorts of cognitive judgements, especially ones that lead to different inferences.

In the experiments in this chapter, we also included subjective measurements, such as participants' confidence level in their judgement, as well as their evaluation of the explanations as helpful. The results indicated that generally, the responses to the subjective measurements were consistent with the responses to the objective measures (i.e., the judgement accuracy), in that participants were more confident in their diagnoses than their predictions, and they rated explanations as more helpful for diagnoses than predictions in Experiment 5. However, there were some nuanced inconsistencies. For example, in Experiment 6, participants were more accurate in diagnosis than prediction, when they were asked to reason about the low-impact causes. Their confidence ratings, however, indicated that they were equally confident for prediction and diagnosis, at least for the causes that were low-impact and had a negative impact. Previous research has also found an inconsistency between subjective and objective measurements (e.g., Celar & Byrne, 2023; Warren et al., 2023). The results imply that it may be important for XAI human user studies to avail of multiple different sorts of measurements, rather than rely on a single one as is often the case, to ensure a solid and replicable assessment of whether the particular sorts of explanations tested genuinely help people to understand AI decisions.

In Chapter 4, we conducted experiments inspired by the algorithmic aversion effect (Burton et al., 2020; Dietvorst et al., 2015; Kelly et al., 2023;

Mahmud et al., 2022), that people are reluctant to follow the advice or decisions of AI systems, and tested whether people are willing to change their choices given a recommendation from an AI system. Experiments 7 and 8 examined the risky choice framing context, where participants were asked to choose between a safe program and a risky program to combat a disease. Participants typically choose the safe program when it is framed as gains (save 400 of 600 lives), and they typically choose the risky program when it is framed as losses (a two-thirds probability that no-one will die and a one-third probability that everyone will die) (Tversky & Kahneman, 1981). Participants followed the non-typical advice from a human expert (to choose the safe program when it was framed as losses, and to follow the risky program when it was framed as gains): given no recommendations, 21% made the non-typical choices in gains and 51% made the non-typical choices in losses; when the human expert recommended choosing the non-typical choice, 39% made the non-typical choice in gains and 63% made the non-typical choices in losses. However, they rarely followed the non-typical advice from an AI decision support system: given no recommendations, 23% made the non-typical choices in gains and 34% made the non-typical choices in losses; when the AI system recommended choosing the non-typical choice, 23% made the non-typical choice in gains and 51% made the non-typical choices in losses. Experiment 9 tested the difference between human and AI advice using a within-participant design. We measured the decision change by asking participants to commit to their own choice of an option before they received the advice to choose the opposite option (unlike Experiments 7 and 8 in which they saw the recommendation, without having to make a decision first). In this experiment,

participants also followed the non-typical advice from a human expert more than from an AI decision support system, 23% vs 50%. The results confirmed the algorithmic aversion in the risky choice context: participants were more willing to change their decision following a human expert's advice. The results extended the algorithmic aversion effect to the risky choice framing problem. The results contribute further to XAI research on AI trustworthiness and add to the evidence that it is important to consider the role of information sources in user acceptance.

Limitations and Future Directions

In the preceding experimental chapters, we discussed the specific limitations of the experiments. In this section, we consider some broad limitations of this thesis and propose future directions for research.

From a cognitive psychology point of view, most of the experiments in this thesis used selection from a set of choices as the measurement task: participants had to choose between two or more options given to them. This choice of experimental task is partially based on existing research and partially because we wanted to test in a very controlled way. There are a few limitations to this task type. Firstly, in the current experiments, we did not ask participants to generate explanations, we only provided them with explanations to choose from or evaluate, so the results we collected may be restrained to the evaluation process alone. Future research can extend the findings of this thesis to other explanatory processes, such as explanation generation, for example, by asking participants to generate explanations for an AI decision in the form of counterfactual or prefactuals. For example, using the mock AI grass growth

management system in Chapter 3, one could ask the participants to generate useful explanations that focus on how things could have turned out differently in the past, or ones that focus on how things could turn out differently in the future. By doing this, one may be able to replicate the findings about the differences between counterfactual and prefactual thinking, such as the focus on controllable or uncontrollable factors (Ferrante et al., 2013). Secondly, a choice task may not have enough capacity to capture complex cognitive processes, as participants could scan for one option that suits the question without thinking explicitly to exhaust all possibilities. Future research could consider other task types that require more cognitive effort, such as asking participants to rank multiple options. For example, in the experiments in Chapter 2, if participants were given competing causal explanations to rank, we could still predict they would rank the explanation with consequent-controllability as highest, and when there is a simplicity cue, they would rank the simple explanation the highest. However, how they rank the other options might give us more insights into what guides them during the process when they consider all the options. For example, after they have ranked the simple explanation as the highest, if they consistently choose the explanation that is complex but has consequent-controllability, it would indicate that consequent-controllability still serves as a heuristic in their selection process but a weaker one compared to simplicity. In Experiments 1-4, we asked participants to choose the “most satisfying explanation, while in Experiment 5, we asked participants to rate “how helpful” the explanations were. We suggest that the result may not be different, however, future research could also consider the

actual questions used to test user satisfaction as they might be interpreted as different goals.

From the XAI point of view, experiments in this thesis tested participants with imaginary scenarios in the disease domain (Chapter 3 and Chapter 4) and in the farming domain (all chapters). In these situations, we assumed that participants had some familiarity with the issues, in that most people have had some experience of having an illness with some symptoms, or with seeing grass and weeds in their gardens or parks, but that none of them were experts. It has been shown that expertise and domain familiarity impact user acceptance of AI systems and counterfactual XAI (Celar & Byrne, 2023; Ford & Keane, 2022; Gaube et al., 2023; Goyal et al., 2019; Muddamsetty et al., 2021; Yacoby et al., 2022). Future research could extend the findings of the experiments in this thesis to users of different levels of expertise. For example, if experienced farmers were to respond to the same tasks we used in our experiments, their responses would show if expertise would impact some heuristics in their decisions and if expertise would influence their judgments of the helpfulness of the explanations provided. To give an example of the consequent-controllability effect we found in Chapter 2, farmers with many years of experience may have a different set of beliefs and heuristics about what could be the cause when there is a grass growth problem. These heuristics may contradict the consequent-controllability information and may serve as a stronger factor when the expert farmers choose between explanations. Their familiarity in this field and their personal goal to manage farms, on the other hand, could lead them to show a stronger preference for the explanation that has consequent-controllability even if it is more complex.

In Chapters 3 and 4, we provided participants with decisions from AI systems, and we measured their understanding in the experiments in Chapter 3 as well as their acceptance of the system's decisions in the experiments in Chapter 4. We did not incorporate an error rate into these experiments, whereas in real-life situations, AI systems can exhibit errors. It has been found that explanations lead people to view errors as being “less incorrect”, and that an increase in error rate undermines user trust (Kenny et al., 2021). Participants' judgments of explanation helpfulness has also been found to be influenced by AI correctness: when AI decisions are correct, participants considered explanations more helpful (Celar & Byrne, 2023). Future studies could incorporate a variable of error with AI decision support systems to test for its influence on our findings. For example, in the disease problem we tested in Chapter 4, we did not inform participants about the correctness of AI recommendations, we simply stated: “An Artificial Intelligence decision support system recommends that you choose program B” in the no explanation conditions and “An Artificial Intelligence decision support system recommends that you choose program B. It says that in previous cases the outcome would have been better if Program B had been chosen” in the counterfactual explanations condition. We did not provide information on the reliability or the accuracy of these AI recommendations, and we had no access to participants' attitudes about the reliability of these recommendations. We observed a decision change of about 15% with no explanation and about 35% with a counterfactual explanation. Future studies could test the decision change with an accuracy variable. For example, in one condition, the participants could be told, “The AI system's recommendation is correct 20% of the time”, in another

condition, “The AI system’s recommendation is correct 90% of the time”.

Through this manipulation, future studies could investigate the impact of system correctness on decision switches. If participants were willing to follow the AI system’s recommendations for more than 15% given an accuracy rate such as “more than 95% correct”, it could also indicate that when no accuracy information was given, participants had a sense that AI recommendations could be wrong.

Conclusion

This thesis aimed to advance cognitive psychological understanding of the explanation evaluation process, investigate the effectiveness of counterfactual explanations, and contribute to research in XAI on how to construct better explanations for human user understanding and acceptance. The results of the 13 experiments suggest that people rely on a range of heuristics when choosing explanations, and that counterfactual explanations can help people to understand AI decisions. We discovered a novel consequent-controllability effect in explanation evaluation, which has the potential to contribute both to the psychological understanding of explanation evaluation, and to the use of “actionability” in explanations in XAI. We found that when participants were committed to a decision, a counterfactual explanation could lead them to consider alternative options, including alternatives proposed by an AI system. The result can contribute both to the psychological understanding of how counterfactual explanations enable people to simulate alternative possibilities, and to the amelioration of algorithmic aversion in XAI. We extended user testing by compared counterfactual explanations to prefactual

ones, and predictive to diagnostic tasks, which may add to psychological understanding of differences in forward and backward inferences, and assist in the expansion of methods in XAI. Our findings provide further support for the heuristics account of explanation, corroborate predictions of the mental models theory that counterfactuals ensure people envisage multiple possibilities, and provide some guidance for XAI methods to better incorporate knowledge from cognitive science.

References

- Abdulrahman, A., Richards, D., & Bilgin, A. A. (2022). Exploring the influence of a user-specific explainable virtual advisor on health behaviour change intentions. *Autonomous Agents and Multi-Agent Systems*, 36(1), Article 25. <https://doi.org/10.1007/s10458-022-09553-x>
- Ahn, W. K., & Bailenson, J. (1996). Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology*, 31(1), 82-123. <https://doi.org/10.1006/cogp.1996.0013>
- Ahn, W. K., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, 54(3), 299-352. [https://doi.org/10.1016/0010-0277\(94\)00640-7](https://doi.org/10.1016/0010-0277(94)00640-7)
- Al-Ansari, N., Al-Thani, D., & Al-Mansoori, R. S. (2024). User-centered evaluation of explainable artificial intelligence (XAI): A systematic literature review. *Human Behavior and Emerging Technologies*, 2024(1), Article 4628855. <https://doi.org/10.1155/2024/4628855>
- Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, 126(4), 556-574. <https://doi.org/10.1037/0033-2909.126.4.556>
- Alotaibi, H., & Singh, R. (2023). Metrics for evaluating actionability in explainable AI. In *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*, 481-487. Springer Nature Singapore. https://doi.org/10.1007/978-981-99-7022-3_44

- Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology*, 99(3), 436-451.
<https://dx.doi.org/10.1037/a0020218>
- Amsalem, E., & Zoizner, A. (2022). Real, but limited: A meta-analytic assessment of framing effects in the political domain. *British Journal of Political Science*, 52(1), 221-237.
<https://dx.doi.org/10.1017/S0007123420000253>
- Anthony, C. I., Cowley, E., & Blaszczynski, A. (2024). For better and for worse: Frequent gamblers use dual counterfactuals to justify continued gambling. *Journal of Gambling Studies*, 40(2), 1-20.
<https://doi.org/10.1007/s10899-023-10221-2>
- Aquino, K., Steisel, V., & Kay, A. (1992). The effects of resource distribution, voice, and decision framing on the provision of public goods. *Journal of Conflict Resolution*, 36(4), 665-687.
<https://doi.org/10.1177/0022002792036004003>
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.
<https://doi.org/10.1016/j.inffus.2019.12.012>
- Artelt, A., Vaquet, V., Velioglu, R., Hinder, F., Brinkrolf, J., Schilling, M., & Hammer, B. (2021). Evaluating robustness of counterfactual explanations. In *Proceedings of the 2021 IEEE Symposium Series on*

Computational Intelligence (SSCI),1-9.

<https://doi.org/10.1109/SSCI50451.2021.9660058>

Atwal, G., Bryson, D., & Williams, A. (2021). An exploratory study of the adoption of artificial intelligence in Burgundy's wine industry. *Strategic Change*, 30(3), 299-306. <https://doi.org/10.1002/jsc.2413>

Baker, A. (2013). Simplicity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition). Stanford University.

<http://plato.stanford.edu/archives/fall2013/entries/simplicity/>

Batzke, M. C., & Ernst, A. (2024). Distinguishing fast change in social norms and slow change in personal norms in cooperative decision-making.

Frontiers in Psychology, 15, Article 1380341.

<https://doi.org/10.3389/fpsyg.2024.1380341>

Beck, S. R., Robinson, E. J., Carroll, D. J., & Apperly, I. A. (2006). Children's thinking about counterfactuals and future hypotheticals as possibilities. *Child Development*, 77(2), 413-426.

<https://doi.org/10.1111/j.1467-8624.2006.00879.x>

Bhattacharya, A. (2024). Towards directive explanations: Crafting explainable AI systems for actionable human-AI interactions. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*, Article 442. <https://doi.org/10.1145/3613905.363817>

Bilgic, M., & Mooney, R. J. (2005). Explaining recommendations: Satisfaction vs. promotion. In *Beyond Personalization Workshop, IUI*.

<http://www.cs.umn.edu/Research/GroupLens/beyond2005>

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & 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*, Paper 337.
<https://doi.org/10.1145/3173574.3173951>

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *Proceedings of IJCAI-17 Workshop on Explainable AI (XAI)*, 8-13.

Blanchard, T., Lombrozo, T., & Nichols, S. (2018a). Bayesian Occam's razor is a razor of the people. *Cognitive Science*, 42(4), 1345-1359.
<https://doi.org/10.1111/cogs.12573>

Blanchard, T., Vasilyeva, N., & Lombrozo, T. (2018b). Stability, breadth and guidance. *Philosophical Studies*, 175, 2263-2283.
<https://doi.org/10.1007/s11098-017-0958-6>

Bogdanov, B., Corbin, J., Dobрева, S., McElroy, T., & Rachev, N. R. (2023). Working memory capacity and the risky-choice framing effect: A preregistered replication and extension of Corbin et al. (2010). *Judgment and Decision Making*, 18, Article e39.
<https://doi.org/10.1017/jdm.2023.40>

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology*, 48(4), 1156-1164. <https://doi.org/10.1037/a0026471>

Bove, C., Lesot, M. J., Tijus, C. A., & Detyniecki, M. (2023). Investigating the intelligibility of plural counterfactual examples for non-expert users: An explanation user interface proposition and user study. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 188-203. <https://doi.org/10.1145/3581641.3584082>

- Branscombe, N. R., Wohl, M. J., Owen, S., Allison, J. A., & N'gbala, A. (2003). Counterfactual thinking, blame assignment, and well-being in rape victims. *Basic and Applied Social Psychology*, 25(4), 265-273.
https://doi.org/10.1207/S15324834BASP2504_1
- Bromme, R., Rambow, R., & Nückles, M. (2001). Expertise and estimating what other people know: The influence of professional experience and type of knowledge. *Journal of Experimental Psychology: Applied*, 7(4), 317-330. <https://doi.org/10.1037/1076-898X.7.4.317>
- Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2), 220-239.
<https://doi.org/10.1002/bdm.2155>
- Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition*, 31(1), 61-83. [https://doi.org/10.1016/0010-0277\(89\)90018-8](https://doi.org/10.1016/0010-0277(89)90018-8)
- Byrne, R. M. J. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Sciences*, 6(10), 426-431.
[https://doi.org/10.1016/S1364-6613\(02\)01974-5](https://doi.org/10.1016/S1364-6613(02)01974-5)
- Byrne, R. M. J. (2005). *The rational imagination: How people create alternatives to reality*. MIT Press.
- Byrne, R. M. J. (2016). Counterfactual thought. *Annual Review of Psychology*, 67, 135–157. <https://doi.org/10.1146/annurev-psych-122414-033249>
- Byrne, R. M. J. (2019). Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 6276–6282. <https://www.ijcai.org/proceedings/2019/0876.pdf>

- Byrne, R. M. J. (2023a). Good explanations in explainable artificial intelligence (XAI): Evidence from human explanatory reasoning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, Article 733, 6536-6544.
<https://doi.org/10.24963/ijcai.2023/733>
- Byrne, R. M. J. (2023b). How people think about possibilities. *Possibility Studies & Society*, 1(1-2), 29-36.
<https://doi.org/10.1177/27538699231171432>
- Byrne, R. M. J., & Egan, S. M. (2004). Counterfactual and prefactual conditionals. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 58(2), 113-120.
<https://doi.org/10.1037/h0085791>
- Byrne, R. M. J., & Johnson-Laird, P. N. (2009). 'If' and the problems of conditional reasoning. *Trends in Cognitive Sciences*, 13(7), 282-287.
<https://doi.org/10.1016/j.tics.2009.04.003>
- Byrne R. M. J., & Johnson-Laird P. N. (2020). If and or: Real and counterfactual possibilities in their truth and probability. *Journal of Experimental Psychology Learning Memory and Cognition*, 46(4), 760-780. <https://doi.org/10.1037/xlm0000756>
- Byrne, R. M. J., & McEleney, A. (2000). Counterfactual thinking about actions and failures to act. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(5), 1318-1331.
<https://doi.org/10.1037/0278-7393.26.5.1318>
- Byrne, R. M. J., Segura, S., Culhane, R., Tasso, A., & Berrocal, P. (2000). The temporality effect in counterfactual thinking about what might have

been. *Memory & Cognition*, 28, 264-281.

<https://doi.org/10.3758/BF03213805>

Byrne, R. M. J., & Tasso, A. (1999). Deductive reasoning with factual, possible, and counterfactual conditionals. *Memory & Cognition*, 27(4), 726-740. <https://doi.org/10.3758/BF03211565>

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42(1), 116-131. <https://doi.org/10.1037/0022-3514.42.1.116>

Catellani, P. (2012). Counterfactuals in the social context: The case of political interviews and their effects. In D. Birke, M. Butter & T. Köppe (Eds.), *Counterfactual Thinking-Counterfactual Writing* (pp.81-94). De Gruyter. <https://doi.org/10.1515/9783110268669.81>

Catellani, P., & Bertolotti, M. (2014). The effects of counterfactual attacks on social judgments. *Social Psychology*, 45(5), 371-381. <https://doi.org/10.1027/1864-9335/a000195>

Catellani, P., & Covelli, V. (2013). The strategic use of counterfactual communication in politics. *Journal of Language and Social Psychology*, 32(4), 480-489. <https://doi.org/10.1177/0261927X13495548>

Celar, L., & Byrne, R. M. J. (2023). How people reason with counterfactual and causal explanations for artificial intelligence decisions in familiar and unfamiliar domains. *Memory & Cognition*, 51(7), 1481-1496. <https://doi.org/10.3758/s13421-023-01407-5>

Chater, N., & Oaksford, M. (2006). Mental mechanisms: Speculations on human causal learning and reasoning. In K. Fiedler & P. Juslin (Eds.),

Information Sampling and Adaptive Cognition (pp.210-236).

Cambridge University Press.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7(1), 19-22.

[https://doi.org/10.1016/S1364-6613\(02\)00005-0](https://doi.org/10.1016/S1364-6613(02)00005-0)

Chick, C. F., Reyna, V. F., & Corbin, J. C. (2016). Framing effects are robust to linguistic disambiguation: A critical test of contemporary theory.

Journal of Experimental Psychology: Learning, Memory, and

Cognition, 42(2), 238-256. <https://doi.org/10.1037/xlm0000158>

Collins, J., Hall, N., & Paul, L. A. (2004). *Causation and counterfactuals*. The MIT Press.

Corbin, J., McElroy, T., & Black, C. (2010). Memory reflected in our decisions: Higher working memory capacity predicts greater bias in risky choice. *Judgment and Decision Making*, 5(2), 110-115.

<https://doi.org/10.1017/S1930297500000966>

Cummins, D. D. (2014). The impact of disablers on predictive inference.

Journal of Experimental Psychology: Learning, Memory, and

Cognition, 40(6), 1638-1655. <https://doi.org/10.1037/xlm0000024>

Dai, X., Keane, M. T., Shalloo, L., Ruelle, E., & Byrne, R. M. J. (2022).

Counterfactual explanations for prediction and diagnosis in XAI. In

Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and

Society, 215-226. <https://doi.org/10.1145/3514094.3534144>

Dai, X., Keane, M. T., Shalloo, L., Ruelle, E., & Byrne, R. M. J. (2025). *How counterfactual explanations lead people to switch from their preferred decision to a non-preferred decision proposed by an AI system*

[Manuscript submitted for publication]. School of Psychology, Trinity College Dublin.

- Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *ArXiv preprint*, arXiv:2006.11371. <https://doi.org/10.48550/arXiv.2006.11371>
- Davis, C. G., & Lehman, D. R. (1995). Counterfactual thinking and coping with traumatic life events. In N .J. Roese & J. M. Olson (Eds)., *What might have been* (pp. 353-374). Psychology Press.
<https://doi.org/10.4324/9781315806419>
- Davis, C. G., Lehman, D. R., Wortman, C. B., Silver, R. C., & Thompson, S. C. (1995). The undoing of traumatic life events. *Personality and Social Psychology Bulletin*, 21(2), 109-124.
<https://doi.org/10.1177/0146167295212002>
- De Brigard, F., Henne, P., & Stanley, M. L. (2021). Perceived similarity of imagined possible worlds affects judgments of counterfactual plausibility. *Cognition*, 209, Article 104574.
<https://doi.org/10.1016/j.cognition.2020.104574>
- De Brigard, F., Szpunar, K.K., Schacter,D.L. (2013). Coming to grips with the past effect of repeated simulation on the perceived plausibility of episodic counterfactual thoughts. *Psychological Science*, 24(7), 1329-1334. <https://doi.org/10.1177/0956797612468163>
- Degen, H., Budnik, C., Gross, R., & Rothering, M. (2023). How to explain it to a model manager? A qualitative user study about understandability, trustworthiness, actionability, and action efficacy. In *Proceedings of*

- International Conference on Human-Computer Interaction*, 209-242. Springer, Cham. https://doi.org/10.1007/978-3-031-35891-3_14
- Dehghani, M., Iliev, R., & Kaufmann, S. (2012). Causal explanation and fact mutability in counterfactual reasoning. *Mind & Language*, 27(1), 55-85. <https://doi.org/10.1111/j.1468-0017.2011.01435.x>
- Delaney, E., Greene, D., Shalloo, L., Lynch, M., & Keane, M. T. (2022). Forecasting for sustainable dairy produce: Enhanced long-term, milk-supply forecasting using k-NN for data augmentation, with prefactual explanations for XAI. In *Proceedings of International Conference on Case-Based Reasoning*, 365-379. Springer, Cham. https://doi.org/10.1007/978-3-031-14923-8_24
- De Martino, B., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain. *Science*, 313(5787), 684-687. <https://doi.org/10.1126/science.1128356>
- Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press
- De Neys, W., Schaeken, W., & D'Ydewalle, G. (2003). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory & Cognition*, 31(4), 581-595. <https://doi.org/10.3758/BF03196099>
- Diab, D. L., Pui, S. Y., Yankelevich, M., & Highhouse, S. (2011). Lay perceptions of selection decision aids in US and non-US samples. *International Journal of Selection and Assessment*, 19(2), 209-216. <https://doi.org/10.1111/j.1468-2389.2011.00548.x>
- Dias, M. D. G., & Harris, P. L. (1990). The influence of the imagination on reasoning by young children. *British Journal of Developmental*

Psychology, 8(4), 305-318.

