

Explanations and Familiarity in XAI:

How users understand predictions and make decisions using an AI support system

Lenart Celar

School of Psychology, Trinity College Dublin

MSc Psychology

Professor Ruth Byrne

February 2023

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:

A handwritten signature in blue ink, consisting of a stylized 'H' followed by a wavy line.

Date: 21/02/2023

Acknowledgements

Looking back, there are too many people to be grateful to and too many things to be grateful for. First and foremost, I would like to thank my supervisor professor Ruth Byrne. In addition to giving me the opportunity to join Trinity College Dublin and the Reasoning and Imagination Lab, her guidance and the opportunity to learn from her have been invaluable. I thank her for her feedback and suggestions on research proposals and experiment designs, countless reviews of the work included in this thesis, for including my research in her conference presentations, and most importantly for her continuous patience on this three year journey.

I would like to thank other researchers of the Reasoning and Imagination Lab, namely Mark Keane, Greta Warren, and Xinyue Dai, for feedback on my presentations, helpful discussions and constant encouragement and support.

My tuition was co-funded by Google and I would like to extend my warmest thanks to my manager Vesna Gredelj to help me secure the funding, encourage me to pursue my passion for research and to support me in balancing work and university responsibilities.

Lastly, a great source of encouragement were also my friends and family. Thank you for your support and for your understanding when I have not replied or returned your calls for weeks at a time.

Table of Contents

Summary	10
Chapter 1: Introduction	15
Explanations of Decisions by Artificial Intelligence Systems	16
Counterfactual Explanations in XAI	20
Counterfactual and causal explanations	22
Familiar and Unfamiliar Domains	26
Chapter 2: Experiments 1 and 2 on Predictions about an AI system's decisions	29
Method	31
Design	31
Participants	33
Materials	35
Procedure	43
Results	44
Explanation helpfulness judgements	44
Prediction Accuracy and Prediction Confidence	45
Explainability and Trust	48
Discussion	50
Chapter 3: Experiments 3 and 4 on Making decisions	53
Method	55
Design, Materials, and Procedure	55
Participants	57

Results	58
Explanation helpfulness judgements	59
Decision Accuracy and Decision Confidence Measures	60
Explainability and Trust	64
Discussion	66
Chapter 4: General Discussion	69
The role of familiarity of domain in explanations for AI decisions	69
The role of counterfactual and causal explanations in how people understand AI decisions	75
References	81
Appendix A: High familiarity cases with correct system outputs	89
Appendix B: Low familiarity cases with correct system outputs	101
Appendix C: High familiarity cases with incorrect system outputs	113
Appendix D: Low familiarity cases with incorrect system outputs	119
Appendix E: DARPA Explainability of an AI system scale	125
Appendix F: DARPA Trust in the AI system scale	126
Appendix G: Information Sheet	127
Appendix H: Memory Check Questions	130
Appendix I: Debriefing Sheet	132
Appendix J: Questions assessing participants' beliefs against ever drinking and driving, or handling potentially unsafe chemicals	133
Appendix K: Ethics Approval	134

List of Tables

Table 1: Number of participants, age, and gender included in Experiments 1 and 2	35
Table 2: Feature conversion chart for cases in high and low familiarity groups	39
Table 3: Number of participants, age, and gender included in Experiments 3 and 4	58
Table 4: A summary of the results of the four experiments	71

List of Figures

Figure 1:	Example of a case presented in the first part of experiments 1 and 2	32
Figure 2:	Example of a case presented in the second part of experiments 1 and 2	33
Figure 3:	Visualisation of the same case presented in the high familiarity group (left) and the low familiarity group (right)	40
Figure 4:	Example of a case with the correct system outputs used in the first experiment (left) and the same case with incorrect system outputs used in the second experiment (right).	43
Figure 5:	Mean explanation helpfulness judgement for high and low familiarity groups in experiment 1 and 2 for prediction judgements.	45
Figure 6:	Percentage Prediction Accuracy in A, and Mean Prediction Confidence in B, for high and low familiarity groups in Experiments 1 and 2 focusing on prediction judgements.	46
Figure 7:	Explainability of the system rating for each of the four groups in Experiments 1 and 2.	49
Figure 8:	Trust in the system rating for each of the four groups in Experiments 1 and 2	50
Figure 9:	Example of a case presented in the second part of experiments 3 and 4 focusing on decisions	56
Figure 10:	Explanation rating for high and low familiarity groups in experiments focusing on decision judgements.	60
Figure 11:	Percentage Decision Accuracy in A, and Mean Decision Confidence in B, for high and low familiarity groups in Experiments 1 and 2 focusing on prediction judgements.	63

Figure 12: Explainability of the system rating for each of the four groups in Experiments 3 and 4. 65

Figure 13: Trust in the system for counterfactual and causal explanation groups in experiments 3 and 4 focusing on decision judgements. 66

Summary

The aim of the four experiments (n=731) in this study was to examine the effect of different sorts of explanations of an Artificial Intelligence (AI) system's decisions, on participants' judgments of the explanation's helpfulness, and on their accuracy in predicting the system's decisions or making their own decisions. We compared counterfactual explanations such as "you would have been under the limit if you had drunk on a full stomach" and causal explanations such as "you are over the limit because you drank on an empty stomach", for a familiar domain (alcohol and driving) and an unfamiliar one (chemical safety).

Participants were shown the information given to an AI system, such as information about a person's gender, weight, the number of units, duration of drinking time, and stomach fullness, for a variety of cases. They were told about the decisions it made for each case, and an explanation for each decision; they then attempted to predict the AI's decisions for a new set of cases (Experiments 1 and 2), or to make their own decisions for a new set of cases (Experiments 3 and 4). The decisions the AI system made were correct (Experiments 1 and 3) or incorrect (Experiments 2 and 4).

The results showed a dissociation between participants' subjective judgments that the counterfactual explanations were more helpful than the causal ones, and participants' objective accuracy in predicting the AI systems' decisions equally given counterfactual or causal explanations, extending previous research to show this dissociation occurred not only for a familiar domain but also an unfamiliar one. The results also show it occurred only for an AI system that made correct decisions, not for one that made incorrect decisions (Experiments 1 and 2). Importantly, the results showed the dissociation was eliminated when participants' made their own decisions rather than predicted the AI systems' decisions, that is, they tended to judge counterfactual explanations more helpful

than causal ones, and also to make more accurate decisions when they had been given counterfactual explanations rather than causal ones. They did so for familiar domains and unfamiliar ones, and once again only for an AI system that made correct decisions, not one that made incorrect decisions (Experiments 3 and 4).

Participants judged the explanations more helpful, and their judgments were more accurate, in the familiar domain than the unfamiliar one, again only when the AI's decisions were correct, i.e., Experiments 1 and 3. Their confidence in their predictions was unaffected by the nature of the explanation they had been given, and they were more confident in their judgments for the familiar domain than the unfamiliar one. The results were observed when explanation helpfulness was measured locally after each explanation, and consistent results were also observed when it was measured globally by assessing overall explainability and trust in the system. We discuss the implications of our results for the field of Explainable AI and for the psychology of counterfactual and causal explanation.

Chapter 1: Introduction

Suppose you have had a few glasses of wine over a long dinner at a family gathering and you are now wondering whether you are legally over or under the limit to drive home safely. You have downloaded an Artificial Intelligence (AI) decision support system onto your smartphone and you decide to use it for the first time. You enter the information it requires: the number of units you have drunk, the duration over which you drank them, and whether you have eaten or not, as well as details such as your weight and gender, and it provides you with a decision: *You are under the limit*. It also provides you with an explanation for its decision: *You would have been over the limit, if you had drunk for a shorter duration*. Others at the gathering want to know what it says about them, and soon you are entering the required details about various cousins, aunts, uncles and grandparents into the software application (app) to tell them its decision, and its explanation for it. Later in the evening another cousin asks what it would say about them; you have left your phone in the other room but after using the app about 15 times you believe you know what it will say. You ask for your cousin's details and tell them your prediction of the app's decision: it will say they are under the limit. Is your prediction accurate about the app's decision? How confident are you in your prediction? Or in a similar way, suppose for example the app you had downloaded about alcohol and driving gave you a causal explanation instead: *You are under the limit, because you drank for a long duration*. Would the accuracy and confidence of your predictions be better given the counterfactual or the causal explanation?

Now suppose that instead of the blood-alcohol app, you have a part-time college job for the summer in a chemical lab, and your employer has provided an AI decision support system to guide you in what chemicals are safe for you to handle. You enter the information it requires: occupational exposure limit, pH, exposure duration, air pollution

rating, and PNEC rating, and it provides you with a decision: Chemical 83220 is safe. It also provides you with an explanation for its decision: *Chemical 83220 would have been deemed unsafe, if it had had a longer exposure duration*. You use the app about 15 times on your first day and then you are faced with another chemical in a different room from where the device with the AI app is located. You think about this chemical's details and come to a prediction of the app's decision: it will say the chemical is safe. Is your prediction accurate about the app's decision? How confident are you in your prediction? Does the task of predicting the chemical safety app's decision seem harder than the task of predicting the alcohol and driving app's decision?

The examples illustrate the experiments we report: we provide participants with experience of seeing an AI system's inputs, its decision and an explanation for its decision, for a series of cases, and we ask them to judge how helpful the explanations are. Then we test how accurately they can predict what the AI's decision would be when they are given only its inputs, for a series of different cases, and we ask them how confident they are in their prediction. Our goal in this study is to examine how users of eXplainable AI (XAI) support systems understand explanations for decisions made by such AI systems. We examine in four experiments some of the boundary conditions for the usefulness of counterfactual explanations of AI decisions for human users, by comparing counterfactual to causal explanations, not only for AI decisions in familiar domains but also in unfamiliar domains.

Explanations of Decisions by Artificial Intelligence Systems

Over the past decades, machine learning and artificial intelligence systems have reached human-level performance in a variety of complex tasks, ranging from object and facial recognition to natural language processing. They have been shown to be able to outperform humans, an example being the strategic game of Go, AlphaGo, an AI system

built by DeepMind which bested Lee Sedol, a top Go player (Bae et al., 2017). While the first machine learning and AI systems were readily interpretable, the sophistication, especially through the increased usage of hard-to-understand deep neural networks, has reached such an extent, that these highly successful systems usually go undescribed (Barredo Arrieta et al., 2020). They are provided in a “black box” manner, without insight into their complexity and without explanations of how they reached a decision. A clear drawback to using a black box AI system is its lack of transparency, which makes it difficult to understand and validate its decisions (Wojciech et al., 2017). Some machine learning algorithms, such as linear models, decision trees or Bayesian classifiers, use a restricted number of components (paths or rules) and allow for a direct inspection of how a decision was reached (Rai, 2020). In this way, machine learning algorithms can provide visibility to enable people to understand what the system’s decisions are based on. In contrast, deep neural networks, even though superior in predictive abilities, sacrifice the transparency and interpretability of the system’s decisions, for the superior performance of the system (Barredo Arrieta et al., 2020; Rai, 2020).

People are increasingly provided with decisions made by AI systems in many areas of daily life, from health to finances, job recruitment to holiday choices. While they are designed to be easy to use, a need to understand how an AI system’s decisions are made arises when those decisions begin affecting human lives - for example as in medicine, law, defence, or finance (Barredo Arrieta et al., 2020). One prominent example that brought forth the importance of explainable artificial intelligence was the General Data Protection Regulation (GDPR) by the European Union, under which users have a “right to an explanation” of how their personal information is being used, stored, and processed by algorithmic systems (Hoffman et al., 2018). EXplainable AI (XAI) develops algorithmic techniques to provide automated explanations, to improve the interpretability

of an AI system and its decisions (for reviews see Karimi et al., 2020; Keane et al., 2021). In addition to people being legally entitled to such explanations, explanations may improve their trust in decisions by AI systems (e.g., Wachter et al., 2017; Hoffman et al., 2018). Some XAI techniques attempt to explain the whole AI system, whereas others attempt to justify the decision (e.g., Verma et al., 2010; Karimi et al., 2020). In line with previous research that indicates that both a lack of understanding of individual decisions a black box system makes, and a lack of transparency of the entire system undercut people's sense of trust (e.g. Guidotti et al., 2018; Deeks, 2019), the present study incorporates multiple measures of quality of explanations – one local and two global. Locally, the study measures helpfulness for each individual explanation provided by the AI system, and globally, the study incorporates the Defence Advanced Research Projects Agency (DARPA) Explanation Satisfaction and Trust scales (Hoffman et al., 2018) to measure explainability and trust in the AI system as a whole.

Methods for explaining black box models can be classified into different families based on their distinctive characteristics (Guidotti et al., 2018). Two of these classifications relating to our research are *black box model explanation*, and *black box outcome explanation*. *Black box model explanation* aims to interpret the black box globally, thus providing an interpretable and transparent model that is understandable by humans. *Black box outcome explanation* aims to provide an interpretable outcome by returning a prediction together with an explanation about the reasons for that prediction. The prediction is only locally interpretable and the explanation is not trying to explain the entire black box system. Similarly, XAI approaches have been referred to as *exogenous* and *decompositional* (Deeks, 2019). Exogenous approaches to XAI can either be model-centric (also referred to as “global”), where an attempt is made to explain the model as a whole, or subject-centric (also referred to as “local”), where individual decisions of the

system are explained. Explanation of one's decisions is often a prerequisite for establishing trust, and like in interpersonal interaction, it has been argued that the same should apply to working with AI systems (Wojciech et al., 2017). The current study will focus on explanations for AI outcomes, that is, the decisions made by an AI system.

The current study examines explanations for AI decisions that may be helpful for ordinary human users. XAI also considers explanations that may be helpful for expert developers to debug AI systems. The main arguments proposed for why we need explainable AI are verification of the system, improvement of the system, learning from the system, and compliance to legislation (Wojciech et al., 2017). As many AI systems function as black boxes, they cannot be trusted by default, and verification of the system needs to take place. Opaque algorithms and their lack of transparency can undercut people's sense of trust, and Deeks (2019) argues in his review of AI and judicial practices - fairness. There is a clear importance of verifying AI outcomes in healthcare, for example, Caruana et al. (2015) show in their research the system trained to predict pneumonia risk came to incorrect conclusions and would increase, not reduce, the number of pneumonia-related deaths. The first step towards improvement of the system is by identifying its shortcomings and that is much harder done on black box models (Wojciech et al., 2017). As AI systems are trained with large quantities of examples, they may observe patterns which are inaccessible to humans. If we wish to learn from the system, explainable AI can try to extract distilled knowledge to show humans new insights and reinforce them with a select number of examples. With AI systems affecting more and more areas of our daily lives, the assignment of responsibility when systems come to a wrong decision or where individual rights are affected will necessarily call for more explainable systems (Wojciech et al., 2017). The current study confines its examination of explanations for AI decisions to those suitable for ordinary human users.

Within DARPA, Hoffman et al. (2018) propose a two-fold measurement of the quality of an XAI system: by measuring explanation satisfaction and by measuring trust in the XAI context. When it comes to evaluating the goodness of an explanation, Hoffman et al. (2018) identify clarity and precision as the two main building blocks of an explanation. By focusing on clarity and precision, any user of a system can see an explanation and make a decontextualized judgement of whether or not the explanation is “good”. However, explanations are rarely presented without context or additional details, hence additional contextual attributes were identified: understandability, feeling of satisfaction, sufficiency of detail, completeness, usefulness, accuracy, and trustworthiness (Hoffman et al., 2018). Hoffman et al. (2018) define “Explanation Satisfaction” as the degree to which users feel that they understand the AI system or process being explained to them, and propose an eight item Explanation Satisfaction Scale that researchers can use to determine whether users of a particular system have understood decisions the system came to. In addition to rating the helpfulness of each individual explanation, and in the same way as in other XAI user studies (e.g. Warren et al., 2022a; Dai et al., 2022), we incorporate DARPA measurements to the present study to assess overall explainability and trust in the system. We examine several sorts of explanations in our experiments, and our goal specifically, is to understand how helpful users perceive counterfactual explanations, compared to causal explanations, as we now outline.

Counterfactual Explanations in XAI

Recently, there has been an explosion of interest in XAI in counterfactual explanations, that is, explanations that describe how the AI system’s decision would have been different, if it had received different input information about some key feature. A frequently used example is an automated decision by a banking AI system to refuse a customer’s loan application, explained by a counterfactual such as, “*if you had asked for*

a lower amount, your loan application would have been approved” (e.g., Dai, et al., 2022; Warren et al., 2022a). Over 100 distinct computational methods for generating counterfactuals using various different automated approaches have been proposed in XAI (see Karimi et al., 2020; Keane et al., 2021). These alternative proposals have claimed that the particular algorithmic method generates “good” counterfactuals for humans, using various criteria such as counterfactuals that are proximal, plausible, actionable, sparse or diverse (e.g., Karimi et al., 2020; Keane & Smyth, 2020; Wachter, et al., 2018). But crucially, such XAI claims about the usefulness of counterfactual explanations for human users are based on intuition rather than psychological evidence; a recent review identified that fewer than 20% of almost 120 XAI counterfactual papers conducted any test of how human users understood the explanations (Keane et al., 2021).

In XAI, explanations are expected to improve people’s understanding of the AI system or domain (e.g., Hoffman et al., 2018), so that they can objectively perform better, for example, be more accurate or faster on a task involving the AI system, such as being able to predict what the system will do next (e.g., Lage et al., 2019; Lucic et al., 2020; van der Waa et al., 2020). For example, if a person uses an app to estimate whether they are over the legal blood-alcohol limit to drive, the system’s predictions should help them to understand the domain so that their predictions also improve, even when the app’s help is not available. In XAI, explanations are also expected to improve people’s trust in the AI system, assessed by participants’ ratings of trust or satisfaction with the system, fairness of its decisions, or subjective judgments of their own understanding (e.g., Förster et al., 2020; Hoffman et al., 2018; Lucic et al., 2020). The few XAI human user studies that have been carried out indicate that counterfactual explanations can help human users to predict what the AI system will do (e.g., Lage et al., 2019; Lucic et al., 2020; van der

Waa et al., 2021), and improve their trust and satisfaction with it (e.g., Forster et al., 2021; Kenny et al., 2021; Lucic et al., 2020).

The XAI interest in counterfactual explanations has been driven partly by their relation to causal explanations (Byrne, 2019; Miller, 2019). For example, consider again the idea that the app you had downloaded about alcohol and driving gave you a causal explanation instead: *You are under the limit, because you drank for a long duration.* People judge counterfactual explanations to be more helpful than causal ones in their subjective evaluations, that is, their overall global scores on the DARPA explanation satisfaction and trust scales, but counterfactual explanations do not improve their objective understanding of an AI system's decision more than causal ones, that is, the accuracy of their predictions (Warren et al., 2022a). Hence, concerns have been raised that if an explanation is subjectively preferred by human users but has no objective impact on their understanding, such a disassociation could lead to unethical explanation strategies that ensure humans subjectively feel more trust in an app's decision without in fact understanding it (e.g., Warren et al., 2022a). Our goal is to examine the usefulness of counterfactual explanations of AI decisions for human users, by comparing counterfactual to causal explanations, not only using the DARPA global evaluations of explanation helpfulness (as in Warren et al., 2022a) but also using local evaluations of each explanation, and not only for AI decisions in familiar domains (as in Warren et al., 2022a) but also in unfamiliar domains.

Counterfactual and causal explanations

Our primary goal is to establish whether people find counterfactual explanations of an AI system's decisions more helpful than causal explanations. People often create counterfactual explanations in daily life, imagining how things could have turned out differently (e.g., Kahneman & Tversky, 1982a), especially after bad outcomes (e.g.,

Roese, 1997). Counterfactuals are also used to justify or defend past actions in many situations, from political discourse to accident safety reports, e.g., *“the collision wouldn’t have happened if the alarm system had worked”* (e.g., Catellani & Covelli, 2013; Markman et al., 2008b; Morris & Moore, 2000). Accordingly, they can affect judgments of blame and moral responsibility, and people are influenced by them in legal arguments about culpability (e.g., Branscombe, et al., 1996; Malle et al., 2014; Parkinson & Byrne, 2017). People often construct counterfactuals in daily life to excuse poor past performance, e.g., *“I would have got high marks in the exam if I had had more time”* (e.g., Markman & Tetlock, 2000; McCrea, 2008). They tend to imagine how things could have been better, rather than worse, e.g., *“I would have failed the exam entirely if the questions had been harder”* (e.g., De Brigard, et al., 2013). Counterfactual explanations about how things could have been better rather than worse can result in emotions of regret rather than relief (e.g., Kahneman & Tversky, 1982b; McCormack et al., 2016). Nonetheless it is counterfactual explanations about how things could have been better that help people to prepare for the future, by providing a “roadmap” for intentions and plans, helping people to learn from mistakes and prevent the reoccurrence of bad outcomes (e.g., Markman et al., 2008a; Smallman & Roese, 2009; Roese & Epstude, 2017). To do so, counterfactuals need to focus on something within an individual’s control, e.g., *“I would have got high marks in the exam if I had worked more quickly”* rather than outside their control (e.g., Giroto et al., 1991; Ferrante et al., 2013).

Counterfactual explanations depend on identifying the relations between events, especially causal, intentional, or deontic relations, and the link between counterfactual and causal reasoning has received particular attention (e.g., Gerstenberg et al., 2021; Meder et al., 2010; Spellman & Mandel, 1999). Counterfactuals amplify judgments that events are causally related, e.g., *“I would have got high marks in the exam if I had had*

extra time” identifies the causal relation, “*I did not get high marks in the exam because I did not have extra time*” and people believe an event caused an outcome when they can imagine the outcome would not have happened if the event had not (e.g., Lagnado, et al., 2013; McCloy & Byrne, 2002). But cognitive processes in counterfactual and causal explanations differ in important ways. People tend to prefer simple causal explanations rather than more complex causal explanations (e.g., Keil, 2006; Lombrozo, 2007; Quinn et al., 2021). Perhaps as a result, when people construct causal explanations, they tend to focus on strong causes that co-vary with an outcome. In contrast when they construct counterfactual explanations they tend to focus on background enabling conditions that could prevent the outcome (e.g., Frosch & Byrne, 2012; Mandel & Lehman, 1996).

