Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Survey Track

Good Explanations in Explainable Artificial Intelligence (XAI): Evidence from Human Explanatory Reasoning

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

Abstract

Insights from cognitive science about how people understand explanations can be instructive for the development of robust, user-centred explanations in eXplainable Artificial Intelligence (XAI). I survey key tendencies that people exhibit when they construct explanations and make inferences from them, of relevance to the provision of automated explanations for decisions by AI systems. I first review experimental discoveries of some tendencies people exhibit when they construct explanations, including evidence on the illusion of explanatory depth, intuitive versus reflective explanations, and explanatory stances. I then consider discoveries of how people reason about causal explanations, including evidence on inference suppression, causal discounting, and explanation simplicity. I argue that central to the XAI endeavor is the requirement that automated explanations provided by an AI system should make sense to human users.

1 Introduction

Artificial Intelligence decision support systems are widely available in many diverse areas, in the public and private sectors, ranging from financial decisions to health care choices, employment recruitment to policing and criminal justice. However, human users may not believe AI decisions are fair or trustworthy, since the reasons for the decisions of AI systems trained on vast arrays of data are usually not transparent [Adadi and Berrada, 2018; Keane *et al.,* 2021]. The development of automated explanations in XAI aims to increase human users' understanding of an AI system, and to explain its decisions, to address issues of interpretability and recourse [Karimi *et al.*, 2020; Kenny *et al.,* 2021]. Some XAI strategies aim to provide a global explanation of the AI system, e.g., by simplifying or mapping it using, say, decision trees; other XAI strategies aim to provide a local explanation, e.g., by justifying the AI system's decision [Guidotti *et al.,* 2018].

Counterfactual explanations have been examined extensively in XAI, that is, explanations that indicate how the AI system's decision would have been different, if some alternative input conditions had occurred. There are over 125 distinct counterfactual explanation algorithms available [for a review, see Keane *et al.,* 2021]. Yet insights from cognitive science about how people understand counterfactuals have called into question some aspects of their use in XAI [Byrne, 2019; see also Miller, 2019]. For example, psychological evidence shows that people tend to make different inferences from causal and counterfactual explanations [Byrne, 2005; Mandel and Lehman, 1996; Orenes *et al.,* 2022]. So too, causal and counterfactual explanations have different effects on users' objective understanding of an AI system's decisions, and their subjective satisfaction with such explanations [Celar and Byrne, 2023; Warren *et al.,* 2023].

In this survey I consider how insights from cognitive science about how people construct explanations and make inferences from them can help the provision of psychologically robust explanations in XAI that are genuinely user-centred. Successful explanations should facilitate understanding and knowledge change in the user [Keil, 2006]. I first review evidence on several key tendencies that people display when they construct explanations, of relevance to automated explanations of decisions by AI systems; I then consider evidence on several tendencies that people exhibit when they reason about causal explanations.

2 Explanation Construction

People show a great variety of preferences and tendencies when they construct explanations [Keil, 2006; Lombrozo, 2016]. In this section, I illustrate some of this variety by focusing on three tendencies, including the illusion of explanatory depth, intuitive versus reflective explanations, and explanatory stances, and I sketch the potential implications of each one for XAI.

2.1 The Illusion of Explanatory Depth

People often believe that they understand how something works when in fact they have little understanding of it. They make an unrealistic judgement that they understand a system very well, but when they must explain how the system works, they become aware of deficiencies in their knowledge, and subsequently judge their understanding more accurately

[Keil, 2006]. For example, experimental participants were asked to think about how various devices work, such as how a helicopter works, or how a cylindrical lock works, or how a zipper works [Rozenblit and Keil, 2002]. They initially tended to judge that their understanding of how such devices worked was very good, and they rated their understanding on a 1-7 scale towards the higher end of the scale, indicating they believed they had a deep understanding or at least a partial understanding of each device. But after they tried to provide a detailed step-by-step explanation of how the device works, they subsequently judged their understanding of it as much lower, and their ratings tended to indicate they had a shallower understanding of it. Moreover, after they tried to answer a diagnostic question about each device, for example, how a helicopter moves from hovering to forward flight, or how to pick a cylindrical lock, their ratings of their understanding dropped even more towards the shallow end of the scale. When participants then received a detailed explanation provided by an expert of how each device works, they judged their initial understanding to have been shallow, and their now-informed understanding to be much deeper [Rozenblit and Keil, 2002]. This illusion of explanatory understanding demonstrates that the very effort of attempting to explain something, to oneself or to someone else, can reveal to a person that they do not understand how something works as well as they had previously thought [Keil, 2006].