<https://doi.org/10.1111/j.2044-835X.1990.tb00847.x>

Diederich, A., Wyszynski, M., & Ritov, I. (2018). Moderators of framing effects in variations of the Asian Disease problem: Time constraint, need, and disease type. *Judgment and Decision Making*, 13(6), 529-546. <https://doi.org/10.1017/S1930297500006574>

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114-126. <https://doi.org/10.1037/xge0000033>

Dijkstra, J. J., Liebrand, W. B., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology*, 17(3), 155-163. <https://doi.org/10.1080/014492998119526>

Dixon, J. E., & Byrne, R. M. J. (2011). "If only" counterfactual thoughts about exceptional actions. *Memory & Cognition*, 39, 1317-1331. <https://doi.org/10.3758/s13421-011-0101-4>

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K., & 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. <https://doi.org/10.1145/3301275.3302310>

Druckman, J. N., & McDermott, R. (2008). Emotion and the framing of risky choice. *Political Behavior*, 30(3), 297-321. <https://doi.org/10.1007/s11109-008-9056-y>

- Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9), Article 194. <https://doi.org/10.1145/3561048>
- Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, 12(2), 168-192. <https://doi.org/10.1177/1088868308316091>
- Epstude, K., Scholl, A., & Roese, N. J. (2016). Prefactual thoughts: Mental simulations about what might happen. *Review of General Psychology*, 20(1), 48-56. <https://doi.org/10.1037/gpr0000064>
- Espino, O., & Byrne, R. M. J. (2020). The suppression of inferences from counterfactual conditionals. *Cognitive Science*, 44(4), Article e12827. <https://doi.org/10.1111/cogs.12827>
- Espino O., Santamaría C., Byrne R. M. J. (2009). People think about what is true for conditionals, not what is false: Only true possibilities prime the comprehension of “if”. *Quarterly Journal of Experimental Psychology*, 62(6), 1072-1078. <https://doi.org/10.1080/17470210802602375>
- Fagley, N.S., Coleman, J.G., Simon, A.F. (2010). Effects of framing, perspective taking, and perspective (affective focus) on choice. *Personality and Individual Differences*, 48(3), 264-269. <https://doi.org/10.1016/j.paid.2009.10.008>
- Fagley, N. S., & Miller, P. M. (1987). The effects of decision framing on choice of risky vs certain options. *Organizational Behavior and Human Decision Processes*, 39(2), 264-277. [https://doi.org/10.1016/0749-5978\(87\)90041-0](https://doi.org/10.1016/0749-5978(87)90041-0)

- Feldman, G. (2020). What is normal? Dimensions of action-inaction normality and their impact on regret in the action-effect. *Cognition and Emotion*, 34(4), 728-742. <https://doi.org/10.1080/02699931.2019.1675598>
- Feldman, G., & Albarracín, D. (2017). Norm theory and the action-effect: The role of social norms in regret following action and inaction. *Journal of Experimental Social Psychology*, 69, 111-120. <https://doi.org/10.1016/j.jesp.2016.07.009>
- Fenker, D. B., Waldmann, M. R., & Holyoak, K. J. (2005). Accessing causal relations in semantic memory. *Memory & Cognition*, 33, 1036-1046. <https://doi.org/10.3758/BF03193211>
- Fernandez, D. K., Gan, H. H., & Chan, A. Y. (2022). Comparing the functional benefits of counterfactual and prefactual thinking: The content-specific and content-neutral pathways. *Thinking & Reasoning*, 28(2), 261-289. <https://doi.org/10.1080/13546783.2021.1989034>
- Fernbach, P.M., Darlow, A., & Sloman, S.A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*, 21(3), 329-336. <https://doi.org/10.1177/0956797610361430>
- Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, 140(2), 168-185. <https://doi.org/10.1037/a0022100>
- Fernbach, P. M., & Rehder, B. (2013). Cognitive shortcuts in causal inference. *Argument & Computation*, 4(1), 64-88. <https://doi.org/10.1080/19462166.2012.682655>

- Ferrante, D., Giroto, V., Stragà, M., & Walsh, C. (2013). Improving the past and the future: A temporal asymmetry in hypothetical thinking. *Journal of Experimental Psychology: General*, 142(1), 23-27.
<https://doi.org/10.1037/a0027947>
- Fillon, A., Kutscher, L., & Feldman, G. (2021). Impact of past behaviour normality: Meta-analysis of exceptionality effect. *Cognition and Emotion*, 35(1), 129-149.
<https://doi.org/10.1080/02699931.2020.1816910>
- Fiok, K., Farahani, F. V., Karwowski, W., & Ahram, T. (2022). Explainable artificial intelligence for education and training. *The Journal of Defense Modeling and Simulation*, 19(2), 133-144.
<https://doi.org/10.1177/15485129211028651>
- Ford, C., & Keane, M. T. (2022). Explaining classifications to non-experts: An XAI user study of post-hoc explanations for a classifier when people lack expertise. In *Proceedings of International Conference on Pattern Recognition*, 246-260. Springer, Cham.
https://doi.org/10.1007/978-3-031-37731-0_15
- Förster, M., Hühn, P., Klier, M., & 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*, 1274-1283. <http://hdl.handle.net/10125/70767>
- Frisch, D. (1993). Reasons for framing effects. *Organizational Behavior and Human Decision Processes*, 54(3), 399-429.
<https://doi.org/10.1006/obhd.1993.1017>

- Frosch, C. A., & Byrne, R. M. J. (2012). Causal conditionals and counterfactuals. *Acta psychologica, 141*(1), 54-66.
<https://doi.org/10.1016/j.actpsy.2012.07.001>
- Frosch, C. A., Egan, S. M., & Hancock, E. N. (2015). The effect of controllability and causality on counterfactual thinking. *Thinking & Reasoning, 21*(3), 317-340.
<https://doi.org/10.1080/13546783.2014.976268>
- Gaube, S., Suresh, H., Raue, M., Lerner, E., Koch, T. K., Hudecek, M. F., Ackery, A. D., Grover, S. C., Coughlin, J. F., Frey, D., Kitamura, F. C., Ghassemi, M., & Colak, E. (2023). Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Scientific Reports, 13*(1), Article 1383.
<https://doi.org/10.1038/s41598-023-28633-w>
- Gavanski, I., & Wells, G. L. (1989). Counterfactual processing of normal and exceptional events. *Journal of Experimental Social Psychology, 25*(4), 314-325. [https://doi.org/10.1016/0022-1031\(89\)90025-5](https://doi.org/10.1016/0022-1031(89)90025-5)
- Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society, 37*, 782-787. <https://escholarship.org/uc/item/74b7k69g>
- Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgements for physical events. *Psychological Review, 128*(5), 936-975.
<https://doi.org/10.1037/rev0000281>

- Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62(1), 451-482.
<https://doi.org/10.1146/annurev-psych-120709-145346>
- Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). (2011). *Heuristics: The foundations of adaptive behavior*. Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780199744282.001.0001>
- Gilovich, T., & Medvec, V. H. (1995). The experience of regret: What, when, and why. *Psychological Review*, 102(2), 379-395.
<http://dx.doi.org.elib.tcd.ie/10.1037/0033-295X.102.2.379>
- Giroto, V., Ferrante, D., Pighin, S., & Gonzalez, M. (2007). Postdecisional counterfactual thinking by actors and readers. *Psychological Science*, 18 (6), 510-515. <https://doi.org/10.1111/j.1467-9280.2007.01931>.
- Giroto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica*, 78(1-3), 111-133.
[https://doi.org/10.1016/0001-6918\(91\)90007-M](https://doi.org/10.1016/0001-6918(91)90007-M)
- Glymour, C. (2015). Probability and the explanatory virtues. *The British Journal for the Philosophy of Science*, 66(3), 591-604.
<https://doi.org/10.1093/bjps/axt051>
- Glymour, C., & Cheng, P. (1999). Causal mechanism and probability: A normative approach. In K. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 295–313). Oxford University Press.
- Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science*, 25(4), 565-610. https://doi.org/10.1207/s15516709cog2504_3

- Gonzalez, C., Dana, J., Koshino, H., & Just, M. (2005). The framing effect and risky decisions: Examining cognitive functions with fMRI. *Journal of Economic Psychology*, 26(1), 1-20.
<https://doi.org/10.1016/j.joep.2004.08.004>
- Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111(1), 3-32.
<https://doi.org/10.1037/0033-295X.111.1.3>
- Gosling, C. J., & Moutier, S. (2017). High but not low probability of gain elicits a positive feeling leading to the framing effect. *Frontiers in Psychology*, 8, Article 81. <https://doi.org/10.3389/fpsyg.2017.00081>
- Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019). Counterfactual visual explanations. In *International Conference on Machine Learning*, PMLR 97, 2376-2384.
<https://proceedings.mlr.press/v97/goyal19a.html>
- Graf, C., Suanet, B., Wiepking, P., & Merz, E. M. (2023). Social norms offer explanation for inconsistent effects of incentives on prosocial behavior. *Journal of Economic Behavior & Organization*, 211, 429-441.
<https://doi.org/10.1016/j.jebo.2023.05.003>
- Guglielmo, S., & Malle, B. F. (2017). Information-acquisition processes in moral judgments of blame. *Personality and Social Psychology Bulletin*, 43(7), 957-971. <https://doi.org/10.1177/0146167217702375>
- Guidotti, R., Monreale, A., Ruggieri, S., Naretto, F., Turini, F., Pedreschi, D., & Giannotti, F. (2022). Stable and actionable explanations of black-box models through factual and counterfactual rules. *Data Mining and*

Knowledge Discovery, 38, 2825-2862.

<https://doi.org/10.1007/s10618-022-00878-5>

Hacker, P., & Passoth, J. H. (2020). Varieties of AI explanations under the law.

From the GDPR to the AIA, and beyond. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, 343-

373. Springer, Cham. https://doi.org/10.1007/978-3-031-04083-2_17

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model

approach. Part I: Causes. *The British Journal for the Philosophy of*

Science, 56(4), 843-887. <https://doi.org/10.1093/bjps/axi147>

Hameleers, M. (2021). Prospect theory in times of a pandemic: The effects of

gain versus loss framing on risky choices and emotional responses

during the 2020 coronavirus outbreak—Evidence from the US and the

Netherlands. *Mass Communication and Society*, 24(4), 479-499.

<https://doi.org/10.1080/15205436.2020.1870144>

Haque, A. B., Islam, A. N., & Mikalef, P. (2023). Explainable artificial

intelligence (XAI) from a user perspective: A synthesis of prior

literature and problematizing avenues for future research. *Technological*

Forecasting and Social Change, 186 (Part A), Article 122120.

<https://doi.org/10.1016/j.techfore.2022.122120>

Harman, G.H. (1965). The inference to the best explanation. *The Philosophical*

Review, 74(1), 88-95. <https://doi.org/10.2307/2183532>

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving

judgment, and sharing responsibility. *Organizational Behavior and*

Human Decision Processes, 70(2), 117-133.

<https://doi.org/10.1006/obhd.1997.2697>

- Henne, P., Kulesza, A., Perez, K., & Houcek, A. (2021). Counterfactual thinking and recency effects in causal judgment. *Cognition*, 212, Article 104708. <https://doi.org/10.1016/j.cognition.2021.104708>
- Henne, P., Niemi, L., Pinillos, Á., De Brigard, F., & Knobe, J. (2019). A counterfactual explanation for the action effect in causal judgment. *Cognition*, 190, 157-164. <https://doi.org/10.1016/j.cognition.2019.05.006>
- Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, 93(1), 75-88. <https://doi.org/10.1037/0033-295X.93.1.75>
- Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *ArXiv preprint*, arXiv:1812.04608. <https://doi.org/10.48550/arXiv.1812.04608>
- Homar, A. R., & Cvelbar, L. K. (2021). The effects of framing on environmental decisions: A systematic literature review. *Ecological Economics*, 183, Article 106950. <https://doi.org/10.1016/j.ecolecon.2021.106950>
- Howlett, J. R., & Paulus, M. P. (2013). Decision-making dysfunctions of counterfactuals in depression: Who might I have been? *Frontiers in Psychiatry*, 4, Article 143. <https://doi.org/10.3389/fpsyt.2013.00143>
- Huang, L., Xie, Y., & Chen, X. (2021). A review of functions of speculative thinking. *Frontiers in Psychology*, 12, Article 728946. <https://doi.org/10.3389/fpsyg.2021.728946>
- Igou, E. R., & Bless, H. (2007). On undesirable consequences of thinking: Framing effects as a function of substantive processing. *Journal of*

Behavioral Decision Making, 20(2), 125-142.

<https://doi.org/10.1002/bdm.543>

John-Mathews, J. M. (2021). Critical empirical study on black-box

explanations in AI. *ArXiv preprint*, arXiv:2109.15067.

<https://doi.org/10.48550/arXiv.2109.15067>

Johnson, S., Johnston, A., Toig, A., & Keil, F. (2014). Explanatory scope

informs causal strength inferences. In *Proceedings of the Annual*

Meeting of the Cognitive Science Society, 36(36), 2453-2458.

<https://escholarship.org/uc/item/91d5t02n>

Johnson, S. G., Valenti, J. J., & Keil, F. C. (2019). Simplicity and complexity

preferences in causal explanation: An opponent heuristic account.

Cognitive Psychology, 113, Article 101222.

<https://doi.org/10.1016/j.cogpsych.2019.05.004>

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of*

language, inference and consciousness. Cambridge University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Lawrence Erlbaum

Associates.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from

inconsistency to consistency. *Psychological Review*, 111(3), 640-661.

<https://doi.org/10.1037/0033-295X.111.3.640>

Johnson-Laird, P. N., & Khemlani, S. S. (2017). Mental models and causation.

In M. R. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp.1-

42). Oxford University Press.

Jones, B. D. (1999). Bounded rationality. *Annual Review of Political Science*,

2(1), 297-321. <https://doi.org/10.1146/annurev.polisci.2.1.297>

- Jou, J., Shanteau, J., & Harris, R. J. (1996). An information processing view of framing effects: The role of causal schemas in decision making. *Memory & Cognition*, 24, 1-15. <https://doi.org/10.3758/BF03197268>
- Jungermann, H., & Fischer, K. (2014). Using expertise and experience for giving and taking advice. In T. Betsch & S. Haberstroh (Eds.), *The routines of decision making* (1st ed., pp. 157-173). Psychology Press.
- Kahneman, D., & Frederick, S. (2007). Frames and brains: Elicitation and control of response tendencies. *Trends in Cognitive Sciences*, 11(2), 45-46. <https://doi.org/10.1016/j.tics.2006.11.007>
- Kahneman, D., & Miller, D.T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93, 136-153. <https://doi.org/10.1037/0033-295X.93.2.136>
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263-291. <https://doi.org/10.2307/1914185>
- Kahneman, D., & Tversky, A. (1982a). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp.201-208). Cambridge University Press
- Kahneman, D., & Tversky, A. (1982b). The psychology of preferences. *Scientific American*, 246(1), 160-173. <http://www.jstor.org/stable/24966506>
- Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, 39(4), 341-350. <https://doi.org/10.1037/0003-066X.39.4.341>

- Karimi, A. H., Barthe, G., Schölkopf, B., & Valera, I. (2020). A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. *ArXiv preprint*, arXiv:2010.04050.
<https://doi.org/10.48550/arXiv.2010.04050>
- Karimi, A. H., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 353-362. <https://doi.org/10.1145/3442188.3445899>
- Keane, M.T., Kenny, E.M., Delaney, E., & Smyth, B. (2021). If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 4466-4474. <https://doi.org/10.48550/arXiv.2103.01035>
- Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, 57(1), 227-254.
<https://doi.org/10.1146/annurev.psych.57.102904.190100>
- Keil, F. C. (2024). Causal explanations and the growth of understanding. *Annual Review of Developmental Psychology*, 6, 163-184.
<https://doi.org/10.1146/annurev-devpsych-010923-103603>
- Keil, F. C., Rozenblit, L., & Mills, C. (2004). What lies beneath? Understanding the limits of understanding. In D. T. Levin (Ed.), *Thinking and seeing: Visual metacognition in adults and children* (pp.227-249). The MIT Press.

- Kelechi, N., & L, Jiao. (2023). Quantifying actionability: Evaluating human-recipient models. *IEEE Access*, *11*, 119811-119823.
<https://doi.org/10.1109/ACCESS.2023.3324906>
- Kelly, S., Kaye, S. A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, *77*, Article 101925.
<https://doi.org/10.1016/j.tele.2022.101925>
- Kenny, E. M., Ford, C., Quinn, M., & 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*, Article 103459. <https://doi.org/10.1016/j.artint.2021.103459>
- Khemlani, S. S., & Johnson-Laird, P. N. (2011). The need to explain. *Quarterly Journal of Experimental Psychology*, *64*(11), 2276-2288.
<https://doi.org/10.1080/17470218.2011.592593>
- Khemlani, S. S., & Johnson-Laird, P. N. (2013). The processes of inference. *Argument & Computation*, *4*(1), 4-20.
<https://doi.org/10.1080/19462166.2012.674060>
- Khemlani, S. S., & Oppenheimer, D. M. (2011). When one model casts doubt on another: A levels-of-analysis approach to causal discounting. *Psychological Bulletin*, *137*(2), 195-210.
<https://doi.org/10.1037/a0021809>
- Kirfel, L., Icard, T., & Gerstenberg, T. (2022). Inference from explanation. *Journal of Experimental Psychology: General*, *151*(7), 1481-1501.
<http://dx.doi.org.elib.tcd.ie/10.1037/xge0001151>

- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., ... Nosek, B. A. (2014). Investigating variation in replicability. *Social psychology*, 45(3), 142-152. <https://doi.org/10.1027/1864-9335/a000178>
- Kment, B. (2020). Counterfactuals and causal reasoning. In E. Siegal & N. Boneh (Eds.), *Perspectives on causation* (pp.463-482). Springer, Cham.
- Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, 137, 196-209. <https://doi.org/10.1016/j.cognition.2015.01.013>
- Krause, J., Perer, A., & Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5686-5697. <https://doi.org/10.1145/2858036.2858529>
- Kühberger, A. (1995). The framing of decisions: A new look at old problems. *Organizational Behavior and Human Decision Processes*, 62(2), 230-240. <https://doi.org/10.1006/obhd.1995.1046>
- Kühberger, A. (1998). The influence of framing on risky decisions: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 75(1), 23-55. <https://doi.org/10.1006/obhd.1998.2781>
- Kühberger, A., & Gradl, P. (2013). Choice, rating, and ranking: Framing effects with different response modes. *Journal of Behavioral Decision Making*, 26(2), 109-117. <https://doi.org/10.1002/bdm.764>

- Kuhl, U., Artelt, A., & Hammer, B. (2023). Let's go to the alien zoo: Introducing an experimental framework to study usability of counterfactual explanations for machine learning. *Frontiers in Computer Science*, 5, Article 1087929.
<https://doi.org/10.3389/fcomp.2023.1087929>
- Kulakova, E., Aichhorn, M., Schurz, M., Kronbichler, M., & Perner, J. (2013). Processing counterfactual and hypothetical conditionals: An fMRI investigation. *NeuroImage*, 72, 265-271.
<https://doi.org/10.1016/j.neuroimage.2013.01.060>
- Kuo, F. Y., Hsu, C. W., & Day, R. F. (2009). An exploratory study of cognitive effort involved in decision under framing-an application of the eye-tracking technology. *Decision Support Systems*, 48(1), 81-91.
<https://doi.org/10.1016/j.dss.2009.06.011>
- Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., & 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. <https://doi.org/10.1609/hcomp.v7i1.5280>
- Lagnado, D. A., Gerstenberg, T., & Zultan, R. I. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, 37(6), 1036-1073.
<https://doi.org/10.1111/cogs.12054>
- Lammert, O., Richter, B., Schütze, C., Thommes, K., & Wrede, B. (2024). Humans in XAI: Increased reliance in decision-making under uncertainty by using explanation strategies. *Frontiers in Behavioral Economics*, 3, Article 1377075.
<https://doi.org/10.3389/frbhe.2024.1377075>

- LeBoeuf, R. A., & Shafir, E. (2003). Deep thoughts and shallow frames: On the susceptibility to framing effects. *Journal of Behavioral Decision Making*, 16(2), 77-92. <https://doi.org/10.1002/bdm.433>
- Lee, M. H., & Chew, C. J. (2023). Understanding the effect of counterfactual explanations on trust and reliance on AI for Human-AI collaborative clinical decision making. In *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2), 1-22. <https://doi.org/10.1145/3610218>
- Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*, 139, Article 107539. <https://doi.org/10.1016/j.chb.2022.107539>
- Levin, I. P., Gaeth, G. J., Schreiber, J., & Lauriola, M. (2002). A new look at framing effects: Distribution of effect sizes, individual differences, and independence of types of effects. *Organizational Behavior and Human Decision Processes*, 88(1), 411-429. <https://doi.org/10.1006/obhd.2001.2983>
- Lewis, D. (1973). Counterfactuals and comparative possibility. *Journal of Philosophical Logic*, 2, 418-446. <https://www.jstor.org/stable/30226074>
- Li, Z., Rau, P. L. P., & Huang, D. (2020). Who should provide clothing recommendation services: Artificial intelligence or human experts? *Journal of Information Technology Research (JITR)*, 13(3), 113-125. <https://doi.org/10.4018/JITR.2020070107>

- Liefgreen, A., & Lagnado, D. A. (2023). Drawing conclusions: Representing and evaluating competing explanations. *Cognition*, 234, Article 105382. <https://doi.org/10.1016/j.cognition.2023.105382>
- Lim, B. Y., Dey, A. K., & 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. <https://doi.org/10.1145/1518701.1519023>
- Lim, J. B., & Oppenheimer, D. M. (2020). Explanatory preferences for complexity matching. *PloS one*, 15(4), Article e0230929. <https://doi.org/10.1371/journal.pone.0230929>
- Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27, 247-266. <https://doi.org/10.1017/S1358246100005130>
- Lipton, P. (2017). Inference to the best explanation. *A Companion to the Philosophy of Science*, 184-193. <https://doi.org/10.1002/9781405164481.ch29>
- Loewenstein, G. F., Weber, E.U., Hsee, C. K. & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, 127(2), 267-286. <http://dx.doi.org.elib.tcd.ie/10.1037/0033-2909.127.2.267>
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90-103. <https://doi.org/10.1016/j.obhdp.2018.12.005>

- Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10), 464-470.
<https://doi.org/10.1016/j.tics.2006.08.004>
- Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3), 232-257.
<https://doi.org/10.1016/j.cogpsych.2006.09.006>
- Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp.260-276). Oxford University Press.
- Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, 20(10), 748-759.
<https://doi.org/10.1016/j.tics.2016.08.001>
- Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4), 629-650.
<https://doi.org/10.1093/jcr/ucz013>
- Lucas, C. G., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review*, 122(4), 700-734.
<https://doi.org/10.1037/a0039655>
- Lucic, A., Haned, H., & 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. <https://doi.org/10.1145/3351095.3372824>
- Macbeth, G. & Razumiejczyk, E. (2019). Implicit facilitation effect on counterfactual and semifactual thinking. *Education Sciences and Psychology*, 54 (4), 15-21.

- Mackonis, A. (2013). Inference to the best explanation, coherence and other explanatory virtues. *Synthese*, 190(6), 975-995.
<https://doi.org/10.1007/s11229-011-0054-y>
- Madhavan, P., & Wiegmann, D. A. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors*, 49(5), 773-785.
<https://doi.org/10.1518/001872007X230154>
- Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175, Article 121390. <https://doi.org/10.1016/j.techfore.2021.121390>
- Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, 25(2), 147-186.
<https://doi.org/10.1080/1047840X.2014.877340>
- Mandel, D. R. (2014). Do framing effects reveal irrational choice? *Journal of Experimental Psychology: General*, 143(3), 1185-1198.
<https://doi.org/10.1037/a0034207>
- Mandel, D. R., & Lehman, D. R. (1996). Counterfactual thinking and ascriptions of cause and preventability. *Journal of Personality & Social Psychology*, 71, 450-463. <https://doi.org/10.1037/0022-3514.71.3.450>
- Mantler, J., Schellenberg, E. G., & Page, J. S. (2003). Attributions for serious illness: Are controllability, responsibility and blame different constructs? *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 35(2), 142-152.
<https://doi.org/10.1037/h0087196>

- Markman, K. D., McMullen, M. N., & Elizaga, R. A. (2008). Counterfactual thinking, persistence, and performance: A test of the reflection and evaluation model. *Journal of Experimental Social Psychology, 44*(2), 421-428. <https://doi.org/10.1016/j.jesp.2007.01.001>
- Markman, K. D., & Miller, A. K. (2006). Depression, control, and counterfactual thinking: Functional for whom? *Journal of Social and Clinical Psychology, 25*(2), 210-227. <https://doi.org/10.1521/jscp.2006.25.2.210>
- Markovits, H., Forgues, H. L., & Brunet, M. L. (2010). Conditional reasoning, frequency of counterexamples, and the effect of response modality. *Memory & Cognition, 38*(4), 485-492. <https://doi.org/10.3758/MC.38.4.485>
- Maule, J., & Villejoubert, G. (2007). What lies beneath: Reframing framing effects. *Thinking & Reasoning, 13*(1), 25-44. <https://doi.org/10.1080/13546780600872585>
- McCloy, R., & Byrne, R. M. J. (2000). Counterfactual thinking about controllable events. *Memory & Cognition, 28*, 1071-1078. <https://doi.org/10.3758/BF03209355>
- McCloy, R., & Byrne, R. M. J. (2002). Semifactual “even if” thinking. *Thinking & Reasoning, 8*, 41-67. <https://doi.org/10.1080/13546780143000125>
- McDonald, K., Graves, R., Yin, S., Weese, T., & Sinnott-Armstrong, W. (2021). Valence framing effects on moral judgments: A meta-analysis. *Cognition, 212*, Article 104703. <https://doi.org/10.1016/j.cognition.2021.104703>

- McEleney, A., & Byrne, R. M. J. (2006). Spontaneous counterfactual thoughts and causal explanations. *Thinking & Reasoning*, 12(2), 235-255.
<https://doi.org/10.1080/13546780500317897>
- McNeil, B. J., Pauker, S. G., Sox Jr, H. C., & Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine*, 306(21), 1259-1262.
<https://doi.org/10.1056/NEJM198205273062103>
- Menzies, P., & Beebe, H. (2001). Counterfactual theories of causation. In E. N. Zalta (Ed.), *The Stanford Encyclopaedia of Philosophy* (Spring 2001 Ed.). Stanford University.
<https://plato.stanford.edu/archives/spr2001/entries/causation-counterfactual/>
- Mercier, H., Rolison, J. J., Stragà, M., Ferrante, D., Walsh, C. R., & Girotto, V. (2017). Questioning the preparatory function of counterfactual thinking. *Memory & Cognition*, 45, 261-269.
<https://doi.org/10.3758/s13421-016-0660-5>
- Meshi, D., Biele, G., Korn, C. W., & Heekeren, H. R. (2012). How expert advice influences decision making. *PLoS One*, 7(11), Article e49748.
<https://doi.org/10.1371/journal.pone.0049748>
- Miller, D., & Gunasegaram, S. (1990). Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of Personality & Social Psychology*, 59(6), 1111-1118.
<https://doi.org/10.1037/0022-3514.59.6.1111>

- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81-97. <https://doi.org/10.1037/h0043158>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1-38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Moreno-Ríos, S., & Byrne, R. M. J. (2018). Inferences from disclosures about the truth and falsity of expert testimony. *Thinking & Reasoning*, 24(1), 41-78. <https://doi.org/10.1080/13546783.2017.1378724>
- Moreno-Ríos, S., Garcia-Madruga, J. A., & Byrne, R. M. J. (2008). Inferences from semifactual ‘even if’ conditionals. *Acta Psychologica*, 128(2), 197-209. <https://doi.org/10.1016/j.actpsy.2007.12.008>
- Muddamsetty, S. M., Jahromi, M. N., & Moeslund, T. B. (2021). Expert-level evaluations for explainable AI (XAI) methods in the medical domain. In *International Conference on Pattern Recognition*, 35-46. Springer, Cham. https://doi.org/10.1007/978-3-030-68796-0_3
- Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92, 289–316.
- Nabi, R. L., Walter, N., Oshidary, N., Endacott, C. G., Love-Nichols, J., Lew, Z. J., & Aune, A. (2020). Can emotions capture the elusive gain-loss framing effect? A meta-analysis. *Communication Research*, 47(8), 1107-1130. <https://doi.org/10.1177/0093650219861256>
- N'gbala, A., & Branscombe, N. R. (1997). When does action elicit more regret than inaction and is counterfactual mutation the mediator of this effect?

Journal of Experimental Social Psychology, 33(3), 324-343.

<https://doi.org/10.1006/jesp.1996.1322>

Okder, H. (2012). The illusion of the framing effect in risky decision making.

Journal of Behavioral Decision Making, 25(1), 63-73.

<https://doi.org/10.1002/bdm.715>

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, 22(4), 390-409. <https://doi.org/10.1002/bdm.637>

Orenes, I., Espino, O., & Byrne, R. M. J. (2022). Similarities and differences in understanding negative and affirmative counterfactuals and causal assertions: Evidence from eye-tracking. *Quarterly Journal of Experimental Psychology*, 75(4), 633-651.

<https://doi.org/10.1177/17470218211044085>

Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, 146(12), 1761-1780. <https://doi.org/10.1037/xge0000318>

Park, H., & Seo, K. (2023). Counterfactual vs. prefactual: Two narrative AIs improve causability for health data by different mechanisms. In *Proceedings of HCI Korea 2023*, 828-835.

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press.

<http://dx.doi.org/10.1017/CBO9780511803161>

Pearl, J. (2010). Causal inference. In *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, PMLR 6, 39-58.