People also tend spontaneously to create fewer counterfactual explanations than causal ones, perhaps because counterfactuals require more cognitive effort (e.g., McEleney & Byrne, 2006). They create a counterfactual by mentally ‘undoing’ some aspects of their mental simulation of what happened, often to add something new, rather than subtract something that happened (e.g., Kahneman & Tversky, 1982a; Roese & Epstude, 2017). They must envisage at least two possibilities, to compare the imagined alternative to reality “*I had extra time and got high marks*” to what actually happened, “*I did not have extra time and I did not get high marks*” (e.g., Byrne, 2005). According to the mental models theory (Johnson-Laird, 1983), humans reason by constructing models to represent states of the world in which the premises are true, reaching conclusions by generating the simplest model with the least assumptions. The core of reasoning is comprehension; people construct mental models based on their understanding of individual premises. This mental representation of premises resembles the way the world would be if the premise was true, (Johnson-Laird, 1983). People mentally simulate possibilities, and they often represent as few possibilities as they can, because of working

memory constraints (e.g., Johnson-Laird & Byrne, 2002; Byrne & Johnson-Laird, 2019). For example, they understand a conditional in the indicative mood “*if I had extra time, I got high marks*” by initially envisaging a single possibility, “*I had extra time and I got high marks*”, and although they know that there may be alternatives to this possibility that are consistent with the conditional, they do not think about them at the outset (Johnson-Laird & Byrne, 2002). Similarly, for a causal assertion, “*Because I had extra time I got high marks*”, people initially envisage a single possibility, and they can flesh out alternative possibilities that are consistent with the causal assertion subsequently (e.g., Frosch & Byrne, 2012). In contrast, they understand a conditional in the subjunctive mood “*if I had had extra time I would have got high marks*” by envisaging the counterfactual conjecture, “*I had extra time and I got high marks*”, but they also recover the known or presupposed facts, “*I did not have extra time and I did not get high marks*” (Byrne, 2005). Hence, when people hear a counterfactual they tend to look at images or words corresponding to both the alternative to reality and the actual reality, whereas when they hear a factual assertion, such as an indicative conditional, or a causal assertion, they tend to look at just images or words corresponding to reality (Orenes et al., 2021). The dual possibilities that people keep in mind ensure they can make more inferences from counterfactuals than factual assertions (e.g., Byrne, & Tasso, 1999).

On this dual possibility theory, counterfactual explanations provide richer information than causal explanations, because they are mentally represented by more possibilities, but they also may require more cognitive resources, again because they are mentally represented by more possibilities. Hence we wish to test whether people subjectively find counterfactual explanations of an AI system’s decisions more helpful than causal explanations, not only when they assess the explanations globally by measures such as the DARPA explanation satisfaction and trust scales, as has been found

(e.g., Warren et al., 2022a), but also when they assess each explanation locally by answering a question such as “how helpful was this explanation?”. We also wish to test whether their understanding of an AI system’s decisions, as measured by their accuracy in predicting its decisions, is improved by counterfactual explanations only as much as by causal ones, as has been found for familiar domains such as the alcohol and driving domain (e.g., Warren et al., 2022a), also for unfamiliar domains such as the chemical safety one. Thus our aim is to examine the relative effects of counterfactual and causal explanations for familiar and unfamiliar domains.

Familiar and Unfamiliar Domains

Our aim is to compare counterfactual and causal explanations of AI decisions not only in a familiar domain but also for an unfamiliar domain. For a familiar domain, it is likely that people will already have beliefs about what prediction the AI system will make, e.g., about whether someone who has drunk 10 units of alcohol will be over the legal limit to drive, whereas in an unfamiliar domain, they may have few prior beliefs about what prediction the AI system will make, e.g., whether a chemical with a pH of 10 units will be unsafe to handle.

People make very different inferences with familiar content, such as alcohol and driving, compared to unfamiliar content, such as chemical safety. For example, many hundreds of experiments have shown that people make more accurate inferences about how to test a rule when it is about a familiar content, such as “*if an envelope is sealed then it has a 90 cent stamp*” compared to unfamiliar content, “*if a card has a vowel on one side then it has an even number on the other side*” (Wason, 1968; for a review see Nickerson, 2015). Familiar content elicits people’s prior beliefs and knowledge (Oaksford & Chater, 2007; Evans & Over, 2004) and hence it can make counterexamples to putative conclusions readily available (Sperber et al., 1995). Familiarity with a domain modulates

people's models of a situation, enabling them to keep in mind more possibilities, and also helping them to eliminate some possibilities from further consideration (e.g., Johnson-Laird & Byrne, 2002; Ragni et al., 2018). Hence we can expect that participants will make more accurate predictions and be more confident about their predictions for a familiar domain such as the alcohol and driving one, compared to an unfamiliar domain such as the chemical safety one. We also wish to examine whether people will judge counterfactual explanations more helpful than causal ones for familiar domains, and also for unfamiliar ones. In a familiar domain people will be readily able to envisage the dual possibilities for a counterfactual explanation, which correspond to their prior beliefs, so they will have access to more information than for a causal explanation; we wish to test whether in an unfamiliar domain people are also as readily able to keep in mind the dual possibilities for a counterfactual explanation. The issue of familiarity of domain is relevant to XAI developments since in some situations, health or holiday choices say, people may have some prior beliefs, whereas in other domains, finance or job recruitment say, they may not; yet familiarity of domain is relatively unexplored in the study of counterfactual explanations in XAI. In the next chapter we describe the rationale for our first two experiments and provide our hypotheses in detail.

Chapter 2: Experiments 1 and 2 on Predictions about an AI system's decisions

Our aim in Experiment 1 was to compare counterfactual and causal explanations of an AI system's decisions. Counterfactual explanations enable people to mentally represent more possibilities explicitly from the outset than causal ones, which ensures that people have access to richer information, but which also requires more cognitive resources. The benefit-at-a-cost nature of this richer representation may go some way to explaining why participants subjectively judge counterfactual explanations for AI decisions to be more helpful than causal explanations; but counterfactual explanations are objectively no more helpful than causal ones in helping them to accurately predict the decisions of an AI system (e.g., Warren et al., 2022a). Our first aim was to compare counterfactual and causal explanations of an AI system's decisions not only in a familiar domain (as in Warren et al., 2022a), but also in an unfamiliar one. Our first hypothesis is that people will find explanations for an AI systems' decisions more helpful in a familiar domain than an unfamiliar one, and they will be more accurate in predicting an AI systems' decisions in a familiar than an unfamiliar domain. Our second hypothesis is that people will find counterfactual explanations more helpful than causal ones in a familiar domain even when we measure explanation helpfulness locally, that is, by judgments after each explanation, rather than only globally by a final explanation satisfaction scale (as in Warren et al., 2022a); we expect to replicate previous findings of no effect of counterfactual and causal explanations on prediction accuracy.

Our final hypothesis was about the effects of the correctness of the AI system's decisions. In Experiment 1 participants attempted to predict an AI system's decisions, and the AI systems decisions were correct. In Experiment 2 participants attempted to predict an AI system's decisions, but its decisions were incorrect. AI decision support systems generally have a high probability of accuracy but nonetheless can make errors, and even

small error rates undermine human trust in their decisions (e.g., Kenny et al., 2021).

When an AI system makes an incorrect decision in a familiar domain, for example, that someone who has drunk 10 units of alcohol over a short duration on an empty stomach is under the legal limit to drive, people are likely to identify the error. We expect that their predictions of the AI systems decisions will be inaccurate, in that they will not align with the AI system's decisions, because its decisions conflict with their own prior beliefs. We also expect that they will judge explanations of such incorrect decisions to be unhelpful in a familiar domain. Hence in Experiment 2, we presented participants with an AI system which made incorrect decisions (that is, it made the opposite decision for each of the 16 cases in Experiment 1). Of course, an AI system that makes wholly incorrect decisions is an extreme situation which would never make it to market. We chose to present an AI system with all incorrect decisions because the cases in Experiment 1 had been selected to be close to the decision boundary, that is, the decision was close to being its opposite. Hence, an AI system with all incorrect decisions provides the best opportunity for participants to detect that the decisions are incorrect and thus enables us to test whether their judgments about explanations and accuracy of their predictions are affected by the correctness of the AI system's decisions. In contrast, when an AI system makes an incorrect decision in an unfamiliar domain, for example, that a chemical with a pH of 10 units with a short exposure duration and a PNEC rating of EC10 is safe to handle, participants will be less likely to spot the error since they have no prior beliefs to guide them. We expect that they will judge explanations of an AI systems decisions to be helpful in an unfamiliar domain even when the decisions are incorrect, and there may be little effect on the accuracy of their predictions. Hence, for an AI system that makes incorrect predictions, we hypothesise that participants will judge explanations to be *less*

helpful in the familiar domain than the unfamiliar one, and they will make *fewer* accurate predictions in the familiar domain than the unfamiliar one.

Method

Design

The first set of two experiments focused on users predicting what the app's decisions would be. The first experiment presented the users with correct system outputs, and the second one presented them with incorrect system outputs. Both experiments had the same design, a 2 (familiarity: high, low) x 2 (explanation type: counterfactual, causal) between-participants design, and in each experiment, participants were randomly assigned to one of the four conditions. The dependent measures were judgments of explanation helpfulness, accuracy of predictions, and confidence in predictions, and also the DARPA explanation satisfaction and trust scales, as described below.

The first part of each experiment, “Understanding the App’s Predictions”, presented participants with a case (output) from the system, with the prediction the system made and the explanation of why the system made that prediction. The case from the system depended on the familiarity condition – the high familiarity groups received a case from the Blood Alcohol Content system, and the low familiarity groups received a case from the Chemical Safety system. The explanation of the system’s decision depended on the condition – the counterfactual explanation groups received a counterfactual explanation while the causal explanation groups received a causal explanation. Participants were asked to rate the helpfulness of the explanation on a 1-5 Likert scale. In the first part, participants were shown a total of 16 cases, see Figure 1.

Figure 1

Example of a case presented in the first part of experiments 1 and 2

Here's information about Stephen,
who's in a bar:

Stephen	
Weight	70 kg
Units of alcohol drank	6
Duration	90 mins
Gender	Male
Stomach	Full

Prediction: The app predicts Stephen is **over the limit**.
Explanation: Stephen's blood alcohol level would have been under the limit, if his weight had been heavier.

This explanation was helpful. *

Strongly disagree
Disagree
Neutral
Agree
Strongly agree

☐
☐
☐
☐
☐

The second part of each experiment, “Predicting the App’s Decisions”, presented participants with cases from the system, again depending on their familiarity group, but without the prediction the system would have made, and without the explanation of how the system made that prediction. Based on the case provided, participants were asked to predict what the app’s decision would be (over/under the limit for the high familiarity group; safe/not safe for the low familiarity group), and to assess how confident they were in their prediction (1-5 Likert scale). In the second part, participants were also shown a total of 16 cases, different from the cases shown in the first part, see Figure 2.

Figure 2

Example of a case presented in the second part of experiments 1 and 2.

Here's information about Michael,
who's in a bar:

Michael	
Weight	85 kg
Units of alcohol drank	6
Duration	90 mins
Gender	Male
Stomach	Full

Based on the information provided I believe this app's prediction for this person will be: *

☐ Under the limit.
☐ Over the limit.

How confident are you in your prediction? *

Not at all confident Not very confident Neither Fairly confident Very confident
☐ ☐ ☐ ☐ ☐

Participants

A power analysis calculated using G*power (Faul, Erdfelder, Buchner, & Lang, 2009) showed that a total of 171 participants across four groups (familiarity: low, high; explanation type: counterfactuals, causals) were required to achieve 90% power for a two-tailed analysis of variance (ANOVA) with a medium size effect ($f = 0.25$) and $\alpha < .05$ for main effects and interactions. Participants were recruited through the online crowdsourcing recruitment platform Prolific, each was paid on average GBP 1.50 for

taking part. Participants were native English speakers from Ireland, the United Kingdom, the United States, Canada, Australia and New Zealand, who had not previously participated in related studies. Participants were optionally invited to provide their age and gender. The number of participants, and their age and gender, included in Experiment 1 ($n=177$) and in Experiment 2 ($n=184$), are provided in Table 1.

Participants were assigned to four groups, in Experiment 1, Familiar Counterfactual ($n=45$), Familiar Causal ($n=48$), Unfamiliar Counterfactual ($n=44$), and Unfamiliar Causal ($n=40$), and in Experiment 2, a different set of participants was assigned to the four groups, Familiar Counterfactual ($n=48$), Familiar Causal ($n=49$), Unfamiliar Counterfactual ($n=43$), and Unfamiliar Causal ($n=44$).

Participants were excluded prior to any data analysis if they failed either of the two attention checks conducted for each experiment (one in Part 1 and one in Part 2), or if they failed to correctly identify at least three out of five features in a memory test conducted after the main parts of the experiment, where from a list of 10 features participants were asked to select those they were presented with earlier in the experiment. Accordingly, a further 24 participants were excluded from Experiment 1 and 24 from Experiment 2 prior to any data analysis. The proportion of eliminations, 11.9% and 11.5% respectively, is commensurate with previous similar studies (e.g., 7.9% in Warren, Keane, & Byrne, 2022). Ethics approval for the study was granted by the School of Psychology Research Ethics Committee, Trinity College Dublin, with the reference code SPREC102020-52 (see Appendix K).

Table 1

Number of participants, age, and gender included in Experiments 1 and 2.

Predictions	<i>Experiment 1</i>	<i>Experiment 2</i>
Participants	<i>n</i> = 177	n = 184
Age	18-56, <i>M</i> = 24.9	18-74, <i>M</i> = 39.1
Gender		
Male	31	63
Female	139	116
Non-binary	5	4
Prefer not to say	2	1

Materials

SafeLimit app. Participants in the high familiarity groups were instructed that they would be testing a new app, named *SafeLimit* (adapted from Warren et al., 2022a). The app was designed to predict whether or not a person would be over the legal limit to drive. This prediction was based on the 5 features the app analysed: the person’s weight, units of alcohol they drank, duration of their drinking, gender, and stomach fullness.

In the first part of the experiments, “*Understanding the App’s Predictions*”, participants were shown a tabular representation of the 5 features for different individuals, with a prediction the system made, and an explanation of why the system came to that prediction. Predictions made by the *SafeLimit* app were either ‘under the limit’ or ‘over the limit’. Participants were asked to rate the helpfulness of the explanation on a 1-5 Likert scale. In the first part, participants were shown a total of 16 cases and one attention check, in a different randomised order for each participant, see Appendix A.

In the second part of the experiments, “*Predicting the Apps’ Decisions*”, participants were shown a tabular representation of the 5 features for different individuals, without the prediction the system would have made, and without the

explanation of how the system would have come to that decision. In the second part, participants were shown a total of 16 cases, which were different to those shown in the first part, and 1 attention check, in a random order.

The attention checks were also presented in a tabular form, visually identical to the target material cases, but the participants were asked to indicate the value of one of the randomly selected features present in the case (see Appendix A).

ChemSafe app. Participants in the low familiarity groups were instructed that they would be testing a new app, called *ChemSafe*. The app was designed to predict whether or not a chemical would be safe to use. This was said to be based on the 5 features the app analysed: occupational exposure limit, pH, exposure duration, air pollution rating, and PNEC rating.

In the first part of the experiments, “*Understanding the App’s Predictions*”, participants were shown a tabular representation of the 5 features for different chemicals, followed by a prediction the system made, and an explanation of why the system came to that prediction. Predictions made by the *ChemSafe app* were either ‘safe’ or ‘not safe’. Participants were asked to rate the helpfulness of the explanation on a 1-5 Likert scale. In the first part, participants were shown a total of 16 cases and one attention check, in a random order (see Appendix B).

In the second part of the experiments, “*Predicting the Apps’s Decisions*”, participants were shown a tabular representation of the 5 features for different chemicals, without the prediction the system would have made, and without the explanation of how the system would have come to that decision. In the second part, participants were shown a total of 16 cases and 1 attention check, in a random order.

Once again, attention checks were presented in a tabular form, visually identical to the target material cases, but the participants were asked to indicate the value of one of the randomly selected features present in the case (see Appendix B).

Cases: high familiarity group. Participants in each group were presented with 34 distinct cases, 16 cases and one attention check in part 1, and a different 16 cases and one attention check in part 2. The high familiarity cases on blood alcohol were taken from an Artificial Intelligence case-base (Warren et al., 2022a; Warren et al., 2022b). Participants in the high-familiarity group were presented with instances of tabular data used to estimate an individual's blood alcohol content (BAC) using the Widmark formula (Posey & Mozayani, 2007).

For the purpose of this study, individuals with $BAC \geq 0.08\%$ were classified as over the legal limit to drive. The cases presented as experimental materials were selected from the Artificial Intelligence case-base generated by randomly perturbing the five features: weight (kg), units of alcohol drank, duration of their drinking (min, 15 min intervals), gender (male/female), and stomach fullness (empty/full), $N = 97,980$ (Warren et al., 2022a). Selection was limited to cases with BAC proximal to the decision boundary ($0.06 < BAC < 0.09$; $n = 10,163$). To ensure a meaningful difference between the query case and the counterfactual case, cases with BAC between 0.076 and 0.084 were excluded from being query cases, resulting in a total of 7,463 cases (based on Warren et al., 2022a).

Out of the remaining pool of cases, 16 cases were randomly selected to be used in the first part of the experiment, "*Understanding the App's Predictions*". To reduce the influence different features might have on participants' perception of system predictions, each case was randomly assigned to one of the four feature groups, resulting in a total of 4 cases in weight category, 4 cases in units of alcohol drank category, 4 cases in duration category, and 4 cases in stomach fullness category. Gender was represented equally, with

50% of cases having the feature gender “Male”, and 50% “Female”. Each case then had the corresponding feature perturbed until the BAC outcome crossed the decision boundary, creating a counterfactual case. For example, a case assigned to the “units of alcohol drank” group had the feature “units of alcohol drank” perturbed, while the other three features (weight, duration, stomach fullness) remained static. Continuous variables were incremented in steps of 5 kg for weight, 1 unit for alcohol, and 15 mins for duration, and the binary variable stomach fullness was assigned the inverse value. If the query case could not be perturbed to cross the decision boundary, a different case was randomly selected, and the same process was applied. If the perturbation was successful, the query case was selected to be included as a material in the experiment, while its counterfactual case was used as the basis for the explanation shown in the first part of the experiment.

The 16 cases that were used in the second part of the experiment, “*Predicting the App’s Decisions*”, were randomly selected from the pool of 7,463 cases, with an additional exclusion of 16 cases that had already been selected as materials in the first part of the experiment, and 16 cases that were used as the basis for the explanation. Two more cases with the same exclusions were randomly selected and they served as attention checks.

Cases: low familiarity group. The low familiarity cases on safe chemicals were created to be analogous to the blood alcohol cases. The same cases as in the high familiarity group were used in the low familiarity group, with certain modifications. To adapt the app and reduce the familiarity, features and units from the BAC system were converted to chemical safety technical terms, while the values and case logic remained the same. The categorical features gender (male/female), and stomach fullness (empty/full) were modified into categorical features, air pollution rating (e-01/e-00), and PNEC rating (EC10/EC50), and the continuous features weight (kg), units drank (units),

and duration (minutes), were modified into continuous features, occupational exposure limit (expressed as ppm), pH (expressed as units), and exposure duration (expressed as minutes). Table 2 indicates how features and units were converted to be used in the low familiarity group.

Table 2

Feature conversion chart for cases in high and low familiarity groups

High familiarity (SafeLimit App)		Low familiarity (ChemSafe App)	
Feature	Unit	Feature	Unit
Weight	kg	Occupational exposure limit	ppm
Units drank	units	pH	units
Duration	min	Exposure duration	min
Gender	Male	Air pollution rating	e-01
	Female		e-00
Stomach fullness	Empty	PNEC rating	EC10
	Full		CE50

Figure 3 shows an example of a case generated by the BAC system used for the high familiarity groups, and the same case modified for the chemical materials used for the low familiarity groups. Unlike the blood alcohol app which provides information based on an accurate formula, the chemical app is entirely fictional.

Figure 3

Visualisation of the same case presented in the high familiarity group (left) and the low familiarity group (right)

Here's information about Zoe, who's in a bar:		Here's information about Chemical 83220.	
Zoe		Chemical 83220	
Weight	68 kg	Occupational exposure limit	68 ppm
Units of alcohol drank	4	pH	4
Duration	105 mins	Exposure duration	105 mins
Gender	Female	Air pollution rating	e-00
Stomach	Empty	PNEC Rating	EC10
<u>Prediction:</u> The app predicts Zoe is under the limit . <u>Explanation:</u> Zoe's blood alcohol level would have been over the limit, if she had drunk more units of alcohol.		<u>Prediction:</u> The app predicts Chemical 83220 is safe . <u>Explanation:</u> Chemical 83220 would have been deemed not safe, if it had a higher pH.	

Explanations. Participants were given either counterfactual or causal explanations in the first part of each experiment. The type of explanation participants received depended on the group they were assigned to. For example, participants in the counterfactual groups were given an explanation in the form, *“Zoe's blood alcohol level would have been over the limit, if she had drunk more units of alcohol.”* Participants in the causal explanations groups were given the explanation in the form, *“Zoe's blood alcohol level was under the limit, because she drank few units of alcohol”*. Participants in the high familiarity groups were given explanations based on the *SafeLimit* App, relating to the blood alcohol level, and participants in the low familiarity group were given explanations based on the *ChemSafe* App, relating to the safety of the chemical. An example of a counterfactual explanation in the low familiarity group is, *“Chemical 83220 would have been deemed not safe, if it had had a higher pH.”* An example of a causal explanation is, *“Chemical 83220 was deemed safe, because its pH was low.”* See Appendices A and B.

To control for the possible effect of different features, participants were presented with four explanations focusing on each of the following four features: weight, units of alcohol drank, duration, and stomach fullness. For the binary feature of stomach fullness, the explanations referred to the opposite of the facts presented about it, e.g., “*Georgina’s blood alcohol level would have been under the limit, if she had drunk on a full stomach*”. For the three continuous features, weight, units of alcohol drank, and duration, the explanations referred to e.g., “more” or “fewer”, e.g., “Zoe’s blood alcohol level would have been over the limit, if she had drunk more units of alcohol”, and “many” and “few”, e.g., “*Zoe’s blood alcohol level was under the limit, because she drank few units of alcohol.*” A baseline was established for each of these variables.

Baseline weight range was 85 ± 5 kg for males and 70 ± 5 kg for females and was determined based on the available data for average weight of countries participants were surveyed in. Weight below the baseline range was referred to in the explanations as “light” and weight above the baseline range was referred to as “heavy”. Baseline units of alcohol was based on the UK National Health Service’s Alcohol Advice in which binge drinking is classified as 8 units of alcohol in a single session for men and 6 for women. Units of 8 or more for men, and 6 or more for women were referred to in the explanations as “many”, and fewer than that were referred to as “few”. Baseline duration was based on the average duration one regular drink (pint of beer / glass of wine) stays in the system, the baseline range for duration was set at 120 ± 30 min. Duration below the baseline range was referred to in the explanations as “short”, and above the baseline range was referred to as “long”. Each explanation focused on a specific feature, and all other features remained within their baseline range; for example, the explanation in Figure 4 focused on units of alcohol, so the weight and duration remained within their baseline ranges (70 ± 5 kg female, 120 ± 30 min). The full set of materials is in Appendices A-D.

Correct and Incorrect system outputs. To further assess the impact of familiarity, the first experiment presented the users with correct system outputs (i.e., the app made an accurate prediction based on the Widmark formula), whereas the second experiment presented them with incorrect system outputs (i.e., the app made an inaccurate prediction based on the Widmark formula).

To create the materials for the incorrect output experiments, the 16 cases that were used in the first experiment were also used in the second experiment. The features and their values remained the same, but as the system output consists of the prediction the system made together with the explanation it provides, both parts of the system output were modified. Incorrect outputs were generated by inverting the original prediction (“under the limit” to “over the limit”, “safe” to “not safe” and vice versa). Explanations provided were modified to be incorrect as well, in a way that the relationship between the feature and the outcome was inverted; for example the explanation focusing on units of alcohol indicated that fewer units of alcohol lead the person to being over the limit Figure 4 demonstrates a case with the correct system outputs used in the first experiment and the same case with incorrect system outputs used in the second experiment. See Appendix C for the high familiarity cases with incorrect outputs, and Appendix D for the low familiarity cases incorrect outputs. The adjustments made to the chemical app to create incorrect outputs were analogous to the adjustments made to the blood alcohol app. Since the chemical app is entirely fictional, there is no true sense in which its predictions are correct or incorrect, and the adjustments were made to ensure it could be used as a control.