The illusion of explanatory depth has been observed for understanding of mechanical devices, such as how a zipper works, or a pen, or a crossbow, and also for natural processes, such as earthquakes, rainbows, or tides. It tends not to be observed for knowledge of facts, procedures, or narratives, although it occurs for social judgments, such as knowledge of political policies [Alter et *al.,* 2010; Fernbach *et al.,* 2013]. It may arise because of cognitive deficits in the encoding or retrieval of relevant information, or from lack of practice in constructing explanations of complex mechanisms, or from an overly simplified model [Alter *et al.,* 2010; Fernbach *et al.,* 2013]. It seems to reflect a tendency to construct mental simulations of mechanical devices that are general and abstract, and focus on their function, rather than mental simulations that are specific and concrete, and focus on their mechanisms [Alter *et al.,* 2010; Rozenblit and Keil, 2002]. An illusion of explanatory depth has been demonstrated when users attempt to understand AI systems, such as a decision support system that provides predictions of the likelihood that a borrower defaults on a loan [Chromik *et al.,* 2021], or predictions about a person's blood alcohol content and whether they are over the legal limit to drive [Warren *et al.,* 2023].

An important implication for XAI is that the illusion of explanatory depth undermines the current reliance in many user studies on the meta-cognitive judgments of users, including scales to measure their introspections about how helpful an explanation is [e.g., Hoffman *et al.,* 2018]. The self-report judgments of users about how satisfying an explanation is, or how much it enables them to trust a system, is an unreliable guide about whether it improves their understanding. Although users may indicate that they understand an AI system's decision after they have been provided with an explanation, their judgment may be an instance of the illusion.

When users are required to carry out a key diagnostic task, such as predicting the AI system's decision, it not only provides an objective measure of their understanding of the system, it may also enable them to appreciate that they do not understand the AI's decision as well as they thought, especially when they are given feedback about the accuracy of their responses [Chromik *et al.,* 2021; Warren *et al.,* 2023]. Accordingly, a diagnostic task may enable users to calibrate their understanding, as can a task requiring them to construct an explanation [Chromik *et al.,* 2021; Hoffman *et al.,* 2018]. People may be better able to assess their understanding when they are encouraged to focus on how the parts of a device enable it to work rather than how or why it works overall [Alter *et al.,* 2010], just as when they are required to provide an explanation for how a policy works rather than provide reasons why they support it [Fernbach *et al.,* 2013]. The illusion of explanatory depth can be dispelled when people construct an explanation of the decision-making process by a human (e.g., how a doctor detects whether a skin blemish is potentially cancerous) and they come to appreciate that they do not understand the process as well as they thought they did. However, when they construct an explanation of the decision-making process of an AI system (e.g., how an AI system detects whether a skin blemish is potentially cancerous), it is less effective at dispelling the illusion [Cadario *et al.,* 2021].

Explanations of an AI system may even potentially mislead users to be overconfident of their understanding [Collaris *et al.,* 2018]. User studies need to be carefully crafted to avoid the illusion of explanatory depth [Sokol and Flach, 2020]. For example, participants who gained experience with an AI decision support system that predicts whether a person's blood alcohol level makes them over the limit to drive, self-reported greater satisfaction and trust in the AI system when they received counterfactual explanations, such as, *"John would have been under the limit if he had drunk fewer units"* compared to those who received matched causal explanations, such as, *"John was over the limit because he drank many units"* [Celar and Byrne, 2023; Warren *et al.,* 2023]. And the explanations did indeed improve their understanding of the AI's decisions, as indicated by their improved accuracy in predicting them. Yet causal explanations were just as effective as counterfactual ones in improving their prediction accuracy. In other words, a dissociation occurred between the subjective self-report measures of the effects of explanations on satisfaction and trust, which indicated users believed counterfactual explanations were better than causal ones, and the objective prediction accuracy measures of the effects of explanations on understanding, which indicated no difference between counterfactual and causal explanations [Warren *et al.,* 2023]. In some situations, counterfactual explanations are objectively better than causal ones, for example, they ensure users' own decisions (to drive or not drive) are aligned with the AI system's recommendations [Celar and Byrne, 2023]. Nonetheless, the dissociation may reflect an illusion of explanatory depth, in that users appear to believe they understand the system given counterfactual

explanations, more than users given causal explanations, yet their actual understanding of the system is affected similarly by either sort of explanation.