<https://proceedings.mlr.press/v6/pearl10a.html>

- Pearl, J. (2011). The algorithmization of counterfactuals. *Annals of Mathematics and Artificial Intelligence*, 61, 29-39.
<http://dx.doi.org/10.1007/s10472-011-9247-9>
- Petrocelli, J. V., Seta, C. E., & Seta, J. J. (2013). Dysfunctional counterfactual thinking: When simulating alternatives to reality impedes experiential learning. *Thinking & Reasoning*, 19(2), 205-230.
<https://doi.org/10.1080/13546783.2013.775073>
- Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in Cognitive Sciences*, 23(12), 1026-1040.
<https://doi.org/10.1016/j.tics.2019.09.007>
- Pighin, S., Byrne, R. M. J., Ferrante, D., Gonzalez, M., & Girotto, V. (2011). Counterfactual thoughts about experienced, observed, and narrated events. *Thinking & Reasoning*, 17(2), 197-211.
<https://doi.org/10.1080/13546783.2011.561598>
- Piñon, A., & Gambara, H. (2005). A meta-analytic review of framing effect: Risky, attribute and goal framing. *Psicothema*, 17(2), 325-331.
- Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020). FACE: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 344-350. <https://doi.org/10.1145/3375627.3375850>
- Quillien, T., & Lucas, C. G. (2023). Counterfactuals and the logic of causal selection. *Psychological Review*, 131(5), 1208-1234.
<https://doi.org/10.1037/rev0000428>

- Quinn, M. S., & Keane, M. T. (2022). Factors affecting “expectations of the unexpected”: The impact of controllability and valence on unexpected outcomes. *Cognition*, 225, Article 105142.
<https://doi.org/10.1016/j.cognition.2022.105142>
- Raab, M., & Gigerenzer, G. (2015). The power of simplicity: A fast-and-frugal heuristics approach to performance science. *Frontiers in Psychology*, 6, Article 1672. <https://doi.org/10.3389/fpsyg.2015.01672>
- Rachev, N. R., Han, H., Lacko, D., Gelpí, R., Yamada, Y., & Lieberoth, A. (2021). Replicating the disease framing problem during the 2020 COVID-19 pandemic: A study of stress, worry, trust, and choice under risk. *PLoS One*, 16(9), Article e0257151.
<https://doi.org/10.1371/journal.pone.0257151>
- Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65(3), 429-447.
<https://doi.org/10.1037/0022-3514.65.3.429>
- Rehder, B. (2003a). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6), 1141-1159.
<https://doi.org/10.1037/0278-7393.29.6.1141>
- Rehder, B. (2003b). Categorization as causal reasoning. *Cognitive Science*, 27(5), 709-748. https://doi.org/10.1207/s15516709cog2705_2
- Rehder, B., & Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios.

Memory & Cognition, 45, 245-260.

<https://doi.org/10.3758/s13421-016-0662-3>

Reyna, V. F., & Brainerd, C. J. (1991). Fuzzy-trace theory and framing effects in choice: Gist extraction, truncation, and conversion. *Journal of Behavioral Decision Making*, 4(4), 249-262.

<https://doi.org/10.1002/bdm.3960040403>

Reyna, V. F., Brainerd, C. J., Chen, Z., & Bookbinder, S. H. (2021). Explaining risky choices with judgments: Framing, the zero effect, and the contextual relativity of gist. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(7), 1037-1053.

<https://doi.org/10.1037/xlm0001016>

Reyna, V. F., Chick, C. F., Corbin, J. C., & Hsia, A. N. (2013). Developmental reversals in risky decision making: Intelligence agents show larger decision biases than college students. *Psychological Science*, 25(1), 76-84. <https://doi.org/10.1177/0956797613497022>

Rim, S., & 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), 391-401.

<https://doi.org/10.1177/0146167213513304>

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, 121, 133-148.

Roese, N. J., & Epstude, K. (2017). The functional theory of counterfactual thinking: New evidence, new challenges, new insights. *Advances in Experimental Social Psychology*, 56, 1-79.

<https://doi.org/10.1016/bs.aesp.2017.02.001>

- Roese, N.J., & Olson, J.M. (Eds.). (1995). *What might have been: The social psychology of counterfactual thinking*. Erlbaum.
- Roese, N. J., Smallman, R., & Epstude, K. (2017). Do episodic counterfactual thoughts focus on controllable action?: The role of self-initiation. *Journal of Experimental Social Psychology*, 73, 14-23.
<https://doi.org/10.1016/j.jesp.2017.05.006>
- Rong, Y., Leemann, T., Nguyen, T. T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., & Kasneci, E. (2023). Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (4), 2104-2122. <https://doi.org/10.1109/TPAMI.2023.3331846>
- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26(5), 521-562.
https://doi.org/10.1207/s15516709cog2605_1
- Russell, C., McGrath, R., & Costabello, L. (2020). Learning relevant explanations. In *2020 ICML Workshop on Human Interpretability in Machine Learning (WHI 2020)*, 296-302.
- Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263, Article 110273.
<https://doi.org/10.1016/j.knosys.2023.110273>
- Santamaría, C., Espino, O., & Byrne, R. M. J. (2005). Counterfactual and semifactual conditionals prime alternative possibilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 1149-1154. <https://doi.org/10.1037/0278-7393.31.5.1149>

Schacter, D. L., Benoit, R. G., De Brigard, F., & Szpunar, K. K. (2015).

Episodic future thinking and episodic counterfactual thinking:

Intersections between memory and decisions. *Neurobiology of Learning and Memory*, 117, 14-21. <https://doi.org/10.1016/j.nlm.2013.12.008>

Schwitzgebel, E., & Cushman, F. (2015). Philosophers' biased judgments

persist despite training, expertise and reflection. *Cognition*, 141, 127-137. <https://doi.org/10.1016/j.cognition.2015.04.015>

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, 1-51. <https://doi.org/10.2307/1165893>

Severes, B., Carreira, C., Vieira, A. B., Gomes, E., Aparício, J. T., & Pereira, I.

(2023). The human side of XAI: Bridging the gap between AI and non-expert audiences. In *Proceedings of the 41st ACM International Conference on Design of Communication*, 126-132.

<https://doi.org/10.1145/3615335.3623062>

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition*, 21(4), 546-556.

<https://doi.org/10.3758/BF03197186>

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: an effort-reduction framework. *Psychological Bulletin*, 134(2), 207-222.

<https://doi.org/10.1037/0033-2909.134.2.207>

Shang, R., Feng, K. K., & Shah, C. (2022). Why am I not seeing it?

Understanding users' needs for counterfactual explanations in everyday recommendations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1330-1340.

<https://doi.org/10.1145/3531146.3533189>

Sherman, S. J., McConnell, A. R., Roese, N. J., & Olson, J. M. (1995).

Dysfunctional implications of counterfactual thinking: When alternatives to reality fail us. In N. J. Roese & J. M. Olson (Eds.), *What might have been: The social psychology of counterfactual thinking* (pp.199-231). Psychology Press.

Shortliffe & Buchanan. (1975). A model of inexact reasoning in medicine.

Mathematical Biosciences, 23(3), 351-379.

[https://doi.org/10.1016/0025-5564\(75\)90047-4](https://doi.org/10.1016/0025-5564(75)90047-4)

Silk-Eglit, G., & Kurtz, K. J. (2011). Types of cognitive content and the role of

relational processing in the illusion of explanatory depth. In

Proceedings of the Annual Meeting of the Cognitive Science Society,

33(33), 643-648. <https://escholarship.org/uc/item/582378mf>

Simon, H. A. (1990a). Invariants of human behaviour. *Annual Review of*

Psychology, 41, 1-19.

<https://doi.org/10.1146/annurev.ps.41.020190.000245>

Simon, H. A. (1990b). Bounded rationality. In A. Eatwell, M. Milgate & P.

Newman (Eds.), *Utility and probability* (pp.15-18). Palgrave

Macmillan. https://doi.org/10.1007/978-1-349-20568-4_5

Singh, R., Miller, T., Lyons, H., Sonenberg, L., Velloso, E., Vetere, F., Howe,

P., & Dourish, P. (2023). Directive explanations for actionable

explainability in machine learning applications. *ACM Transactions on*

Interactive Intelligent Systems, 13(4), Article 23.

<https://doi.org/10.1145/3579363>

- Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgements of likelihood. *Cognition*, 52(1), 1-21.
[https://doi.org/10.1016/0010-0277\(94\)90002-7](https://doi.org/10.1016/0010-0277(94)90002-7)
- Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, 66, 223-247.
<https://doi.org/10.1146/annurev-psych-010814-015135>
- Smith, S. M., & Levin, I. P. (1996). Need for cognition and choice framing effects. *Journal of Behavioral Decision Making*, 9(4), 283-290.
[https://doi.org/10.1002/\(SICI\)1099-0771\(199612\)9:4<283::AID-BDM241>3.0.CO;2-7](https://doi.org/10.1002/(SICI)1099-0771(199612)9:4<283::AID-BDM241>3.0.CO;2-7)
- Snizek, J. A., Schrah, G. E., & Dalal, R. S. (2004). Improving judgement with prepaid expert advice. *Journal of Behavioral Decision Making*, 17(3), 173-190. <https://doi.org/10.1002/bdm.468>
- Sober, E. (2015). *Ockham's razors*. Cambridge University Press.
- Sokol, K., & Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 56-67.
<https://doi.org/10.1145/3351095.3372870>
- Spirtes, P., Glymour, C., & Scheines, R. (1991). From probability to causality. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 64(1), 1-36.
<https://www.jstor.org/stable/4320244>
- Spellman, B. A., & Mandel, D. R. (1999). When possibility informs reality: Counterfactual thinking as a cue to causality. *Current Directions*

in Psychological Science, 8, 120-123.

<https://doi.org/10.1111/1467-8721.00028>

Spence, A., & Pidgeon, N. (2010). Framing and communicating climate change: The effects of distance and outcome frame manipulations. *Global Environmental Change*, 20(4), 656-667.

<https://doi.org/10.1016/j.gloenvcha.2010.07.002>

Steiger, A., & Kühberger, A. (2018). A meta-analytic re-appraisal of the framing effect. *Zeitschrift für Psychologie*, 226 (1), 45-55.

<https://doi.org/10.1027/2151-2604/a000321>

Stevenson, R. J., & Over, D. E. (2001). Reasoning from uncertain premises: Effects of expertise and conversational context. *Thinking & Reasoning*, 7(4), 367-390. <https://doi.org/10.1080/13546780143000080>

Sweeny, K., & Vohs, K. D. (2012). On near misses and completed tasks: The nature of relief. *Psychological Science*, 23(5), 464-468.

<https://doi.org/10.1177/0956797611434590>

Taylor, J. E. T., & Taylor, G. W. (2021). Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review*, 28(2), 454-475.

<https://doi.org/10.3758/s13423-020-01825-5>

Teigen, K. H., & Jensen, T. K. (2010). Unlucky victims or lucky survivors? *European Psychologist*, 16(1), 48-57.

<https://doi.org/10.1027/1016-9040/a000033>

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, 12(3), 435-467. <https://doi.org/10.1017/S0140525X00057046>

Thagard, P. (2000). Probabilistic networks and explanatory coherence.

Cognitive Science Quarterly, 1(1), 91-114.

Thagard, P., & Verbeurgt, K. (1998). Coherence as constraint satisfaction.

Cognitive Science, 22(1), 1-24. [https://doi.org/10.1016/S0364-0213\(99\)80033-0](https://doi.org/10.1016/S0364-0213(99)80033-0)

Thompson, V. A., & Byrne, R. M. J. (2002). Reasoning counterfactually:

Making inferences about things that didn't happen. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 1154-1170. <https://doi.org/10.1037//0278-7393.28.6.1154>

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis

of loss aversion in decision-making under risk. *Science*, 315(5811), 515-518. <https://doi.org/10.1126/science.1134239>

Trafimow, D., & Uhalt, J. (2015). The alleged tradeoff between explanatory

breadth and predictive power. *Theory & Psychology*, 25(6), 833-840. <https://doi.org/10.1177/0959354315591052>

Tran, K., & Nguyen, T. (2021). Preliminary research on the social attitudes

toward AI's involvement in Christian education in Vietnam: Promoting AI technology for religious education. *Religions*, 12(3), Article 208. <https://doi.org/10.3390/rel12030208>

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics

and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157), 1124-1131. <https://doi.org/10.1126/science.185.4157.1124>

- Tversky, A., & Kahneman, D. (1980). Causal schemata in judgment under uncertainty. In Fishbein, M. (Ed.) *Progress in social psychology*, (Vol.1, pp. 49-72). Lawrence Erlbaum Associates.
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453-458.
<https://doi.org/10.1126/science.7455683>
- Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10-19.
<https://doi.org/10.1145/3287560.3287566>
- van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291, Article 103404.
<https://doi.org/10.1016/j.artint.2020.103404>
- Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual explanations for machine learning: A review. *ArXiv preprint*, ArXiv:2010.10596.
<https://doi.org/10.48550/arXiv.2010.10596>
- Vrantsidis, T. H., & Lombrozo, T. (2022). Simplicity as a cue to probability: Multiple roles for simplicity in evaluating explanations. *Cognitive Science*, 46(7), Article e13169. <https://doi.org/10.1111/cogs.13169>
- Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology (Harvard JOLT)*, 31(2), 841-888.
<https://heinonline-org.elib.tcd.ie/HOL/Page?public=true&handle=hein.>

journals/hjlt31&div=29&collection=journals

- Walker, C. M., Bonawitz, E., & Lombrozo, T. (2017). Effects of explaining on children's preference for simpler hypotheses. *Psychonomic Bulletin & Review*, 24, 1538-1547. <https://doi.org/10.3758/s13423-016-1144-0>
- Walsh, C. R., & Sloman, S. A. (2004). Revising causal beliefs. *In Proceedings of the Annual Meeting of the Cognitive Science Society*, 26(26), 1423-1427. <https://escholarship.org/uc/item/6vk9g665>
- Walsh, C. R., & Sloman, S. A. (2004). Revising causal beliefs. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp.1423-1427). Lawrence Erlbaum Associates.
- Warren, G., Byrne, R. M. J., & Keane, M. T. (2023). Categorical and continuous features in counterfactual explanations of AI systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 171-187. <https://doi.org/10.1145/3581641.3584090>
- Waytz, A., & Norton, M. I. (2014). Botsourcing and outsourcing: Robot, British, Chinese, and German workers are for thinking—not feeling—jobs. *Emotion*, 14(2), 434-444. <https://doi.org/10.1037/a0036054>
- Wells, G.L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology*, 56, 161–169. <https://doi.org/10.1037/0022-3514.56.2.161>
- White, P. A. (1995). Use of prior beliefs in the assignment of causal roles: Causal powers versus regularity-based accounts. *Memory & Cognition*, 23, 243-254. <https://doi.org/10.3758/BF03197225>

- Whitney, P., Rinehart, C. A., & Hinson, J. M. (2008). Framing effects under cognitive load: The role of working memory in risky decisions. *Psychonomic Bulletin & Review*, 15, 1179-1184.
<https://doi.org/10.3758/PBR.15.6.1179>
- Wien, A. H., & Peluso, A. M. (2021). Influence of human versus AI recommenders: The roles of product type and cognitive processes. *Journal of Business Research*, 137, 13-27.
<https://doi.org/10.1016/j.jbusres.2021.08.016>
- Williamson, J. (2009). Probabilistic theories. In H. Beebe, C. Hitchcock & P. Menzies (Eds.), *The Oxford handbook of causation* (pp.185-212). Oxford University Press.
- Wilkenfeld, D. A., & Lombrozo T. (2015). Inference to the best explanation (IBE) versus explaining for the best inference (EBI). *Science & Education*, 24 (9-10), 1059-1077.
<https://doi.org/10.1007/s11191-015-9784-4>
- Wilkinson, M., Ball, L., & Cooper, R. (2010). Arbitrating between theory-theory and simulation theory: Evidence from a think-aloud study of counterfactual reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 32(32), 1008-1013.
<https://escholarship.org/uc/item/5sz300xz>
- Wyszynski, M., & Diederich, A. (2022). Keep your budget together! Investigating determinants on risky decision-making about losses. *PLoS one*, 17(3), Article e0265822.
<https://doi.org/10.1371/journal.pone.0265822>

- Xie, Y., & Beck, S. R. (2022). The bidirectional relation between counterfactual thinking and closeness, controllability, and exceptionality. *Frontiers in Psychology, 13*, Article 732870. <https://doi.org/10.3389/fpsyg.2022.732870>
- Yacoby, Y., Green, B., Griffin Jr, C. L., & Doshi-Velez, F. (2022). “If it didn’t happen, why would I change my decision?”: How judges respond to counterfactual explanations for the public safety assessment. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 10*, 219-230. <https://doi.org/10.1609/hcomp.v10i1.22001>
- Yang, Y., Zheng, J., Yu, Y., Qiu, Y., & Wang, L. (2024). The role of recommendation sources and attribute framing in online product recommendations. *Journal of Business Research, 174*, Article 114498. <https://doi.org/10.1016/j.jbusres.2024.114498>
- Yaniv, I. (2004). Receiving other people’s advice: Influence and benefit. *Organizational Behavior and Human Decision Processes, 93*(1), 1-13. <https://doi.org/10.1016/j.obhdp.2003.08.002>
- Ye, L. R., & Johnson, P. E. (1995). The impact of explanation facilities on user acceptance of expert systems advice. *MIS Quarterly, 19*(2), 157-172. <https://doi.org/10.2307/249686>
- Yeung, S. K., & Feldman, G. (2022). Revisiting the temporal pattern of regret in action versus inaction: Replication of Gilovich and Medvec (1994) with extensions examining responsibility. *Collabra: Psychology, 8*(1), Article 37122. <https://doi.org/10.1525/collabra.37122>

Zemla, J. C., Sloman, S. A., Bechlivanidis, C., & Lagnado, D. A. (2017).

Evaluating everyday explanations. *Psychonomic Bulletin & Review*, 24,

1488-1500. <https://doi.org/10.3758/s13423-017-1258-z>

Zemla, J. C., Sloman, S. A., Bechlivanidis, C., & Lagnado, D. A. (2023). Not

so simple! Causal mechanisms increase preference for complex

explanations. *Cognition*, 239, Article 105551.

<https://doi.org/10.1016/j.cognition.2023.105551>

Zhang, Y., & Slovic, P. (2019). Much ado about nothing: The zero effect in

life-saving decisions. *Journal of Behavioral Decision Making*, 32(1),

30-37. <https://doi.org/10.1002/bdm.2089>

Appendices

Table of Contents

Appendix 2.1 – Ethics Approval, Data Protection and Consent Information (for Experiments 1 to 11)	323
Appendix 2.2 – Information and Debrief Sheets for Chapter 2.....	330
Appendix 2.3 – Experimental Materials for Chapter 2	334
Appendix 2.4 – Supplementary Analysis for Chapter 2	350
Appendix 2.5 – Materials Check for Experiment 2b	359
Appendix 3.1 – Information and Debrief Sheets for Chapter 3	363
Appendix 3.2 – Pilot Study.....	366
Appendix 3.3 – Experimental Materials for Chapter 3.....	382
Appendix 3.4 - Supplementary Analysis for Chapter 4.....	405
Appendix 4.1 – Information and Debrief Sheets for Chapter 4.....	410
Appendix 4.2 – Experimental Materials for Chapter 4.....	413
Appendix 4.3 - Supplementary Analysis for Chapter 4.....	415

Appendix 2.1 – Ethics Approval, Data Protection, and Consent Information (for Experiments 1 to 11)



Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

F.A.O. Xinyue Dai

Approval ID: SPREC112020-22

School of Psychology Research Ethics Committee

20th January 2021

Dear Xinyue,

The School of Psychology Research Ethics Committee has reviewed your application entitled “Counterfactual Explanations for AI Decision Support Systems in Smart Agriculture”, and I am pleased to inform you that it was approved.

Please note that you will be required to submit a completed **Project Annual Report Form** on each anniversary of this approval, until such time as the research is complete and the thesis is submitted. The form is available for download from the Ethics section of the School website.

Adverse events associated with the conduct of this research must be reported immediately to the Chair of the Ethics Committee.

Yours sincerely,

A handwritten signature in black ink, appearing to read 'Richard Carson'.

Richard Carson
Chair,
School of Psychology Research Ethics Committee

**DATA PROTECTION, ETHICAL APPROVAL AND RELATED
MATTERS:**

Name of Study: X

Principal Investigator: Xinyue Dai, PhD student, School of Psychology, Trinity College Dublin, University of Dublin, Ireland. Email: daix@tcd.ie, Tel: +353 1 896 4054.

Supervised by: Ruth Byrne, Professor of Cognitive Science, School of Psychology and Institute of Neuroscience, Trinity College Dublin, University of Dublin, Ireland, Email: rmbyrne@tcd.ie, Tel: +353 1 896 4054.

Data Controllers: Trinity College Dublin; Data Protection Officer, Secretary's Office, Trinity College Dublin, Dublin 2, Ireland.

You are being invited to take part in a research study that is being presented online through Prolific by Xinyue Dai. Before you decide whether you wish to take part, please read this information screen carefully and ask any questions you have by contacting Xinyue Dai (Email: daix@tcd.ie, Tel: +353 1 896 4054). This information screen has five parts: (1)The Study; (2) Data Protection; (3) Costs, Funding and Approval; (4) Future Research; (5) Further Information

PART 1-THE STUDY. We are doing this study to investigate how people understand various processes. You are invited if you are over 18 years old. We intend to invite about X participants at the initial stage. Participation in the study is completely voluntary. Your decision not to consent will have no adverse consequences. Your consent can be withdrawn at any time up to the point when data analysis has been completed, i.e., two months from today's date. Should you wish to withdraw during the experiment you can simply quit

the webpage. All incomplete responses will be eliminated. If you decide you wish to withdraw after you have submitted your answers you can do so at any point during the next two months. On a subsequent screen you will be asked to generate a unique random code known only to you and if you wish to withdraw you can contact Xinyue Dai to reveal your code and delete your data (Email: daix@tcd.ie, Tel:+353 1 896 4054). The study will take place starting in X (month & year). It is programmed using the software package Alchemer and participants are recruited through Prolific. If you decide to participate, you will need to read the consent form and give your consent before starting the experiment. The experiment will take approximately X minutes to complete. You will be asked to make decisions in different scenarios. After completing the experiment you will be provided with further information on its aims. Your responses will be stored in an excel file which is completely anonymised and does not hold any personal data about you. The anonymised data will be analysed statistically. The fully anonymised data set will be made available online at the Open Science Forum in line with best international practice for publication of scientific results. Taking part in the study will not directly benefit you. However, your data may help us to better understand how people make inferences. This is a long-term research project, so the benefits of the research may not be seen for several years. It is not expected that taking part will cause any risk to you or others. The results of the study will be reported in academic publications and at academic meetings and conferences. No information which reveals your identity will be disclosed.

PART 2- DATA PROTECTION. This study will NOT collect any personal data. You are entitled to: The right to access to your data and receive a copy of

it; The right to restrict or object to processing of your data; The right to object to any further processing of the information we hold about you (except where it is de-identified); The right to have inaccurate information about you corrected or deleted; The right to receive your data in a portable format and to have it transferred to another data controller; The right to request deletion of your data.

By law you can exercise these rights in relation to your personal data, unless the request would make it impossible or very difficult to conduct the research.

You can exercise these rights by contacting the study researcher (Xinyue Dai, daix@tcd.ie) or the Trinity College Data Protection Officer, Secretary's Office, Trinity College Dublin, Dublin 2, Ireland, dataprotection@tcd.ie, www.tcd.ie/privacy.

PART 3 - COSTS, FUNDING AND APPROVAL. The study has been approved by Trinity College Dublin School of Psychology Research Ethics Committee. Approval was granted on 20/02/2021 (SPREC112020-22). The study is a part of a PhD project conducted by Xinyue Dai and supervised by Professor Ruth Byrne. The project is funded by a PhD scholarship from Teagasc, the agriculture and food development authority of Ireland. Results will be published in scientific publications and in closed academic settings and will not be used for commercial purposes. You will be paid X UK sterling on Prolific once you have successfully completed the experiment which will take about X minutes.

PART 4 - FUTURE RESEARCH. This study will NOT collect any personal data. We do not intend to contact you again following your participation.

PART 5 - FURTHER INFORMATION. If you have any concerns or questions, you can contact:

Principal Investigator: Xinyue Dai, PhD student, School of Psychology, Trinity College Dublin, University of Dublin, Ireland. Email: daix@tcd.ie, Tel: +353 1 896 4054.

Supervisor: Ruth Byrne, Professor of Cognitive Science, School of Psychology and Institute of Neuroscience, Trinity College Dublin, University of Dublin, Ireland, Email: rmbyrne@tcd.ie, Tel: +353 1 896 4054.

Data Protection Officer, Trinity College Dublin: Data Protection Officer, Secretary's Office, Trinity College Dublin, Dublin 2, Ireland. Email: dataprotection@tcd.ie. Website: www.tcd.ie/privacy.

Under GDPR, if you are not satisfied with how your data is being processed, you have the right to lodge a complaint with the Office of the Data Protection Commission, 21 Fitzwilliam Square South, Dublin 2, Ireland. Website: www.dataprotection.ie.

You can take a screen shot of this information to keep a copy if you wish. If you would like to take part in this study, you will be asked to provide CONSENT on the next screen.

CONSENT FORM

There are three sections in this form. Each section has statements and you are asked to tick a box if you agree. Please ask any questions you may have when reading each of the statements. Thank you for participating.

GENERAL. I confirm I have read and understood the Information leaflet (the previous screen) for this study. The information has been fully explained to me, and I have been able to ask questions, all of which have been answered to my satisfaction. I understand that this study is entirely voluntary,

and if I decide that I do not want to take part, I can stop taking part at any time without giving a reason. I can contact the researchers with my self-generated unique participant code if I wish to withdraw my information after submitting, up until the date the data are combined for processing (two months from today's data). I understand that my data will be analysed by the research team. I agree that these individuals in the team can access my records. I understand that I will be paid X sterling on Prolific. I know how to contact the research team if I need to. I agree to take part in this research study, having been fully informed of the risks, benefits and alternatives which are set out in full in the information which I have been provided with.

DATA PROCESSING. I understand that I have the right to provide my personal information on a voluntary basis. I understand my data will be anonymized. I understand that the random code that I generate will be stored with my data and the research team will only be able to identify my data through this code if I choose to reveal it. I agree to allow my data to be shared with third parties including: academic research institutions and online Open Science Forum for the purpose of best international practice for replicability of published work in psychology, as described on the Information screen. I understand that personal information about me, including the transfer of this personal information about me outside of the EU, will be protected in accordance with the General Data Protection Regulation. I understand that there are no direct benefits to me from participating in this study. I understand that results from analysis of my data will not be given to me.