Figure 4

Example of a case with the correct system outputs used in the first experiment (left) and the same case with incorrect system outputs used in the second experiment (right).

Here's information about Zoe, who's in a bar:		Here's information about Zoe, who is in a bar.	
Zoe		Zoe	
Weight	68 kg	Weight	68 kg
Units of alcohol drank	4	Units	4
Duration	105 mins	Duration	105 mins
Gender	Female	Gender	Female
Stomach	Empty	Stomach content	Empty
<u>Prediction:</u> The app predicts Zoe is under the limit . <u>Explanation:</u> Zoe's blood alcohol level would have been over the limit, if she had drunk more units of alcohol.		<u>Prediction:</u> The app predicts Zoe is over the limit . <u>Explanation:</u> Zoe's blood alcohol level would have been under the limit, if she had drunk more units of alcohol.	

DARPA Scales. At the end of each experiment, participants were also presented with two scales designed to measure explanation satisfaction and trust in an AI system (Hoffman et al., 2018). The Explanation Satisfaction Scale (see Appendix E) consists of 8 statements assessing the satisfaction with the explanations provided by the system, while the Trust Scale (see also Appendix F) comprises 8 statements relating to trust in the system. Agreement with each statement is indicated on a 5-point Likert scale.

Procedure

Prolific users who decided to participate in the study were provided with a link to the online experiment, presented via Alchemer. Participants read a detailed information sheet (see Appendix G) and upon giving consent they completed one example trial for the first part of the study, and a second example trial for the second part of the study. In addition to the two attention checks interspersed during the experiments, participants also completed a memory check question at the end: they were asked to identify 5 features

used by the app, by selecting them from a list of 10 features (Appendix H). After submitting their responses, participants read a debriefing sheet and were thanked for their participation (see Appendix I).

Results

The data for all of the experiments are available at the Open Science Framework at <https://osf.io/fg9xw/>. For each of the two experiments focusing on prediction judgements a 2 (familiarity: high vs low) x 2 (explanation type: counterfactual vs causal) fully between-participants ANOVA was performed on each of the five dependent variables: explanation helpfulness, prediction accuracy, prediction confidence, DARPA explanation satisfaction judgements, and DARPA trust judgements.

Explanation helpfulness judgements

When presented with correct system outputs in Experiment 1, participants found the explanations more helpful in the familiar domain than the unfamiliar domain, as shown by the significant main effect found for familiarity, corroborating our first hypothesis, $F(1, 173) = 14.73, p < .001, \eta_p^2 = .08$, and they found counterfactual explanations more helpful than causal explanations, as shown by the main effect of explanation type, corroborating our second hypothesis, $F(1, 173) = 22.04, p < .001, \eta_p^2 = .11$. There was no interaction of the two variables, $F(1, 173) = 0.10, p = .75^1$.

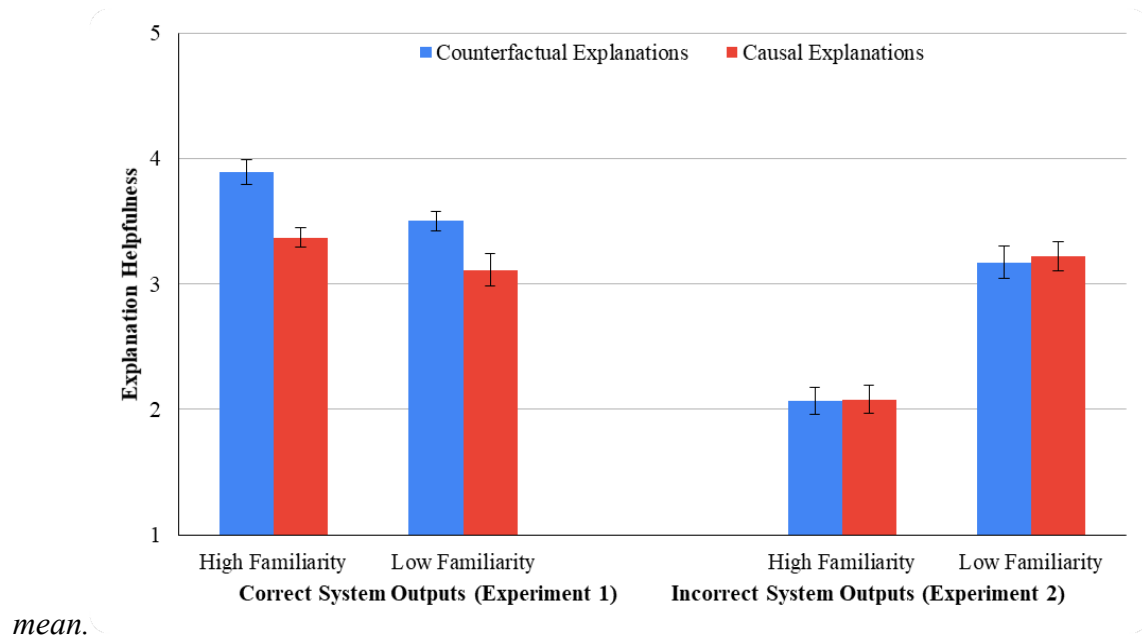
In Experiment 2, when presented with incorrect system outputs, participants instead judged explanations as *less* helpful for the familiar domain than the unfamiliar domain, as shown by the significant main effect of familiarity, corroborating our third hypothesis, $F(1, 180) = 92.54, p < .001, \eta_p^2 = .34$; and this time, there was no main effect

¹ Levene's test was significant for the explanation helpfulness measure in Experiment 1 indicating violation of the assumption of homogeneity of variance for this measure and so we followed the recommended procedure for unequal n's (Gignac, 2019) of repeating the ANOVA using equal numbers in each group (reducing all groups to the smallest n, i.e., n=40, by random selection for each group via SPSS). The same results were found i.e., familiarity, $F(1, 156) = 13.756, p < .001, \eta_p^2 = .08$; explanation type, $F(1, 156) = 20.238, p < .001, \eta_p^2 = .12$; interaction, $F(1, 156) = .005, p = .95$.

of explanation type, $F(1, 180) = .02, p = .89$. There was no interaction, $F(1, 180) = .06, p = .81$. See Figure 5.

Figure 5

Mean explanation helpfulness judgement for high and low familiarity groups in experiment 1 and 2 for prediction judgements. Error bars represent standard error of mean.



Prediction Accuracy and Prediction Confidence

In the second part of the prediction judgement experiments participants were asked to predict what the app's decision would be. The prediction accuracy measure indicates the rate at which participants' predictions aligned with the app's predictions.

When presented with correct system outputs in Experiment 1, participants made more correct predictions for the familiar domain than the unfamiliar domain, as shown by a significant main effect of familiarity for prediction accuracy, again corroborating our first hypothesis, $F(1, 173) = 28.68, p < .001, \eta_p^2 = .14$. They made as many correct predictions whether they had been given counterfactual or causal explanations, as shown by no main effect of explanation type, replicating previous findings (Warren et al, 2022a),

$F(1,173) = .14, p = .71$. There was no interaction of the two variables, $F(1,173) = .91, p = .34$.

For Prediction Confidence, participants were more confident about their predictions for the familiar domain than the unfamiliar domain, as shown by a significant main effect for familiarity, $F(1, 173) = 8.54, p = .004, \eta_p^2 = .05$. There was no significant difference between being given counterfactual or causal explanations, as shown by no main effect of explanation type, $F(1, 173) = .91, p = .34$. There was no interaction of the two variables, $F(1, 173) = .04, p = .84$.²

When presented with incorrect system outputs in Experiment 2, participants instead made fewer “correct” predictions, that is, in this case, predictions that were the same as the app’s predictions, for the familiar domain than the unfamiliar one, as shown by a significant main effect of familiarity, again corroborating our third hypothesis, $F(1, 180) = 34.95, p < .001, \eta_p^2 = .16$. There was no main effect of explanation type on prediction accuracy, $F(1, 180) = .16, p = .69$, and no interaction of the two variables, $F(1, 180) = 2.48, p = .12$.

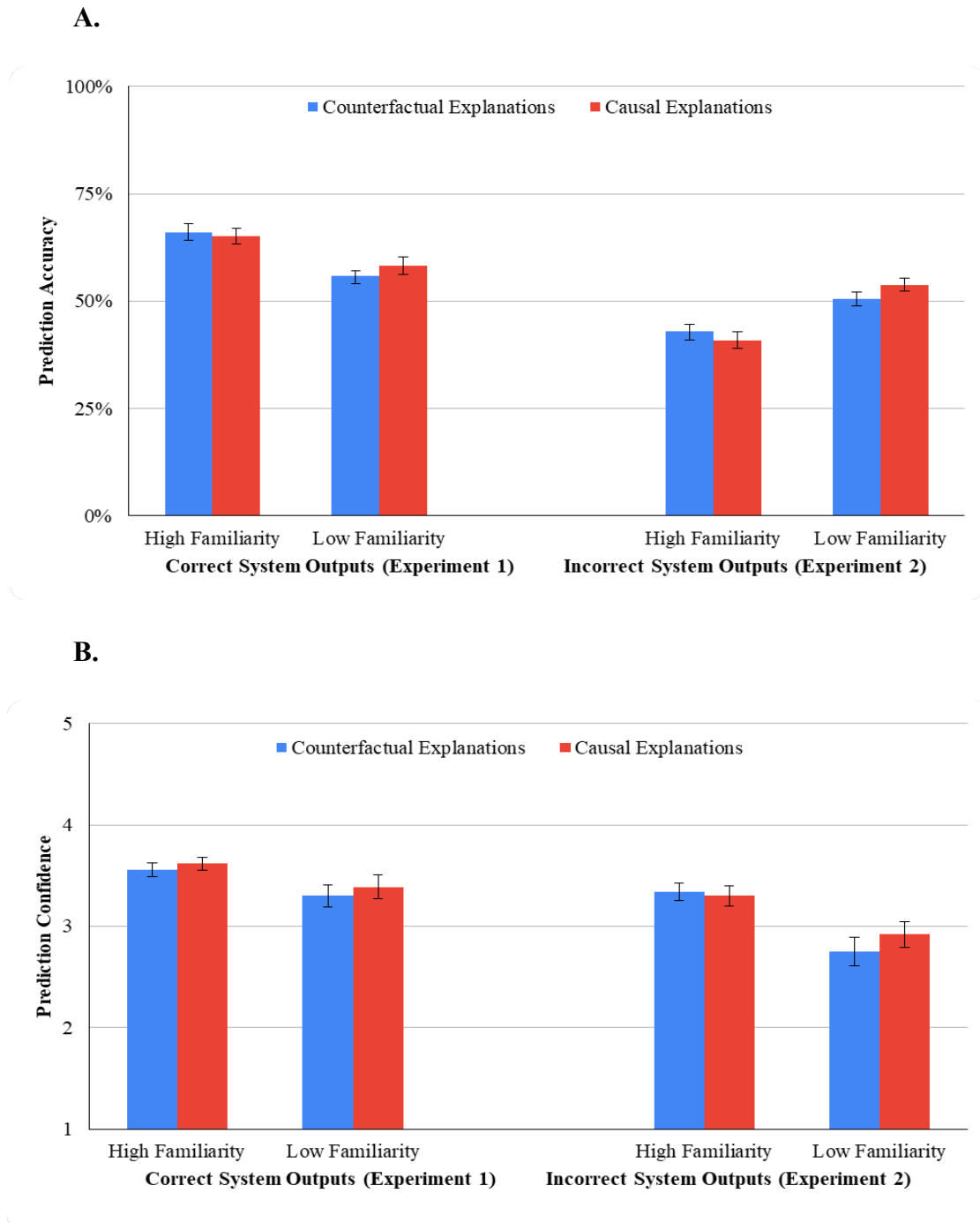
Participants were nevertheless more confident about their predictions for the familiar domain than the unfamiliar domain, as shown by a significant main effect of familiarity for prediction confidence, $F(1,180) = 17.26, p < .001, \eta_p^2 = .09$. There was no significant difference in participants’ prediction confidence whether they had been given counterfactual or causal explanations, as shown by no main effect of explanation type, $F(1, 180) = .33, p = .57$. There was no interaction of the two variables, $F(1, 180) = .86, p = .36$.³ See Figure 6.

² Levene’s test was significant for the confidence measure in Experiment 1 and so once again we followed the recommended procedure for unequal n’s (Gignac, 2019). The same results were found i.e., familiarity, $F(1, 156) = 8.048, p < .001, \eta_p^2 = .05$; explanation type, $F(1, 156) = .884, p = .35$; interaction, $F(1, 156) = .00, p = .99$.

³ Levene’s test was significant for the confidence measure in Experiment 2 and so once again we followed the recommended procedure for unequal n’s (Gignac, 2019). The same results were found when we repeated the ANOVA (reducing all groups to the smallest n, i.e., n=43 by random selection via SPSS), i.e., familiarity, $F(1, 168) = 9.527, p = .001, \eta_p^2 = .08$; explanation type, $F(1, 168) = .295, p = .59$; interaction, $F(1, 168) = .613, p = .44$.

Figure 6

Percentage Prediction Accuracy in A, and Mean Prediction Confidence in B, for high and low familiarity groups in Experiments 1 and 2 focusing on prediction judgements. Error bars represent standard error of mean.



DARPA Explainability and Trust

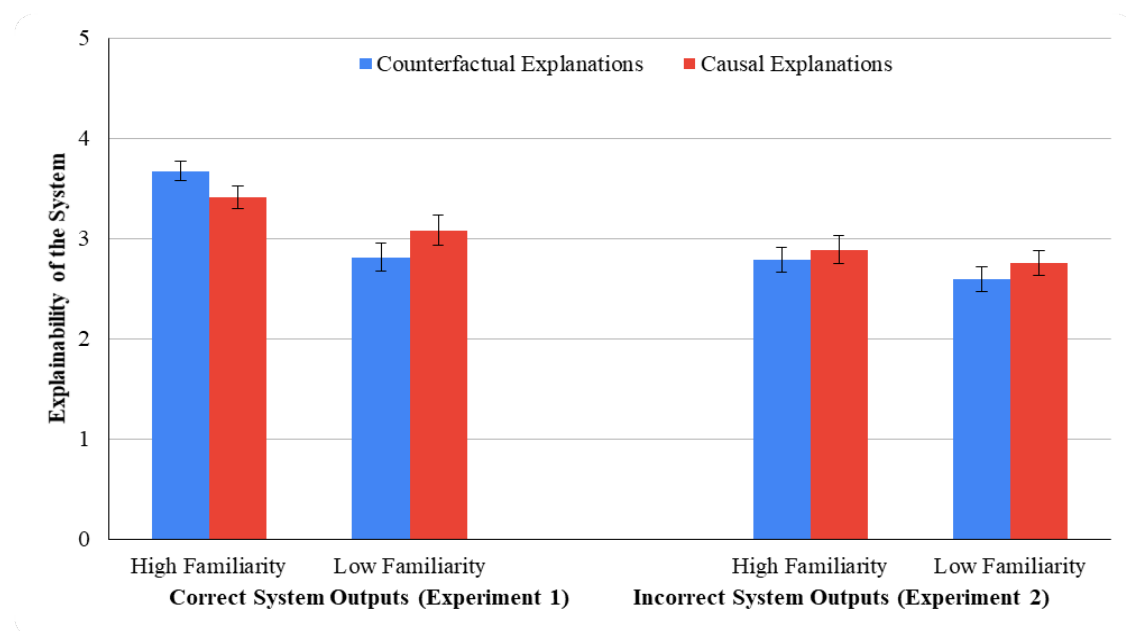
All groups completed the DARPA Explanation Satisfaction and Trust scales after completing the two main parts of the experiment. Participants in Experiment 1 given correct system outcomes were more satisfied with explanations given for the familiar domain than the unfamiliar domain, as shown by the main effect of familiarity, consistent with the local explanation helpfulness judgments, ($F(1, 173) = 15.10, p < .001, \eta_p^2 = .11$). There was no significant difference between counterfactual explanations and causal explanations, as shown by no main effect of explanation type, $F(1, 173) = .03, p = .88$, however, there was a significant interaction effect of familiarity and explanation type, $F(1, 173) = 4.41, p = .04, \eta_p^2 = .03$.

To decompose the interaction effect, a series of paired comparisons using independent sample t-tests was carried out, with a Bonferroni correction for six comparisons of $p = .008$. Given high familiarity materials, there was no significant difference in judgements between participants in the counterfactual group and the causal group, $t(92) = -1.61, p = .11$. Similarly, given low familiarity materials, there was again no significant difference in judgements between participants in the counterfactual group again judged explainability equally as the causal explanations group, $t(82) = 1.39, p = .17$. The interaction arises mainly because given counterfactual explanations, participants in the high familiarity group judged explainability higher than the low familiarity group, $t(87) = 5.02, p < .001$, Cohen's $d = .80$; given causal explanations, there was no difference between the high familiarity group and the low familiarity group, $t(86) = 1.74, p = .09$. Explainability was judged higher by the high familiarity counterfactual group than the low familiarity causal group, $t(83) = 3.22, p = .002$, Cohen's $d = .81$, and by the high familiarity causal group than the low familiarity counterfactual group, $t(90) = 3.39, p = .001$, Cohen's $d = .86$. Differences between groups are indicated in Figure 7.

When participants were presented with incorrect system outcomes in Experiment 2, there was no significant difference in explainability rating between the participants in the familiar domain and the unfamiliar domain, as shown by no main effect of familiarity $F(1, 180) = 1.63, p = .20$, unlike the local explanation helpfulness measure, and given counterfactual explanations as causal explanations, as shown by no main effect of explanation type, $F(1, 180) = .95, p = .33$; there was no interaction, $F(1, 180) = .05, p = .82$. The results suggest that the global explanation satisfaction measure may not be as sensitive as the local explanation helpfulness judgment.

Figure 7

Explainability of the system rating for each of the four groups in Experiments 1 and 2. Error bars represent standard error of mean.

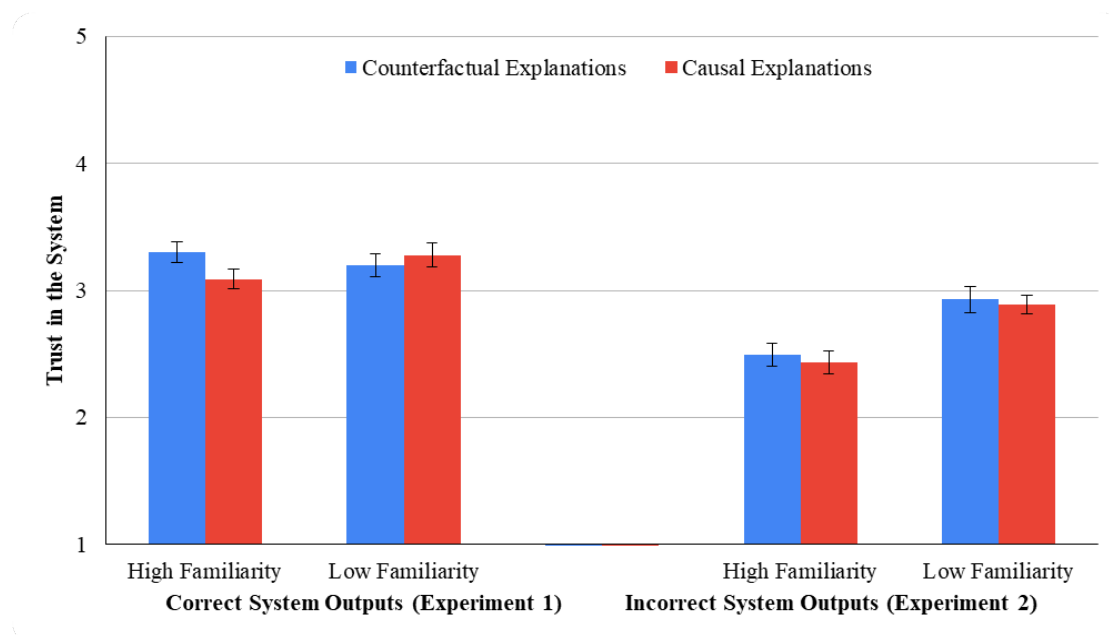


For Trust ratings in Experiment 1 with correct system outputs, participants rated the system as trustworthy equally given the familiar domain as the unfamiliar domain, as shown by no main effect of familiarity $F(1, 177) = .88, p = .35$, and given counterfactual explanations as causal explanations, as shown by no main effect of explanation type, $F(1, 177) = .45, p = .50$; there was no interaction, $F(1, 177) = 1.08, p = .30$.

When presented with incorrect system outputs in Experiment 2 however, participants rated the system as less trustworthy given the familiar domain than the unfamiliar domain, as indicated by a significant main effect of familiarity, $F(1, 184) = 24.64$, $p < .001$, $\eta_p^2 = .12$. They rated the system as equally trustworthy given counterfactual explanations as causal explanations, as shown by no main effect of explanation type, $F(1, 184) = .29$, $p = .59$; there was no interaction, $F(1, 184) = .01$, $p = .91$. Once again, the results suggest that the global trust measure may not be as sensitive as the local explanation helpfulness judgment. Differences between groups are shown in Figure 8.

Figure 8

Trust in the system rating for each of the four groups in Experiments 1 and 2. Error bars represent standard error of mean.



Discussion

The results show that participants judged explanations to be more helpful in a familiar domain than an unfamiliar one for the decisions of a correct AI system in Experiment 1, in line with the first hypothesis, but to be less helpful in a familiar domain

than an unfamiliar one for decisions of an incorrect AI system in Experiment 2, in line with the third hypothesis. The results also show that participants considered counterfactual explanations to be more helpful than causal ones when they considered the decisions of a correct AI system in Experiment 1, in line with the second hypothesis, and they did so for both the familiar and the unfamiliar domain, but when they considered the decisions of an incorrect AI system in Experiment 2, they judged both sorts of explanations to be equally unhelpful for the familiar domain, and equally helpful for the unfamiliar domain.

Their predictions were more accurate in the familiar domain than the unfamiliar one when the AI's decisions were correct, again corroborating our first hypothesis, and less accurate in the familiar domain than the unfamiliar one when the AI's decisions were incorrect, i.e., less aligned with its decisions, again corroborating our third hypothesis. Hence, in the high familiarity domain their prediction accuracy was above chance when the AI system's predictions were correct but below chance when its decisions were incorrect, as Figure 6A shows, whereas for the low familiarity domain their predictions remained around chance level whether the AI system's predictions were correct or incorrect.

The results replicate previous findings of a dissociation between, on the one hand, subjective preferences for counterfactual explanations over causal ones, extending it to local judgments of individual explanation helpfulness as well as global judgments of overall explanation satisfaction, and on the other hand no increase in objective accuracy given counterfactual explanations rather than causal ones, in familiar domains (Warren et al., 2022a); they further extend it to. Moreover, the results show that for incorrect decisions of an AI system there is no subjective preference for counterfactual explanations over causal ones, and no increase in accuracy given counterfactual

explanations rather than causal ones, that is, there was no dissociation between subjective and objective measures.