FUTURE USE OF INFORMATION. I give permission for my anonymised data, to be shared with the scientific community and the general

public e.g., via the Open Science Forum database on the internet. I understand that my data will be fully anonymised before sharing - no personally identifying data will be shared. I understand that anonymised data shared in this way will be accessible to researchers and members of the public anywhere in the world, not just the EU. I understand that by sharing anonymised data in this way, my anonymised data might be used for other, future research projects in addition to the study I am currently participating in. Those future projects can focus on any topic and might be completely unrelated to the goals of this study. I understand that once the data are shared, they cannot be destroyed, withdrawn, or recalled, because they can no longer be linked with me. I give permission for my anonymised data to be stored for possible future research related to the current study without further consent being required but only if the research is approved by a Research Ethics Committee. I understand I will not be paid for any future use of my data or products derived from it.

Please tick to indicate you have read and understand the information above and that you have given your consent to participate.

Appendix 2.2 – Information and Debrief Sheets for Chapter 2***Experiments 1 to 4: Information***

Thank you for agreeing to participate in this study.

Participation in the study will take about X minutes and you will be paid a sum of X in recognition of your participation.

Please take part in this study only if you are willing to consider the tasks seriously and take some time to work out your answer. Please also only take part in the study in a quiet place where you will be uninterrupted for its duration.

The study examines how you reason about explanations, you will be asked to answer questions about several imaginary situations on an alien planet.

Please do not take part in the study if you have previously taken part in any study that has asked you to do a similar task.

Please click the "Next" button for more information about participation.

Experiments 1 to 4: Debrief

All experiments: Thank you very much for your participation! You have now reached the end of the study.

Experiment 1: The aim of the study is to examine how people choose explanations based on information provided - when the treatability¹ of causes is being manipulated. Some participants received scenarios where no treatability information was provided, other participants received scenarios where treatability information was provided. We hope to investigate whether there is a preference when the explanation contains only treatable causes.

Experiments 2a and 2b: The aim of the study is to examine how people choose explanations based on information provided - when the controllability² of causes is being manipulated. Some participants received scenarios where no controllability information was provided, other participants received scenarios where controllability information was provided. We hope to investigate whether there is a preference when the explanation contains only controllable causes.

¹ During the experimental design and data collection stage, we referred to consequent-controllability as treatability in our experimental materials in Experiments 1 and 3a.

² During the experimental design and data collection stage, we referred to antecedent-controllability as controllability in our experimental materials in Experiments 2a, 2b, and 3b.

Experiment 3a: The aim of the study is to examine how people choose explanations based on the information provided. People generally choose the simpler explanations, and we aim to examine whether people still choose the simpler explanation when treatability is being manipulated.

Experiment 3b: The aim of the study is to examine how people choose explanations based on the information provided. People generally choose the simpler explanations, and we aim to examine whether people still choose the simpler explanation when the controllability is being manipulated.

Experiment 4: The aim of the study is to examine how people choose explanations based on information provided - when the simplicity and type of explanations are being manipulated. Some participants received explanations without counterfactual explanations (i.e., how the outcome would have been different), e.g., "Treda has Trichet's syndrome", while other participants received explanations with counterfactual statements, e.g., "If Treda had not had Trichet's syndrome, Treda would not have had the symptoms". We hope to investigate whether the type of explanations influences people's choices of explanations.

All experiments: We do not anticipate that anything in the questions we presented will cause any distress to participants. However, if you find that any of the tasks or your participation in the study does cause you some distress, you may wish to contact a support service such as the Samaritans, www.samaritans.org. Their website provides details of their local telephone numbers and addresses, for example, in Ireland, Samaritans Ireland is available at: telephone- 01 872 7700; address: 151 Marlborough Street, Dublin 1.

If you have any questions or concerns, please contact:

Xinyue Dai, PhD student, School of Psychology, Trinity College Dublin,
University of Dublin, Ireland. Email: daix@tcd.ie, Tel: +353 1 896 4054.

Ruth Byrne, Professor of Cognitive Science, School of Psychology and
Institute of Neuroscience, Trinity College Dublin, University of Dublin,
Ireland, Email: rmbyrne@tcd.ie, Tel: +353 1 896 4054.

The next page will redirect you back to the Prolific site.

Appendix 2.3 – Experimental Materials for Chapter 2 (Farming Domain)***Experiment 1***

Participants were given the information below for the farming domain:

There is a population of 750 aliens that lives on planet Zorg. You are a farming expert trying to understand an alien's farm problem. The alien farm X has two problems: grass growth is low and weed growth is high.

Both micro-organisms Excie and micro-organism Rirtan always cause low grass growth, but they never cause high weed growth. (For consequent-controllability information condition only: Micro-organisms Excie can be remedied with fertiliser. There is no remedy for micro-organism Rirtan.)

Both micro-organism Cartia and micro-organism Pacis always cause high weed growth, but they never cause low grass growth. (For consequent-controllability information condition only: There is no remedy for Micro-organism Cartia. Micro-organisms Pacis can be remedied with herbicide.)

Nothing else is known to cause an alien farm to have low grass growth or high weed growth.

What do you think is the most satisfying explanation for the problems that the alien farm X is exhibiting?

- (A) Alien farm X has Micro-organism Excie and Micro-organism Pacis.
- (B) Alien farm X has Micro-organism Excie and Micro-organism Cartia.
- (C) Alien farm X has Micro-organism Excie and Micro-organism Rirtan.
- (D) Alien farm X has Micro-organism Pacis and Micro-organism Rirtan.

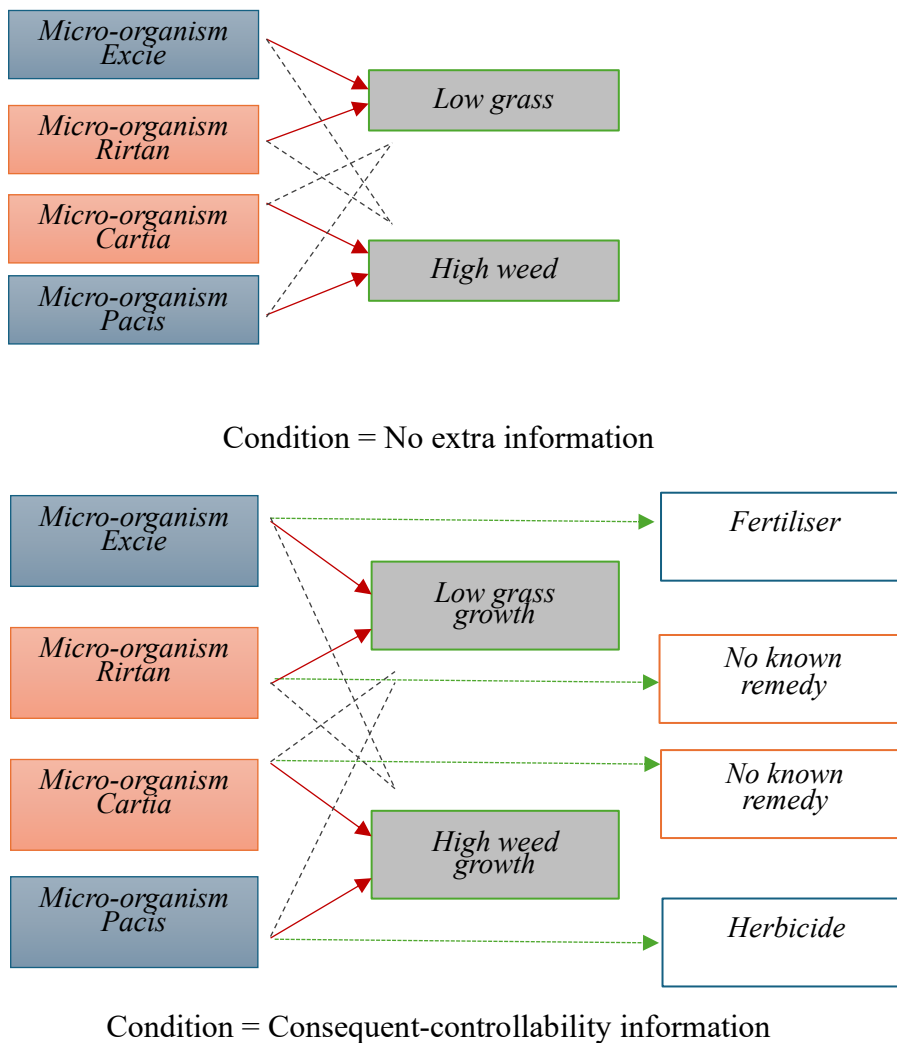
(E) Alien farm X has Micro-organism Pacis and Micro-organism Cartia.

(F) Alien farm X has Micro-organism Rirtan and Micro-organism
Cartia.

See **Figure A1** for a diagram of the causal relations of the farming
domain.

Figure A1

Diagrams for the Causal Structure of the Farming Domain in Experiment 1



Note. Red arrows indicate a cause-effect relation. Dotted lines indicate that the diseases do not lead to the symptoms. These diagrams were for illustration purposes only, participants were not provided with diagrams during the experiment.

Experiment 2a

Participants were given the information below for the farming domain:

There is a population of 750 aliens that lives on planet Zorg. You are a farming expert trying to understand an alien's farm problem. The alien farm X has two problems: grass growth is low and weed growth is high.

Both micro-organisms Excie and micro-organism Rirtan always cause low grass growth, but they never cause high weed growth. (For antecedent-controllability information condition only: The micro-organism Excie is the result of fertiliser misuse. There is no known cause for the micro-organism Rirtan.)

Both micro-organism Cartia and micro-organism Pacis always cause high weed growth, but they never cause low grass growth. (For antecedent-controllability information condition only: There is no known cause for the micro-organism Cartia. The micro-organism Pacis is the result of herbicide misuse.)

Nothing else is known to cause an alien farm to have low grass growth or high weed growth.

What do you think is the most satisfying explanation for the problems that the alien farm X is exhibiting?

- (A) Alien farm X has Micro-organism Excie and Micro-organism Pacis.
- (B) Alien farm X has Micro-organism Excie and Micro-organism Cartia.
- (C) Alien farm X has Micro-organism Excie and Micro-organism Rirtan.

(D) Alien farm X has Micro-organism Pacis and Micro-organism Rirtan.

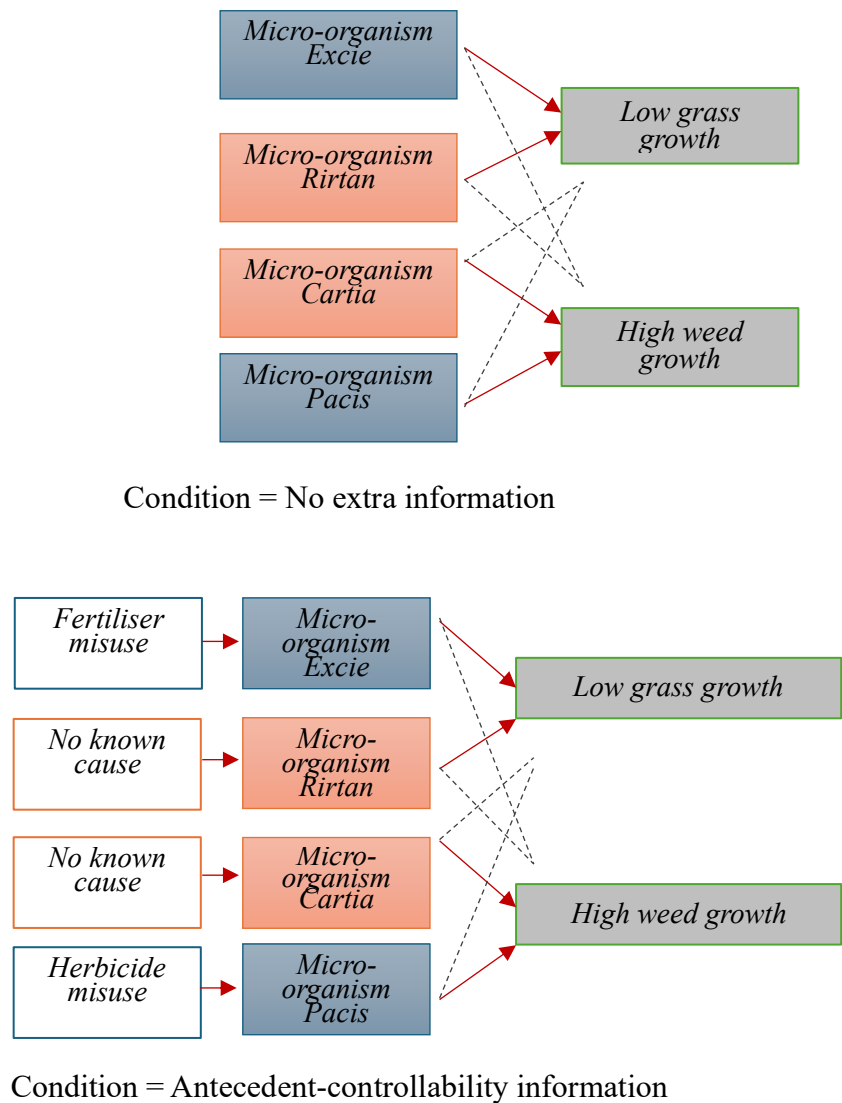
(E) Alien farm X has Micro-organism Pacis and Micro-organism Cartia.

(F) Alien farm X has Micro-organism Rirtan and Micro-organism Cartia.

See **Figure A2** for a diagram of the causal relations of the farming domain.

Figure A2

Diagrams for the Causal Structure of the Farming Domain in Experiment 2a



Note. Red arrows indicate a cause-effect relation. Dotted lines indicate that the diseases do not lead to the symptoms. These diagrams were for illustration purposes only, participants were not provided with diagrams during the experiment.

Experiment 2b

Participants were given the information below for the farming domain:

There is a population of 750 aliens that lives on planet Zorg. You are a farming expert trying to understand an alien's farm problem. The alien farm X has two problems: grass growth is low and weed growth is high.

Both micro-organisms Excie and micro-organism Rirtan always cause low grass growth, but they never cause high weed growth. (For antecedent-controllability information condition only: Micro-organisms Excie is the result of fertiliser misuse. Micro-organism Rirtan is the result of day-length.)

Both micro-organism Cartia and micro-organism Pacis always cause high weed growth, but they never cause low grass growth. (For antecedent-controllability information condition only: Micro-organism Cartia is the result of mountainous terrain. Micro-organisms Pacis is the result of herbicide misuse.)

Nothing else is known to cause an alien farm to have low grass growth or high weed growth.

What do you think is the most satisfying explanation for the problems that the alien farm X is exhibiting?

- (A) Alien farm X has Micro-organism Excie and Micro-organism Pacis.
- (B) Alien farm X has Micro-organism Excie and Micro-organism Cartia.
- (C) Alien farm X has Micro-organism Excie and Micro-organism Rirtan.
- (D) Alien farm X has Micro-organism Pacis and Micro-organism Rirtan.

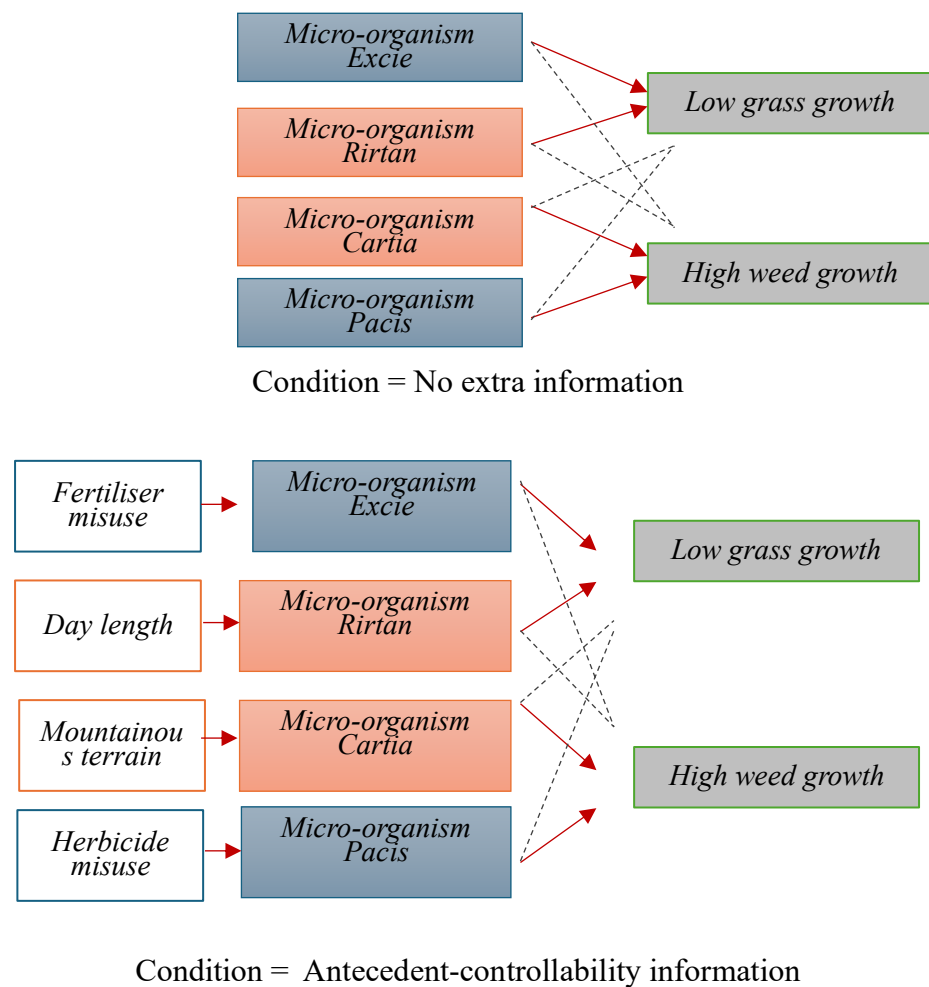
(E) Alien farm X has Micro-organism Pacis and Micro-organism Cartia.

(F) Alien farm X has Micro-organism Rirtan and Micro-organism Cartia.

See **Figure A3** for a diagram of the causal relations of the farming domain.

Figure A3

Diagrams for the Causal Structure of the Farming Domain in Experiment 2b



Note. Red arrows indicate a cause-effect relation. Dotted lines indicate that the diseases do not lead to the symptoms. These diagrams were for illustration purposes only, participants were not provided with diagrams during the experiment.

Experiment 3a

Participants were given the information below for the farming domain:

There is a population of 750 aliens that lives on planet Zorg. You are a farming expert trying to understand an alien's farm problem. The alien farm X has two problems: grass growth is low and weed growth is high.

No extra information: Micro-organism Excie in the soil always causes both low grass growth and high weed growth. Micro-organism Rirtan in the soil always causes low grass growth, but it never causes high weed growth. When an alien farm soil has micro-organism Pacis, that farm will always have high weed growth, but this soil type will never cause low grass growth. Nothing else is known to cause an alien farm to have low grass growth or high weed growth.

Simple cause uncontrollable: Micro-organism Excie in the soil always causes both low grass growth and high weed growth. There is no remedy for micro-organism Excie. Micro-organism Rirtan in the soil always causes low grass growth, but it never causes high weed growth. Micro-organism Rirtan can be remedied with fertiliser. When an alien farm soil has micro-organism Pacis, that farm will always have high weed growth, but this soil type will never cause low grass growth. Micro-organism Pacis can be remedied with herbicide. Nothing else is known to cause an alien farm to have low grass growth or high weed growth.

Simple cause controllable: Micro-organism Excie in the soil always causes both low grass growth and high weed growth. Micro-organism Excie can be remedied with specialised seed choice. Micro-organism Rirtan in the soil always causes low grass growth, but it never causes high weed growth.

There is no remedy for Micro-organism Rirtan. When an alien farm soil has micro-organism Pacis, that farm will always have high weed growth, but this soil type will never cause low grass growth. There is no remedy for Micro-organism Pacis. Nothing else is known to cause an alien farm to have low grass growth or high weed growth.

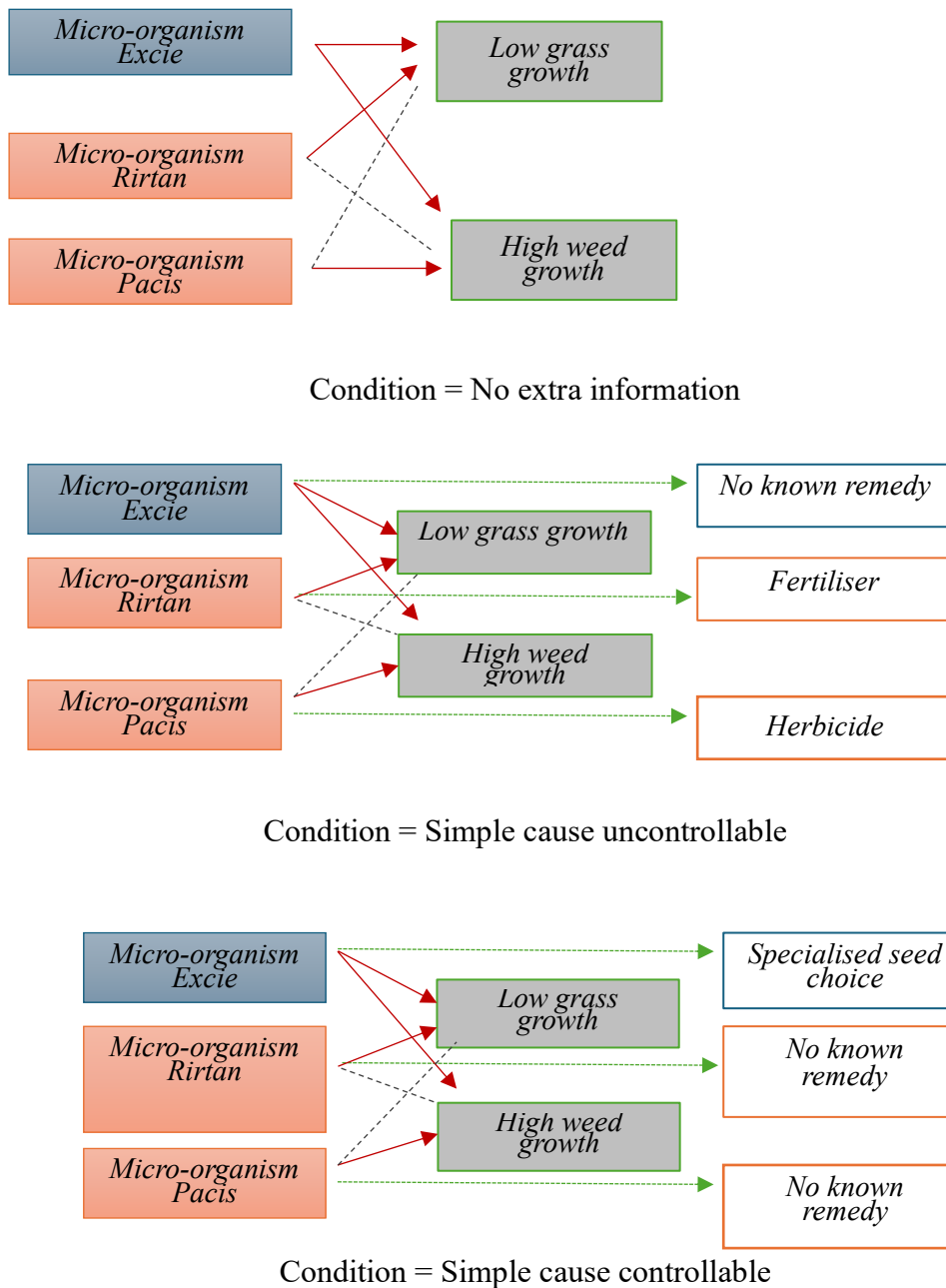
What do you think is the most satisfying explanation for the problems that the alien farm X is exhibiting?

- (A) Alien farm X has Micro-organism Excie in the soil.
- (B) Alien farm X has Micro-organism Rirtan in the soil.
- (C) Alien farm X has Micro-organism Pacis in the soil.
- (D) Alien farm X has Micro-organism Excie and Micro-organism Rirtan in the soil.
- (E) Alien farm X has Micro-organism Excie and Micro-organism Pacis in the soil.
- (F) Alien farm X has Micro-organism Rirtan and Micro-organism Pacis in the soil.

See **Figure A4** for a diagram of the causal relations of the farming domain.

Figure A4

Diagrams for the Causal Structure of the Farming Domain in Experiment 3a



Note. Red arrows indicate a cause-effect relation. Dotted lines indicate that the diseases do not lead to the symptoms. These diagrams were for illustration purposes only, participants were not provided with diagrams during the experiment.

Experiment 3b

Participants were given the information below for the farming domain:

There is a population of 750 aliens that lives on planet Zorg. You are a farming expert trying to understand an alien's farm problem. The alien farm X has two problems: grass growth is low and weed growth is high. Micro-organism Excie in the soil always causes both low grass growth and high weed growth.

No extra information: Micro-organism Rirtan in the soil always causes low grass growth, but it never causes high weed growth. When an alien farm soil has micro-organism Pacis, that farm will always have high weed growth, but this soil type will never cause low grass growth. Nothing else is known to cause an alien farm to have low grass growth or high weed growth.

Simple cause uncontrollable: Micro-organism Excie in the soil, which is the result of day-length, always causes both low grass growth and high weed growth. Micro-organism Rirtan in the soil, which is the result of herbicide misuse, always causes low grass growth, but it never causes high weed growth. When an alien farm soil has micro-organism Pacis, which is the result of poor seed choice, that farm will always have high weed growth, but this soil type will never cause low grass growth. Nothing else is known to cause an alien farm to have low grass growth or high weed growth.

Simple cause controllable: Micro-organism Excie in the soil, which is the result of fertiliser misuse, always causes both low grass growth and high weed growth. Micro-organism Rirtan in the soil, which is the result of mountainous terrain, always causes low grass growth, but it never causes high weed growth. When an alien farm soil has micro-organism Pacis, which is the

result of its ground temperature, that farm will always have high weed growth, but this soil type will never cause low grass growth. Nothing else is known to cause an alien farm to have low grass growth or high weed growth.