There were many more female than male participants in the experiments, as is commonly the case with online experiments. Few studies have examined gender differences in counterfactual thinking, and their focus has been on potential differences in counterfactual thinking about moral and emotional issues, with mixed results (e.g., Johnson, Kim, Blue, Summerville, & Kirkmeyer, 2021; Levy & Ben-David, 2015; Migliore, Curcio, Mancini, & Cappa, 2014). Hence, there is little a priori reason to suppose gender differences may underlie the prediction accuracy or explanation preference results of Experiments 1 and 2.

A final observation is that participants' confidence in their predictions was generally around the midpoint regardless of the type of explanation, familiarity of the domain, or correctness of the AI system. The results also suggest that the local explanation helpfulness measure may be more sensitive than the global explanation satisfaction and trust measures.

Our next experiments examine the helpfulness of counterfactual and causal explanations further, this time examining their effects on the accuracy of one's own decisions rather than the accuracy of predictions of an AI system.

Chapter 3: Experiments 3 and 4 on Making decisions

Suppose once again that you are at a family gathering and have had a few glasses of wine. Once again you have gained experience with the alcohol and driving app, and now you no longer have access to it. Suppose now you must decide whether you are prepared to drive home. Note that your task in this case is not simply predicting what the app would decide about you, instead your task is to make your own decision about what is safe for you to do. The aim of Experiments 3 and 4 was to compare counterfactual and causal explanations for an AI system's decisions in familiar and unfamiliar domains when participants attempted to make their own decisions, rather than to predict the decision the AI system would make. Our interest in people's own decisions arose for two reasons, first, AI systems have been shown to be able to outperform humans (e.g. Bae et al., 2017), but in a variety of situations, for example, in finance, medicine, or legislation, human users are the ones interacting with the system and ultimately deciding whether to follow the decision of the system or not. It is important for XAI to gain an understanding of how experience with AI systems' decisions and explanations for them might affect people's own decisions. Hence it is crucial to test the effects of counterfactual explanations not only by examining their effects on accuracy in predicting an AI system's decision, but also in making one's own decisions. Second, counterfactuals are often personal, that is, people create counterfactual thoughts about how events in their own lives could have turned out differently, and counterfactuals are also often centred on goals, that is, they help people prepare for the future by enabling them to form intentions, for example, to prevent bad outcomes from occurring again (Roese & Epstude, 2017; Ferrante et al., 2013). They can be particularly influential in helping people to make decisions (O'Connor et al., 2014). Accordingly, our new hypothesis was that when participants are required to make their own decisions, they will be more accurate when they have

previously received counterfactual explanations than causal explanations. In other words, our hypothesis is that the dissociation between subjective and objective measures will be eliminated in that participants will not only subjectively judge counterfactual explanations as more helpful than causal explanations, but their own decisions will also be more accurate, that is, aligned with the AI system's decisions, when they have been given counterfactual explanations rather than causal ones.

In the same way as in Experiments 1 and 2, participants in Experiment 3 were asked to assess the helpfulness of explanations provided by the system, and the system's decisions were correct; and in Experiment 4, they were asked to rate the helpfulness of explanations provided by the system, but its decisions were incorrect. In the second part of the experiments, they were asked to imagine that the information entered into the app was about them. They were asked to make a decision based on this information about them, about whether they would drive or not drive.

Our other hypotheses are similar to those in Experiments 1 and 2, that is, we hypothesise that the decisions participants make will be accurate, that is, they will align with the decisions the system would make, more when they are familiar with the domain, than when they are unfamiliar with it, for an AI system that produces correct outputs. In contrast, for an AI system that produces incorrect outputs, we hypothesise that participants will make fewer accurate decisions when they are familiar with the domain, than when they are unfamiliar with it, because they will be able to identify that the system was making incorrect decisions, and elect to make decisions based on their personal beliefs and their understanding of the domain. For example, if they are asked to imagine they had just drunk 7 units of alcohol over a short period of time and on an empty stomach, they will decide not to drive based on their prior knowledge of blood alcohol

content, even though they have previously observed that the (incorrect) system would predict they are under the limit.

Method

Design, Materials, and Procedure

The second set of two experiments focused on how users make decisions based on the information provided, rather than how they predict the app's decisions. The only difference between these experiments and the previous ones was in the second part of each experiment, "*Making a Decision*", where participants were asked to imagine the data provided was about them and to make a decision about whether they would be prepared to drive/not drive (high familiarity) or handle/not handle the chemical (low familiarity), see Figure 9. The first part of each experiment, "*Understanding the App's Predictions*", was identical to the first part of the previous experiments on predictions.

Figure 9

Example of a case presented in the second part of experiments 3 and 4 focusing on decisions.

Imagine this is information about you in a bar.

Weight	73 kg
Units	5
Duration	60 mins
Gender	Female
Stomach content	Full

You are now asked to make a **decision** based on the information provided, and assess how confident you are in your decision.

Based on the information provided I would be prepared to *

- ☐ Drive.
- ☐ Not drive.

How confident are you in your decision? *

- Not at all confident Not very confident Neither Fairly confident Very confident
- ☐ ☐ ☐ ☐ ☐

As in the previous experiments, to assess the impact of familiarity, the first experiment of each set presented the users with correct system outputs, and the second one presented them with incorrect system outputs. Again, a 2 (familiarity: high, low) x 2 (explanation type: counterfactual, causal) between-participants design was employed. The dependent measures were identical to the ones in the previous experiments: judgments of

explanation helpfulness, accuracy of decisions, and confidence in decisions, and the two DARPA scales of explainability and trust.

The materials and procedure were identical to the previous experiments, with the exception of the addition of two additional questions on their beliefs, see Appendix J, included as part of the exclusion criteria. Given that these experiments focused on participants own beliefs, the additional questions were included in case participants with strong beliefs, e.g., that someone should never drink and drive, may simply respond with the same decision (do not drive) to every case.

Participants

Participants for Experiments 3 and 4 focusing on decisions were recruited in the same way as participants for Experiments 1 and 2. The number of participants, and their age and gender, included in Experiment 3 (n=173) and Experiment 4 (n=186) are provided in Table 3. Participants were assigned to four groups, in Experiment 3, Familiar Counterfactual (n=41), Familiar Causal (n=44), Unfamiliar Counterfactual (n=43), and Unfamiliar Causal (n=45), and in Experiment 4, a different set of participants were assigned to four groups, Familiar Counterfactual (n=49), Familiar Causal (n=43), Unfamiliar Counterfactual (n=47), and Unfamiliar Causal (n=47). The participants in Experiments 3 and 4 were a new set of participants who had not participated in Experiments 1 and 2.

Based on the exclusion criteria described for the previous experiments, that is, failing the attention or memory checks, a further 36 participants were excluded from Experiment 3 and 55 from Experiment 4. Once again, the proportion of eliminations, 13.9% and 18% respectively, is consistent with previous studies, although somewhat higher. In addition, for Experiments 3 and 4 focusing on participants' own decisions, an *a priori* decision was made to exclude any participants if they indicated in response to two

additional questions on their beliefs, that they had strong beliefs against ever drinking and driving, or handling potentially unsafe chemicals. Strong beliefs were indicated by agreeing strongly (5 on a 1-5 scale) with the first statement (e.g., I believe that people who have had even one drink should not drink and drive) or disagreeing strongly (1 on a 1-5 scale) with the second statement (e.g., I believe that it is acceptable for people who are under the legal limit to drive), on assessing participants' beliefs against ever drinking and driving, or handling potentially unsafe chemicals (see Appendix J). Accordingly, a further 49 were excluded from Experiment 3, and 59 from Experiment 4.

Table 3

Number of participants, age, and gender included in Experiments 3 and 4.

Decisions	<i>Experiment 3</i>	<i>Experiment 4</i>
Participants	$n = 173$	$n = 186$
Age	18-77, $M = 36.2$	19-72, $M = 39$
Gender		
Male	46	63
Female	122	122
Non-binary	4	-
Prefer not to say	1	1

Results

For each of the two experiments focusing on decision judgements a 2 (familiarity: high vs low) x 2 (explanation type: counterfactual vs causal) fully between-participants ANOVA was performed on each of the five dependent variables: explanation helpfulness, decision accuracy, decision confidence, DARPA explainability judgements, and DARPA trust judgements.

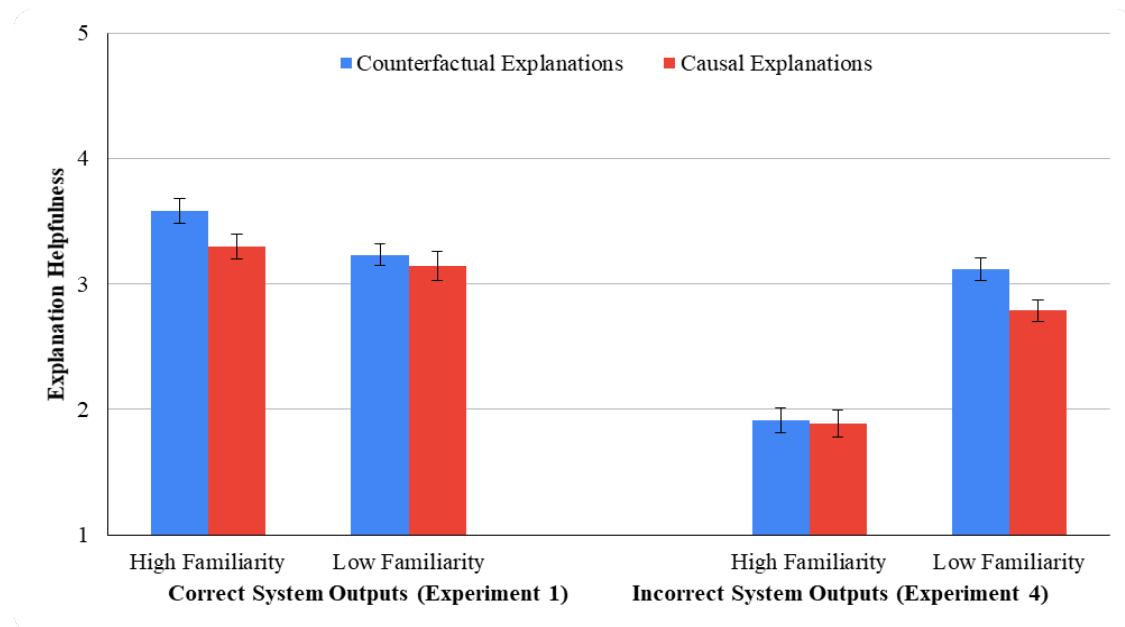
Explanation helpfulness judgements

The first part of experiments 3 and 4 was identical to the first part of experiments 1 and 2 respectively, and so we expect to replicate the findings for explanation helpfulness judgments. When presented with correct system outputs in Experiment 3, participants found the explanations more helpful in the familiar domain than the unfamiliar domain, as shown by the significant main effects found for familiarity, $F(1, 169) = 6.19$, $p = .014$, $\eta_p^2 = .04$, just as they did in Experiment 1. They found the counterfactual explanations somewhat more helpful than causal ones, but the difference did not reach significance $F(1, 169) = 3.38$, $p = .07$, $\eta_p^2 = .02$, unlike in Experiment 1, as Figure 10 shows. There was no significant interaction effect, $F(1, 169) = 0.88$, $p = .35$. The explanation helpfulness results of Experiment 3 replicate those of Experiment 1 for familiarity, and are in the same direction although not significantly so for explanations.

In contrast, when presented with incorrect system outputs in Experiment 4, participants instead found the explanations less helpful for the familiar domain than the unfamiliar domain, as indicated by a significant main effect of familiarity, $F(1, 182) = 119.94$, $p < .001$, $\eta_p^2 = .40$, just as they did in Experiment 2. There were no significant differences between explanation types, $F(1, 182) = 3.46$, $p = .06$, $\eta_p^2 = .02$, and no interaction effect, $F(1, 182) = 2.46$, $p = .12$. See Figure 10. The explanation helpfulness results of Experiment 4 replicate those for Experiment 2.

Figure 10

Explanation rating for high and low familiarity groups in experiments focusing on decision judgements. Error bars represent standard error of mean.



Decision Accuracy and Decision Confidence Measures

In the second part of the decision judgement experiments participants were asked to imagine the data provided was about them and to make a decision about whether they would be prepared to drive/not drive (high familiarity) or handle/not handle the chemical (low familiarity). The decision accuracy measure indicates the rate at which participants' decisions aligned with the system output.

When presented with correct system outputs in Experiment 3, participants' decisions were more accurate, that is, they aligned more with the AI system when they had been previously presented with counterfactual explanations compared to causal explanations, as indicated by a significant main effect of explanation type for decision accuracy, $F(1, 169) = 4.61$, $p = .03$, $\eta_p^2 = .03$, corroborating our new hypothesis. Whether participants had been given counterfactual or causal explanations, there was no significant difference in confidence in their predictions, as shown by no main effect of explanation type, $F(1, 169) = .75$, $p = .39$, as in Experiment 1. To further test the consistency of the

items in relation to decision accuracy, a 2 (explanation) x 2 (familiarity) x 2 (outcome: good – under the limit/safe to handle, vs bad - over the limit/unsafe to handle) ANOVA with repeated measures on the third factor was performed. The same direction of difference between counterfactual and causal explanations was not observed for all items. Participants' decisions were more accurate when they had been given counterfactual rather than causal explanations (more than 5% difference) for 8 of the 16 items, but more accurate when they had been given causal rather than counterfactual explanations for 6 items, with no difference (greater than 5%) for 2 items. Their decisions were more accurate for the 8 items with bad outcomes than the 8 items with good outcomes, $F(1, 169) = 48.923, p < .0001, \eta^2 = .224$. Their decisions about items with good outcomes were more accurate when they had received counterfactual explanations rather than causal ones, $t(171) = 3.482, p < .001, CI [-.23773, -.0657]$, but more accurate about items with bad outcomes when they had received causal explanations rather than counterfactual ones, $t(171) = 2.196, p < .029, CI [.00822, .15413]$, in the decomposition of the interaction of outcome with explanation, $F(1, 169) = 10.594, p < .001, \eta^2 = .059$

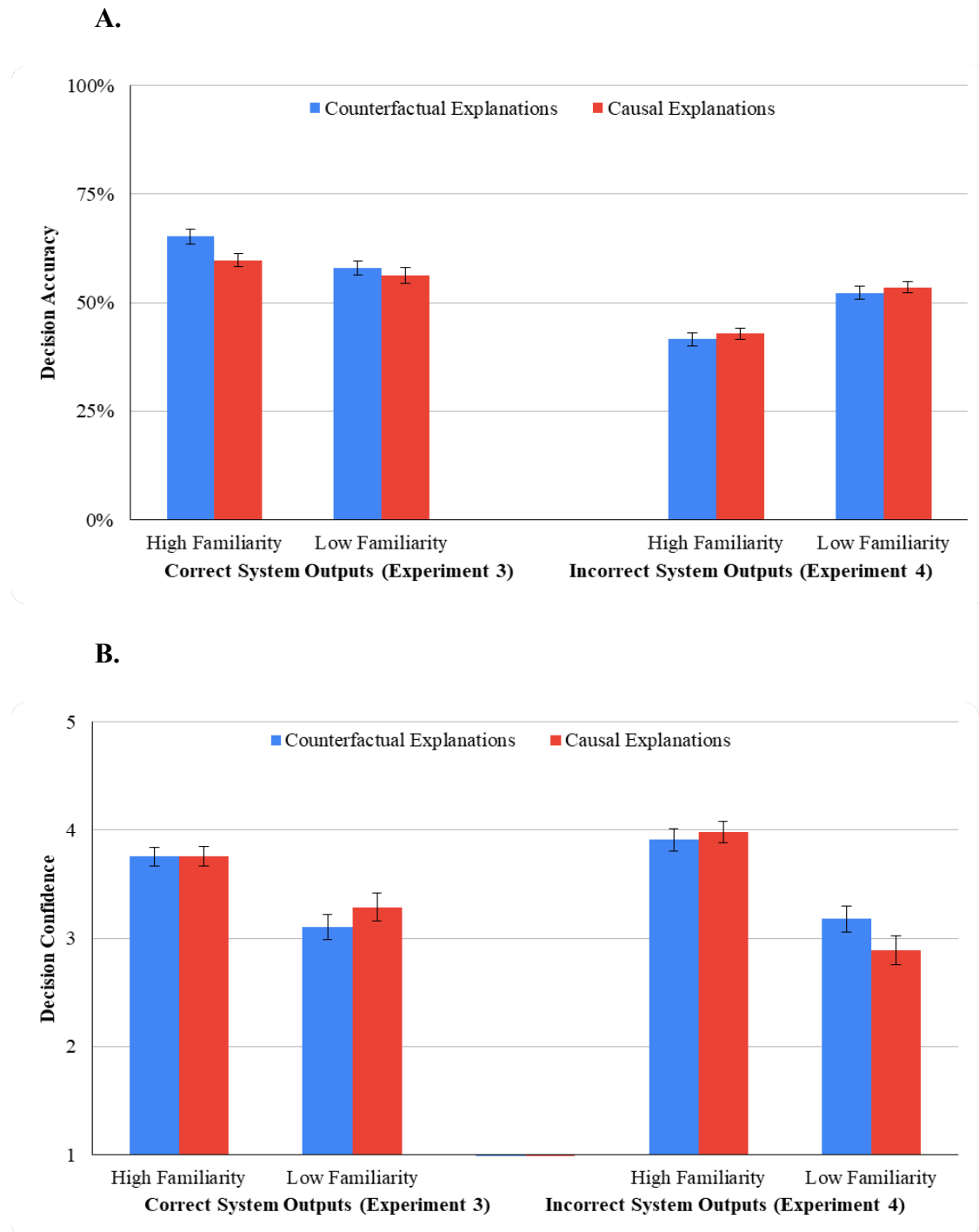
Participants' decisions were more accurate that is, they aligned more with the system's decisions, in the familiar domain than the unfamiliar domain, as shown by a significant main effect of familiarity for decision accuracy, $F(1, 169) = 10.41, p < .002, \eta_p^2 = .06$, just as in Experiment 1. Participants in the high familiarity group were also more confident in the decisions they had made, indicated by a significant main effect of familiarity for decision confidence, $F(1, 169) = 27.37, p < .001, \eta_p^2 = .14$. The results of the effects of familiarity on decision accuracy and confidence are similar to those for prediction accuracy and confidence in Experiment 1 and extends them to participants' own decisions. There was also no interaction effect for decision accuracy, $F(1, 169) = 1.22, p = .27$, and no interaction effect for decision confidence $F(1, 169) = .66, p = .42$.

When presented with incorrect system outputs in Experiment 4, participants' decisions were *less* accurate, that is, that is, they aligned less with the AI system's outputs, in the familiar domain than the unfamiliar domain, as shown by the significant main effect of familiarity, $F(1, 182) = 55.62$, $p < .001$, $\eta_p^2 = .23$, just as they did for predictions in Experiment 2. Participants in the high familiarity group were nevertheless more confident in the decisions they had made, indicated by a significant main effect of familiarity for decision confidence, $F(1, 182) = 63.41$, $p < .001$, $\eta_p^2 = .26$, again similar to the results of Experiment 2 for predictions.

When it comes to participants' decisions aligning with the system, there was no significant difference between being given a counterfactual or causal explanation, as shown by no main effect of familiarity for decision accuracy, $F(1, 182) = .84$, $p = .36$, in the same way as in Experiment 2. There was also no significant difference between the groups with regards to confidence in their decisions, as shown by no main effect of explanation type for decision confidence, $F(1, 186) = .89$, $p = .35$. The results for Experiment 4 are similar to those for Experiment 2. Additionally, there was also no interaction effect for decision accuracy, $F(1, 182) = .00$, $p = .99$, and no interaction effect for decision confidence $F(1, 182) = 2.51$, $p = .12$. See Figure 11.

Figure 11

Percentage Decision Accuracy in A, and Mean Decision Confidence in B, for high and low familiarity groups in Experiments 1 and 2 focusing on prediction judgements. Error bars represent standard error of mean.



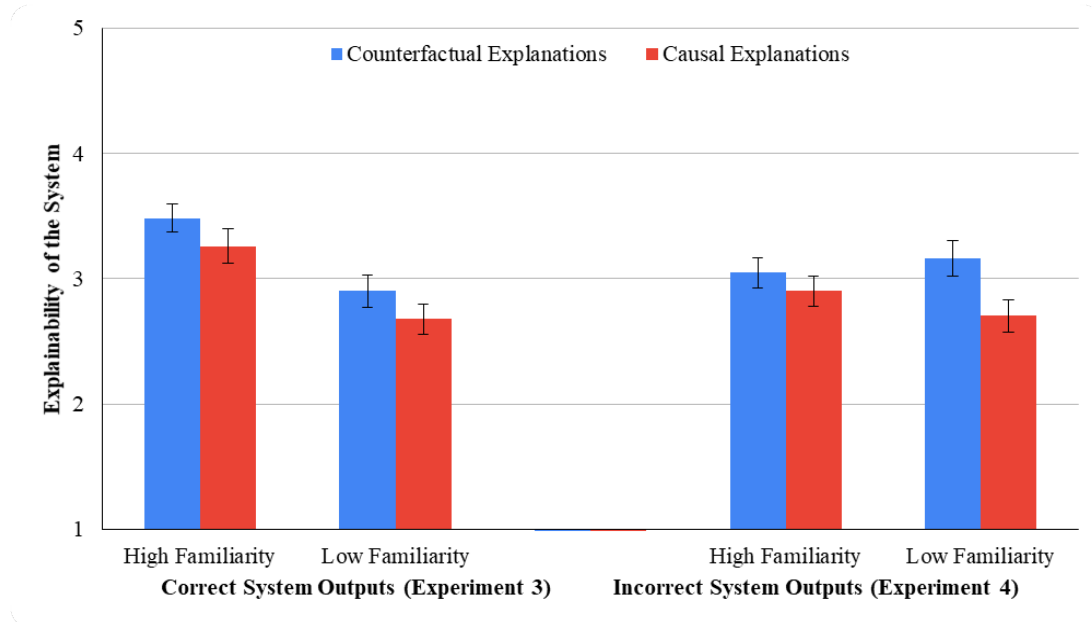
DARPA Explainability and Trust

Participants in Experiment 3 given correct system outputs rated the explainability of the system higher in the familiar domain than in the unfamiliar domain, shown by a significant main effect of familiarity on explanation satisfaction of the system, $F(1, 173) = 14.71$, $p < .001$, $\eta_p^2 = .12$. Participants were somewhat more- satisfied with the explainability of the system given counterfactual than ~~or~~ causal explanations, although the main effect of explanation type was not significant, $F(1, 173) = 3.23$, $p = .07$, $\eta_p^2 = .02$. There was no interaction effect, $F(1, 173) = .00$, $p = 1.00$. The results of the global explanation satisfaction measure are consistent with those of the local explanation helpfulness measure.

Participants in Experiment 4 given incorrect system outputs rated explainability of the system equally in the familiar domain as in the unfamiliar domain, shown by no main effect of familiarity, $F(1, 186) = .11$, $p = .75$. In Experiment 4, participants presented with counterfactual explanations rated the explainability of the system higher than those presented with causal explanations, $F(1, 186) = 14.71$, $p = .02$, $\eta_p^2 = .03$, unlike Experiment 2. There was no interaction effect, $F(1, 186) = 1.48$, $p = .23$. See Figure 12.