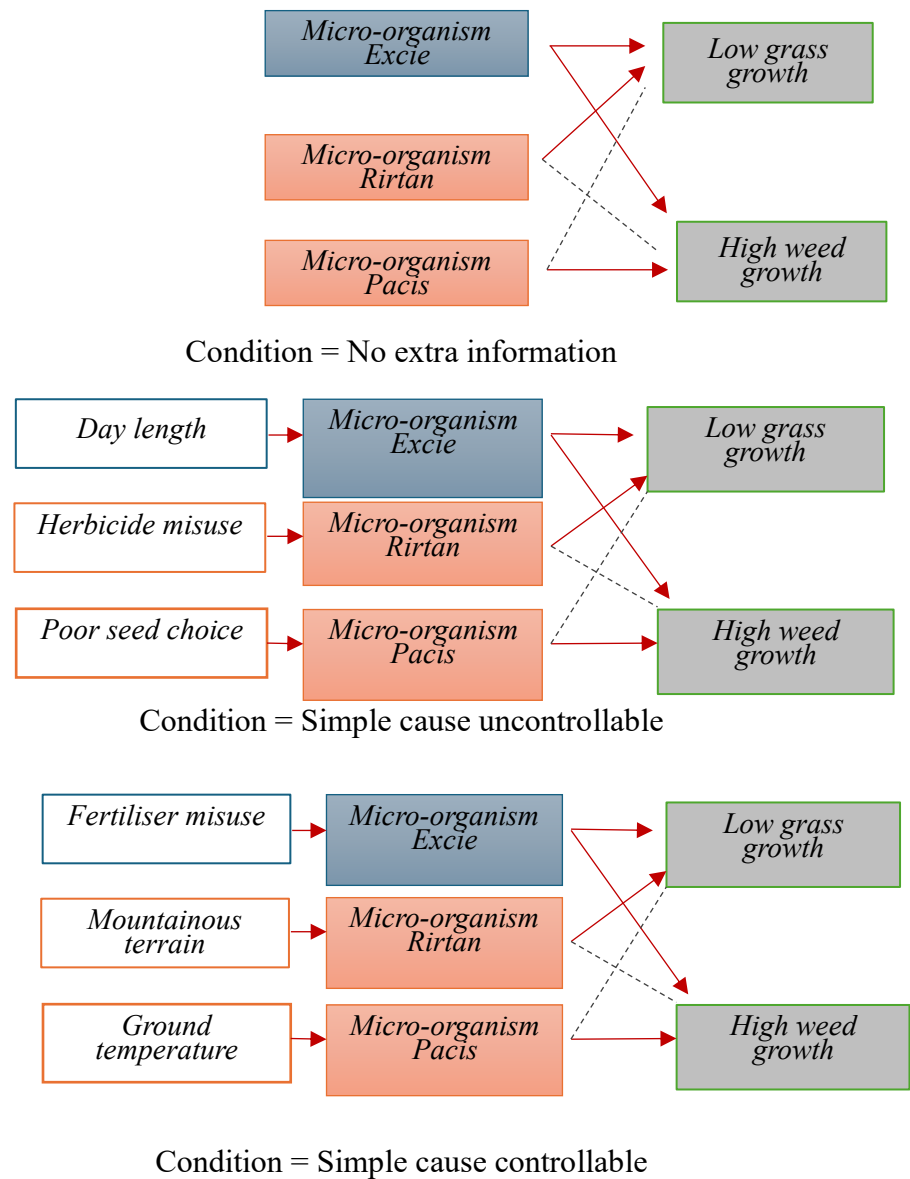
What do you think is the most satisfying explanation for the problems that the alien farm X is exhibiting?

- (A) Alien farm X has Micro-organism Excie in the soil.
- (B) Alien farm X has Micro-organism Rirtan in the soil.
- (C) Alien farm X has Micro-organism Pacis in the soil.
- (D) Alien farm X has Micro-organism Excie and Micro-organism Rirtan in the soil.
- (E) Alien farm X has Micro-organism Excie and Micro-organism Pacis in the soil.
- (F) Alien farm X has Micro-organism Rirtan and Micro-organism Pacis in the soil.

See **Figure A5** for a diagram of the causal relations of the farming domain.

Figure A5

Diagrams for the Causal Structure of the Farming Domain in Experiment 3b



Note. Red arrows indicate a cause-effect relation. Dotted lines indicate that the diseases do not lead to the symptoms. These diagrams were for illustration purposes only, participants were not provided with diagrams during the experiment.

Experiment 4

Participants were given the information below for the farming domain:

There is a population of 750 aliens that lives on planet Zorg. You are a farming expert trying to understand an alien's farm problem. The alien farm had two problems: grass growth was low and weed growth was high.

Micro-organism Excie in the soil always causes both low grass growth and high weed growth. Micro-organism Pacis in the soil always causes low grass growth, but it never causes high weed growth. When an alien farm soil has micro-organism Rirtan, that farm will always have high weed growth, but this micro-organism will never cause low grass growth.

Nothing else is known to cause an alien farm to have low grass growth or high weed growth. What do you think is the most satisfying explanation for the problems that the alien farm was exhibiting?

Causal explanation condition:

- A) Because the alien's farm had micro-organism Excie, it had the problems.
- B) Because the alien's farm had micro-organism Pacis, it had the problems.
- C) Because the alien's farm had micro-organism Rirtan, it had the problems.
- D) Because the alien's farm had micro-organism Excie and Pacis, it had the problems.
- E) Because the alien's farm had micro-organism Excie and Rirtan, it had the problems.

- F) Because the alien's farm had micro-organism Pacis and Rirtan, it had the problems.

Counterfactual explanation condition:

- A) If the alien's farm had not had micro-organism Excie, the farm would not have had the problems.
- B) If the alien's farm had not had micro-organism Pacis, the farm would not have had the problems.
- C) If the alien's farm had not had micro-organism Rirtan, the farm would not have had the problems.
- D) If the alien's farm had not had micro-organism Excie and Pacis, the farm would not have had the problems.
- E) If the alien's farm had not had micro-organism Excie and Rirtan, the farm would not have had the problems.
- F) If the alien's farm had not had micro-organism Pacis and Rirtan, the farm would not have had the problems.

Appendix 2.4 - Supplementary Analysis for Chapter 2

Experiment 1

Domain and Order Checks

Four chi-square tests were planned to test the difference of domain, and the effect of order, with a Bonferroni corrected alpha level of $p < .0125$.

We tested the effect of the order in which the domains were presented by testing the responses from participants' first responses to their second responses in a 6X2 chi-square test for each of the conditions, and it was found that the distribution of choices between the first received domain and the second received domain was not different in the no extra information condition, $\chi^2(5) = 4.348, p = .5005$, and in the consequent-controllability information condition, $\chi^2(5) = 4.608, p = .4656$.

We carried out a 6X2 chi-square test on responses between the farming domain and the disease domain for each of the two conditions separately, and it was found that the distribution of choices between the two domains was not different in the no extra information condition, $\chi^2(5) = 9.854, p = .0795$, and in the consequent-controllability information condition, $\chi^2(5) = 5.574, p = .3499$.

Comparison to the Equal Distribution of Choices

To examine the effect of consequent-controllability further, we compared the choices of each of the four explanations to a chance level of 25%, by carrying out four Wilcoxon signed rank tests on participants' mean responses to the combined domains, with a Bonferroni corrected alpha for four tests of $p < .0125$.

In the consequent-controllability information condition, 47% of participants chose the explanation with both consequent-controllable causes, Trichet and Humel, and it was higher than chance, $z = 3.738$, $p = .0002$, $r = 0.40$; 19% of participants chose the explanation with neither causes consequent-controllable, Morad and Sommet, and it was lower than chance, $z = 2.568$, $p = .010$, $r = 0.28$; 10% of participants chose the option with the first-mentioned cause in each pair in the scenario, Trichet and Sommet, and it was lower than chance, $z = 5.424$, $p < .0001$, $r = 0.58$; 8% of participants chose the option with the second-mentioned cause in each pair in the scenario, Morad and Humel and it was lower than chance, $z = 6.250$, $p < .0001$, $r = 0.67$. To note, the percentages do not add to 100% because there are filler explanations. Overall, we saw a strong preference for the explanation with consequent-controllable factors in this experiment, as we predicted. Almost half of all responses chose this explanation. We note again that participants could choose only one answer option.

It was found that when no extra information was given, 19% of participants chose the option of Trichet and Humel, and it was less than by chance, $z = 2.698$, $p = .007$, $r = 0.27$; 9% of participants chose the option of Morad and Sommet) and it was also lower than chance, $z = 6.468$, $p < .0001$, $r = 0.65$; 36% of participants chose the option with the first-mentioned cause in each pair in the scenario, Trichet and Sommet, and it was higher than chance, $z = 2.543$, $p = .011$, $r = 0.26$; 25% of participants chose the option with the second-mentioned cause in each pair in the scenario, Morad and Humel, and it was the same as chance, $z = 0.836$, $p = .493$. Overall, in this condition, we observed that people chose the first causes in each pair that were mentioned in

the scenario. Note the order of explanation options was randomised, and the order in which the causes appeared in the scenarios was not.

Experiment 2a

Domain and Order Checks

The same tests were carried out to test domain and order effects. We tested the effect of the order in which the domains were presented, it was found that the distribution of choices was not different in the no extra information condition, $\chi^2(5) = 6.986, p = .2217$, and in the antecedent-controllability condition, $\chi^2(5) = 2.858, p = .7219$.

The 6X2 chi-square tested responses between the farming domain and the disease domain for both information conditions, and it was found that the distribution of choices between the two domains was different in the no extra information condition, $\chi^2(5) = 15.394, p = .0088, w = 0.28$, and no difference between domains was found in the antecedent-controllability information condition, $\chi^2(5) = 1.386, p = .9258$. However, the differences were not systematic across experiments the relationship between the experimental conditions and responses was significant for both the disease domain, $\chi^2(5) = 14.89, p = .01, w = 0.28$, and the farming domain, $\chi^2(5) = 31.47, p < .001, w = 0.41$.

Comparison to the Equal Distribution of Choices

In the antecedent-controllability information condition, the four Wilcoxon signed rank tests with a Bonferroni corrected alpha level of $p < .0125$ showed that 31% of participants chose the antecedent-controllable explanation, Trichet and Humel, and it was the same as chance, $z = 0.892$, $p = .372$; 29% of participants chose the antecedent-uncontrollable explanation, Morad and Sommet, and it was the same as chance, $z = 0.398$, $p = .691$; 15% of participants chose the first mentioned option, Trichet and Sommet, and it was lower than chance, $z = 4.036$, $p < .0001$, $r = 0.44$; 13% of participants chose the second mentioned, Morad and Humel, and it was lower than chance, $z = 4.529$, $p < .0001$, $r = 0.49$. Overall, we did not observe a preference for the antecedent-controllable explanation that was stronger than chance. Nonetheless, the provision of antecedent-information reduced the preference for choices in the first mentioned causes option.

In the no extra information condition, we tested whether participants would choose one of the four causally eligible explanations by chance, i.e. 25%. Four Wilcoxon signed rank tests compared the choices of these four explanations, each to a chance of 25%, with a Bonferroni corrected alpha level of $p = .0125$. It was found that when no extra information was given, 18% of participants chose Trichet and Humel, and it was lower than choices by chance, $z = 3.196$, $p = .001$, $r = 0.32$; 14% of participants chose Morad and Sommet, and it was also lower than chance, $z = 4.347$, $p < .0001$, $r = 0.43$.; 38% of participants chose Trichet and Sommet, and it was higher than chance, $z = 2.726$, $p = .006$, $r = 0.27$; 18% of participants chose Morad and Humel, and it

was lower than chance, $z = 2.994$, $p = .003$, $r = 0.30$. Overall, we observed a tendency to choose the explanation that contained the first-mentioned causes.

Experiment 2b

Domain and Order Checks

The same tests were carried out to test domain and order effects. We tested the effect of the order in which the domains were presented, and it was found that the distribution of choices between the first domain and the second domain was not different in the no extra information condition, $\chi^2(5) = 6.196$, $p = .2876$, and in the antecedent-controllability condition, $\chi^2(5) = 12.706$, $p = .0263$ on the corrected alpha-level.

The 6X2 chi-square tested the responses of the farming domain and the disease domain for both information conditions, and it was found that the distribution of choices was not different in the no extra information condition, $\chi^2(5) = 8.854$, $p = .1150$, and in the antecedent-controllability information condition, $\chi^2(5) = 3.16$, $p = .6753$.

Comparison to the Equal Distribution of Choices

In the antecedent-controllability information condition, the four Wilcoxon signed rank tests with a Bonferroni corrected alpha level of $p < .0125$ showed that when controllability information was available, 25% of participants chose the antecedent-controllable explanation, Trichet and Humel, and it was the same as chance, $z = 0.686$, $p = .493$; 23% of participants chose the antecedent-uncontrollable explanation, Morad and Sommet, and it was the same as chance, $z = 1.194$, $p = .232$; 13% of participants chose the primacy

effect option, Trichet and Sommet, and it was significantly lower than chance, $z = 4.374, p < .0001, r = 0.47$; 21% of participants chose the recency option, Morad and Humel, and it was same as chance, $z = 1.96, p = .050$. Overall, the primacy effect was no longer observed. The result replicates the finding of Experiment 2a of no above-chance preference for the antecedent-controllable explanations in this experiment.

In the no extra information condition, it was found that when no extra information was given, 19% of participants chose the antecedent-controllable explanation (Trichet and Humel), and it was the same as choices by chance, $z = 2.379, p = .017$; 10% of participants chose the antecedent-uncontrollable explanation (Morad and Sommet), and it was significantly lower than chance, $z = 5.771, p < .0001, r = 0.62$; 42% of participants chose the primacy effect option (Trichet and Sommet), and it was significantly higher than chance, $z = 3.597, p = .0003, r = 0.38$; 13% of participants chose the recency effect option (Morad and Humel), and it was significantly lower than chance, $z = 4.534, p < .0001, r = 0.48$. Overall, we observed that people showed a tendency to choose the causes that were mentioned first in the materials. The result replicates the finding in Experiment 2a. The lower-than-chance choices of some of the options could be due to a higher percentage of choices in the first-mentioned causes option.

Experiment 3a

Domain and Order Checks

Six chi-square tests were planned to test the difference of the order of receipt on people's choice distribution and domain, with a Bonferroni corrected

alpha level of $p < .0083$. It was found that the distribution of choices was not different in the no extra information condition, $\chi^2(5) = 2.874, p = .7194$, in the simple cause uncontrollable condition, $\chi^2(5) = 1.198, p = .9451$, and in the simple cause controllable condition, $\chi^2(5) = 3.854, p = .5706$.

It was found that the distribution of choices between the disease domain and the farming domain was not different in the no extra information condition, $\chi^2(5) = 6.652, p = .2478$, no difference in the simple cause uncontrollable condition, $\chi^2(5) = 2.27, p = .8107$, and no difference in the simple cause controllable condition, $\chi^2(5) = 2.952, p = .7074$.

Analysis of the Complex Explanation

We used 1 to label responses of choice F (the complex explanation), and 0 to label all responses of choices that are not F (choices A-E). In the no extra information condition, 1% of participants chose the complex explanation, in the simple cause uncontrollable condition, 9% of participants chose the complex explanation, and in the simple cause controllable condition, 6% of participants chose the complex explanation. A Kruskal-Wallis test indicated that there was no significant difference between the three experimental conditions, $H(2) = 6.936, p = .031$, on the corrected alpha of $p < .025$. In other words, participants also chose the complex explanation as infrequently as in all experimental conditions.

Experiment 3b

Domain and Order Checks

We carried out the same tests as in Experiment 3a, with a Bonferroni corrected alpha level of $p < .0083$.

We tested the effect of the order in which the domains were presented. It was found that the distribution of choices was not different in the no extra information condition, $\chi^2(5) = 9.234, p = .1001$, in the simple cause uncontrollable condition, $\chi^2(5) = 8.692, p = .122$, and in the simple cause controllable condition, $\chi^2(5) = 8.286, p = .1412$.

We also carried out chi-square tests on the response distribution between the disease domain and the farming domain for each of the experimental conditions separately. It was found that the distribution of choices was not different in the no extra information condition, $\chi^2(5) = 3.968, p = .554$, no difference in the simple cause uncontrollable condition, $\chi^2(5) = 11.492, p = .0425$, and no difference in the simple cause controllable condition, $\chi^2(5) = 8.97, p = .1103$.

Analysis of the Complex Explanation

In the no extra information condition, 4% of participants chose the complex explanation, in the simple cause uncontrollable condition, 6% of participants chose the complex explanation, and in the simple cause controllable condition, 5% of participants chose the complex explanation. A Kruskal-Wallis test indicated that there was no significant difference between the three experimental conditions, $H(2) = 0.412, p = .814$. In other words,

participants also chose the complex explanation as infrequently in all experimental conditions.

Experiment 4

Domain and Order Checks

Four chi-square tests were planned to test the difference of domain, and the effect of the order of receipt on people's choice distribution, with a Bonferroni corrected alpha level of $p < .0125$. For order, it was found that the distribution of choices was not different in the causal explanation condition, $\chi^2(5) = 2.332, p = .8016$, and in the counterfactual explanation condition, $\chi^2(5) = 7.038, p = .2178$.

For domains, it was found that the distribution of choices was not different between the two domains in the causal explanation condition, $\chi^2(5) = 2.018, p = .847$, and no difference in the counterfactual explanation condition, $\chi^2(5) = 4.388, p = .495$.

Analysis of the Complex Explanation

In the causal explanation condition, 3% of participants chose the complex explanation, and in the counterfactual explanation condition, 10% of participants chose the complex explanation. A Mann-Whitney U test indicated that there was a significant difference as people chose the complex explanation more when the explanations were stated as counterfactual explanations, $U = 4751.5, z = 2.381, p = .017, r = 0.175$.

Appendix 2.5 – Materials Check for Experiment 2b

Since six factors (3 uncontrollable factors and 3 controllable factors) are needed for the actual experimental materials, we listed 12 factors for each domain and listed 24 items in total for participants to rate (See **Table A1**). To avoid repetition of the names and confusion for participants in this manipulation check, we used the letters A to L to name all the diseases (in the disease domain) and micro-organisms (in the farming domain). We stated the factors as a causal statement that were to be used in the later experiment, e.g., “Disease A is the result of *unhealthy diet*”, “Disease B is the result of *old age*”, and the participants were asked to rate the controllability of each factor item by answering the question “Is *unhealthy diet* a controllable or uncontrollable cause of the disease” on a 7-point scale anchored at highly uncontrollable (1) and highly controllable (7). Each participant received all 24 items to rate, some participants rated items in the disease domain first, and the other ones rated items in the farming domain first. The order of items was randomized for each participant.

The participants were 10 members of the general public recruited from the platform Prolific. There were 7 women and 3 men, aged between 29 and 67 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. The experiment took approximately 5 minutes to complete, and each participant was paid £0.75 UK sterling for their participation. Participants provided informed consent by ticking a box, they were informed that their participation was anonymous and that they could withdraw their data at any time.

Participants accessed the experiment online and the materials were presented via Alchemer to which participants were directed by a link from Prolific.

The average rating (see **Table A1**) of all 24 items was computed across 10 participants and we selected the highest three rated items to be used as controllable and the lowest three rated items to be used as uncontrollable causes in the experimental materials. In the disease domain, the following items were rated as most controllable: *Smoking*, *Physical inactivity*, and *Unhealthy diet*; and the following items were rated as most uncontrollable: *Gender*, *Ethnicity*, and *Genetic predisposition*. In the farming domain, the following items were rated as the most controllable causes of micro-organisms in the soil: *Fertiliser misuse*, *Herbicide misuse*, and *Poor seed choice*; and the following items were rated as most uncontrollable: *Day-length*, *Mountainous terrain*, and *Ground temperature*.

Table A1.*Average Ratings of Controllability of All Test Items*

Domain = Disease		Domain = Farming	
Item	Average rating	Item	Average rating
Smoking	6.8 +	Fertiliser misuse	6.9 +
Physical inactivity	6.5 +	Herbicide misuse	6.6 +
Unhealthy diet	6.2 + *	Poor seed choice	6.5 +
Excessive alcohol consumption	6.2 *	Herd over-grazing	6.3
Irregular sleep habits	5.4	Bad management of white clover growth	5.8
Stress	5.1	Inadequate watering	5.8
High blood pressure in pregnancy	4.8	Soil erosion	4.5
Economic class at birth	2.9	Flooding	3.5
Old age	1.8	Sandy soil	3.3
Genetic predisposition	1.7 -	Ground temperature	2.4 -
Ethnicity	1.1 -	Mountainous terrain	1.7 -
Gender	1.0 -	Day-length	1.0 -

Note. All test items were shown to the participants in a randomised order

during the material check and ordered here in a descending order based on their rating scores. Three items with the highest controllability ratings in each domain were marked with +, and the three items with the lowest controllability ratings were marked with -. Unhealthy diet and Excessive alcohol consumption (*) had the same rating, we chose to include an unhealthy diet for the final material to match the length of other causes.

Following the materials check, we constructed materials using the most controllable and most uncontrollable antecedents selected.

Appendix 3.1 – Information and Debrief Sheets for Chapter 3

Information Screen for Experiments 5 and 6

Thank you for agreeing to participate in this study.

Participation in the study will take up to X minutes and you will be paid a sum of X in recognition of your participation.

Please take part in this study only if you are willing to consider the tasks seriously and take some time to work out your answer. Please also only take part in the study in a quiet place where you will be uninterrupted for its duration.

The study can ONLY be completed on a desktop or laptop computer - the information will NOT display correctly on a smartphone or similar device. Please do NOT attempt to take part in the study on a smartphone, as you will not be able to complete the study.

Please ensure your browser window is set to full-screen while completing the study.

We want to study how people think about decisions and different types of explanation.

You will read information about grass growth and a smart app that helps farmers to make decisions. Your task is to figure out how the app works. On each screen you will see an example of the app's decision. You will be asked to provide your judgement about the decision and the explanation.

Please do not take part in the study if you have previously taken part in any study that has asked you to do a similar task.

Please click the "Next" button for more information about participation.

Debrief

Both Experiments: Thank you very much for your participation! You have now reached the end of the study.

Experiment 5: The aim of this study is to understand how different types of explanations impact people's understanding of an Artificial Intelligence app. Some participants received counterfactual explanations about how things could have turned out differently in the past, e.g., "If micro-organisms had been absent last month, your grass growth would have been low" Some participants received prefactual explanations about how things could turn out differently in the future, e.g., "If micro-organisms were to be absent next month, your grass growth would be low", while some received explanation just as a description of the app's decision. The study examines whether the different sorts of explanations affect people's understanding of the app and their satisfaction with the explanation.

The study also examines whether the explanations have different effects on different sorts of tasks, and so we tested both prediction tasks and diagnosis tasks. We hope that the results will help us to understand how different types of explanation may help people when they use Artificial Intelligence decision support systems.

Experiment 6: The aim of this study is to understand how different types of explanations impact people's understanding of an Artificial Intelligence app. By asking you to choose between counterfactual explanations about how things could have turned out differently in the past, e.g., "If Besloor had been absent last week, your Elos growth would have been low" and prefactual explanations about how things could turn out differently in

the future, e.g., "If Besloor were to be absent next week, your Elos growth would be low", we aim to see whether the different sorts of explanation affect people's understanding of the app and their satisfaction with the explanation.

The study examines whether the explanations have different effects on different sorts of tasks, and so we tested both prediction tasks and diagnosis tasks in a random order. We hope that the results will help us to understand how different types of explanation may help people when they use Artificial Intelligence decision support systems.

Both Experiments: We do not anticipate that anything in the sentences we presented will cause any distress to participants. However, if you find that any of the stories or your participation in the study does cause you some distress, you may wish to contact a support service such as the Samaritans, www.samaritans.org. Their website provides details of their local telephone numbers and addresses, for example, in Ireland, Samaritans Ireland is available by telephone- 01 872 7700; address: 151 Marlborough Street, Dublin 1.

If you have any questions or concerns, please contact:

Xinyue Dai, PhD student, School of Psychology, Trinity College Dublin, University of Dublin, Ireland. Email: daix@tcd.ie, Tel: +353 1 896 4054.

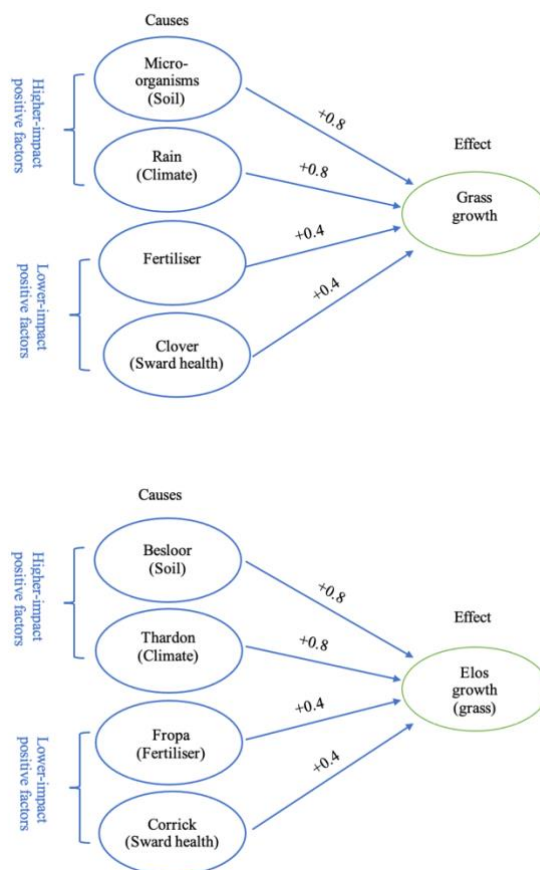
Ruth Byrne, Professor of Cognitive Science, School of Psychology and Institute of Neuroscience, Trinity College Dublin, University of Dublin, Ireland, Email: rmbyrne@tcd.ie, Tel: +353 1 896 4054.

Appendix 3.2 – Pilot Study

This study was carried out before Experiments 5 and 6. We aimed to compare predictions to diagnoses, and the effects of counterfactual explanations to prefactual ones, in an alien domain. We used materials that were the same as described in Experiment 6, with four positive causes (Besloor, Thardon, Fropa, Corrick). We hypothesised that participants would show different levels of accuracy between prediction tasks and diagnosis tasks and that prefactual explanations would be rated as more satisfying due to their emphasis on future planning and ease of cognitive process. Descriptions of these factors and background information were the same as in Experiment 6. See **Figure A6** for the causal structure of the factors and effects.

Figure A6

A Simplified Farming Scenario About Four Causes and an Outcome in Pilot Study



Note. The diagram on top is matched with a realistic farming setting, and the diagram on bottom is adapted from the diagram on top and the names were matched with non-English words with 2 syllables.

Same as Experiment 6, the participants were asked to read about farming activities on planet Kronus and make a judgement about an app that predicts/diagnoses plant growth. In this experiment, task type (prediction vs diagnosis) was manipulated between groups, the participants were either informed that the app predicts the Elos growth based on the four factors or that

the app diagnoses the condition of one factor based on Elos growth and the other three factors. After this, we showed an example of the tasks to be completed and asked the participants to confirm they understood the tasks before proceeding to the experiment.