Figure 12

Explainability of the system rating for each of the four groups in Experiments 3 and 4. Error bars represent standard error of mean.



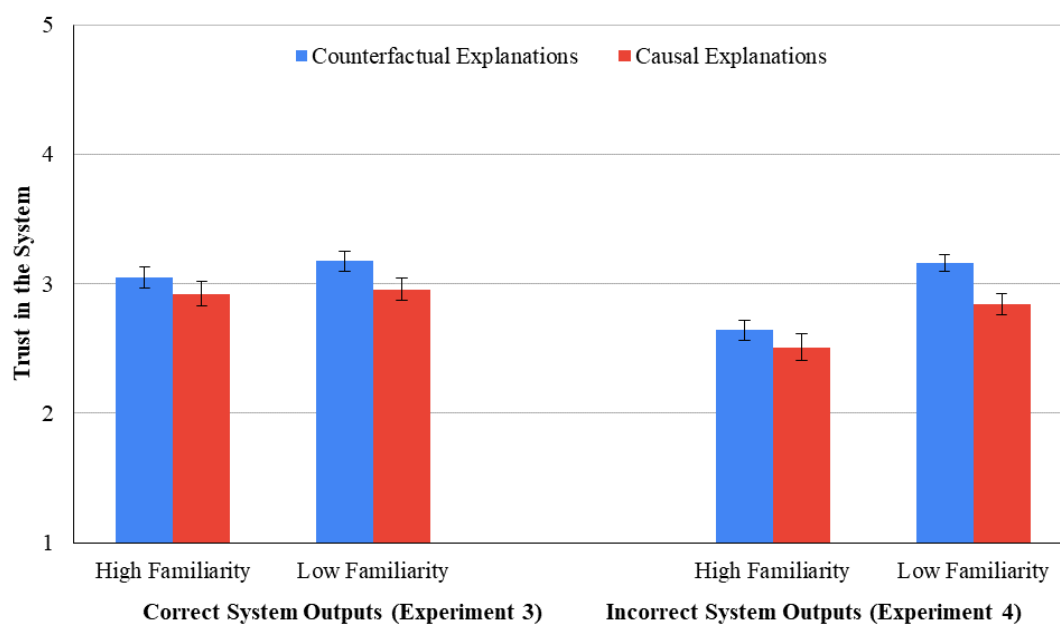
In Experiment 3 with correct system outputs there was no significant difference between participants rating the trustworthiness of the system in the familiar domain and the unfamiliar domain, as shown by no main effect of familiarity, $F(1, 173) = .32$, $p = .57$. However, participants presented with counterfactual explanations rated trust in the system higher than those presented with causal explanations, evident from a significant main effect of explanation type, $F(1, 173) = 3.96$, $p = .048$, $\eta_p^2 = .02$. There was no interaction effect, $F(1, 173) = .32$, $p = .57$.

When presented with incorrect system outputs in Experiment 4, participants rated the system as less trustworthy given the familiar domain than the unfamiliar domain, as indicated by a significant main effect of familiarity, $F(1, 186) = 26.25$, $p < .001$, $\eta_p^2 = .13$. Additionally, participants presented with counterfactual explanations rated trust in the system higher than those presented with causal explanations $F(1, 186) = 7.37$, $p = .007$, $\eta_p^2 = .04$, unlike Experiment 2. There was no interaction effect, $F(1, 186) = 1.27$, $p = .26$. See Figure 13. The results for explainability and trust in both Experiments 3 and 4

show that when people must make their own decisions, they subsequently judge counterfactual explanations more highly than causal ones, whether for correct or incorrect decisions.

Figure 13

Trust in the system for counterfactual and causal explanation groups in experiments 3 and 4 focusing on decision judgements. Error bars represent standard error of mean.



Discussion

The first parts of Experiments 3 and 4 were identical to the first parts of Experiments 1 and 2 respectively, and the explanation helpfulness results in Experiments 3 and 4 in general replicate those of Experiments 1 and 2. Participants judged explanations to be more helpful in a familiar domain than an unfamiliar one for the decisions of a correct AI system in Experiment 3, just as they did in Experiment 1, but to be less helpful in a familiar domain than an unfamiliar one for decisions of an incorrect AI system in Experiment 4, just as they did in Experiment 2. They also considered counterfactual explanations to be somewhat more helpful than causal ones when they considered the decisions of a correct AI system in Experiment 3, as they did in

Experiment 1, although not significantly so in Experiment 3; but when they considered the decisions of an incorrect AI system, i.e., in Experiment 2 and Experiment 4, there was no significant difference between the perceived helpfulness of counterfactuals and casuals.

In the second parts of Experiments 3 and 4, participants made their own decisions, rather than predicting the decisions of the AI system. Participants were more accurate in their decisions, that is, their own decisions aligned with the decisions of a correct AI system, when they were given counterfactual explanations rather than causal ones, in line with our new hypothesis. The results are different from those of prediction accuracy in Experiment 1. They suggest that counterfactual explanations improve participants understanding of the AI system domain more than causal explanations, in that they improve the accuracy of their own decisions, that is, their own decisions align with the AI system's decisions more when they have received counterfactual explanations rather than causal explanations. The result is consistent with proposals that counterfactuals serve both an explanatory and a preparatory function, in that people engage in counterfactual thinking about past events partly so that they can make decisions about what to do in the future (e.g., Roese & Epstude, 2017; Ferrante et al., 2013; O'Connor et al., 2014). Importantly, the result somewhat ameliorates the worrying dissociation between subjective preferences for counterfactual explanations over causal ones, yet no objective improvements in accuracy given counterfactual explanations rather than causal ones. Instead, our result indicates that the dissociation between subjective measures and objective measures may be limited to certain types of objective measures (prediction accuracy), given that in Experiment 3, participants show a tendency in their subjective judgements to consider the counterfactual explanations to be more helpful than the causal ones (although not significantly so at $p = .07$), and they also show an objective increase in

accuracy in that they made more accurate decisions given the counterfactual explanations than the causal ones, in other words, their subjective helpfulness judgments and their objective accuracy were aligned.

Their decisions were more accurate, that is, aligned with the AI system's decisions, in the familiar domain than the unfamiliar one when the AI's decisions were correct in Experiment 3, and less accurate when the AI's decisions were incorrect, i.e., less aligned with its decisions, in Experiment 4, just as their predictions were in Experiments 1 and 2.

Once again their confidence in their decisions was the same whether they were given a familiar or unfamiliar domain, or counterfactual or causal explanations, and whether the AI system made correct or incorrect decisions. The results for explainability and trust in both Experiments 3 and 4 show that when people must make their own decisions, they subsequently judge counterfactual explanations more highly than causal ones, whether for correct or incorrect decisions.

Chapter 4: General Discussion

In this study we examined how people understand and reason with counterfactual and causal explanations of an AI system’s decision, in the familiar domain of alcohol and driving, and the unfamiliar domain of chemical safety, and in circumstances when the AI system made correct decisions, or incorrect decisions. In this chapter we first consider the results for familiarity, summarising the findings for the familiar alcohol and driving domain and the unfamiliar chemical safety domain. We discuss their implications, both for psychological theories and for eXplainable AI, and we consider some of the limitations of the experiments and possible directions for future research on domain familiarity in XAI. Next we consider the results for explanations, summarising the findings for counterfactual explanations such as “you would have been under the limit if you had drunk on a full stomach” and causal explanations such as “you were over the limit because you drank on an empty stomach”. We discuss their implications, again both for psychological theories and for eXplainable AI, and we consider some of the limitations of the experiments and possible directions for future research on explanations in XAI.

The role of familiarity of domain in explanations for AI decisions

The results for familiarity are clear, as shown by Table 4 (the first lines in each section in green), which summarises the results from all four experiments. Participants judged explanations as more helpful, and they made more accurate predictions or decisions, that is, predictions or decisions aligned with the AI system’s decisions, when they reasoned about an AI system in the familiar alcohol and driving domain, than an AI system in the unfamiliar chemical safety domain, when the AI system made correct decisions.

Conversely, they rated explanations in the familiar domain as *less* helpful, and they made *fewer* accurate predictions or decisions, when the AI system made incorrect decisions.

There was no difference in confidence about the accuracy of their predictions or decisions for familiar and unfamiliar domains. Participants judged the global explainability of the AI system higher for the familiar domain than the unfamiliar domain, when it made correct but not incorrect decisions. Conversely, they trust its decisions less in a familiar than an unfamiliar domain when the AI system made incorrect, but not correct decisions.

The results for familiarity are consistent with psychological findings that familiarity of domain content leads to more accurate inferences (see Nickerson, 2015 for a review). Familiarity with a domain elicits people's prior beliefs. According to the mental model theory, familiarity enables people to keep in mind more possibilities, to flesh out their models to be explicit, to search for counterexamples (e.g. Johnson-Laird & Byrne, 2002). It also helps people to eliminate some possibilities from further consideration, which is difficult if not impossible to do, when they have fewer prior beliefs about what leads to a decision an AI system in an unfamiliar domain will make. Our interest in familiarity in our studies was motivated by such proposals in the mental model theory. Nonetheless, alternative psychological theories have proposed different cognitive mechanisms by which familiarity may have an effect, for example, familiarity may alter the probabilities people assign to particular beliefs (e.g., Evans & Over, 2004; Oaksford & Chater, 2007). An interesting direction for future research may be to assess whether such probabilistic theories make predictions about how people reason about explanations for AI decisions that differ from the predictions of the mental model theory in some circumstances, to enable a test of the alternative theories.

Table 4

A summary of the results of the four experiments.

	Experiment 1	Experiment 2	Experiment 3	Experiment 4
<i>Judgements:</i>	Predictions	Predictions	Decisions	Decisions
<i>System outputs:</i>	Correct	Incorrect	Correct	Incorrect
Explanation Helpfulness	Familiar	Familiar	Familiar	Familiar
	>	<	>	<
	Unfamiliar	Unfamiliar	Unfamiliar	Unfamiliar
	Counterfactual	Counterfactual	Counterfactual	Counterfactual
	>	=	$\geq *$	=
Accuracy	Causal	Causal	Causal	Causal
	Familiar	Familiar	Familiar	Familiar
	>	<	>	<
	Unfamiliar	Unfamiliar	Unfamiliar	Unfamiliar
	Counterfactual	Counterfactual	Counterfactual	Counterfactual
Confidence	=	=	>	=
	Causal	Causal	Causal	Causal
	Familiar	Familiar	Familiar	Familiar
	>	>	>	>
	Unfamiliar	Unfamiliar	Unfamiliar	Unfamiliar
Explainability	Counterfactual	Counterfactual	Counterfactual	Counterfactual
	=	=	=	=
	Causal	Causal	Causal	Causal
	Familiar	Familiar	Familiar	Familiar
	>	=	>	=
Trust	Unfamiliar	Unfamiliar	Unfamiliar	Unfamiliar
	Counterfactual	Counterfactual	Counterfactual	Counterfactual
	=	=	$\geq *$	>
	Causal+	Causal	Causal	Causal
	Familiar	Familiar	Familiar	Familiar
	=	<	=	<
	Unfamiliar	Unfamiliar	Unfamiliar	Unfamiliar
	Counterfactual	Counterfactual	Counterfactual	Counterfactual
	=	=	>	>
	Causal	Causal	Causal	Causal

$\geq *$ p = .07

+ Interaction of familiarity and explanation type: see text.

These findings about familiarity have important implications for the field of explainable AI, where familiarity, for example, domain knowledge or expertise, or a lack of it, has not been widely explored (for a review see Ford & Keane, 2022). At its simplest, these findings suggest that explanations in XAI will need to be calibrated to the domain knowledge of the user, a calibration that is currently not undertaken in most AI systems. One limitation of the current study is that it included only two domains, the familiar alcohol and driving domain, and the unfamiliar chemical safety domain. Future research is needed to extend the results to other examples of familiar domains and other examples of unfamiliar domains, such as finance, legislature, or medicine should be explored. In particular, even though the systems were designed to *prevent* decisions that could cause harm (such as being over the limit and driving, or handling an unsafe chemical) the possibility of potential harm might have affected participants' judgments. We mitigated that potential issue by excluding participants with strong beliefs against ever making such decisions. Nonetheless, future studies could overcome this potential limitation by examining less 'sensitive' domains, to test whether decisions with smaller personal implications might yield different results. It may also be fruitful to examine different levels of familiarity with a domain. Finally, we note a further limitation with the content used in the experiments is that the alcohol and driving AI system was partially chosen also because it is designed with a limited number (5) of easy-to-understand features (gender, weight, units, duration, stomach fullness). In reality, AI systems often consist of many more features, therefore future research could also consider utilising more complex systems.

Familiarity plays a role in how people evaluate individual, local explanations that an AI system provides, in other words, if a user has prior beliefs about the domain of the system, explanations are considered more helpful. However, additionally, people are able

to identify incorrect decisions made by an AI system in a familiar domain and hence, individual explanations are considered less helpful for familiar domains in which an AI system makes incorrect decisions. An important implication is that intolerance for incorrect AI decisions may be particularly pronounced in familiar domains, and explanations may be particularly unhelpful in such circumstances.

Moreover, when presented with correct system outputs, global judgments about the explainability of the system align with local judgments of the helpfulness of explanations. However, when presented with incorrect system outputs, even though individual explanations were rated as *less* helpful, the entire system was perceived as *more* explainable. The result suggests that assessments of individual local explanations may provide a more fine-grained sensitive measure of explainability than assessments of global explanation satisfaction which may provide a broader impressionistic measure.

Intriguingly, our findings suggest that users of an XAI system appear to believe that the system is trying to reach the correct decision and is capable of doing so. This belief is evident from the finding that when users are familiar with the system's domain and are presented with a sequence of incorrect system outputs, their prediction about what the next decision the system will make does not correspond to the previous, entirely incorrect decisions the system made, but rather they predict the system will make a decision based on their own understanding of the domain. The incorrect system outputs in the present study were designed in a way that both *prediction* and the *explanation* part of the output were entirely incorrect. Compared to the correct system output, the relationship between the feature and the outcome of an incorrect system output was inverted, leading to predictions and explanations a user of the XAI system familiar with the domain would recognize as incorrect (for example a high number of units of alcohol drank would lead to the prediction that the person is under the limit). One potential limitation of these

materials is the implied direction of one feature in the unfamiliar domain, the “occupational exposure limit”, potentially indicating a higher amount is safer than a lower amount, and thus making the system more informative for its users. A potential future direction for unfamiliar domain materials would be to make features entirely undescriptive (for example Feature A-E), however it is important to note that in a real world scenario, depending on the feature, some features of an XAI system would be descriptive by design to help its users understand the direction of the influence on the outcome. This admittedly varies from system to system and is entirely feature-dependent, as it is also the case in the present study – the relationship between the outcome and the two binary features (Air pollution rating and PNEC rating) or the outcome and the direction of the two continuous features (pH and Exposure duration) could not be discernible from the feature name only and the users had to rely on the system output to understand the prediction the system would make or make a decision. The present study explored multiple boundary conditions, one being our app providing users with entirely correct or entirely incorrect outputs. In reality, it is highly unlikely that an XAI system would make decisions in such a way, and research suggests that users generally have a very low tolerance for error in AI systems, with some studies suggesting that even a low error rate of 3-4% undermines people’s trust in the system (Kenny et al., 2021). One avenue for future research in XAI is thus the detailed examination of how people understand explanations in AI systems with different levels of incorrect decisions, and for different levels of familiarity with a domain. It is important to note that the use of an XAI system often precedes a human decision, for example, *“if you had asked for a lower amount, your loan application would have been approved”* (Dai, et al., 2022; Warren et al., 2022a) can lead to a bank clerk rejecting a loan. Our findings suggest that users of an XAI system will elect to make decisions based on their prior beliefs rather than follow the

AI support system, especially when actions might affect them personally. This implication arises from our findings that users in familiar domains make decisions that would align with correct AI system decisions, rather than incorrect AI system decisions.

The role of counterfactual and causal explanations in how people understand AI decisions

The results for explanations are also clear, as shown by Table 4. Participants judged counterfactual explanations such as “he would have been over the limit if he had drunk on an empty stomach”, as more helpful than causal ones “he was under the limit because he drank on a full stomach”, for correct AI decisions (significantly in Experiment 1 and somewhat so in Experiment 3), whereas there was no difference in helpfulness ratings for incorrect AI decisions.

Despite their subjective judgments that counterfactual explanations were more helpful than causal ones, on the objective measure of the accuracy of their predictions of the AI’s decisions (in Experiment 1), there was no difference in accuracy whether they were given counterfactual explanations or causal ones. The result replicates the previous findings of a dissociation of subjective and objective measures, extends it to unfamiliar domains, and shows it occurs even when subjective measures rely on local individual explanation helpfulness judgments rather than global ones (Warren et al., 2022a).

However, when participants made their own decisions (in Experiment 3), there was no such dissociation between subjective and objective measures. Participants tended to judge subjectively the counterfactual explanations to be somewhat more helpful than the causal ones (although the difference was not significant at $p = .07$), and importantly, objectively the accuracy of their own decisions increased given the counterfactual explanations more than the causal ones. Their decisions were more accurate about good outcomes when they had been given a counterfactual, e.g., “*He would have been over the limit if he had drunk*

on an empty stomach”, rather than a causal, e.g., “*He was under the limit because he drank on a full stomach*”; but their decisions were more accurate about bad outcomes when they had been given a causal, e.g., “*He was over the limit because he drank on an empty stomach*”, rather than a counterfactual, e.g., “*He would have been under the limit if he had drunk on a full stomach*”. Explanations that focus on a bad outcome, e.g., *over the limit*, whether as a potential alternative (“*he would have been over the limit if...*”) or as a real occurrence (“*he was over the limit because...*”) improve the accuracy of their decisions. Of note also is that for incorrect decisions of an AI system (Experiments 2 and 4) there was no subjective preference for counterfactual explanations over causal ones, and no objective difference in accuracy of predictions or decisions.

There was no difference in participants’ confidence about the accuracy of their predictions or decisions when they had been given counterfactual or causal explanations. When participants had attempted to predict the decisions of the AI system (in Experiments 1 and 2), their subsequent global judgments of explainability and trust in the system were the same given counterfactual explanations and causal ones; but when they had attempted to make their own decisions (in Experiments 3 and 4), their subsequent global judgments of explainability and trust in the system were higher given counterfactual explanations than causal ones.

The results on explanations have implications for XAI. The present study replicated the findings of previous XAI research (Warren et al., 2022a) showing that counterfactual explanations are judged subjectively to be more helpful than causal explanations, and it extends the results by showing the difference occurs for local explanation helpfulness judgments (after each explanation), not only for the global explanation satisfaction judgments (at the end of the study) studied previously (Warren et al., 2022a). The present study shows that such findings of a preference for counterfactual

over causal explanations occurs for unfamiliar domains as well as familiar domains, and it also adds the information that such findings are limited to correct system outputs.

Importantly, the study adds to the findings of a dissociation between subjective preferences for counterfactual explanations over causal ones, and objective improvements in accuracy given counterfactual explanations rather than causal ones, by amending it somewhat, in that participants show a tendency towards a subjective preference for counterfactual explanations over causal ones for the correct AI system's decisions. They also show an objective increase in their own decision accuracy when they had been given a counterfactual explanation, but only when deciding about a good outcome. Their decisions were more accurate about bad outcomes when they had been given a causal. Hence our results may be considered to somewhat ameliorate the ethical concerns over the dissociation between subjective preferences for counterfactual explanations over causal ones, with no objective improvements in accuracy given counterfactual explanations rather than causal ones. Instead, our results indicate that the dissociation between subjective measures and objective measures may be limited to certain types of objective measures (prediction accuracy), and, to some extent, not others (decision accuracy). If in some circumstances counterfactual explanations for AI decisions are not only preferred by people, but also help them to improve the accuracy of their own decisions, then they may indeed be better than causal explanations in such circumstances.

One limitation of the current study is that we compared counterfactual explanations only to causal explanations, and one fruitful direction for future research would be to compare counterfactual explanations to other sorts of explanations, including even to non-explanation descriptions. The results also indicate that the use of counterfactual explanations in XAI may need to be calibrated to the error potential of the AI decision support system. However, the incorrect AI decisions that we examined were

accompanied by explanations that were also incorrect, which may have made the incorrectness of the AI system particularly noticeable, and so future research to examine incorrect AI decisions accompanied by correct explanations may also be useful.

The results also have important implications for the psychology of counterfactual and causal explanations. The effects of counterfactual explanations on improving accuracy in making one's own decisions, but not on improving accuracy in predicting the decisions of an AI system, are consistent with proposals that counterfactuals are personal and goal-related. For example, Roese and Epstude (2017, p.17) suggest that counterfactual thoughts "*focus on what one personally could have done to have achieved the goal, or to have bypassed some obstacle so as to achieve a more desirable outcome*" and that they are "*personal (as opposed to focused on others)*". Since counterfactual thoughts are personal, it makes sense that they would lead to higher accuracy only when people are asked to make a decision; asking about what prediction a system will make removes the user from the decision and focuses their attention away from the personal goal. People engage in counterfactual thinking about past events partly so that they can make decisions about what to do in the future (e.g., Roese & Olson, 2017; Ferrante et al., 2013; O'Connor et al., 2014). Our finding that when users are presented with correct system outputs and asked to make a decision, counterfactual explanations lead to more accurate decisions that are closer to the system's decision, may reflect the tendency for counterfactual thoughts to serve both an explanatory and a preparatory function. The findings are consistent with the proposal that people understand counterfactuals by envisaging dual possibilities at the outset, whereas they understand causal assertions initially by envisaging a single possibility (e.g., Byrne, 2005; Orenes et al., 2022). The findings suggest that although the representation of dual possibilities for counterfactuals may require more cognitive resources than the representation of a single possibility for

causal assertions, this cost is outweighed by the benefits of the richer representation of counterfactuals. The richer representation of counterfactuals provides more information explicitly which appears to be helpful to people in forming their own decisions, aligned with those of an AI decision support system.