We gave each participant 24 tasks in a random order (each of the eight tasks was shown three times), asking them to judge what kind of prediction/diagnosis the app would make by choosing the condition of the factor in question (e.g., “What do you think the app will predict about Elos growth on this Delc?” “High/Low”) and rated how confident they were in their judgement using a 5-point scale anchored at strongly disagree (1) and strongly agree (5). After having provided their response to these two questions, the participants saw the app’s actual prediction/diagnosis on display with either a counterfactual or prefactual explanation. Then they were asked to rate their satisfaction with the explanation using a 5-point scale anchored at strongly disagree (1) and strongly agree (5). At the end of the experiment, the participants were asked to complete the DARPA Explanation Satisfaction Scale (Hoffman et al., 2019, see **Table 3.1**).

Methods

Participants

The participants were 133 volunteers (55 men, 76 women, 1 non-binary individual, and 1 preferred not to say) recruited from the online platform Prolific. Their ages ranged from 18 to 61 years, with a mean of 32 years. We implemented two attention checks during the experiment and one memory check task at the end of the experiment (see Appendix 3.2), and we eliminated

participants who failed more than one of these three tasks. As a result, 10 participants were eliminated. A priori power test with G*Power with a medium effect size and 80% power suggested we needed 128 participants. The selection criteria were the same as in Experiments 6 and 7, and we did not control for participants' self-reported occupation in this experiment. Each participant was paid £1.50 UK sterling for participation based on a 10-minute estimated completion time.

Design

A between-groups design was adopted. Both explanation type (counterfactual vs. prefactual) and task type (prediction vs. diagnosis) were manipulated between groups. The participants were assigned to one of the four groups: counterfactual prediction ($n = 34$), counterfactual diagnosis ($n = 31$), prefactual prediction ($n = 34$), and prefactual diagnosis ($n = 34$). The objective measurement was the accuracy of people's judgement of the app's decision, the subjective measurements were how confident they were about their judgements, how satisfied they were with the explanations provided afterwards and the DARPA explanation satisfaction scale at the end of the experiment.

Materials & Procedure

The materials were the same as in Experiment 6, without the Umbrat factor. There were 16 possible combinations of the presence or absence of the four causes, and we selected 8 of them to test (see Appendix 3.2), for which a change to one factor would result in a change to the outcome, four with a positive outcome (high Elos growth) and four with a negative outcome (low

Elos growth). We ensured that there were two cases related to each one of the four causes (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 positive impact on the cause (for those factors set at 0.8) or a low positive impact (for those set at 0.4), they were not told the specific numeric values associated with these terms. The experiment took about 10 minutes to complete.

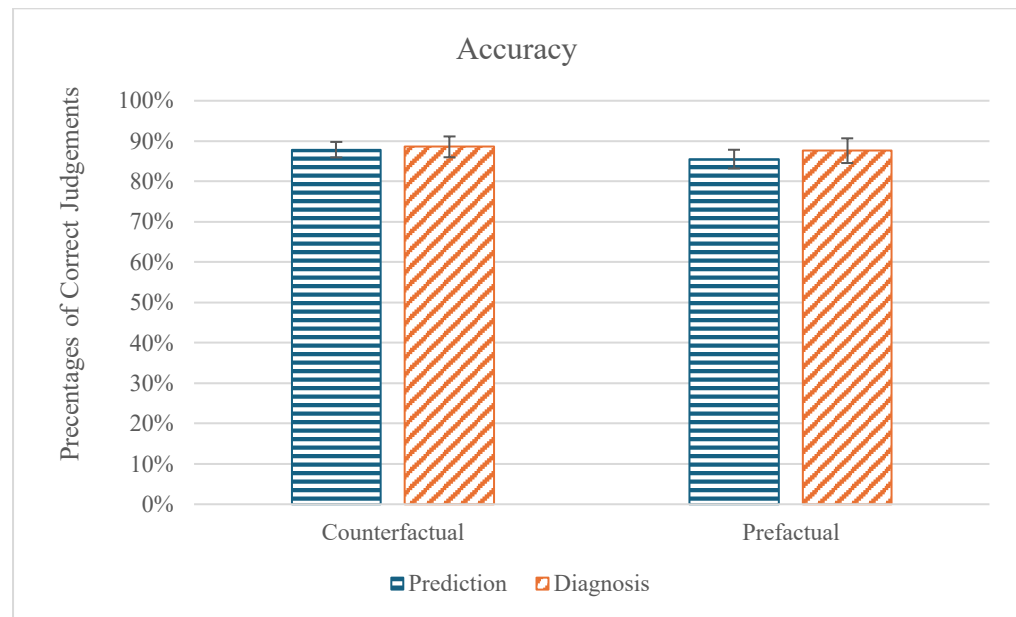
Results and Discussion

Accuracy. Overall, participants were very accurate in their prediction and diagnosis, having an 87% accuracy rate across experimental conditions, see **Figure A7**. Participants made the same number of correct decisions for predictions (88%) and for diagnoses (87%), as shown by the absence of the main effect of task type, $F(1,132) = 0.311, p = .578$, in a 2 (task type: prediction vs diagnosis) X 2 (explanation type: counterfactual vs prefactual) analysis of variance (ANOVA) on accuracy of participants' decisions. People made the same number of correct judgements given counterfactual explanations, 88%, and given prefactual explanation, 87%, as shown by the absence of the main effect of explanations, $F(1,132) = .429, p = .513$. There was no interaction between explanation type and task type, $F(1,132) = 0.075, p = .784$. The results on participants' objective performance were against our hypothesis that prediction tasks would differ from diagnosis tasks. We conjectured that the high accuracy levels reflected in the results showed a ceiling effect that could mask a difference in the performance. To deal with

this, we added a fifth negative low-impact factor in Experiments 5 and 6, with the aim to make the tasks more difficult for participants.

Figure A7

The Percentages of Correct Decisions in Pilot Study



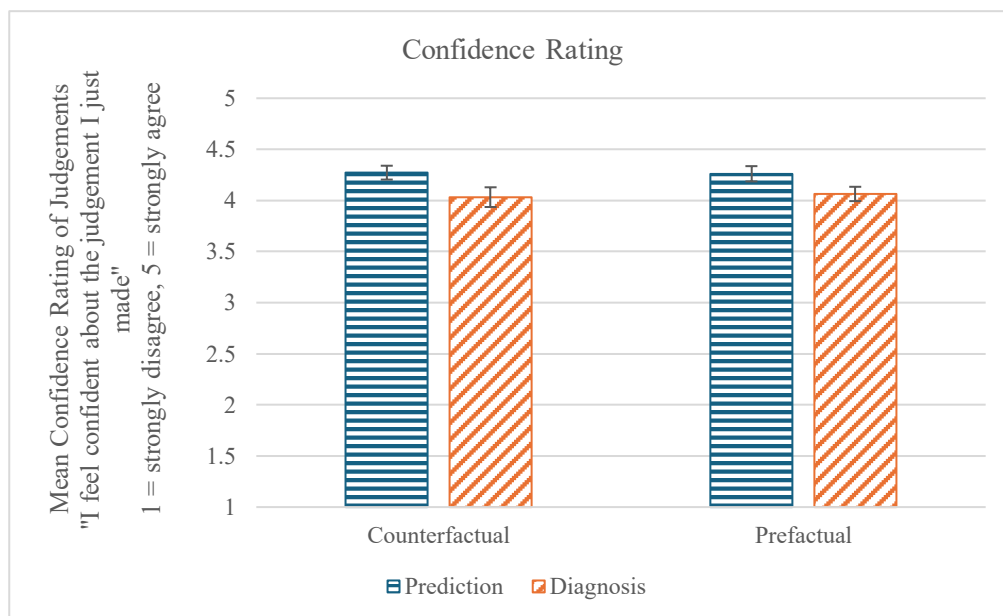
Note. The percentages of correct responses for each of the four experimental conditions in this experiment. Error bars are standard errors of the mean.

Confidence ratings. Participants were overall very confident in their choice ($M = 4.15$, $SD = 0.45$) on the 1-5 Likert scale, see **Figure A8**. A 2 (task type) X 2 (explanation type) ANOVA showed that there was no main effect of explanation type, $F(1,129) = .017$, $p = .895$, a main effect of task, $F(1,129) = 7.989$, $p = .005$, $\eta^2 = 0.058$, and no interaction between explanation type and task, $F(1,129) = 0.076$, $p = .784$. Confidence rating was higher in the prediction group ($M = 4.27$, $SD = .41$) than in the diagnosis group ($M = 4.05$, $SD = .47$). Since we did not find a difference in people's objective accuracy

between the prediction and diagnosis tasks, the results for confidence rating firstly suggested an inconsistency between people's subjective confidence level and their objective performance. There is also a reliability issue with self-rated confidence as we did not replicate this difference in confidence level in Experiments 5 and 6.

Figure A8

The Confidence Ratings in Pilot Study



Note. The mean confidence rating for each of the four experimental conditions in this experiment. Error bars are standard errors of the mean.

Explanation Satisfaction Rating. Participants were overall satisfied with the explanations they received ($M = 3.93$, $SD = 0.61$) on a 1-5 Likert scale. A 2 (task type) X 2 (explanation type) ANOVA showed that there was no main effect of explanation type, $F(1,129) = 0.625$, $p = .431$, no main effect of

task, $F(1,129) = 2.685, p = .104$, and no interaction between explanation type and task, $F(1,129) = 0.216, p = .643$, see **Figure A9**.

Figure A9

The Explanation Satisfaction Ratings in Pilot Study



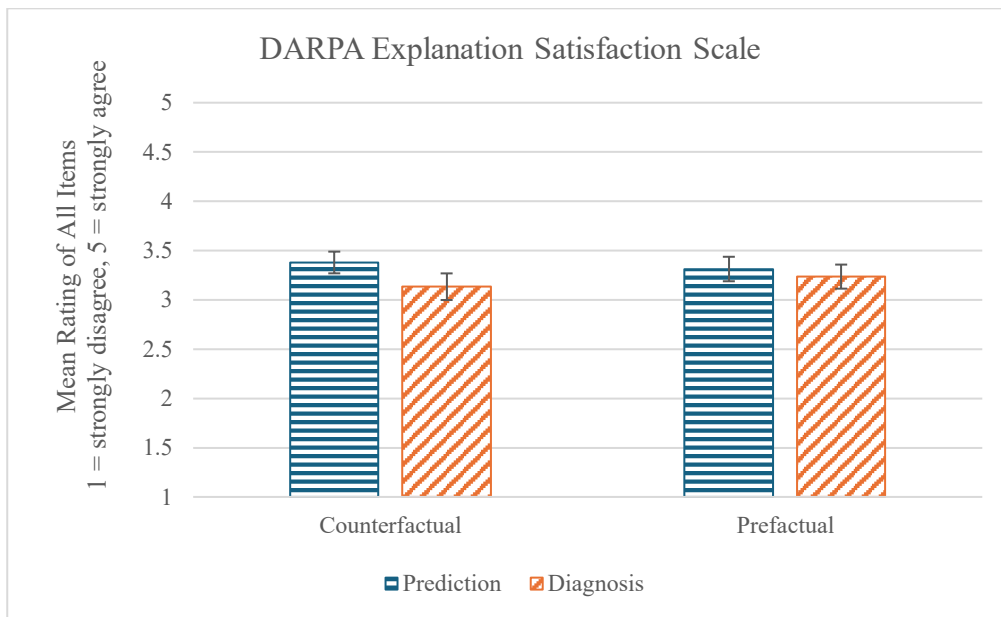
Note. The mean explanation satisfaction rating for each of the four experimental conditions in this experiment. Error bars are standard errors of the mean.

DARPA Explanation Satisfaction Scale. Participants overall scored neutrally on the DARPA scale ($M = 3.27, SD = 0.71$) on a 1-5 Likert scale. A 2 (task type) X 2 (explanation type) ANOVA showed that there was no main effect of explanation type, $F(1,132) = 0.022, p = .884$, no main effect of task, $F(1,132) = 1.728, p = .191$, and no interaction between explanation type and task, $F(1,132) = 0.470, p = .494$, see **Figure A10**. We suggest that due to the

simplicity of tasks, the participants might not feel the need for explanation in this experiment.

Figure A10

The DARPA Scale Ratings in Pilot Study



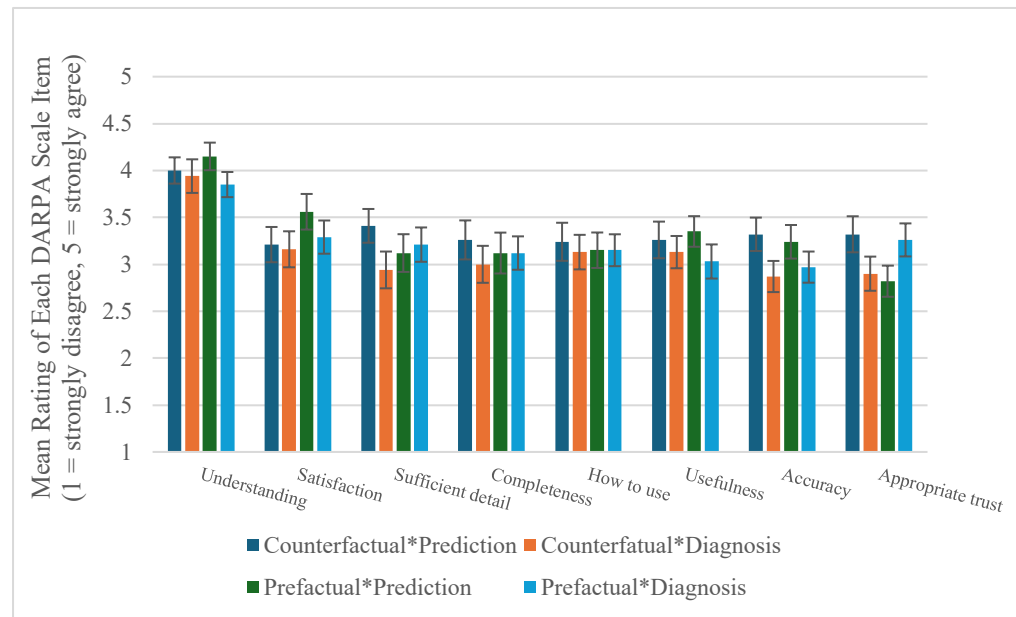
Note. The mean rating of all DARPA scale items for each of the four experimental conditions in this experiment. Error bars are standard errors of the mean.

For each of the DARPA scale items, as can be seen in **Figure A 11**, a 2 (explanation: counterfactual vs prefactual) X 2 (task: prediction vs diagnosis) X 8 (each DARPA scale item) ANOVA showed there was no significant effect of explanation type, $F(1,129) = 0.022$, $p = .884$, no effect of task, $F(1,129) = 1.728$, $p = .191$, and no interaction between explanation type and task, $F(1,129) = 0.470$, $p = .494$. DARPA item had a main effect, $F(7,129) = 17.520$, $p < .001$, $\eta^2 = 0.12$, there was no interaction between DARPA item and

explanation type, $F(7,129) = 0.448, p = .872$, no interaction between DARPA item and task, $F(7,129) = 0.611, p = .747$, and no interaction between the three, $F(7,129) = 1,906, p = .066$.

Figure A11

The DARPA Scale Ratings by Each Item in Pilot Study



Note. The mean rating of each DARPA scale item for each of the four experimental conditions in this experiment. Error bars are standard errors of the mean.

Exploratory Analysis

Since we observed a ceiling effect in accuracy, we ran an exploratory analysis of the different causes, to probe further the effects of different features. We carried out a 2 (task type: prediction vs diagnosis) x 2 (explanation type: Counterfactual vs Prefactual) X 2 (cause type: high-impact vs low-impact) ANOVA on the accuracy of participants' decisions.

For accuracy, there was no main effect of factor, $F(1,132) = 3.813, p = .053$, no interaction between factor and explanation type, $F(1,132) = 0.588, p = .444$, no interaction between factor and task type, $F(1,132) = 0.588, p = .444$. The three did not interact, $F(1,132) = 3.230, p = .075$.

For confidence ratings, there was no main effect of factor, $F(1,132) = 0.687, p = .409$, no interaction between factor and explanation type, $F(1,132) = 0.108, p = .743$. There was a significant interaction between factor and task type, $F(1,132) = 12.988, p < .001, \eta^2 = .091$. The three did not interact, $F(1,132) = 0.002, p = .961$. A decomposition of the interaction between task type and factor with four comparisons and a Bonferroni corrected alpha-level of $p = .0125$ showed that for prediction tasks, participants were more confident about low-impact factors ($M = 4.335, SD = 0.418$) than high-impact factors ($M = 4.199, SD = 0.482$) than, $t(67) = 3.099, p = .003, d = 0.362$; for diagnosis tasks, there was no difference in confidence levels between the factors, $t(64) = 2.011, p = .049$.

For explanation satisfaction rating, there was a main effect of factor, $F(1,132) = 4.124, p = .044, \eta^2 = 0.031$, explanation satisfaction was higher in high-impact factors ($M = 3.950, SD = 0.612$) than in low-impact factors ($M = 3.906, SD = 0.641$), $t(132) = 2.089, p = .039, d = 0.298$. There was no interaction between factor and explanation type, $F(1,132) = 3.845, p = .052$, and no interaction between factor and task type, $F(1,132) = 0.007, p = .932$. The three did not interact, $F(1,132) = 2.797, p = .097$.

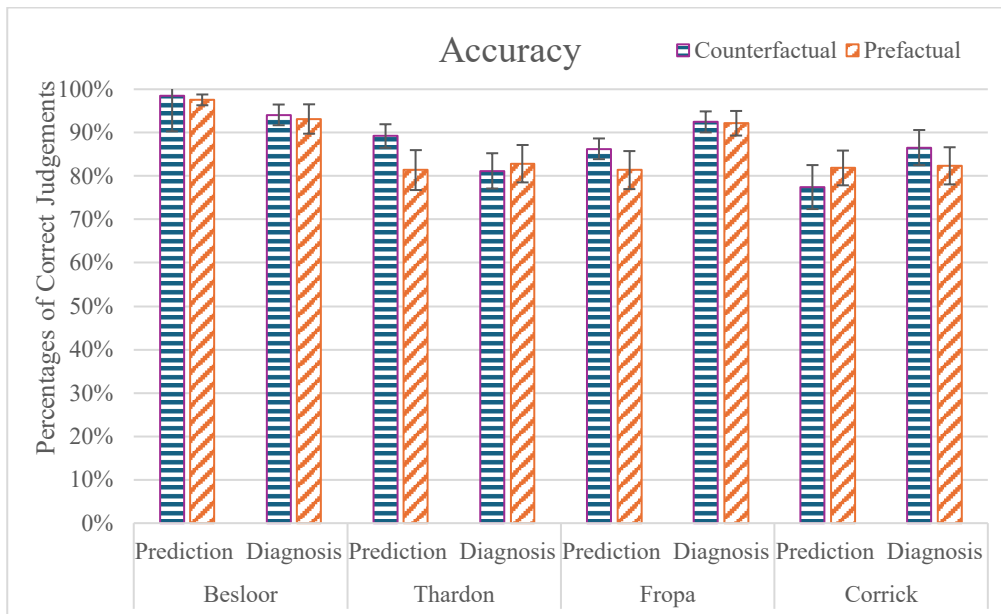
We also carried out a 2 (task type: prediction vs diagnosis) x 2 (explanation type: Counterfactual vs Prefactual) X 4 (causes: Besloor vs

Thardon vs Fropa vs Corrick) ANOVA on accuracy, confidence, and explanation satisfaction ratings.

Accuracy. Participants made more correct decisions for some of the causes than for others, as shown by the main effect of causes, $F(1,129) = 34.079, p < .001, \eta^2 = 0.209$. Explanation type and causes did not interact, $F(1,129) = 0.085, p = .771$. However, task type and causes interacted, $F(1,129) = 9.744, p = .002, \eta^2 = 0.70$. See **Figure A12**. The decomposition of the interaction, by independent samples t-tests with a Bonferroni corrected alpha for 4 tests of $p < .0125$, revealed that participants were more accurate in diagnoses than predictions for the cause Fropa, 92% vs 84%, $t(131) = 2.706, p = .008, d = 0.18$; and participants were equally accurate in diagnoses and predictions for all the other causes, for Corrick, 84% vs 80%, $t(131) = 1.076, p = .284$; Besloor, 94% vs 98%, $t(131) = 2.042, p = .043$, and Thardon, 82% vs 85%, $t(131) = 0.812, p = .418$. The three did not interact, $F(3, 130) = 1.790, p = .160$. The difference in accuracy between task types for the high-impact factor Fropa, however, was not replicated in our follow-up experiments and we suggest that it may have occurred by chance.

Figure A12

The Percentages of Correct Decisions for Exploratory Analysis in Pilot Study



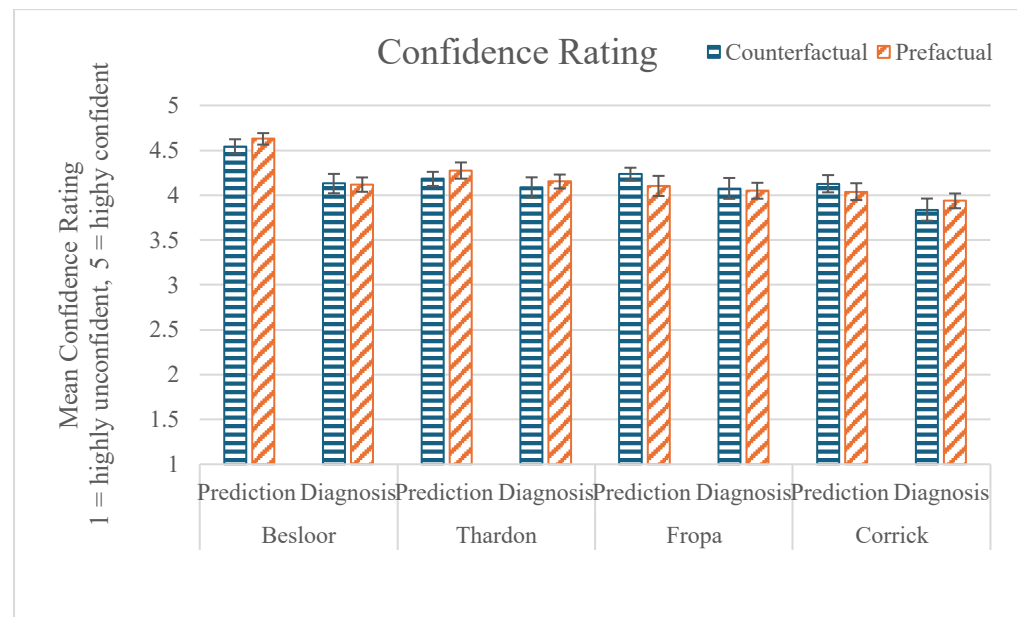
Note. The percentages of correct responses for each of the four experimental conditions in this experiment. Error bars are standard errors of the mean.

Confidence judgements. A 2 (task) X 2 (explanation) X 4 (causes) ANOVA showed that there was a main effect of causes, $F(1,129) = 26.693$, $p < .001$, $\eta^2 = 0.171$. Causes and explanation type did not interact, $F(1,129) = 1.28$, $p = .281$, but there was a significant interaction between causes and task, $F(1,129) = 8.00$, $p < .001$, $\eta^2 = 0.058$. The three did not interact, $F(1,129) = 1.143$, $p = .332$. The decomposition of the interaction between causes and task, by independent samples t-tests with a Bonferroni corrected alpha for 4 tests of $p < .0125$, revealed that participants were more confident in their judgment for prediction than for diagnosis for Besloor, $t(131) = 5.596$, $p < .001$, $d = 0.48$; and they were equally confident in prediction tasks and diagnosis tasks for the other factors, Thardon, $t(131) = 1.199$, $p = .234$; Fropa, $t(131) = 1.103$, p

= .272; and Corrick, $t(131) = 1.962, p = .052$. See **Figure A 13** for the distribution of averaged confidence rating across conditions. The difference in self-reported confidence level between task types for the high-impact factor Besloor, however, was not replicated in our follow-up experiments and we suggest that it may have occurred by chance.

Figure A13

Mean Confidence Ratings for Exploratory Analysis in Pilot Study



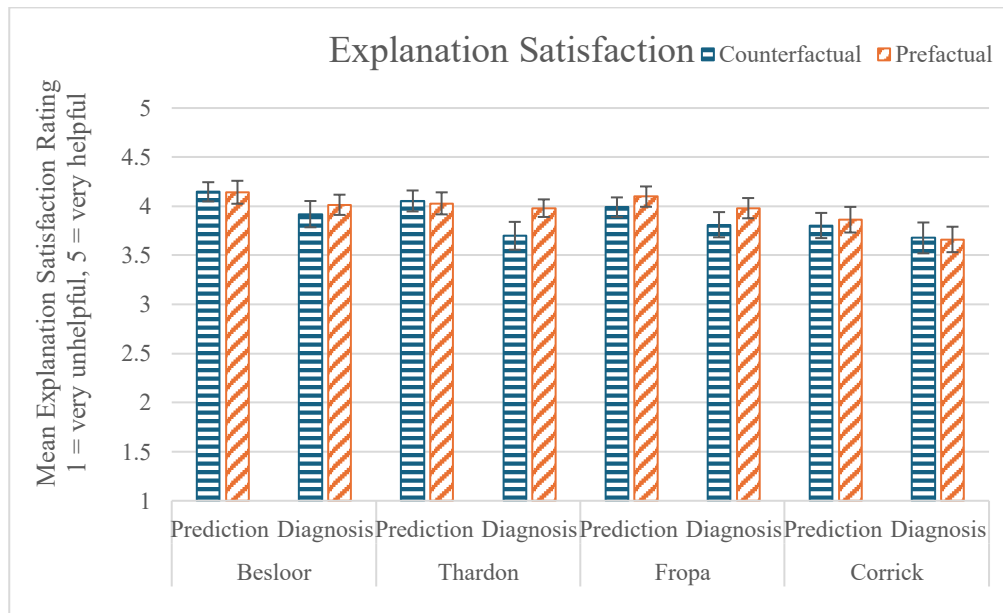
Note. The mean confidence rating for each of the four experimental conditions in this experiment. Error bars are standard errors of the mean.

Explanation Satisfaction Rating. A 2 (task) X 2 (explanation) X 4 (causes) ANOVA showed that there was a main effect of causes, $F(1,129) = 26.632, p < .001, \eta^2 = 0.171$. Causes and explanation type did not interact, $F(1,129) = 0.032, p = .858$, and there was no interaction between causes and task, $F(1,129) = 0.078, p = .781$. The three did not interact, $F(1,129) = 1.262, p$

= .263. **Figure A14** displays the distribution of explanation satisfaction ratings. Pair-wise comparisons with a corrected alpha-level of $p = .0083$ indicated that there was a difference in people's satisfaction rating between Besloor and Thardon, $t(132) = 2.923$, $p = .004$, $d = 0.445$, between Besloor and Corrick, $t(132) = 5.446$, $p < .001$, $d = 0.648$, between Thardon and Corrick, $t(132) = 4.379$, $p < .001$, $d = 0.508$, and between Fropa and Corrick, $t(132) = 4.813$, $p < .001$, $d = .528$.