References

- Bae, J., Cha, Y.-J., Lee, H., Lee, B., Baek, S., Choi, S., & Jang, D. (2017). Social networks and inference about unknown events: A case of the match between Google's AlphaGo and Sedol Lee. *PLOS ONE*, 12(2).
doi:10.1371/journal.pone.0171472
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*. 58(1), 82-115. doi:10.1016/j.inffus.2019.12.012
- Branscombe, N. R., Owen, S., Garstka, T. A., & Coleman, J. (1996). Rape and accident counterfactuals: Who might have done otherwise and would it have changed the outcome? *Journal of Applied Social Psychology*, 26(12), 1042–1067.
- Byrne, R. M. J. (2005). *The rational imagination: How people create alternatives to reality*. Cambridge, MA: MIT Press.
- Byrne, R. M. J. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. *IJCAI*, 6276 – 6282.
- Byrne, R.M.J. & Johnson-Laird, P.N. (2019). *If and or: real and counterfactual possibilities in their truth and probability*. *Journal of Experimental Psychology: Learning, Memory & Cognition*. 46 (4), 760-780.
- Byrne, R. M. J., & Tasso, A. (1999) Deductive reasoning with factual, possible, and counterfactual conditionals. *Memory & Cognition*, 27(4), 726–740.
- Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare. *KDD '15: The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
<https://doi.org/10.1145/2783258.2788613>

- Catellani, P., & Covelli, V. (2013). The strategic use of counterfactual communication in politics. *Journal of Language and Social Psychology*, 32(4), 480–489.
- Dai, X., Keane, M.T., Shalloo, L., Ruelle, E., & Byrne, R.M.J. (2022). Counterfactual explanations for prediction and diagnosis in XAI. *Proceedings of the AIES conference, AIES'22*. Oxford, UK.
- De Brigard, F., Addis, D. R., Ford, J. H., Schacter, D. L., & Giovanello, K. S. (2013). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia*, 51(12), 2401–2414.
- Deeks, A. (2019). The Judicial Demand for Explainable Artificial Intelligence. *Virginia Public Law and Legal Theory Research Paper*, 119(51). Retrieved from <https://ssrn.com/abstract=3440723>
- Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford, UK: Oxford University Press.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4), 1149-1160.
- Ferrante, D., Giroto, V., Strag., 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.
- Ford, C., & Keane, M.T. (2022, August). *Explaining Classifications to Non-Experts: An XAI User Study of Post-Hoc Explanations for a Classifier When People Lack Expertise* [Paper presentation]. International conference in pattern recognition 2022, Montreal, Canada.
- Förster, M., Hühn, P., Klier, M. and Kluge, K. (2021). Capturing Users' Reality: A Novel Approach to Generate Coherent Counterfactual Explanations. *Proceedings of the 54th Hawaii International Conference on System Sciences*, 1274.

- Frosch, C., & Byrne, R. M. J. (2012). Causal conditionals and counterfactuals. *Acta Psychologica*, 14, 54–66.
- Gerstenberg, T., Goodman, N.D., Lagnado, D.A., and Tenenbaum, J.B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*. 128, 5 (2021), 936–975.
- Gignac, G. E. (2019). *How2statsbook* (online Edition 1). Perth, WA: Gignac G
- Giroto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica*, 78, 111–133.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A Survey Of Methods For Explaining Black Box Models (Version 3). *arXiv*.
<https://doi.org/10.48550/ARXIV.1802.01933>
- Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018) Metrics for Explainable AI: Challenges and Prospects. *XAI Metrics*. Retrieved from arXiv:1708.08296.
- Johnson, B. N., Kim, W. J., Blue, J., Summerville, A., & Kirkmeyer, B. P. (2021). Gender differences in the functionality of regret on academic performance. In 2021 CoNECD.
- Johnson-Laird, P. N. (1983). *Mental Models. Towards a Cognitive Science of Language, Inference and Consciousness*. Cambridge, UK: Cambridge University Press.
- Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, 109, 646–678.
- 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). New York: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1982b). The psychology of preferences. *Scientific American*, 246(1), 160–173.

- Karimi, A.H., Barthe, G., Schölkopf, B. and Valera, I. 2020. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv:2010.04050*. Retrieved from <https://arxiv.org/abs/2010.04050>
- Keane, M.T., & Barry Smyth. 2020. Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI). In *International Conference on Case-Based Reasoning*, Springer, Cham, 163–178.
- Keane, M.T., Kenny, E.M., Delaney, E., and Smyth, B. (2021). If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. *IJCAI-21*.
- Keil, F. C. (2006). Explanation and understanding. *Annual review of psychology*, 57, 227.
- Kenny, E. M., Ford, C., Quinn, M. and Keane, M. T. (2021). Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, 294, 103459.
- Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J. and Doshi-Velez, F. 2019. Human evaluation of models built for interpretability. *Proceedings of the AAAI conference on Human Computation and Crowdsourcing*, 7, 59-67.
- Lagnado, D. A., Gerstenberg, T., & Zultan, R. I. (2013). Causal responsibility and counterfactuals. *Cognitive science*, 37(6), 1036-1073.
- Levy, I., & Ben-David, S. (2015). Mechanism of bystander-blaming: Defensive attribution, counterfactual thinking, and gender. *International journal of offender therapy and comparative criminology*, 59(1), 96-113.
- Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive psychology*, 55(3), 232-257.

- Lucic, A., Haned, H. and de Rijke, M. (2020). Why does my model fail? Contrastive local explanations for retail forecasting. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 90-98.
- Malle, B. F., Monroe, A. E., & Guglielmo, S. (2014). A theory of blame. *Psychological Inquiry*, 25(2), 147–186.
- Mandel, D. R., & Lehman, D. R. (1996). Counterfactual thinking and ascriptions of cause and preventability. *Journal of Personality and Social Psychology*, 71(3), 450–463.
- Markman, K. D., & Tetlock, P. E. (2000). I couldn't have known: Accountability, foreseeability, and counterfactual denials of responsibility. *British Journal of Social Psychology*, 39, 313–325.
- Markman, K. D., McMullen, M. N., & Elizaga, R. A. (2008a). Counterfactual thinking, persistence, and performance: A test of the reflection and evaluation model. *Journal of Experimental Social Psychology*, 44(2), 421–428.
- Markman, K. D., Mizoguchi, N., & McMullen, M. N. (2008b). "It would have been worse under Saddam:" Implications of counterfactual thinking for beliefs regarding the ethical treatment of prisoners of war. *Journal of Experimental Social Psychology*, 44(3), 650-654.
- McCormack, T., O'Connor, E., Beck, S., & Feeney, A. (2016). The development of regret and relief about the outcomes of risky decisions. *Journal of experimental child psychology*, 148, 1-19.
- McCloy, R., & Byrne, R. M. J. (2002). Semifactual "even if" thinking. *Thinking & Reasoning*, 8, 41–67.

- McCrea, S. M. (2008). Self-handicapping, excuse making, and counterfactual thinking: Consequences for self-esteem and future motivation. *Journal of Personality and Social Psychology*, 95, 274–292.
- McEleney, A., & Byrne, R. M. J. (2006). Spontaneous causal and counterfactual thoughts. *Thinking and Reasoning*, 12, 235–255.
- Meder, B., Gerstenberg, T., Hagmayer, Y., & Waldmann, M. R. (2010). Observing and intervening: Rational and heuristic models of causal decision making. *The Open Psychology Journal*, 3, 119–135.
- Migliore, S., Curcio, G., Mancini, F., & Cappa, S. F. (2014). Counterfactual thinking in moral judgment: an experimental study. *Frontiers in psychology*, 5, 451.
- Miller, T. (2019). *Explanation in artificial intelligence: Insights from the social sciences*. *Artificial intelligence*, 267, 1-38.
- Morris, M. N., & Moore, P. C. (2000). The lessons we (don't) learn: Counterfactual thinking and organizational accountability after a close call. *Administrative Science Quarterly*, 45, 737–765.
- Nickerson, R. (2015). *Conditional reasoning*. Oxford: Oxford University Press.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. New York, NY: Oxford University Press
- O'Connor, E., McCormack, T. & Feeney, A. (2014). Do Children Who Experience Regret Make Better Decisions? A Developmental Study of the Behavioral Consequences of Regret. *Child Development*, 85, 1995-2010.
<https://doi.org/10.1111/cdev.12253>
- Orenes, I., Espino, O., & Byrne, R. M. (2021). 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.
- Parkinson, M. & Byrne, R. M. J. (2017). Counterfactual and semifactual thoughts in moral judgments about failed attempts to harm. 23(4), 409-448
- Quinn, M. S., Campbell, K., & Keane, M. T. (2021). Do we “fear for the worst” or “Hope for the best” in thinking about the unexpected: Factors affecting the valence of unexpected outcomes reported for everyday scenarios. *Cognition*, 208, 104520.
- Ragni, M., Kola, I., & Johnson-Laird, P. N. (2018). On selecting evidence to test hypotheses: A theory of selection tasks. *Psychological Bulletin*, 144(8), 779.
- Rai, A. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, 48, 137-141. doi:10.1007/s11747-019-00710-5
- Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, 121(1), 133.
- Roese, N. J., & Epstude, K. (2017). The functional theory of counterfactual thinking: New evidence, new challenges, new insights. In J. M. Olson (Ed.), *Advances in experimental social psychology*, 1–79. Elsevier Academic Press.
- Smallman, R., & Roese, N. J. (2009). Counterfactual thinking facilitates behavioral intentions. *Journal of experimental social psychology*, 45(4), 845-852.
- Spellman, B. A., & Mandel, D. R. (1999). When possibility informs reality: Counterfactual thinking as a cue to causality. *Current Directions in Psychological Science*, 8(4), 120–123.
- Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, 57(1), 31-95.
- van der Waa, J., Nieuwburg, E., Cremers, A. and Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291.

- Verma, S., Dickerson, J. and Hines, K. (2020). Counterfactual explanations for machine learning: A review. Retrieved from <https://arxiv.org/abs/2010.10596>
- Wachter, S., Mittelstadt, B. and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31 841.
- Warren, G., Keane, M.T. & Byrne, R.M.J. (2022a). Features of explainability: how users understand counterfactual and causal explanations for categorical and continuous features in XAI. *Cognitive Aspects of Knowledge Representation Workshop, IJCAI-22*
- Warren, G., Smyth, B., Keane, M.T. (2022b). “Better” Counterfactuals, Ones People Can Understand: Psychologically-Plausible Case-Based Counterfactuals Using Categorical Features for Explainable AI (XAI). *Lecture Notes in Computer Science* 134(5). https://doi.org/10.1007/978-3-031-14923-8_5
- Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20, 273–281.
- Wojciech, S. & Wiegand, T.s & Müller, K. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ITU Journal: ICT Discoveries*, 1(1). Retrieved from arXiv:1708.08296.

Appendix A

High familiarity cases with correct system outputs

Cases with counterfactual explanations

Here's information about Walter,
who's in a bar:

Walter	
Weight	80 kg
Units of alcohol drank	6
Duration	30 mins
Gender	Male
Stomach	Full
<u>Prediction:</u> The app predicts Walter is over the limit. <u>Explanation:</u> Walter's blood alcohol level would have been under the limit, if he had drunk for a longer time.	

Here's information about Jonny,
who's in a bar:

Jonny	
Weight	82 kg
Units of alcohol drank	7
Duration	210 mins
Gender	Male
Stomach	Empty
<u>Prediction:</u> The app predicts Jonny is under the limit. <u>Explanation:</u> Jonny's blood alcohol level would have been over the limit, if he had drunk for a shorter time.	

Here's information about Lillian,
who's in a bar:

Lillian	
Weight	71 kg
Units of alcohol drank	5
Duration	60 mins
Gender	Female
Stomach	Full
<u>Prediction:</u> The app predicts Lillian is over the limit. <u>Explanation:</u> Lillian's blood alcohol level would have been under the limit, if she had drunk for a longer time.	

Here's information about Violet,
who's in a bar:

Violet	
Weight	71 kg
Units of alcohol drank	5
Duration	180 mins
Gender	Female
Stomach	Empty
<u>Prediction:</u> The app predicts Violet is under the limit. <u>Explanation:</u> Violet's blood alcohol level would have been over the limit, if she had drunk for a shorter time.	

Here's information about Jarred,
who's in a bar:

Jarred	
Weight	84 kg
Units of alcohol drank	7
Duration	105 mins
Gender	Male
Stomach	empty
<u>Prediction:</u> The app predicts Jarred is over the limit. <u>Explanation:</u> Jarred's blood alcohol level would have been under the limit, if he had drunk on a full stomach.	

Here's information about Elliot, who's
in a bar:

Elliot	
Weight	83 kg
Units of alcohol drank	7
Duration	120 mins
Gender	Male
Stomach	full
<u>Prediction:</u> The app predicts Elliot is under the limit. <u>Explanation:</u> Elliot's blood alcohol level would have been over the limit, if he had drunk on an empty stomach.	

Here's information about Loreen, who's in a bar:

Loreen	
Weight	72 kg
Units of alcohol drank	5
Duration	90 mins
Gender	Female
Stomach	full
Prediction: The app predicts Loreen is under the limit . Explanation: Loreen's blood alcohol level would have been over the limit, if she had drunk on a empty stomach.	

Here's information about Georgina, who's in a bar:

Tyler	
Weight	66 kg
Units of alcohol drank	5
Duration	120 mins
Gender	Female
Stomach	Empty
Prediction: The app predicts Georgina is over the limit . Explanation: Georgina's blood alcohol level would have been under the limit, if she had drunk on a full stomach.	

Here's information about Noah, who's in a bar:

Noah	
Weight	90 kg
Units of alcohol drank	8
Duration	90 mins
Gender	Male
Stomach	Full
Prediction: The app predicts Noah is over the limit . Explanation: Noah's blood alcohol level would have been under the limit, if he had drunk fewer units of alcohol.	

Here's information about Bryan, who's in a bar:

Bryan	
Weight	74 kg
Units of alcohol drank	5
Duration	120 mins
Gender	Male
Stomach	Empty
Prediction: The app predicts Bryan is under the limit . Explanation: Bryan's blood alcohol level would have been over the limit, if he had drunk more units of alcohol.	

Here's information about Zoe, who's in a bar:

Zoe	
Weight	68 kg
Units of alcohol drank	4
Duration	105 mins
Gender	Female
Stomach	Empty
Prediction: The app predicts Zoe is under the limit . Explanation: Zoe's blood alcohol level would have been over the limit, if she had drunk more units of alcohol.	

Here's information about Betty, who's in a bar:

Betty	
Weight	74 kg
Units of alcohol drank	6
Duration	120 mins
Gender	Female
Stomach	Full
Prediction: The app predicts Betty is over the limit . Explanation: Betty's blood alcohol level would have been under the limit, if she had drunk fewer units of alcohol.	

Here's information about Leo, who's in a bar:

Leo	
Weight	97 kg
Units of alcohol drank	7
Duration	105 mins
Gender	Male
Stomach	Empty
<u>Prediction:</u> The app predicts Leo is under the limit . <u>Explanation:</u> Leo's blood alcohol level would have been over the limit, if his weight had been lighter.	

Here's information about Stephen, who's in a bar:

Stephen	
Weight	70 kg
Units of alcohol drank	6
Duration	90 mins
Gender	Male
Stomach	Full
<u>Prediction:</u> The app predicts Stephen is over the limit . <u>Explanation:</u> Stephen's blood alcohol level would have been under the limit, if his weight had been heavier.	

Here's information about Angelica, who's in a bar:

Angelica	
Weight	53 kg
Units of alcohol drank	4
Duration	90 mins
Gender	Female
Stomach	Full
<u>Prediction:</u> The app predicts Angelica is over the limit . <u>Explanation:</u> Angelica's blood alcohol level would have been under the limit, if her weight had been heavier.	

Here's information about Anne, who's in a bar:

Anne	
Weight	79 kg
Units of alcohol drank	5
Duration	120 mins
Gender	Female
Stomach	Empty
<u>Prediction:</u> The app predicts Anne is under the limit . <u>Explanation:</u> Anne's blood alcohol level would have been over the limit, if her weight had been lighter.	

Attention Check

Here's information about Tyler, who's in a bar:

Tyler	
Weight	88 kg
Units of alcohol drank	8
Duration	120 mins
Gender	Male
Stomach	Full

Cases with causal explanations

Here's information about Walter,
who's in a bar:

Walter	
Weight	80 kg
Units of alcohol drank	6
Duration	30 mins
Gender	Male
Stomach	Full
<u>Prediction:</u> The app predicts Walter is over the limit . <u>Explanation:</u> Walter's blood alcohol level was over the limit, because he drank for a short time.	

Here's information about Jonny,
who's in a bar:

Jonny	
Weight	82 kg
Units of alcohol drank	7
Duration	210 mins
Gender	Male
Stomach	Empty
<u>Prediction:</u> The app predicts Jonny is under the limit . <u>Explanation:</u> Jonny's blood alcohol level was under the limit, because he drank for a long time.	

Here's information about Lillian,
who's in a bar:

Lillian	
Weight	71 kg
Units of alcohol drank	5
Duration	60 mins
Gender	Female
Stomach	Full
<u>Prediction:</u> The app predicts Lillian is over the limit . <u>Explanation:</u> Lillian's blood alcohol level was over the limit, because she drank for a short time.	

Here's information about Violet,
who's in a bar:

Violet	
Weight	71 kg
Units of alcohol drank	5
Duration	180 mins
Gender	Female
Stomach	Empty
<u>Prediction:</u> The app predicts Violet is under the limit . <u>Explanation:</u> Violet's blood alcohol level was under the limit, because she drank for a long time.	

Here's information about Jarred,
who's in a bar:

Jarred	
Weight	84 kg
Units of alcohol drank	7
Duration	105 mins
Gender	Male
Stomach	empty
<u>Prediction:</u> The app predicts Jarred is over the limit . <u>Explanation:</u> Jarred's blood alcohol level was over the limit, because he drank on an empty stomach.	

Here's information about Elliot, who's
in a bar:

Elliot	
Weight	83 kg
Units of alcohol drank	7
Duration	120 mins
Gender	Male
Stomach	full
<u>Prediction:</u> The app predicts Elliot is under the limit . <u>Explanation:</u> Elliot's blood alcohol level was under the limit, because he drank on a full stomach.	

Here's information about Loreen, who's in a bar:

Loreen	
Weight	72 kg
Units of alcohol drank	5
Duration	90 mins
Gender	Female
Stomach	full
<u>Prediction:</u> The app predicts Loreen is under the limit . <u>Explanation:</u> Loreen's blood alcohol level was under the limit, because she drank on a full stomach.	

Here's information about Georgina, who's in a bar:

Georgina	
Weight	66 kg
Units of alcohol drank	5
Duration	120 mins
Gender	Female
Stomach	empty
<u>Prediction:</u> The app predicts Georgina is over the limit . <u>Explanation:</u> Georgina's blood alcohol level was over the limit, because she drank on an empty stomach.	

Here's information about Noah, who's in a bar:

Noah	
Weight	90 kg
Units of alcohol drank	8
Duration	90 mins
Gender	Male
Stomach	Full
<u>Prediction:</u> The app predicts Noah is over the limit . <u>Explanation:</u> Noah's blood alcohol level was over the limit, because he drank many units of alcohol.	

Here's information about Bryan, who's in a bar:

Bryan	
Weight	74 kg
Units of alcohol drank	5
Duration	120 mins
Gender	Male
Stomach	Empty
<u>Prediction:</u> The app predicts Bryan is under the limit . <u>Explanation:</u> Bryan's blood alcohol level was under the limit, because he drank few units of alcohol.	

Here's information about Zoe, who's in a bar:

Zoe	
Weight	68 kg
Units of alcohol drank	4
Duration	105 mins
Gender	Female
Stomach	Empty
<u>Prediction:</u> The app predicts Zoe is under the limit . <u>Explanation:</u> Zoe's blood alcohol level was under the limit, because she drank few units of alcohol.	

Here's information about Betty, who's in a bar:

Betty	
Weight	74 kg
Units of alcohol drank	6
Duration	120 mins
Gender	Female
Stomach	Full
<u>Prediction:</u> The app predicts Betty is over the limit . <u>Explanation:</u> Betty's blood alcohol level was over the limit, because she drank many units of alcohol.	

Here's information about Leo, who's in a bar:

Leo	
Weight	97 kg
Units of alcohol drank	7
Duration	105 mins
Gender	Male
Stomach	Empty
Prediction: The app predicts Leo is under the limit . Explanation: Leo's blood alcohol level was under the limit, because his weight was heavy.	

Here's information about Stephen, who's in a bar:

Stephen	
Weight	70 kg
Units of alcohol drank	6
Duration	90 mins
Gender	Male
Stomach	Full
Prediction: The app predicts Stephen is over the limit . Explanation: Stephen's blood alcohol level was over the limit, because his weight was light.	

Here's information about Angelica, who's in a bar:

Angelica	
Weight	53 kg
Units of alcohol drank	4
Duration	90 mins
Gender	Female
Stomach	Full
Prediction: The app predicts Angelica is over the limit . Explanation: Angelica's blood alcohol level was over the limit, because her weight was light.	

Here's information about Anne, who's in a bar:

Anne	
Weight	79 kg
Units of alcohol drank	5
Duration	120 mins
Gender	Female
Stomach	Empty
Prediction: The app predicts Anne is under the limit . Explanation: Anne's blood alcohol level was under the limit, because her weight was heavy.	

Attention Check

Here's information about Tyler, who's in a bar:

Tyler	
Weight	88 kg
Units of alcohol drank	8
Duration	120 mins
Gender	Male
Stomach	Full

Cases used for prediction judgements

Here's information about Charles, who's in a bar:

Charles	
Weight	84 kg
Units of alcohol drank	7
Duration	75 mins
Gender	Male
Stomach	Full

Here's information about Oliver, who's in a bar:

Oliver	
Weight	80 kg
Units of alcohol drank	6
Duration	135 mins
Gender	Male
Stomach	Empty

Here's information about Holly, who's in a bar:

Holly	
Weight	68 kg
Units of alcohol drank	4
Duration	30 mins
Gender	Female
Stomach	Empty

Here's information about Millie, who's in a bar:

Millie	
Weight	65 kg
Units of alcohol drank	5
Duration	180 mins
Gender	Female
Stomach	Full

Here's information about Chris, who's in a bar:

Chris	
Weight	83 kg
Units of alcohol drank	7
Duration	105 mins
Gender	Male
Stomach	Empty

Here's information about Richard, who's in a bar:

Richard	
Weight	80 kg
Units of alcohol drank	6
Duration	105 mins
Gender	Male
Stomach	Full

Here's information about Amelia,
who's in a bar:

Amelia	
Weight	66 kg
Units of alcohol drank	5
Duration	135 mins
Gender	Female
Stomach	Full

Here's information about Molly, who's
in a bar:

Molly	
Weight	69 kg
Units of alcohol drank	5
Duration	105 mins
Gender	Female
Stomach	Empty

Here's information about Michael,
who's in a bar:

Michael	
Weight	85 kg
Units of alcohol drank	6
Duration	90 mins
Gender	Male
Stomach	Full

Here's information about Thomas,
who's in a bar:

Thomas	
Weight	84 kg
Units of alcohol drank	8
Duration	150 mins
Gender	Male
Stomach	Empty

Here's information about Sophie,
who's in a bar:

Sophie	
Weight	65 kg
Units of alcohol drank	4
Duration	90 mins
Gender	Female
Stomach	Empty

Here's information about Chloe,
who's in a bar:

Chloe	
Weight	72 kg
Units of alcohol drank	6
Duration	120 mins
Gender	Female
Stomach	Full

Here's information about David,
who's in a bar:

David	
Weight	85 kg
Units of alcohol drank	6
Duration	135 mins
Gender	Male
Stomach	Empty

Here's information about Liam, who's
in a bar:

Liam	
Weight	80 kg
Units of alcohol drank	7
Duration	90 mins
Gender	Male
Stomach	Full

Here's information about Amber,
who's in a bar:

Amber	
Weight	65 kg
Units of alcohol drank	5
Duration	90 mins
Gender	Female
Stomach	Full

Here's information about Mary, who's
in a bar:

Mary	
Weight	74 kg
Units of alcohol drank	5
Duration	120 mins
Gender	Female
Stomach	Empty

Attention Check

Here's information about Benjamin,
who's in a bar:

Benjamin	
Weight	55 kg
Units of alcohol drank	5
Duration	195 mins
Gender	Male
Stomach	Empty

Cases used in decision judgements

Imagine this is information about you in a bar.

Weight	84 kg
Units	7
Duration	75 mins
Gender	Male
Stomach content	Full

Imagine this is information about you in a bar.

Weight	80 kg
Units	6
Duration	135 mins
Gender	Male
Stomach content	Empty

Imagine this is information about you in a bar.

Weight	68 kg
Units	4
Duration	30 mins
Gender	Female
Stomach content	Empty

Imagine this is information about you in a bar.

Weight	65 kg
Units	5
Duration	180 mins
Gender	Female
Stomach content	Full

Imagine this is information about you in a bar.

Weight	83 kg
Units	7
Duration	105 mins
Gender	Male
Stomach content	Empty

Imagine this is information about you in a bar.

Weight	80 kg
Units	6
Duration	105 mins
Gender	Male
Stomach content	Full

Imagine this is information about you in a bar.

Weight	66 kg
Units	5
Duration	135 mins
Gender	Female
Stomach content	Full

Imagine this is information about you in a bar.

Weight	69 kg
Units	5
Duration	105 mins
Gender	Female
Stomach content	Empty

Imagine this is information about you in a bar.

Weight	85 kg
Units	6
Duration	90 mins
Gender	Male
Stomach content	Full

Imagine this is information about you in a bar.

Weight	84 kg
Units	8
Duration	150 mins
Gender	Male
Stomach content	Empty

Imagine this is information about you in a bar.

Weight	65 kg
Units	4
Duration	90 mins
Gender	Female
Stomach content	Empty

Imagine this is information about you in a bar.

Weight	72 kg
Units	6
Duration	120 mins
Gender	Female
Stomach content	Full

Imagine this is information about you in a bar.

Weight	85 kg
Units	6
Duration	135 mins
Gender	Male
Stomach content	Empty

Imagine this is information about you in a bar.

Weight	80 kg
Units	7
Duration	90 mins
Gender	Male
Stomach content	Full

Imagine this is information about you in a bar.

Weight	65 kg
Units	5
Duration	90 mins
Gender	Female
Stomach content	Full

Imagine this is information about you in a bar.

Weight	55 kg
Units	5
Duration	195 mins
Gender	Male
Stomach content	Empty

Attention Check

Imagine this is information about you in a bar.

Weight	55 kg
Units	5
Duration	195 mins
Gender	Male
Stomach content	Empty

Appendix B

Low familiarity cases with correct system outputs

Cases with counterfactual explanations

Here's information about Chemical 13916.

Chemical 13916	
Occupational exposure limit	80 ppm
pH	6
Exposure duration	30 mins
Air pollution rating	e-01
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 13916 is not safe. <u>Explanation:</u> Chemical 13916 would have been deemed safe, if it had a longer exposure duration.	

Here's information about Chemical 39136.

Chemical 39136	
Occupational exposure limit	82 ppm
pH	7
Exposure duration	210 mins
Air pollution rating	e-01
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 39136 is safe. <u>Explanation:</u> Chemical 39136 would have been deemed not safe, if it had a shorter exposure duration.	

Here's information about Chemical 59780.

Chemical 59780	
Occupational exposure limit	71 ppm
pH	5
Exposure duration	60 mins
Air pollution rating	e-00
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 59780 is not safe. <u>Explanation:</u> Chemical 59780 would have been deemed safe, if it had a longer exposure duration.	

Here's information about Chemical 84283.

Chemical 84283	
Occupational exposure limit	71 ppm
pH	5
Exposure duration	180 mins
Air pollution rating	e-00
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 84283 is safe. <u>Explanation:</u> Chemical 84283 would have been deemed not safe, if it had a shorter exposure duration.	

Here's information about Chemical 39819.

Chemical 39819	
Occupational exposure limit	84 ppm
pH	7
Exposure duration	105 mins
Air pollution rating	e-01
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 39819 is not safe. <u>Explanation:</u> Chemical 39819 would have been deemed safe, if it had a PNEC rating of CE50.	

Here's information about Chemical 14980.

Chemical 14980	
Occupational exposure limit	83 ppm
pH	7
Exposure duration	120 mins
Air pollution rating	e-01
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 14980 is safe. <u>Explanation:</u> Chemical 14980 would have been deemed not safe, if it had a PNEC rating of EC10.	

Here's information about Chemical 60127.

Chemical 60127	
Occupational exposure limit	72 ppm
pH	5
Exposure duration	90 mins
Air pollution rating	e-00
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 60127 is safe. <u>Explanation:</u> Chemical 60127 would have been deemed not safe, if it had a PNEC rating of EC10.	

Here's information about Chemical 82554.

Chemical 82554	
Occupational exposure limit	66 ppm
pH	5
Exposure duration	120 mins
Air pollution rating	e-00
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 82554 is not safe. <u>Explanation:</u> Chemical 82554 would have been deemed safe, if it had a PNEC rating of CE50.	

Here's information about Chemical 17416.

Chemical 17416	
Occupational exposure limit	90 ppm
pH	8
Exposure duration	90 mins
Air pollution rating	e-01
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 17416 is not safe. <u>Explanation:</u> Chemical 17416 would have been deemed safe, if it had a lower pH.	

Here's information about Chemical 36324.

Chemical 36324	
Occupational exposure limit	74 ppm
pH	5
Exposure duration	120 mins
Air pollution rating	e-01
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 36324 is safe. <u>Explanation:</u> Chemical 36324 would have been deemed not safe, if it had a higher pH.	

Here's information about Chemical 83220.

Chemical 83220	
Occupational exposure limit	68 ppm
pH	4
Exposure duration	105 mins
Air pollution rating	e-00
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 83220 is safe. <u>Explanation:</u> Chemical 83220 would have been deemed not safe, if it had a higher pH.	

Here's information about Chemical 60842.

Chemical 60842	
Occupational exposure limit	74 ppm
pH	6
Exposure duration	120 mins
Air pollution rating	e-00
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 60842 is not safe. <u>Explanation:</u> Chemical 60842 would have been deemed safe, if it had a lower pH.	

Here's information about Chemical 44304.

Chemical 44304	
Occupational exposure limit	97 ppm
pH	7
Exposure duration	105 mins
Air pollution rating	e-01
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 44304 is safe. <u>Explanation:</u> Chemical 44304 would have been deemed not safe, if its occupational exposure limit was lower.	

Here's information about Chemical 10470.

Chemical 10470	
Occupational exposure limit	70 ppm
pH	6
Exposure duration	90 mins
Air pollution rating	e-01
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 10470 is not safe. <u>Explanation:</u> Chemical 10470 would have been deemed safe, if its occupational exposure limit was higher.	

Here's information about Chemical 53549.

Chemical 53549	
Occupational exposure limit	53 ppm
pH	4
Exposure duration	90 mins
Air pollution rating	e-00
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 53549 is not safe. <u>Explanation:</u> Chemical 53549 would have been deemed safe, if its occupational exposure limit was higher.	

Here's information about Chemical 87039.

Chemical 87039	
Occupational exposure limit	79 ppm
pH	5
Exposure duration	120 mins
Air pollution rating	e-00
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 87039 is safe. <u>Explanation:</u> Chemical 87039 would have been deemed not safe, if its occupational exposure limit was lower.	

Attention Check

Here's information about Chemical 881208.

Chemical 881208	
Occupational exposure limit	88 ppm
pH	8
Exposure duration	120 mins
Air pollution rating	e-01
PNEC Rating	CE50

Cases with causal explanations

Here's information about Chemical 13916.

Chemical 13916	
Occupational exposure limit	80 ppm
pH	6
Exposure duration	30 mins
Air pollution rating	e-01
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 13916 is not safe. <u>Explanation:</u> Chemical 13916 was deemed not safe, because its exposure duration was short.	

Here's information about Chemical 39136.

Chemical 39136	
Occupational exposure limit	82 ppm
pH	7
Exposure duration	210 mins
Air pollution rating	e-01
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 39136 is safe. <u>Explanation:</u> Chemical 39136 was deemed safe, because its exposure duration was long.	

Here's information about Chemical 59780.

Chemical 59780	
Occupational exposure limit	71 ppm
pH	5
Exposure duration	60 mins
Air pollution rating	e-00
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 59780 is not safe. <u>Explanation:</u> Chemical 59780 was deemed not safe, because its exposure duration was short.	

Here's information about Chemical 84283.

Chemical 84283	
Occupational exposure limit	71 ppm
pH	5
Exposure duration	180 mins
Air pollution rating	e-00
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 84283 is safe. <u>Explanation:</u> Chemical 84283 was deemed safe, because its exposure duration was long.	

Here's information about Chemical 39819.

Chemical 39819	
Occupational exposure limit	84 ppm
pH	7
Exposure duration	105 mins
Air pollution rating	e-01
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 39819 is not safe. <u>Explanation:</u> Chemical 39819 was deemed not safe, because its PNEC rating was EC10.	

Here's information about Chemical 14980.

Chemical 14980	
Occupational exposure limit	83 ppm
pH	7
Exposure duration	120 mins
Air pollution rating	e-01
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 14980 is safe. <u>Explanation:</u> Chemical 14980 was deemed safe, because its PNEC rating was CE50.	

Here's information about Chemical 60127.

Chemical 60127	
Occupational exposure limit	72 ppm
pH	5
Exposure duration	90 mins
Air pollution rating	e-00
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 60127 is safe. <u>Explanation:</u> Chemical 60127 was deemed safe, because its PNEC rating was CE50.	

Here's information about Chemical 82554.

Chemical 82554	
Occupational exposure limit	66 ppm
pH	5
Exposure duration	120 mins
Air pollution rating	e-00
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 82554 is not safe. <u>Explanation:</u> Chemical 82554 was deemed not safe, because its PNEC rating was EC10.	

Here's information about Chemical 17416.

Chemical 17416	
Occupational exposure limit	90 ppm
pH	8
Exposure duration	90 mins
Air pollution rating	e-01
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 17416 is not safe. <u>Explanation:</u> Chemical 17416 was deemed not safe, because its pH was high.	

Here's information about Chemical 36324.

Chemical 36324	
Occupational exposure limit	74 ppm
pH	5
Exposure duration	120 mins
Air pollution rating	e-01
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 36324 is safe. <u>Explanation:</u> Chemical 36324 was deemed safe, because its pH was low.	

Here's information about Chemical 83220.

Chemical 83220	
Occupational exposure limit	68 ppm
pH	4
Exposure duration	105 mins
Air pollution rating	e-00
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 83220 is safe. <u>Explanation:</u> Chemical 83220 was deemed safe, because its pH was low.	

Here's information about Chemical 60842.

Chemical 60842	
Occupational exposure limit	74 ppm
pH	6
Exposure duration	120 mins
Air pollution rating	e-00
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 60842 is not safe. <u>Explanation:</u> Chemical 60842 was deemed not safe, because its pH was high.	

Here's information about Chemical 44304.

Chemical 44304	
Occupational exposure limit	97 ppm
pH	7
Exposure duration	105 mins
Air pollution rating	e-01
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 44304 is safe. <u>Explanation:</u> Chemical 44304 was deemed safe, because its occupational exposure limit was high.	

Here's information about Chemical 10470.

Chemical 10470	
Occupational exposure limit	70 ppm
pH	6
Exposure duration	90 mins
Air pollution rating	e-01
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 10470 is not safe. <u>Explanation:</u> Chemical 10470 was deemed not safe, because its occupational exposure limit was low.	

Here's information about Chemical 53549.

Chemical 53549	
Occupational exposure limit	53 ppm
pH	4
Exposure duration	90 mins
Air pollution rating	e-00
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 53549 is not safe. <u>Explanation:</u> Chemical 53549 was deemed not safe, because its occupational exposure limit was low.	

Here's information about Chemical 87039.

Chemical 87039	
Occupational exposure limit	79 ppm
pH	5
Exposure duration	120 mins
Air pollution rating	e-00
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 87039 is safe. <u>Explanation:</u> Chemical 87039 was deemed safe, because its occupational exposure limit was high.	

Attention Check

Here's information about Chemical 201548.

Chemical 201548	
Occupational exposure limit	55 ppm
pH	5
Exposure duration	195 mins
Air pollution rating	e-01
PNEC Rating	EC10

Cases used for prediction judgements

Here's information about Chemical 15322.

Chemical 15322	
Occupational exposure limit	84 ppm
pH	7
Exposure duration	75 mins
Air pollution rating	e-01
PNEC Rating	CE50

Here's information about Chemical 38418.

Chemical 38418	
Occupational exposure limit	80 ppm
pH	6
Exposure duration	135 mins
Air pollution rating	e-01
PNEC Rating	EC10

Here's information about Chemical 83215.

Chemical 83215	
Occupational exposure limit	68 ppm
pH	4
Exposure duration	30 mins
Air pollution rating	e-00
PNEC Rating	EC10

Here's information about Chemical 57718.

Chemical 57718	
Occupational exposure limit	65 ppm
pH	5
Exposure duration	180 mins
Air pollution rating	e-00
PNEC Rating	CE50

Here's information about Chemical 39474.

Chemical 39474	
Occupational exposure limit	83 ppm
pH	7
Exposure duration	105 mins
Air pollution rating	e-01
PNEC Rating	EC10

Here's information about Chemical 13921.

Chemical 13921	
Occupational exposure limit	80 ppm
pH	6
Exposure duration	105 mins
Air pollution rating	e-01
PNEC Rating	CE50

Here's information about Chemical 58060.

Chemical 58060	
Occupational exposure limit	66 ppm
pH	5
Exposure duration	135 mins
Air pollution rating	e-00
PNEC Rating	CE50

Here's information about Chemical 83588.

Chemical 83588	
Occupational exposure limit	69 ppm
pH	5
Exposure duration	105 mins
Air pollution rating	e-00
PNEC Rating	EC10

Here's information about Chemical 15645.

Chemical 15645	
Occupational exposure limit	85 ppm
pH	6
Exposure duration	90 mins
Air pollution rating	e-01
PNEC Rating	CE50

Here's information about Chemical 39845.

Chemical 39845	
Occupational exposure limit	84 ppm
pH	8
Exposure duration	150 mins
Air pollution rating	e-01
PNEC Rating	EC10

Here's information about Chemical 82184.

Chemical 82184	
Occupational exposure limit	65 ppm
pH	4
Exposure duration	90 mins
Air pollution rating	e-00
PNEC Rating	EC10

Here's information about Chemical 60152.

Chemical 60152	
Occupational exposure limit	72 ppm
pH	6
Exposure duration	120 mins
Air pollution rating	e-00
PNEC Rating	CE50

Here's information about Chemical 40143.

Chemical 40143	
Occupational exposure limit	85 ppm
pH	6
Exposure duration	135 mins
Air pollution rating	e-01
PNEC Rating	EC10

Here's information about Chemical 13943.

Chemical 13943	
Occupational exposure limit	80 ppm
pH	7
Exposure duration	90 mins
Air pollution rating	e-01
PNEC Rating	CE50

Here's information about Chemical 57712.

Chemical 57712	
Occupational exposure limit	65 ppm
pH	5
Exposure duration	90 mins
Air pollution rating	e-00
PNEC Rating	CE50

Here's information about Chemical 85314.

Chemical 85314	
Occupational exposure limit	74 ppm
pH	5
Exposure duration	120 mins
Air pollution rating	e-00
PNEC Rating	EC10

Attention Check

Here's information about Chemical 201548.

Chemical 201548	
Occupational exposure limit	55 ppm
pH	5
Exposure duration	195 mins
Air pollution rating	e-01
PNEC Rating	EC10

Cases used for decision judgements

Imagine this is a chemical in front of you.

Occupational exposure limit	84 ppm
pH	7
Exposure duration	75 mins
Air pollution rating	e-01
PNEC rating	CE50

Imagine this is a chemical in front of you.

Occupational exposure limit	80 ppm
pH	6
Exposure duration	135 mins
Air pollution rating	e-01
PNEC rating	EC10

Imagine this is a chemical in front of you.

Occupational exposure limit	68 ppm
pH	4
Exposure duration	30 mins
Air pollution rating	e-00
PNEC rating	EC10

Imagine this is a chemical in front of you.

Occupational exposure limit	65 ppm
pH	5
Exposure duration	180 mins
Air pollution rating	e-00
PNEC rating	CE50

Imagine this is a chemical in front of you.

Occupational exposure limit	83 ppm
pH	7
Exposure duration	105 mins
Air pollution rating	e-01
PNEC rating	EC10

Imagine this is a chemical in front of you.

Occupational exposure limit	80 ppm
pH	6
Exposure duration	105 mins
Air pollution rating	e-01
PNEC rating	CE50

Imagine this is a chemical in front of you.

Occupational exposure limit	66 ppm
pH	5
Exposure duration	135 mins
Air pollution rating	e-00
PNEC rating	CE50

Imagine this is a chemical in front of you.

Occupational exposure limit	69 ppm
pH	5
Exposure duration	105 mins
Air pollution rating	e-00
PNEC rating	EC10

Imagine this is a chemical in front of you.

Occupational exposure limit	85 ppm
pH	6
Exposure duration	90 mins
Air pollution rating	e-01
PNEC rating	CE50

Imagine this is a chemical in front of you.

Occupational exposure limit	84 ppm
pH	8
Exposure duration	150 mins
Air pollution rating	e-01
PNEC rating	EC10

Imagine this is a chemical in front of you.

Occupational exposure limit	65 ppm
pH	4
Exposure duration	90 mins
Air pollution rating	e-00
PNEC rating	EC10

Imagine this is a chemical in front of you.

Occupational exposure limit	72 ppm
pH	6
Exposure duration	120 mins
Air pollution rating	e-00
PNEC rating	CE50

Imagine this is a chemical in front of you.

Occupational exposure limit	85 ppm
pH	6
Exposure duration	135 mins
Air pollution rating	e-01
PNEC rating	EC10

Imagine this is a chemical in front of you.

Occupational exposure limit	80 ppm
pH	7
Exposure duration	90 mins
Air pollution rating	e-01
PNEC rating	CE50

Imagine this is a chemical in front of you.

Occupational exposure limit	65 ppm
pH	5
Exposure duration	90 mins
Air pollution rating	e-00
PNEC rating	CE50

Imagine this is a chemical in front of you.

Occupational exposure limit	74 ppm
pH	5
Exposure duration	120 mins
Air pollution rating	e-00
PNEC rating	EC10

Attention Check

Imagine this is a chemical in front of you.

Occupational exposure limit	55 ppm
pH	5
Exposure duration	195 mins
Air pollution rating	e-01
PNEC rating	EC10

Appendix C

High familiarity cases with incorrect system outputs

Cases with counterfactual explanations

Here's information about Walter, who is in a bar.

Walter	
Weight	80 kg
Units	6
Duration	30 mins
Gender	Male
Stomach content	Full
<u>Prediction:</u> The app predicts Walter is under the limit. <u>Explanation:</u> Walter's blood alcohol level would have been over the limit, if he had drunk for a longer time.	

Here's information about Jonny, who is in a bar.

Jonny	
Weight	82 kg
Units	7
Duration	210 mins
Gender	Male
Stomach content	Empty
<u>Prediction:</u> The app predicts Jonny is over the limit. <u>Explanation:</u> Jonny's blood alcohol level would have been under the limit, if he had drunk for a shorter time.	

Here's information about Lillian, who is in a bar.

Lillian	
Weight	71 kg
Units	5
Duration	60 mins
Gender	Female
Stomach content	Full
<u>Prediction:</u> The app predicts Lillian is under the limit. <u>Explanation:</u> Lillian's blood alcohol level would have been over the limit, if she had drunk for a longer time.	

Here's information about Violet, who is in a bar.

Violet	
Weight	71 kg
Units	5
Duration	180 mins
Gender	Female
Stomach content	Empty
<u>Prediction:</u> The app predicts Violet is over the limit. <u>Explanation:</u> Violet's blood alcohol level would have been under the limit, if she had drunk for a shorter time.	

Here's information about Jarred, who is in a bar.

Jarred	
Weight	84 kg
Units	7
Duration	105 mins
Gender	Male
Stomach content	empty
<u>Prediction:</u> The app predicts Jarred is under the limit. <u>Explanation:</u> Jarred's blood alcohol level would have been over the limit, if he had drunk on a full stomach.	

Here's information about Elliot, who is in a bar.

Elliot	
Weight	83 kg
Units	7
Duration	120 mins
Gender	Male
Stomach content	full
<u>Prediction:</u> The app predicts Elliot is over the limit. <u>Explanation:</u> Elliot's blood alcohol level would have been under the limit, if he had drunk on an empty stomach.	

Here's information about Loreen, who is in a bar.

Loreen	
Weight	72 kg
Units	5
Duration	90 mins
Gender	Female
Stomach content	full
<u>Prediction:</u> The app predicts Loreen is over the limit. <u>Explanation:</u> Loreen's blood alcohol level would have been under the limit, if she had drunk on a empty stomach.	

Here's information about Georgina, who is in a bar.

Georgina	
Weight	66 kg
Units	5
Duration	120 mins
Gender	Female
Stomach content	empty
<u>Prediction:</u> The app predicts Georgina is under the limit. <u>Explanation:</u> Georgina's blood alcohol level would have been over the limit, if she had drunk on an full stomach.	

Here's information about Noah, who is in a bar.

Noah	
Weight	90 kg
Units	8
Duration	90 mins
Gender	Male
Stomach content	Full
<u>Prediction:</u> The app predicts Noah is under the limit. <u>Explanation:</u> Noah's blood alcohol level would have been over the limit, if he had drunk fewer units of alcohol.	

Here's information about Bryan, who is in a bar.

Bryan	
Weight	74 kg
Units	5
Duration	120 mins
Gender	Male
Stomach content	Empty
<u>Prediction:</u> The app predicts Bryan is over the limit. <u>Explanation:</u> Bryan's blood alcohol level would have been under the limit, if he had drunk more units of alcohol.	

Here's information about Zoe, who is in a bar.

Zoe	
Weight	68 kg
Units	4
Duration	105 mins
Gender	Female
Stomach content	Empty
<u>Prediction:</u> The app predicts Zoe is over the limit. <u>Explanation:</u> Zoe's blood alcohol level would have been under the limit, if she had drunk more units of alcohol.	

Here's information about Betty, who is in a bar.

Betty	
Weight	74 kg
Units	6
Duration	120 mins
Gender	Female
Stomach content	Full
<u>Prediction:</u> The app predicts Betty is under the limit. <u>Explanation:</u> Betty's blood alcohol level would have been over the limit, if she had drunk fewer units of alcohol.	

Here's information about Leo, who is in a bar.

Leo	
Weight	97 kg
Units	7
Duration	105 mins
Gender	Male
Stomach content	Empty
<u>Prediction:</u> The app predicts Leo is over the limit. <u>Explanation:</u> Leo's blood alcohol level would have been under the limit, if his weight had been lighter.	

Here's information about Stephen, who is in a bar.

Stephen	
Weight	70 kg
Units	6
Duration	90 mins
Gender	Male
Stomach content	Full
<u>Prediction:</u> The app predicts Stephen is under the limit. <u>Explanation:</u> Stephen's blood alcohol level would have been over the limit, if his weight had been heavier.	

Here's information about Angelica, who is in a bar.

Angelica	
Weight	53 kg
Units	4
Duration	90 mins
Gender	Female
Stomach content	Full
<u>Prediction:</u> The app predicts Angelica is under the limit. <u>Explanation:</u> Angelica's blood alcohol level would have been over the limit, if her weight had been heavier.	