Figure A14

Mean Explanation Satisfaction Ratings for Exploratory Analysis in Pilot Study



Note. The mean explanation satisfaction rating for each of the four experimental conditions in this experiment. Participants were given four causes in an alien planet domain. Error bars are standard errors of the mean.

In summary, we aimed to test whether the explanation types and task types examined here influence people's performance (judgement accuracy),

self-reported confidence in judgement and self-reported satisfaction level of explanations provided. The results obtained, however, showed a strong ceiling effect and we found that people's confidence levels were significantly higher in the prediction groups than in the diagnosis group even though their accuracy did not vary. We then carried out Experiments 5 and 6, where we firstly adjusted the causal structure to be a five-factor one with the aim to increase task difficulty, secondly modified the way to test people's judgements on explanation, and thirdly tested with realistic farming scenarios as well as an alien one.

Appendix 3.3 – Experimental Materials for Chapter 3

The instruction materials of Experiments 5 and 6 were the same, in the appendix below we provide an example of Experiment 5, and in Experiment 6 the corresponding descriptions were matched to be on the alien planet.

Background information screen (following consent and voluntary demographic information collection)

Please do not press the back arrow on your browser or you will lose access to the experiment.

On each of the following screens, you will see a description about grass growth. Please read these descriptions carefully, because they give important information about grass growth. Your task will be to make judgments about grass growth under different conditions.

Please consider the following five factors that have an impact on grass growth. For the purpose 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.

Fertiliser is an element that can be added to soil. The growth of grass increases if fertiliser is present. Fertiliser 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 categorises the five factors that influence grass growth for each individual farm into either “Present” or “Absent”.

On the next three pages, you will be asked three questions about grass growth and the five factors mentioned above to test your understanding of the information provided, before you start the experiment yourself.

Please click "Next" if you have fully understood the five factors described above.

First comprehension check screen

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. They were not later eliminated if they have failed these questions.

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

(Once you have chosen two options, a feedback will reveal. If you fail to choose the correct ones, please feel free to keep trying selecting different ones until you get them correct, and then move on to the next question.) *

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

If correct response:

Good job!

The correct answers are:

1. Micro-organisms, rain, fertiliser 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.

2. Cows have a negative impact on grass growth, which means that the presence of cows will lead to a decrease in grass growth.

If incorrect response:

Please try again!

Second comprehension check screen

Which **three** of the following factors have a **relatively low** impact on grass growth?
(Once you have chosen three options, a feedback will reveal. If you fail to choose the correct ones, please feel free to keep trying selecting different ones until you get them correct, and then move on to the next question.) *

- ☐ Cows
- ☐ Clover
- ☐ Rain
- ☐ Micro-organisms
- ☐ Fertiliser

If correct response:

Good job!

The correct answers are:

Clover, Fertiliser

You are ready, now let's start our tasks!

If incorrect response:

Please try again!

App Introduction and task example (one before both prediction and diagnosis block)

Prediction block screen 1:

Let's go! In this block, we will focus on the **predictive function** of the app. Based on the presence or absence of the five factors, the app predicts whether a farm will have high or low grass growth that month. Below is a screenshot of what the app does.

Farm NK - Prediction	
Micro-organisms	Present
Rain	Present
Fertiliser	Absent
Clover	Present
Cows	Present
Grass growth	?

The app's prediction is: High.

You have been invited as a test user for this smart app. You will see information about conditions on different farms in the same form of a table.

On the next page you will see an example of the task you need to complete.

Prediction block screen 2:

You are asked to decide what kind of prediction the app will make and to say how confident you are in your judgment. On each screen you will see a description of a different farm in a given month. Each screen will have a reminder of the five factors as follows:

REMINDER:

- Micro-organisms live in the soil that grass is planted in, they have a relatively high positive impact on grass growth.
- Rain is produced in various climates, it has a relatively high positive impact on grass growth.
- Fertiliser is an element that can be added to soil, it has a relatively low positive impact on grass growth.
- Clover is often found growing with grass, it has a relatively low positive impact on grass growth.
- Cows feed on grass, they have a relatively low negative impact on grass growth.

In this example of a prediction task, micro-organisms, rain, clover and cows are present, while fertiliser is absent. Your task is to judge whether the prediction of the app will be that the grass growth will be "High" or "Low". Please note, these judgement questions will disappear after you have chosen your answer, so please choose carefully.

Farm NK - Prediction	
Micro-organisms	Present
Rain	Present
Fertiliser	Absent
Clover	Present
Cows	Present
Grass growth	?

What prediction will the app make about grass growth on this farm? *

☐ High
 ☐ Low

How confident are you in your judgement of the app's prediction? *

☐ Highly unconfident
 ☐ Unconfident
 ☐ Neutral
 ☐ Confident
 ☐ Highly confident

(These questions are for illustration only here, you are not required to answer them)

If allocated to counterfactual explanation group:

The app's prediction is: High.

Explanation: If micro-organisms had been absent last month, your grass growth would have been low.

Please rate how helpful the explanation above is in helping you to understand how the app works.

☐ Very unhelpful
 ☐ Unhelpful
 ☐ Neutral
 ☐ Helpful
 ☐ Very helpful

(This question is for illustration only here, you are not required to answer it)

If allocated to prefactual explanation group:

The app's prediction is: High.

Explanation: If micro-organisms were to be absent next month, your grass growth would be low.

Please rate how helpful the explanation above is in helping you to understand how the app works.

Very unhelpful
Unhelpful
Neutral
Helpful
Very helpful

☐
☐
☐
☐
☐

(This question is for illustration only here, you are not required to answer it)

If allocated to control description group:

Explanation: The app's prediction is : High.

Please rate how helpful the explanation above is in helping you to understand how the app works.

Very unhelpful
Unhelpful
Neutral
Helpful
Very helpful

☐
☐
☐
☐
☐

(This question is for illustration only here, you are not required to answer it)

Note: For Experiment 6, participants were asked to choose between a counterfactual and a prefactual explanation.

For all groups:

Let's start! You will be given **10** prediction tasks on the next screens. Please read the information in the table on each screen carefully before you make your judgements. The information in the tables may sometimes seem similar to information in other tables you have seen, but there may be some differences so please pay close attention to the tables. Please note, there will be some attention check tasks intermingled among the regular prediction tasks throughout the experiment to make sure you are paying attention, so please make sure you get them correct.

Diagnosis block screen 1:

Let's go! In this block, we will focus on the **diagnostic function** of the app. Based on the condition of grass growth and the presence or absence of four of the factors, the app diagnoses whether a fifth factor is present or absent that month.
Below is a screenshot of what the app does.

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

The app's diagnosis is: **Present**.

You have been invited as a test user for this smart app. You will see information about conditions on different farms in the same form of a table.

On the next page you will see an example of the task you need to complete.

Diagnosis block screen 2:

You are asked to decide what kind of diagnosis the app will make and to say how confident you are in your judgement. On each screen you will see a description of a different farm in a given month. Each screen will have a reminder of the five factors as follows:

REMINDER:

- Micro-organisms live in the soil that grass is planted in, they have a relatively high positive impact on grass growth.
- Rain is produced in various climates, it has a relatively high positive impact on grass growth.
- Fertiliser is an element that can be added to soil, it has a relatively low positive impact on grass growth.
- Clover is often found growing with grass, it has a relatively low positive impact on grass growth.
- Cows feed on grass, they have a relatively low negative impact on grass growth.

In this example of a diagnosis task, rain and cows are present, while Fertiliser and Clover are absent, and grass growth is high. Your task is to judge whether the diagnosis of the app will be that the micro-organisms will be "Present" or "Absent". Please note, these judgement questions will disappear after you have chosen your answer, so please choose carefully.

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

What diagnosis will the app make about micro-organisms on this farm? *

☐ Present ☐ Absent

How confident are you in your judgement of the app's diagnosis? *

☐ Highly unconfident
 ☐ Unconfident
 ☐ Neutral
 ☐ Confident
 ☐ Highly confident

(These questions are for illustration only here, you are not required to answer them)

If allocated to counterfactual explanation group:

The app's diagnosis is: Present.

Explanation: If micro-organisms had been absent last month, your grass growth would have been low.

Please rate how helpful the explanation above is in helping you to understand how the app works.

Very unhelpful ☐
 Unhelpful ☐
 Neutral ☐
 Helpful ☐
 Very helpful ☐

(This question is for illustration only here, you are not required to answer it)

If allocated to prefactual explanation group:

The app's diagnosis is: Present.

Explanation: If micro-organisms were to be absent next month, your grass growth would be low.

Please rate how helpful the explanation above is in helping you to understand how the app works.

Very unhelpful	Unhelpful	Neutral	Helpful	Very helpful
☐	☐	☐	☐	☐

If allocated to control description group:

Explanation: The app's diagnosis is: Present.

Please rate how helpful the explanation above is in helping you to understand how the app works.

Very unhelpful	Unhelpful	Neutral	Helpful	Very helpful
☐	☐	☐	☐	☐

(This question is for illustration only here, you are not required to answer it)

Note: For Experiment 6, participants were asked to choose between a counterfactual and a prefactual explanation.

For all groups:

Let's start! You will be given **10** diagnosis tasks on the next screens. Please read the information in the table on each screen carefully before you make your judgements. The information in the tables may sometimes seem similar to information in other tables you have seen, but there may be some differences so please pay close attention to the tables. Please note, there will be some attention check tasks intermingled among the regular diagnosis tasks throughout the experiment to make sure you are paying attention, so please make sure you get them correct.

Attention checks screens (randomised to show during tasks, one during prediction block, one during diagnosis block)

REMINDER:

- Micro-organisms live in the soil that grass is planted in, they have a relatively high positive impact on grass growth.
- Rain is produced in various climates, it has a relatively high positive impact on grass growth.
- Fertiliser is an element that can be added to soil, it has a relatively low positive impact on grass growth.
- Clover is often found growing with grass, it has a relatively low positive impact on grass growth.
- Cows feed on grass, they have a relatively low negative impact on grass growth.

Farm UU - Prediction

Micro-organisms	Present
Rain	Absent
Fertiliser	Present
Clover	Absent
Cows	Present
Grass growth	?

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

☐ Present

☐ Absent

How confident are you in your judgement of the app's prediction? *

Highly unconfident

☐

Unconfident

☐

Neutral

☐

Confident

☐

Highly confident

☐

REMINDER:

- Micro-organisms live in the soil that grass is planted in, they have a relatively high positive impact on grass growth.
- Rain is produced in various climates, it has a relatively high positive impact on grass growth.
- Fertiliser is an element that can be added to soil, it has a relatively low positive impact on grass growth.
- Clover is often found growing with grass, it has a relatively low positive impact on grass growth.
- Cows feed on grass, they have a relatively low negative impact on grass growth.

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

What is the condition of clover on this farm? *

- ☐ Present ☐ Absent

How confident are you in your judgement of the app's diagnosis? *

Highly unconfident ☐ Unconfident ☐ Neutral ☐ Confident ☐ Highly confident ☐

Experimental tasks*Tasks for Experiment 6 (realistic farming scenarios)*

We provide the 10 cases used in Experiment 6, for prediction and diagnosis.

The tasks were presented in a randomised order. Explanations were also provided below each task, the counterfactual explanations for the counterfactual condition were listed here first, prefactual second.

Case 1:

Farm PW - Prediction	
Micro-organisms	Present
Rain	Present
Fertiliser	Present
Clover	Absent
Cows	Present
Grass growth	?

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

Explanations:

Prediction:

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.

Diagnosis:

If your grass growth had been low last month, micro-organisms would have been absent.

If your grass growth were to be low next month, micro-organisms would be absent.

Case 2:

Farm RH - Prediction	
Micro-organisms	Present
Rain	Absent
Fertiliser	Present
Clover	Present
Cows	Present
Grass growth	?

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

Explanations:

Prediction:

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.

Diagnosis:

If your grass growth had been low last month, fertiliser would have been absent.

If your grass growth were to be low next month, fertiliser would be absent.

Case 3:

Farm UU - Prediction	
Micro-organisms	Present
Rain	Absent
Fertiliser	Present
Clover	Absent
Cows	Present
Grass growth	?

Farm SW - Diagnosis	
Micro-organisms	Present
Rain	Absent
Fertiliser	Present
Clover	Absent
Cows	?
Grass growth	Low

Explanation:

Prediction:

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.

Diagnosis:

If your grass growth had been high last month, cows would have been absent.

If your grass growth were to be high next month, cows would be absent.

Case 4:

Farm EO - Prediction	
Micro-organisms	Absent
Rain	Present
Fertiliser	Present
Clover	Absent
Cows	Present
Grass growth	?

Farm XT - Diagnosis	
Micro-organisms	Absent
Rain	Present
Fertiliser	Present
Clover	?
Cows	Present
Grass growth	Low

Explanation:

Prediction:

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.

Diagnosis:

If your grass growth had been high last month, clover would have been present.

If your grass growth were to be high next month, clover would be present.

Case 5:

Farm JC - Prediction	
Micro-organisms	Absent
Rain	Present
Fertiliser	Absent
Clover	Absent
Cows	Present
Grass growth	?

Farm CE - Diagnosis	
Micro-organisms	?
Rain	Present
Fertiliser	Absent
Clover	Absent
Cows	Present
Grass growth	Low

Explanation:

Prediction:

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.

Diagnosis:

If your grass growth had been high last month, micro-organisms would have been present.

If your grass growth were to be high next month, micro-organisms would be present.

Case 6:

Farm KD - Prediction	
Micro-organisms	Present
Rain	Absent
Fertiliser	Absent
Clover	Present
Cows	Absent
Grass growth	?

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

Explanation:

Prediction:

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.

Diagnosis:

If your grass growth had been low last month, cows would have been present.

If your grass growth were to be low next month, cows would be present.

Case 7:

Farm XM - Prediction	
Micro-organisms	Present
Rain	Absent
Fertiliser	Absent
Clover	Absent
Cows	Absent
Grass growth	?

Farm WJ - Diagnosis	
Micro-organisms	Present
Rain	Absent
Fertiliser	?
Clover	Absent
Cows	Absent
Grass growth	Low

Explanation:

Prediction:

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.

Diagnosis:

If your grass growth had been high last month, fertiliser would have been present.

If your grass growth were to be high next month, fertiliser would be present.

Case 8:

Farm DX - Prediction	
Micro-organisms	Absent
Rain	Present
Fertiliser	Present
Clover	Present
Cows	Absent
Grass growth	?

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

Explanation:

Prediction:

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.

Diagnosis:

If your grass growth had been low last month, rain would have been absent.

If your grass growth were to be low next month, rain would be absent.

Case 9:

Farm TB - Prediction	
Micro-organisms	Absent
Rain	Present
Fertiliser	Absent
Clover	Present
Cows	Absent
Grass growth	?

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

Explanation:

Prediction:

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.

Diagnosis:

If your grass growth had been low last month, clover would have been absent.

If your grass growth were to be low next month, clover would be absent.

Case 10:

Farm SI - Prediction	
Micro-organisms	Absent
Rain	Absent
Fertiliser	Present
Clover	Absent
Cows	Absent
Grass growth	?

Farm UZ - Diagnosis	
Micro-organisms	Absent
Rain	?
Fertiliser	Present
Clover	Absent
Cows	Absent
Grass growth	Low

Explanation:

Prediction:

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.

Diagnosis:

If your grass growth had been high last month, rain would have been present.

If your grass growth were to be high next month, rain would be present.

Memory check screen (following prediction and diagnosis blocks)

You have reached the end of your tasks. Now please choose **ALL FIVE** factors that have an impact on grass growth from the options below: *

- ☐ Fertiliser
- ☐ Rain
- ☐ Sunlight
- ☐ Clover
- ☐ Sand
- ☐ Micro-organisms
- ☐ Cows

(Correct answers are Micro-organisms, Rain, Fertiliser, Clover, Cows; all the options were presented to participants in a randomised order)

Appendix 3.4 – Supplementary analysis for Chapter 3

Experiment 5

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-sample 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.

As an exploratory analysis, we investigated the effect of the three types of factors in our causal structure by grouping them into three categories, high-impact positive causes (Micro-organisms, Rain), low-impact positive causes (Fertiliser, Clover), and low-impact negative causes (Cows). We ran a 2 (task type: prediction vs diagnosis) X 3 (explanation type: counterfactual vs prefactual vs control) X 3 (causal type: high-impact positive, low-impact positive, low-impact negative) ANOVA on accuracy, confidence ratings and explanation helpfulness ratings.

For accuracy, there was a main effect of causal type, $F(1, 192) = 113.293, p < .001, \eta^2 = 0.374$. There was no interaction between causal type and explanation type, $F(2, 191) = 0.576, p = .563$. Causal type and task type interacted, $F(1, 192) = 63.234, p < .001, \eta^2 = 0.250$. The three did not interact, $F(2, 191) = 2.091, p = .126$. A decomposition of the interaction with six paired-samples t-tests with a Bonferroni corrected alpha level of $p = .0083$ showed that there was no difference in accuracy between low-impact positive and low-impact negative causal types in prediction tasks, $t(192) = 0.298, p = .766$; and in diagnosis tasks, $t(192) = 1.114, p = .266$. Accuracy was higher

in high-impact positive than low-impact positive causal types than low-impact positive causal types in prediction tasks, $t(192) = 16.090, p < .001, d = 0.254$; and in diagnosis tasks, $t(192) = 4.097, p < .001, d = 0.233$, Accuracy was higher in high-impact positive factors than low-impact negative factors for prediction tasks, $t(192) = 12.253, p < .001$, but there was no difference in diagnosis tasks, $t(192) = 2.071, p = .040$.

For confidence ratings, there was a main effect of causal type, $F(1,192) = 84.872, p < .001, \eta^2 = 0.309$. There was no interaction between causal type and explanation type, $F(2,191) = 0.203, p = .937$. Causal type and task type interacted, $F(1, 192) = 13.682, p < .001, \eta^2 = 0.067$. The three did not interact, $F(2,191) = 0.204, p = .936$. A decomposition of the interaction with six paired-samples t-tests with a Bonferroni corrected alpha level of $p = .0083$ showed that in prediction tasks, confidence was higher in high-impact positive causes than low-impact positive causes, $t(192) = 7.929, p < .001, d = 0.545$, the same difference was observed in diagnosis tasks, $t(192) = 4.587, p < .001, d = 0.565$. Confidence was higher for high-impact positive causes than for low-impact negative causes for prediction tasks, $t(192) = 8.145, p < .001, d = 0.619$; this difference was not observed in diagnosis tasks, $t(192) = 1.454, p = .148$. Confidence was higher for low-impact negative factors than for low-impact positive factors in diagnosis tasks, $t(192) = 2.989, p = .003, d = 0.578$; this difference was not observed in prediction tasks, $t(192) = 1.358, p = .176$.

For explanation helpfulness ratings, there was a main effect of causal type, $F(1,192) = 20.997, p < .001, \eta^2 = 0.100$. There was no interaction between causal type and explanation type, $F(2,191) = 1.377, p = .241$. Causal type and task type interacted, $F(1, 192) = 8.853, p < .001, \eta^2 = 0.045$. The

three did not interact, $F(2, 191) = 0.237, p = .917$. A decomposition of the interaction with six paired-samples t-tests with a Bonferroni corrected alpha level of $p = .0083$ showed that helpfulness rating was higher in high-impact positive causes than low-impact positive causes in prediction tasks, $t(192) = 5.892, p < .001, d = 0.491$; and in diagnosis tasks, $t(192) = 4.120, p < .001, d = 0.367$. Helpfulness rating was higher for high-impact positive causes than for low-impact negative causes, $t(192) = 4.484, p < .001, d = 0.626$; this difference was not observed in diagnosis tasks, $t(192) = 0.517, p = .606$. Helpfulness rating was higher for low-impact negative causes than for low-impact positive causes, $t(192) = 3.332, p = .001, d = 0.394$, this difference was not observed in prediction tasks, $t(192) = 0.179, p = .858$.

Results from exploratory analysis indicated that firstly the difference of factors was mediated by task types, suggesting that the participants' objective performance was affected by task type, as there were more differences between factors in prediction than diagnosis tasks. We suggested that, as discussed in Chapter 3, this could be closely linked to the task designs and the relative simplicity of diagnosis tasks compared to prediction. Secondly, the subjective difference in confidence ratings and explanation helpfulness ratings for prediction tasks were: high-impact positive > low-impact positive = low-impact negative, and for diagnosis tasks: high-impact positive = low-impact negative > low-impact positive. We could conclude that for prediction tasks, participants' subjective experiences were aligned with the difficulty of the tasks, while for diagnosis tasks when there was a clear indication of the focal cause, participants' subjective experiences were sensitive to the negative one (disabler) as it was not significantly different from the high-impact positive

cause and was significantly higher than the low-impact positive factor. In other words, when the disabler became the focal cause, participants felt more confident and more satisfied with the explanation.

Experiment 6

There were no differences between whether participants completed the prediction trials first ($n = 28$) or the diagnosis trials first ($n = 22$), by independent-sample t -tests for accuracy, $t(49) = 0.34$, $p = .74$, confidence judgments, $t(49) = 0.16$, $p = .88$, and average DARPA scale rating, $t(48) = 0.219$, $p = .83$. Hence, order of task had no effect.

As an exploratory analysis, we investigated the effect of the three types of factors in our causal structure by grouping them into three categories, high-impact positive causes (Besloor, Thardon), low-impact positive causes (Fropa, Corrick), and low-impact negative causes (Umbrat). We ran a 2 (task type: prediction vs diagnosis) $\times$ 3 (cause type: high-impact positive, low-impact positive, low-impact negative) ANOVA on accuracy and confidence ratings.

For accuracy, there was a main effect of causal type, $F(1,49) = 16.948$, $p < .001$, $\eta^2 = 0.257$. Causal type and task type interacted, $F(2,48) = 8.823$, $p < .001$, $\eta^2 = 0.153$. A decomposition of the interaction with six paired-sample t -tests with a Bonferroni corrected alpha level of $p = .0083$ showed that accuracy was higher in high-impact positive causes than low-impact positive causes in prediction tasks, $t(49) = 7.456$, $p < .001$, $d = 0.266$; the same difference was observed in diagnosis tasks, $t(49) = 2.830$, $p = .007$, $d = 0.300$. There was no difference in accuracy between low-impact positive and low-impact negative causes in prediction tasks, $t(49) = 0.612$, $p = .543$; and in diagnosis tasks, t

(49) = 1.531, $p = .132$. Accuracy was higher in high-impact positive factors than low-impact negative factors in prediction tasks, $t(49) = 5.751$, $p < .001$, $d = 0.387$, but there was no difference in diagnosis tasks, $t(49) = 0.622$, $p = .537$. The results were consistent with exploratory results in Experiment 5.

For confidence ratings, there was a main effect of causal type, $F(1,49) = 8.355$, $p < .001$, $\eta^2 = 0.146$. Causal type and task type interacted, $F(1, 49) = 4.939$, $p = .009$, $\eta^2 = 0.092$. A decomposition of the interaction with six paired-samples t -tests with a Bonferroni corrected alpha level of $p = .0083$ showed that for prediction tasks, confidence ratings in high-impact positive factors were higher than low-impact negative factors, $t(49) = 2.933$, $p = .005$, $d = 0.663$. There was no difference between high-impact positive and low-impact negative causes, $t(49) = 2.291$, $p = .026$; and no difference between low-impact positive and low-impact negative causes, $t(49) = 1.467$, $p = .149$. For diagnosis tasks, confidence ratings in high-impact positive causes were higher than low-impact positive causes, $t(49) = 4.204$, $p < .001$, $d = 0.437$. There was no difference between high-impact positive and low-impact negative causes, $t(49) = 0.724$, $p = .472$; and no difference between low-impact positive and low-impact negative causes, $t(49) = 2.581$, $p = .013$.

To summarize, the exploratory analysis indicated that accuracy was generally higher for high-impact causes than low-impact causes and that there were more differences in prediction tasks, suggesting that the objective accuracy levels were mediated by the differences between task types. The results for confidence ratings were consistent with the exploratory analysis of Experiment 5 and again indicated that subjective reporting of confidence levels did not always align with objective performance levels.

Appendix 4.1 – Information and Debrief Sheets for Chapter 4***Experiments 7 to 11: Information***

Thank you for agreeing to participate in this study.

Participation in the study will take about X minutes and you will be paid a sum of X in recognition of your participation.

Please take part in this study only if you are willing to consider the tasks seriously and take some time to work out your answer. Please also only take part in the study in a quiet place where you will be uninterrupted for its duration.

The study examines how you make decisions in different situations.

Please do not take part in the study if you have previously taken part in any study that has asked you to do a similar task.

Please click the "Next" button for more information about participation.

Experiments 7 to 11: Debrief

All experiments: Thank you very much for your participation! You have now reached the end of the study.

Experiment 7: The aim of the study is to examine how people make different when the same decision is framed in different ways. People tend to make risk-averse choices when the decisions are framed as gain (lives saved) and make risk-seeking choices when the decisions are framed as loss (lives lost). We also wanted to examine whether an AI's decision will have an effect on people's decision, in situations when the AI's decision is in line with people's decision, and also in situations when AI's decisions are contradictory to people's decision.

Experiment 8: The aim of the study is to examine how people make different when the same decision is framed in different ways. People tend to make risk-averse choices when the decisions are framed as gain (lives saved) and make risk-seeking choices when the decisions are framed as loss (lives lost). We also wanted to examine whether an expert's decision will have an effect on people's decision, in situations when the expert's decision is in line with people's decision, and also in situations when expert's decisions are contradictory to people's decision.

Experiment 9: The aim of the study is to examine how people make different when the same decision is framed in different ways. People tend to make risk-averse choices when the decisions are framed as gain (lives saved) and make risk-seeking choices when the decisions are framed as loss (lives lost). We also wanted to examine whether an AI's vs human expert's decision will have an effect on people's decision, in situations when the AI's and the human expert's decisions are contradictory to people's decision.

Experiment 10: The aim of the study is to examine how people make different choices when the same decision is framed in different ways. People tend to make risk-averse choices when the decisions are framed as gains (lives saved) and make risk-seeking choices when the decisions are framed as losses (lives lost). We also wanted to examine whether an AI's decision will affect people's decision, in situations when the AI's decisions are contradictory to people's decisions. Last but not least, we wanted to examine whether counterfactual explanations such as "It says that in previous cases the outcome would have been better if Program B had been chosen" will increase the effect of AI decisions.