Here's information about Anne, who is in a bar.

Anne	
Weight	79 kg
Units	5
Duration	120 mins
Gender	Female
Stomach content	Empty
<u>Prediction:</u> The app predicts Anne is over the limit. <u>Explanation:</u> Anne's blood alcohol level would have been under the limit, if her weight had been lighter.	

Attention Check

Here's information about Tyler, who's in a bar:

Tyler	
Weight	88 kg
Units of alcohol drank	8
Duration	120 mins
Gender	Male
Stomach	Full

Cases with causal explanations

Here's information about Walter, who is in a bar.

Walter	
Weight	80 kg
Units	6
Duration	30 mins
Gender	Male
Stomach content	Full
<u>Prediction:</u> The app predicts Walter is under the limit. <u>Explanation:</u> Walter's blood alcohol level was under the limit, because he drank for a short time.	

Here's information about Jonny, who is in a bar.

Jonny	
Weight	82 kg
Units	7
Duration	210 mins
Gender	Male
Stomach content	Empty
<u>Prediction:</u> The app predicts Jonny is over the limit. <u>Explanation:</u> Jonny's blood alcohol level was over the limit, because he drank for a long time.	

Here's information about Lillian, who is in a bar.

Lillian	
Weight	71 kg
Units	5
Duration	60 mins
Gender	Female
Stomach content	Full
<u>Prediction:</u> The app predicts Lillian is under the limit. <u>Explanation:</u> Lillian's blood alcohol level was under the limit, because she drank for a short time.	

Here's information about Violet, who is in a bar.

Violet	
Weight	71 kg
Units	5
Duration	180 mins
Gender	Female
Stomach content	Empty
<u>Prediction:</u> The app predicts Violet is over the limit. <u>Explanation:</u> Violet's blood alcohol level was over the limit, because she drank for a long time.	

Here's information about Jarred, who is in a bar.

Jarred	
Weight	84 kg
Units	7
Duration	105 mins
Gender	Male
Stomach content	empty
<u>Prediction:</u> The app predicts Jarred is under the limit. <u>Explanation:</u> Jarred's blood alcohol level was under the limit, because he drank on an empty stomach.	

Here's information about Elliot, who is in a bar.

Elliot	
Weight	83 kg
Units	7
Duration	120 mins
Gender	Male
Stomach content	full
<u>Prediction:</u> The app predicts Elliot is over the limit. <u>Explanation:</u> Elliot's blood alcohol level was over the limit, because he drank on a full stomach.	

Here's information about Loreen, who is in a bar.

Loreen	
Weight	72 kg
Units	5
Duration	90 mins
Gender	Female
Stomach content	full
<u>Prediction:</u> The app predicts Loreen is over the limit . <u>Explanation:</u> Loreen's blood alcohol level was over the limit, because she drank on a full stomach.	

Here's information about Georgina, who is in a bar.

Georgina	
Weight	66 kg
Units	5
Duration	120 mins
Gender	Female
Stomach content	empty
<u>Prediction:</u> The app predicts Georgina is under the limit . <u>Explanation:</u> Georgina's blood alcohol level was under the limit, because she drank on an empty stomach.	

Here's information about Noah, who is in a bar.

Noah	
Weight	90 kg
Units	8
Duration	90 mins
Gender	Male
Stomach content	Full
<u>Prediction:</u> The app predicts Noah is under the limit . <u>Explanation:</u> Noah's blood alcohol level was under the limit, because he drank many units of alcohol.	

Here's information about Bryan, who is in a bar.

Bryan	
Weight	74 kg
Units	5
Duration	120 mins
Gender	Male
Stomach content	Empty
<u>Prediction:</u> The app predicts Bryan is over the limit . <u>Explanation:</u> Bryan's blood alcohol level was over the limit, because he drank few units of alcohol.	

Here's information about Zoe, who is in a bar.

Zoe	
Weight	68 kg
Units	4
Duration	105 mins
Gender	Female
Stomach content	Empty
<u>Prediction:</u> The app predicts Zoe is over the limit . <u>Explanation:</u> Zoe's blood alcohol level was over the limit, because she drank few units of alcohol.	

Here's information about Betty, who is in a bar.

Betty	
Weight	74 kg
Units	6
Duration	120 mins
Gender	Female
Stomach content	Full
<u>Prediction:</u> The app predicts Betty is under the limit . <u>Explanation:</u> Betty's blood alcohol level was under the limit, because she drank many units of alcohol.	

Here's information about Leo, who is in a bar.

Leo	
Weight	97 kg
Units	7
Duration	105 mins
Gender	Male
Stomach content	Empty
Prediction: The app predicts Leo is over the limit . Explanation: Leo's blood alcohol level was over the limit, because his weight was heavy.	

Here's information about Stephen, who is in a bar.

Stephen	
Weight	70 kg
Units	6
Duration	90 mins
Gender	Male
Stomach content	Full
Prediction: The app predicts Stephen is under the limit . Explanation: Stephen's blood alcohol level was under the limit, because his weight was light.	

Here's information about Angelica, who is in a bar.

Angelica	
Weight	53 kg
Units	4
Duration	90 mins
Gender	Female
Stomach content	Full
Prediction: The app predicts Angelica is under the limit . Explanation: Angelica's blood alcohol level was under the limit, because her weight was light.	

Here's information about Anne, who is in a bar.

Anne	
Weight	79 kg
Units	5
Duration	120 mins
Gender	Female
Stomach content	Empty
Prediction: The app predicts Anne is over the limit . Explanation: Anne's blood alcohol level was over the limit, because her weight was heavy.	

Attention Check

Here's information about Tyler, who's in a bar:

Tyler	
Weight	88 kg
Units of alcohol drank	8
Duration	120 mins
Gender	Male
Stomach	Full

Appendix D

Low familiarity cases with incorrect system outputs

Cases with counterfactual explanations

Here's information about Chemical 13916.

Chemical 13916	
Occupational exposure limit	80 ppm
pH	6
Exposure duration	30 mins
Air pollution rating	e-01
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 13916 is safe. <u>Explanation:</u> Chemical 13916 would have been deemed not safe, if it had a longer exposure duration.	

Here's information about Chemical 39136.

Chemical 39136	
Occupational exposure limit	82 ppm
pH	7
Exposure duration	210 mins
Air pollution rating	e-01
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 39136 is not safe. <u>Explanation:</u> Chemical 39136 would have been deemed safe, if it had a shorter exposure duration.	

Here's information about Chemical 59780.

Chemical 59780	
Occupational exposure limit	71 ppm
pH	5
Exposure duration	60 mins
Air pollution rating	e-00
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 59780 is safe. <u>Explanation:</u> Chemical 59780 would have been deemed not safe, if it had a longer exposure duration.	

Here's information about Chemical 84283.

Chemical 84283	
Occupational exposure limit	71 ppm
pH	5
Exposure duration	180 mins
Air pollution rating	e-00
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 84283 is not safe. <u>Explanation:</u> Chemical 84283 would have been deemed safe, if it had a shorter exposure duration.	

Here's information about Chemical 39819.

Chemical 39819	
Occupational exposure limit	84 ppm
pH	7
Exposure duration	105 mins
Air pollution rating	e-01
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 39819 is safe. <u>Explanation:</u> Chemical 39819 would have been deemed not safe, if it had a PNEC rating of CE50.	

Here's information about Chemical 14980.

Chemical 14980	
Occupational exposure limit	83 ppm
pH	7
Exposure duration	120 mins
Air pollution rating	e-01
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 14980 is not safe. <u>Explanation:</u> Chemical 14980 would have been deemed safe, if it had a PNEC rating of EC10.	

Here's information about Chemical 60127.

Chemical 60127	
Occupational exposure limit	72 ppm
pH	5
Exposure duration	90 mins
Air pollution rating	e-00
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 60127 is not safe. <u>Explanation:</u> Chemical 60127 would have been deemed safe, if it had a PNEC rating of EC10.	

Here's information about Chemical 82554.

Chemical 82554	
Occupational exposure limit	66 ppm
pH	5
Exposure duration	120 mins
Air pollution rating	e-00
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 82554 is safe. <u>Explanation:</u> Chemical 82554 would have been deemed not safe, if it had a PNEC rating of CE50.	

Here's information about Chemical 17416.

Chemical 17416	
Occupational exposure limit	90 ppm
pH	8
Exposure duration	90 mins
Air pollution rating	e-01
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 17416 is safe. <u>Explanation:</u> Chemical 17416 would have been deemed not safe, if it had a lower pH.	

Here's information about Chemical 36324.

Chemical 36324	
Occupational exposure limit	74 ppm
pH	5
Exposure duration	120 mins
Air pollution rating	e-01
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 36324 is not safe. <u>Explanation:</u> Chemical 36324 would have been deemed safe, if it had a higher pH.	

Here's information about Chemical 83220.

Chemical 83220	
Occupational exposure limit	68 ppm
pH	4
Exposure duration	105 mins
Air pollution rating	e-00
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 83220 is not safe. <u>Explanation:</u> Chemical 83220 would have been deemed safe, if it had a higher pH.	

Here's information about Chemical 60842.

Chemical 60842	
Occupational exposure limit	74 ppm
pH	6
Exposure duration	120 mins
Air pollution rating	e-00
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 60842 is safe. <u>Explanation:</u> Chemical 60842 would have been deemed not safe, if it had a lower pH.	

Here's information about Chemical 44304.

Chemical 44304	
Occupational exposure limit	97 ppm
pH	7
Exposure duration	105 mins
Air pollution rating	e-01
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 44304 is not safe. <u>Explanation:</u> Chemical 44304 would have been deemed safe, if its occupational exposure limit was lower.	

Here's information about Chemical 10470.

Chemical 10470	
Occupational exposure limit	70 ppm
pH	6
Exposure duration	90 mins
Air pollution rating	e-01
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 10470 is safe. <u>Explanation:</u> Chemical 10470 would have been deemed not safe, if its occupational exposure limit was higher.	

Here's information about Chemical 53549.

Chemical 53549	
Occupational exposure limit	53 ppm
pH	4
Exposure duration	90 mins
Air pollution rating	e-00
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 53549 is safe. <u>Explanation:</u> Chemical 53549 would have been deemed not safe, if its occupational exposure limit was higher.	

Here's information about Chemical 87039.

Chemical 87039	
Occupational exposure limit	79 ppm
pH	5
Exposure duration	120 mins
Air pollution rating	e-00
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 87039 is not safe. <u>Explanation:</u> Chemical 87039 would have been deemed safe, if its occupational exposure limit was lower.	

Attention Check

Here's information about Chemical 881208.

Chemical 881208	
Occupational exposure limit	88 ppm
pH	8
Exposure duration	120 mins
Air pollution rating	e-01
PNEC Rating	CE50

Cases with causal explanations

Here's information about Chemical 13916.

Chemical 13916	
Occupational exposure limit	80 ppm
pH	6
Exposure duration	30 mins
Air pollution rating	e-01
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 13916 is safe. <u>Explanation:</u> Chemical 13916 was deemed safe, because its exposure duration was short.	

Here's information about Chemical 39136.

Chemical 39136	
Occupational exposure limit	82 ppm
pH	7
Exposure duration	210 mins
Air pollution rating	e-01
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 39136 is not safe. <u>Explanation:</u> Chemical 39136 was deemed not safe, because its exposure duration was long.	

Here's information about Chemical 59780.

Chemical 59780	
Occupational exposure limit	71 ppm
pH	5
Exposure duration	60 mins
Air pollution rating	e-00
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 59780 is safe. <u>Explanation:</u> Chemical 59780 was deemed safe, because its exposure duration was short.	

Here's information about Chemical 84283.

Chemical 84283	
Occupational exposure limit	71 ppm
pH	5
Exposure duration	180 mins
Air pollution rating	e-00
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 84283 is not safe. <u>Explanation:</u> Chemical 84283 was deemed not safe, because its exposure duration was long.	

Here's information about Chemical 39819.

Chemical 39819	
Occupational exposure limit	84 ppm
pH	7
Exposure duration	105 mins
Air pollution rating	e-01
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 39819 is safe. <u>Explanation:</u> Chemical 39819 was deemed safe, because its PNEC rating was EC10.	

Here's information about Chemical 14980.

Chemical 14980	
Occupational exposure limit	83 ppm
pH	7
Exposure duration	120 mins
Air pollution rating	e-01
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 14980 is not safe. <u>Explanation:</u> Chemical 14980 was deemed not safe, because its PNEC rating was CE50.	

Here's information about Chemical 60127.

Chemical 60127	
Occupational exposure limit	72 ppm
pH	5
Exposure duration	90 mins
Air pollution rating	e-00
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 60127 is not safe. <u>Explanation:</u> Chemical 60127 was deemed not safe, because its PNEC rating was CE50.	

Here's information about Chemical 82554.

Chemical 82554	
Occupational exposure limit	66 ppm
pH	5
Exposure duration	120 mins
Air pollution rating	e-00
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 82554 is safe. <u>Explanation:</u> Chemical 82554 was deemed safe, because its PNEC rating was EC10.	

Here's information about Chemical 17416.

Chemical 17416	
Occupational exposure limit	90 ppm
pH	8
Exposure duration	90 mins
Air pollution rating	e-01
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 17416 is safe. <u>Explanation:</u> Chemical 17416 was deemed safe, because its pH was high.	

Here's information about Chemical 36324.

Chemical 36324	
Occupational exposure limit	74 ppm
pH	5
Exposure duration	120 mins
Air pollution rating	e-01
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 36324 is not safe. <u>Explanation:</u> Chemical 36324 was deemed not safe, because its pH was low.	

Here's information about Chemical 83220.

Chemical 83220	
Occupational exposure limit	68 ppm
pH	4
Exposure duration	105 mins
Air pollution rating	e-00
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 83220 is not safe. <u>Explanation:</u> Chemical 83220 was deemed not safe, because its pH was low.	

Here's information about Chemical 60842.

Chemical 60842	
Occupational exposure limit	74 ppm
pH	6
Exposure duration	120 mins
Air pollution rating	e-00
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 60842 is safe. <u>Explanation:</u> Chemical 60842 was deemed safe, because its pH was high.	

Here's information about Chemical 44304.

Chemical 44304	
Occupational exposure limit	97 ppm
pH	7
Exposure duration	105 mins
Air pollution rating	e-01
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 44304 is not safe. <u>Explanation:</u> Chemical 44304 was deemed not safe, because its occupational exposure limit was high.	

Here's information about Chemical 10470.

Chemical 10470	
Occupational exposure limit	70 ppm
pH	6
Exposure duration	90 mins
Air pollution rating	e-01
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 10470 is safe. <u>Explanation:</u> Chemical 10470 was deemed safe, because its occupational exposure limit was low.	

Here's information about Chemical 53549.

Chemical 53549	
Occupational exposure limit	53 ppm
pH	4
Exposure duration	90 mins
Air pollution rating	e-00
PNEC Rating	CE50
<u>Prediction:</u> The app predicts Chemical 53549 is safe. <u>Explanation:</u> Chemical 53549 was deemed safe, because its occupational exposure limit was low.	

Here's information about Chemical 87039.

Chemical 87039	
Occupational exposure limit	79 ppm
pH	5
Exposure duration	120 mins
Air pollution rating	e-00
PNEC Rating	EC10
<u>Prediction:</u> The app predicts Chemical 87039 is not safe. <u>Explanation:</u> Chemical 87039 was deemed not safe, because its occupational exposure limit was high.	

Attention Check

Here's information about Chemical 881208.

Chemical 881208	
Occupational exposure limit	88 ppm
pH	8
Exposure duration	120 mins
Air pollution rating	e-01
PNEC Rating	CE50

Appendix E

DARPA Explainability of an AI system scale

1. From the explanation, I understand how the app works.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

2. This explanation of how the app works is satisfying.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

3. This explanation of how the app works has sufficient detail.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

4. This explanation of how the app works seems complete.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

5. This explanation of how the app works tells me how to use it.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

6. This explanation of how the app works is useful to my goals.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

7. This explanation of the app shows me how accurate the app is.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

8. This explanation lets me judge when I should trust and not trust the app

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

Appendix F

DARPA Trust in the AI system scale

1. I am confident in the app. I feel that it works well

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

2. The outputs of the app are very predictable.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

3. The tool is very reliable. I can count on it to be correct all the time.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

4. I feel safe that when I rely on the app I will get the right answers.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

5. The app is efficient in that it works very quickly.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

6. I am wary of the app.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

7. The app can perform the task better than a novice human user.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

8. I like using the system for decision making.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

Appendix G

Information Sheet

Hi and thank you for your interest in participating in this study!

My name is Lenart Celar and I am a masters student at the School of Psychology. This study is supervised by Ruth Byrne, Professor of Cognitive Science at the School of Psychology, Trinity College Dublin.

Before you decide whether or not you wish to take part, please read this information sheet carefully. Don't feel rushed or under pressure to make a quick decision. You should understand the risks and benefits of taking part in this study so that you can make a decision that is right for you.

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. It is best to complete the study on a desktop computer or laptop; if you are doing it on a smartphone or similar device, please make sure to scroll to the end of each page so you see all the information.

If you would like to participate, the next screen provides the CONSENT form when you press 'next' below.

First, here is some necessary information about data protection and other matters.

DATA PROTECTION, ETHICAL APPROVAL AND RELATED MATTERS:

Name of Study: *Human Reasoning in Explainable AI*. Principal Investigator: Lenart Celar, School of Psychology, Trinity College Dublin, Email: celarl@tcd.ie. Supervisor: Professor Ruth Byrne, School of Psychology, Trinity College Dublin, Email: rmbyrne@tcd.ie. 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 Lenart Celar. Before you decide whether you wish to take part, please read this information screen carefully and ask any questions you have by email (rmbyrne@tcd.ie). 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 interested in how people understand the predictions made by Artificial Intelligence (AI) decision support systems. Sometimes it is hard for users to know what drove a system to make a specific prediction and which factors were taken into account. The current study aims to examine how people understand why AI systems come to certain predictions and how different sorts of explanations can help people understand AI decisions. You are invited if you are over 18 years old. We intend to invite about 170 participants. 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., 3 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 three 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 Lenart Celar to reveal your code and delete your data (celarl@tcd.ie). The study will take place starting in January 2021. 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. This study will consist approximately 20 scenarios in which you will be presented with a certain prediction an AI system came to. You will then be asked to assess what the outcome would be if certain features were changed. You will provide your judgments on a scale from 1 to 5 by clicking a tickbox. The study should take approximately 10-15 minutes to complete. 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 (Lenart Celar, celarl@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 21/11/2020. The study is a part of a research project conducted by Lenart Celar. Results will be published in scientific publications and will not be used for commercial purposes. You will be paid 1.5 EUR on Prolific once you have successfully completed the experiment which will take about 10-15 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: Lenart Celar, School of Psychology, Trinity College Dublin, Email: celarl@tcd.ie.

Supervisor: Prof Ruth Byrne, School of Psychology, Aras an Phairsaigh, Trinity College Dublin, Dublin 2,

Ireland. Email: rmbyrne@tcd.ie +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.

Appendix H

Memory Check Questions

High Familiarity Groups

In this study, you were shown predictions about people being over or under the legal alcohol limit to drive, made from 5 different features about them.

Can you recall what these 5 features were? Select all that you remember.

- ☐ Body Mass Index (BMI)
- ☐ Duration of drinking period
- ☐ Alcohol tolerance
- ☐ Height
- ☐ Gender
- ☐ Age
- ☐ Units of alcohol consumed
- ☐ Stomach fullness
- ☐ Weight
- ☐ Alcohol strength

Low Familiarity Groups

In this study, you were shown predictions about chemicals being either safe or not safe, made from 5 different features about them.

Can you recall what these 5 features were? Select all that you remember.

- ☐ PNEC rating
- ☐ Chemical Index
- ☐ Chemical Age
- ☐ Chemical tolerance
- ☐ Air pollution rating
- ☐ Occupational exposure limit
- ☐ Exposure duration
- ☐ Chemical Barrier Height
- ☐ pH
- ☐ Chemical strength

Appendix I

Debriefing Sheet

STUDY DEBRIEFING SHEET

Research Topic: Do familiarity and different types of explanations affect the trust and evaluation of an explanation of an AI system?

Dear Participant,

Thank you for your participation in this study. We're interested in examining how people assess explanations in systems that they have some familiarity with compared to those that have no familiarity with.

You were assigned to one of the four groups and presented with cases from a fictional AI system. The AI system of the first group predicted whether a person's blood alcohol level is over or under the limit, based on features such as weight, units of alcohol drank, time since last meal, etc. This system's features are descriptive and were predicted to be understandable to participants - this would be the 'high familiarity' group. The fictional AI system the second group was presented with predicted whether an unknown fictional chemical solution will be deemed safe, based on chemical features occupational exposure limit, pH, exposure duration, air pollution rating, and PNEC rating. This system's features were predicted to be less understandable to participants ('low familiarity' group).

In addition to that, you were presented with one type of explanations, counterfactual ("Walter's blood alcohol level would have been under the limit, if he had drunk for a longer time.") or causal explanations ("Walter's blood alcohol level was over the limit, because he drank for a short time."). As you can see from the statements they differ a bit, and we're examining whether that affects how users evaluate explanations and trust of AI systems.

For further information on the study, please feel free to contact me at any time. Participants are also able to contact researchers to obtain the results of the study. As part of the Freedom of Information Act and General Data Protection Regulation participants can gain access to any data taken from them during the study or request erasure of the data by contacting the researcher or the supervisor.

This study was not expected to cause distress. In the event it did, feel free to contact [samaritans.org](mailto:jo@samaritans.org) at jo@samaritans.ie or on 116 123.

Thank you again for your participation!

Researcher: Lenart Celar, *MSc student*, celarl@tcd.ie, *School of Psychology, Áras an Phiarsaigh, Trinity College, Dublin 2*

Supervisor: Ruth Byrne, *Professor of Cognitive Science*, rmbyrne@tcd.ie, +353 1 896 4054, *School of Psychology, Áras an Phiarsaigh, Trinity College, Dublin 2*

Appendix J

Questions assessing participants' beliefs against ever drinking and driving, or handling potentially unsafe chemicals

High Familiarity Groups

I believe that people who have had even one drink should not drink and drive.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

I believe that it is acceptable for people who are under the legal limit to drive.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

Low Familiarity Groups

I believe you should never handle chemicals you are familiar with.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

I believe that it is acceptable to handle chemicals that are deemed safe.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

Appendix K

Ethics Approval



Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

F.A.O. Lenart Celar

Approval ID: SPREC102020-52

School of Psychology Research Ethics Committee

10th December 2020

Dear Lenart,

The School of Psychology Research Ethics Committee has reviewed your application entitled "Effects of domain knowledge on explainability in AI systems", 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

Scoil na Síceolaíochta
Dámh na nEolalochtaí Sóisialta agus Daonna,
Áras an Phiarsaigh,
Coláiste na Tríonóide,
Baile Átha Cliath,
Ollscoil Átha Cliath,
Baile Átha Cliath 2, Éire.

School of Psychology
Faculty of Arts, Humanities and Social Sciences,
Trinity College Dublin,
The University of Dublin,
Dublin 2, Ireland.

+353 1 896 1886
psychology@tcd.ie
www.tcd.ie/psychology