Experiment 11: The aim of the study is to examine how people make different choices when the same decision is framed in different ways. People tend to make risk-averse choices when the decisions are framed as gains (lives saved) and make risk-seeking choices when the decisions are framed as loss (lives lost). We also wanted to examine whether an human expert's decision will affect people's decision, in situations when the expert's decisions are contradictory to people's decisions. Last but not least, we wanted to examine whether counterfactual explanations such as "It says that in previous cases the outcome would have been better if Program B had been chosen" will increase the effect of expert decisions.

All experiments: We do not anticipate that anything in the questions we presented will cause any distress to participants. However, if you find that any of the tasks or your participation in the study does cause you some distress, you may wish to contact a support service such as the Samaritans, www.samaritans.org. Their website provides details of their local telephone numbers and addresses, for example, in Ireland, Samaritans Ireland is available at: telephone- 01 872 7700; address: 151 Marlborough Street, Dublin 1.

If you have any questions or concerns, please contact:

Xinyue Dai, PhD student, School of Psychology, Trinity College Dublin, University of Dublin, Ireland. Email: daix@tcd.ie, Tel: +353 1 896 4054.

Ruth Byrne, Professor of Cognitive Science, School of Psychology and Institute of Neuroscience, Trinity College Dublin, University of Dublin, Ireland, Email: rmbyrne@tcd.ie, Tel: +353 1 896 4054.

The next page will redirect you back to the Prolific site.

Appendix 4.2 – Experimental Materials for Chapter 4

Disease domain

Imagine that the US is preparing for an outbreak of an unusual disease which is expected to kill 600 people.

Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows: ...

	Gain	Loss
No Recommendation	If program A is adopted, 200 people will be saved. If program B is adopted, there is a 1/3rd probability that 600 people will be saved, and a 2/3rd probability that no one will be saved.	If program A is adopted, 400 people will die. If program B is adopted, there is a 1/3rd probability that no one will die, and a 2/3rd probability that 600 will die.
Typical ¹ Recommendation	If program A is adopted, 200 people will be saved. If program B is adopted, there is a 1/3rd probability that 600 people will be saved, and a 2/3rd probability that no one will be saved. An Artificial Intelligence decision support system ² recommends that you choose program A. ³	If program A is adopted, 400 people will die. If program B is adopted, there is a 1/3rd probability that no one will die, and a 2/3rd probability that 600 will die. An Artificial Intelligence decision support system recommends that you choose program B.
Non-typical Recommendation	If program A is adopted, 200 people will be saved. If program B is adopted, there is a 1/3rd probability that 600 people will be saved, and a 2/3rd probability that no one will be saved. An Artificial Intelligence decision support system recommends that you choose program B.	If program A is adopted, 400 people will die. If program B is adopted, there is a 1/3rd probability that no one will die, and a 2/3rd probability that 600 will die. An Artificial Intelligence decision support system recommends that you choose program A.

Which of the two programs would you choose?

¹ The typical recommendation condition was only tested in Experiment 9 and 10, following experiments only tested using the no recommendation condition and the non-typical recommendation condition

² In Experiment 10, “An Artificial Intelligence decision support system” was replaced by either “A medical expert” or “An agricultural expert” as appropriate.

³ In Experiment 12 participants in the counterfactual explanation conditions received an explanation here that says “An Artificial Intelligence decision support system recommends that you choose program A/B. It says that in previous cases the outcome would have been better if Program A/B had been chosen”. In Experiment 13 participants in the counterfactual explanation condition received an explanation here that says “A medical/farming expert recommends that you choose program A/B. The expert says that in previous cases the outcome would have been better if Program A/B had been chosen”.

Farming domain

Imagine that your farm is preparing for an outbreak of an unusual disease which is expected to kill 60 cows.

Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:

	Gain	Loss
No Recommendation	If program A is adopted, 20 cows will be saved. If program B is adopted, there is a 1/3rd probability that 60 cows will be saved, and a 2/3rd probability that no cows will be saved.	If program A is adopted, 40 cows will die. If program B is adopted, there is a 1/3rd probability that no cows will die, and a 2/3rd probability that 60 cows will die.
Typical Recommendation	If program A is adopted, 20 cows will be saved. If program B is adopted, there is a 1/3rd probability that 60 cows will be saved, and a 2/3rd probability that no cows will be saved. An Artificial Intelligence decision support system recommends that you choose program A.	If program A is adopted, 40 cows will die. If program B is adopted, there is a 1/3rd probability that no cows will die, and a 2/3rd probability that 60 cows will die. An Artificial Intelligence decision support system recommends that you choose program B.
Non-typical Recommendation	If program A is adopted, 20 cows will be saved. If program B is adopted, there is a 1/3rd probability that 60 cows will be saved, and a 2/3rd probability that no cows will be saved. An Artificial Intelligence decision support system recommends that you choose program B.	If program A is adopted, 40 cows will die. If program B is adopted, there is a 1/3rd probability that no cows will die, and a 2/3rd probability that 60 cows will die. An Artificial Intelligence decision support system recommends that you choose program A.

Which of the two programs would you choose?

Appendix 4.3 - Supplementary Analysis for Chapter 4

4.3.1. Domain differences and order effect check

To test whether there were differences between the disease and farming domains, and to test whether there were differences in the order in which participants received the two domains, we carried out non-parametric comparisons.

Experiment 7

A Wilcoxon signed ranks test indicated there were no differences between the responses in the farming domain and the disease domain, $Z = 1.250, p = .211$. A Mann-Whitney U test indicated there was no difference in responses when it comes to the order of the domains (those who received the disease domain first vs those who received the farming domain first), $U = 10927, z = 0.140, p = .888$.

Experiment 8

A Wilcoxon signed ranks test indicated there were no differences between the responses in the farming domain and the disease domain, $Z = 1.254, p = .210$. A Mann-Whitney U test indicated there was no difference in responses when it comes to the order of the domains (those who received the disease domain first vs those who received the farming domain first), $U = 12099, z = 0.077, p = .939$.

Experiment 9

First presented domain only. Since we only analysed the first presented domain responses in the main analysis, we tested there was no order effect by carrying out a Mann-Whitney U test on the difference score between participants who received the disease domain first and those who received the farming domain first, there was no significant difference, $U = 7139$, $z = 0.401$, $p = .688$.

Both domains. A Wilcoxon signed ranks test indicated there were no differences between the responses in the farming domain and the disease domain, $z = 0.928$, $p = .353$. A Mann-Whitney U test indicated there was no difference in responses when it comes to the order of the domains, $U = 3351$, $z = 0.495$, $p = .621$.

Experiment 10

First presented domain only. As in Experiment 9, we carried out a Mann-Whitney U test on the difference score between participants who received the disease domain first and those who received the farming domain first, there was no significant difference, $U = 7540$, $z = 0.206$, $p = .837$.

Both domains. A Wilcoxon signed ranks test indicated there were no differences between the responses in the farming domain and the disease domain, $z = 0.229$, $p = .819$. A Mann-Whitney U test indicated there was no difference in responses when it comes to the order of the domains, $U = 3660$, $z = 0.356$, $p = .722$.

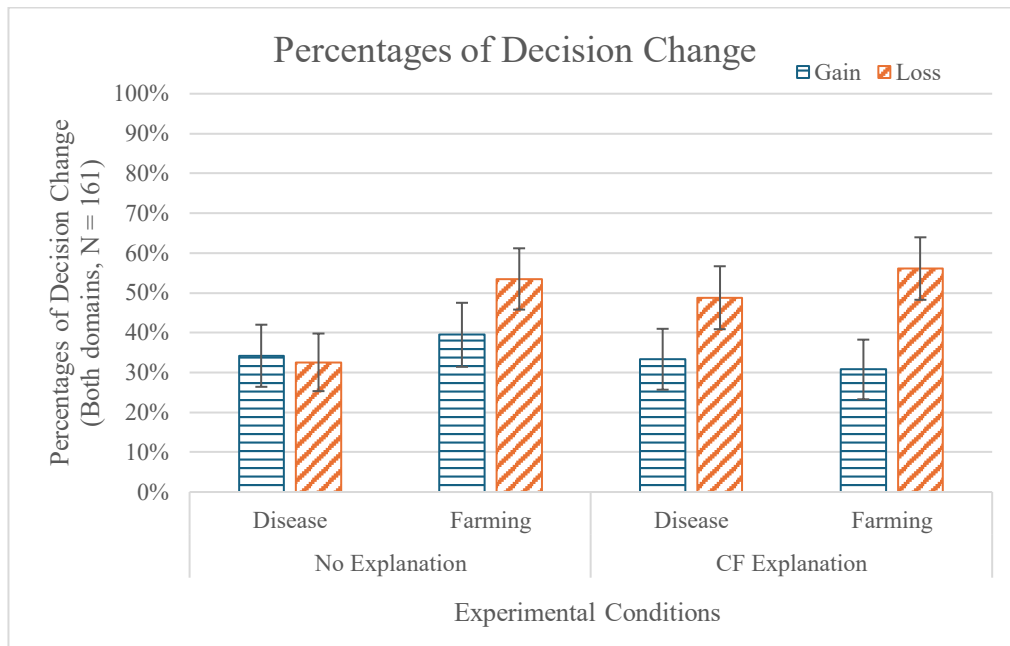
Experiment 11

First presented domain only. As in Experiments 9 and 10, we carried out a Mann-Whitney U test on the difference score between participants who received the disease domain first and those who received the farming domain first, there was no significant difference, $U = 7165$, $z = 0.443$, $p = .658$.

Both domains. A Wilcoxon signed ranks test indicated there was a significant difference between the responses in the farming domain and the disease domain, $z = 2.082$, $p = .037$, $r = 0.16$. Participants switched their decisions more in the farming domain (average switch of 45%) than in the disease domain (average switch of 37%), as shown in **Figure A15**. A Mann-Whitney U test indicated there was no difference in responses when it comes to the order of the domains, $U = 3182$, $z = 0.211$, $p = .833$.

Figure A15

Percentages of Decision Change in Experiment 11 Divided by Domain



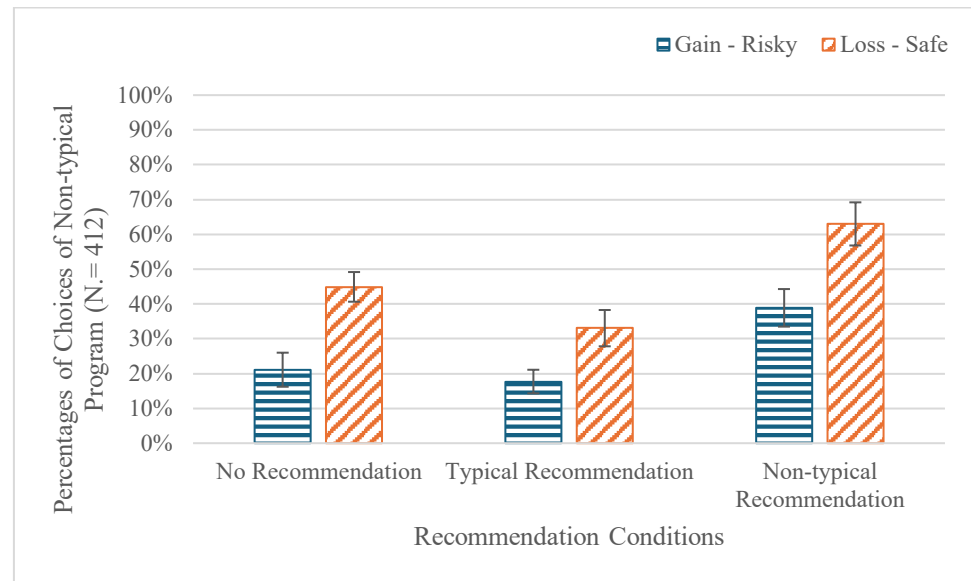
Note. The percentages of choice switch percentages from both domains in Experiment 11, divided by domains. Error bars are standard errors of the mean.

4.3.2 Analysis of all 412 data points for Experiment 8

As described in the report of Experiment 8, due to an administrative error in randomisation, we recruited more participants for two of the groups, and so we eliminated 100 participants at random from those groups for the main analysis. Results from the analysis of all 412 participants were consistent with the results of the selected sample of 312. There were six groups: gain & no recommendation ($n = 45$), loss & no recommendation ($n = 108$), gain & typical recommendation ($n = 96$), loss & typical recommendation ($n = 59$), gain & non-typical recommendation ($n = 54$), and loss & non-typical recommendation ($n = 50$). **Figure A16** shows the distribution of choices for this data set.

Figure A16

Percentages of Choices of Non-typical Programs in Experiment 8 (N = 412)



Note. The percentages of choices of the safe program A and the risky program B when participants (N = 412) were given no recommendation or a typical recommendation, that is, program A in the Gain frame, or program B in the Loss frame, or the non-typical choice, that is, program B in the Gain frame, or program A in the Loss frame in Experiment 8.

When received no recommendation, in the gain frame, 79% of participants made the typical choice of the safe Program A and 21% of participants made the non-typical choice of risky Program B, and in the loss frame, 55% of participants made the typical choice of the risky-Program B and 45% of participants made the non-typical choice of the safe Program A. When they received recommendations that were the same as typical choices, 82% of participants made the typical choice of the safe Program A, and 18% of participants made the non-typical choice of the risky Program B; and in the loss frame, 67% of participants made the typical choice of the risky Program B

and 33% of participants made the non-typical choice of the safe Program A.

When they received recommendations that were opposite typical choices, 61% of participants made the typical choice of the safe Program A, 39% of participants made the non-typical choice of the risky Program B, and in the loss frame, 37% of participants made the typical choice of the risky Program B, 63% of participants made the non-typical choice of the safe Program A.

First to examine domain and order effects for the full set of participants, A Wilcoxon signed ranks test indicated there was no differences between the responses in the farming domain and the disease domain, $Z = 1.251$, $p = .211$. A Mann-Whitney U test indicated there was no difference in responses when it comes to the order of the domains, $U = 20262$, $z = 0.856$, $p = .392$.

A Mann Whitney's U test indicated there was a significant difference between gain and loss frames, $U = 26581$, $z = 4.951$, $p < .001$, $r = 0.24$; for information conditions, Kruskal-Wallis test by ranks was significant, $H(2) = 25.339$, $p < .001$, $\eta^2 = 0.06$. We carried out 4 comparisons, with a Bonferroni-adjusted alpha level of $p < 0.0125$. We compared the choices of non-typical programs (program B in gains and program A in losses) between non-typical recommendation condition and no recommendation condition, as well as between non-typical recommendation condition and typical recommendation condition. We compared the non-typical choices between the non-typical recommendation condition and the no recommendation condition, as well as between the non-typical recommendation condition and the typical recommendation condition. In the gain frame, the risky program B was chosen as often by participants in the non-typical recommendation condition, as in the no recommendation condition, $U = 915$, $z = 2.345$, $p = .019$, and it was chosen

more when compared to the typical recommendation condition, $U = 1827$, $z = 3.500$, $p = .0005$, $r = 0.29$. In the loss condition, the safe program A was chosen as often by participants in the non-typical recommendation condition as in the no recommendation condition, $U = 2142$, $z = 2.240$, $p = .025$, and more often than the typical recommendation condition, $U = 953$, $z = 3.402$, $p < .001$, $r = 0.33$. The results were broadly consistent with the dataset that contained 312 participants in Experiment 8, as the differences in recommendations for both gain frame and loss frame were due to participants choosing the non-typical program more often in the non-typical recommendation condition than in the typical recommendation condition.

4.3.3. Supplementary Analysis of Responses from Both Domains

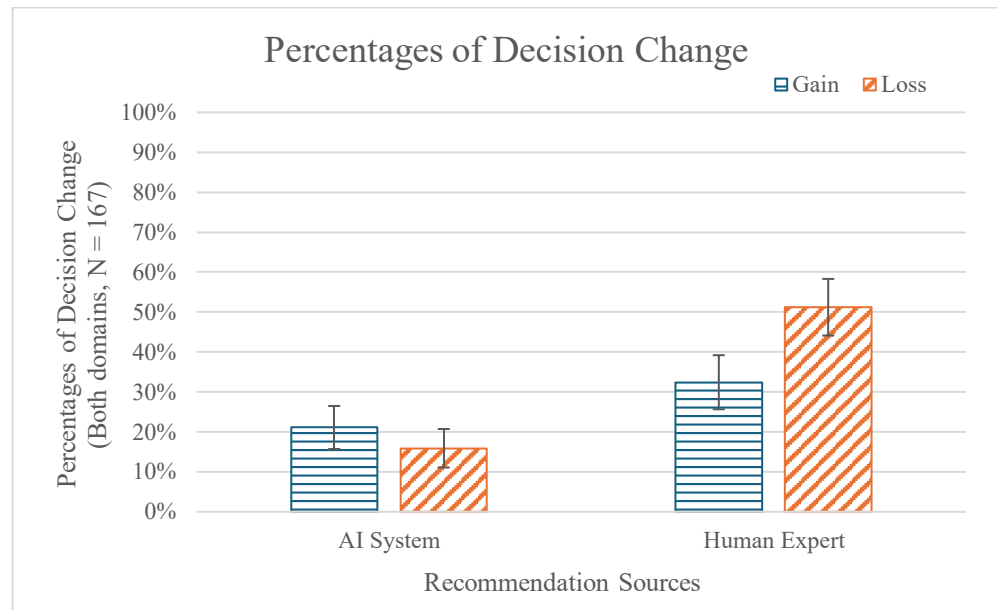
Experiment 9

Figure A17 shows the responses of the 167 participants who made the typical response in the pre-recommendation phase to both scenarios. When they were recommended to make the opposite choice to the typical one that they had made, by an AI system, in the gain frame, 79% of them continued to choose the safe option and 21% changed to the recommended risky one instead; in the loss frame, 84% continued to choose the risky option and 16% changed to the recommended safe one instead. Given a human recommendation, in the gain frame, 68% of participants continued to choose the safe option, and 32% changed to the recommended risky one; in the loss frame, 49% continued to choose the risky option, and 51% changed to the recommended safe one instead. Non-parametric tests indicated that there was a significant effect for source, $U = 4481$, $z = 3.739$, $p < .0001$, $r = 0.289$, and no

effect for framing, $U = 3747$, $z = 0.968$, $p = .333$. Four comparisons with a Bonferroni adjusted p -value of $p < .00125$ showed that for gains, participants switched their choices from the safe option to a risky one as often following a human expert's recommendation and an AI system's recommendations, but for losses, people switched their decisions more following a human expert recommendation, $U = 522$, $z = 3.791$, $p = .0001$, $r = 0.411$. There were no differences between gains and losses for AI system recommendation conditions, $U = 925$, $z = 0.696$, $p = .487$; or human expert recommendation conditions, $U = 588$, $z = 1.849$, $p = .065$. The results for both analyses are broadly consistent and show that people are influenced more by the recommendations of a human expert than by an AI system.

Figure A17

Percentages of Decision Change for Experiment 9 (N = 167)



Note. Percentages of switches from a safe choice to a risky one (for gains) and from a risky choice to a safe one (for losses) following a recommendation from an AI system or a human expert, summed for both domains in Experiment 9. Error bars are standard errors of the mean.

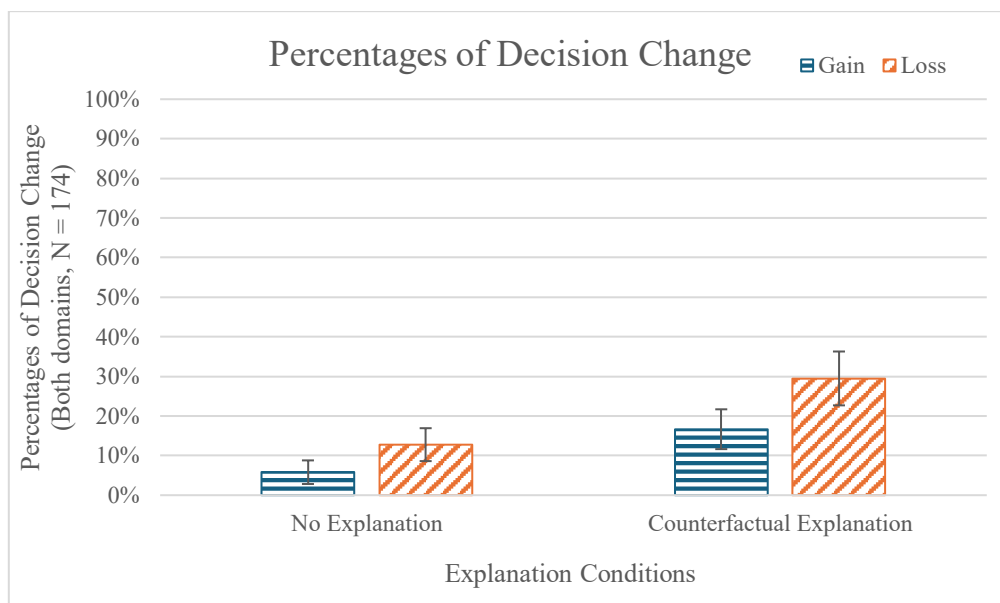
Experiment 10

Figure A18 shows the responses of the 174 participants who made the typical response in the pre-recommendation phase to both scenarios. When they were recommended to make the opposite choice to the typical one that they had made, with no explanation, in the gain frame, 94% of them continued to choose the safe option and 6% changed to the recommended risky one instead; in the loss frame, 87% continued to choose the risky option and 13%

changed to the recommended safe one instead. Given a counterfactual explanation, in the gain frame, 83% of participants continued to choose the safe option, and 17% changed to the recommended risky one; in the loss frame, 71% continued to choose the risky option, and 29% changed to the recommended safe one instead.

Figure A18

Percentages of Decision Change for Experiment 10 (N = 174)



Note. Percentages of switches from a safe choice to a risky one (for gains) and from a risky choice to a safe one (for losses) following a recommendation with no explanation or a counterfactual explanation, summed for both scenarios in Experiment 10. Error bars are standard errors of the mean.

Non-parametric tests showed that there was a significant effect of explanation condition, $U = 4357$, $z = 1.435$, $p = .015$, $r = 0.185$, and no effect of framing, $U = 4209$, $z = 1.792$, $p = .073$. As for the main analysis reported in the main text, the results showed that people are influenced more by the AI

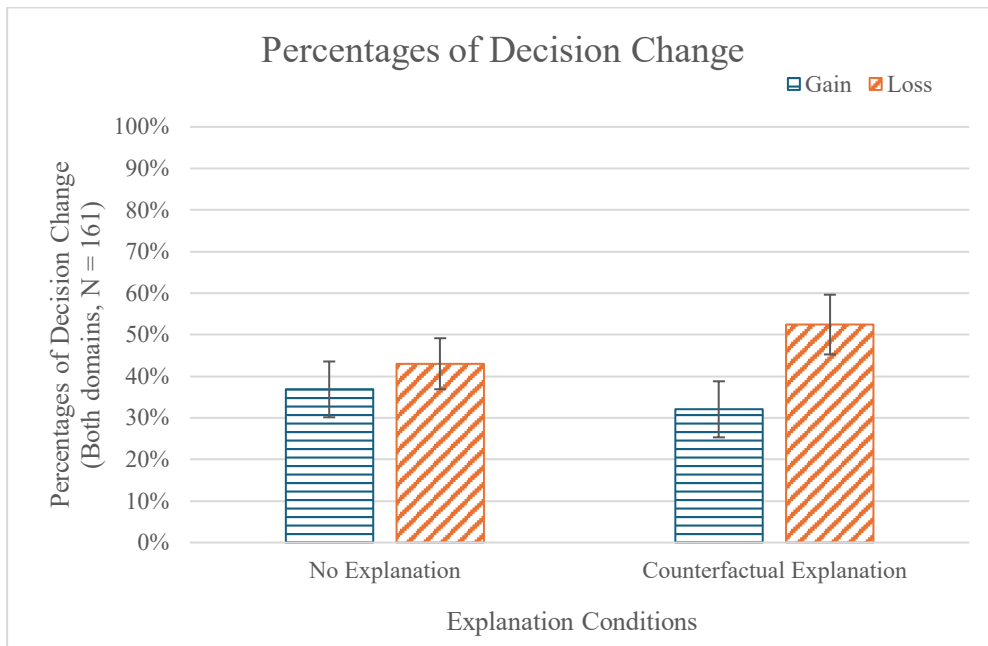
system's recommendations with a counterfactual explanation than without an explanation. Pairwise comparisons showed there was no difference in switches between explanation conditions for gains, $U = 1097.5$, $z = 1.707$, $p = .088$; and the result was marginal for losses, $U = 1089.5$, $z = 1.934$, $p = .053$, $r = 0.21$. There were no differences between gains and losses for no explanation conditions, $U = 1111.5$, $z = 1.336$, $p = .181$; or counterfactual explanation conditions, $U = 1007.5$, $z = 1.470$, $p = .142$. The results are consistent with the main analysis.

Experiment 11

Figure A19 shows the responses of the 161 participants who made the typical response in the pre-recommendation phase to both scenarios. When they were recommended to make the opposite choice to the typical one that they had made, with no explanation, in the gain frame, 63% of them continued to choose the safe option and 37% changed to the recommended risky one instead; in the loss frame, 57% continued to choose the risky option and 43% changed to the recommended safe one instead. Given a counterfactual explanation, in the gain frame, 68% of participants continued to choose the safe option, and 32% changed to the recommended risky one; in the loss frame, 48% continued to choose the risky option, and 52% changed to the recommended safe one instead.

Figure A19

Percentages of Decision Change in Experiment 11 (N = 161)



Note. Percentages of switches from a safe choice to a risky one (for gains) and from a risky choice to a safe one (for losses) following a recommendation with no explanation or a counterfactual explanation, summed for both scenarios, in Experiment 11. Error bars are standard errors of the mean.

Non-parametric tests indicated that there was no significant effect for the explanation condition, $U = 3298$, $z = 0.211$, $p = .833$, and an effect for framing, $U = 3776$, $z = 1.978$, $p = .048$, $r = 0.156$. We carried out Mann-Whitney U tests using a Bonferroni corrected alpha of $p < .0125$ for four post-hoc comparisons. Pairwise comparisons showed there was no difference in switches between explanation conditions for gains, $U = 668$, $z = 0.598$, $p = .550$; and for losses, $U = 982.5$, $z = 0.963$, $p = .335$. There were no differences between gains and losses for no explanation conditions, $U = 890$, z

= 0.742, $p = .458$; or counterfactual explanation conditions, $U = 990.5$, $z = 2.005$, $p = .045$. The result is consistent with the main analysis.