Dissociations between subjective and objective measures are of consequence for ethics and fairness in AI as an explanation that makes a user feel they understand a system, yet does not change their understanding, could be misleading [Warren *et al.,* 2023; see also Buçinca *et al.,* 2020]. A key implication for XAI studies of the efficacy of explanations is that rigorous experimental methods using objective measures such as accuracy of prediction are required to test users' understanding of an AI system's decisions and explanations, rather than a reliance on meta-cognitive, self-report measures of explanation satisfaction and trust. Aids to enable users to calibrate their awareness of their own understanding of an AI system may be required, such as ensuring users' attempt to explain an AI system's decision to themselves or others, or that they are required to answer diagnostic questions about the AI system's domain or decisions.

2.2 Intuitive and Reflective Explanations

People generally manage to get by with incomplete and partial explanations [Keil, 2006]. They appear to construct initial explanations based on fast and immediate intuitive thinking, which are then subsequently revised after slower and deliberative reflective thinking [Thorstad and Wolff, 2016; see also Kahneman, 2011]. People may construct explanations, for example, of how a person's salary amount contributed to an AI system's recommendation of refusal of their loan application, by relying on different sorts of empirical information, such as covariation of an input feature with an output decision, and factors such as the temporal order of events, their contiguity, or similarity [Einhorn and Hogart, 1986]. But they also seem to rely on conceptual beliefs about how an input can cause or produce the outcome. People appear to have beliefs about various sorts of causal dependencies, processes, or capacities, for example, they may have beliefs about power that explain, say, how electricity causes an engine to turn, or beliefs about force that explain, say, how the wind causes a tree to fall down, or beliefs about mechanisms that explain, say, how converting electromagnetic waves in radio causes sounds [for a review, see Johnson-Laird and Khemlani, 2017]. Their knowledge of causal relations cannot be captured simply by causal structures and strengths characterised as probabilistic dependencies [e.g., Pearl 2009], since such dependencies cannot account for the inferences people make [Sloman and Lagnado, 2015; Stephan *et al.,* 2023]. Their explanations reflect not only the current available data, but also their beliefs, activated automatically as heuristics to guide or even constrain their analytic evaluation of causal information [Fugelsang and Thompson, 2003; Verschueren *et al.,* 2005].

The initial reliance on intuitive explanatory beliefs can give rise to remarkable illusions. For example, participants were brought by elevator to a laboratory to take part in an experiment [Thorstad and Wolff, 2016]. When they were in the elevator, as the doors closed, a man standing at the back moved his hands apart, and the elevator doors opened. The man was far from the doors and made no physical contact with them. As the elevator doors started to close a second time, the man moved his hands again and the doors opened again; and as the doors began to close a third time, the man moved his hands again, and the doors opened again. The man was of course a confederate of the experiment, and a second confederate, unseen outside the elevator, pressed the button to open the doors each time they began to close. When participants arrived in the laboratory they were asked whether anything had occurred in the elevator, and most of their explanations tended to indicate that the man caused the doors to open, e.g., *"A man was controlling the doors of the elevator with his hand"*, and *"The man in the elevator kept causing the door to stay open on the wrong floor, like magic"* [Thorstad and Wolff, 2016, p. 920]. If the participants had been in the elevator without the man and the doors had repeatedly opened, perhaps they would have looked out to see whether somebody was pressing the elevator call button, or maybe they might have inferred that there was an electrical fault with the doors. But when another person in the elevator moved their arms apart and the doors opened, they did not seem to think about these other possibilities. Nonetheless, their subsequent analytic evaluation of their empirical experience led some of them to revise their explanations. Participants were asked to what extent they felt *"for a moment"* that the man caused the elevator doors to open, and their immediate hunch tended to be, "somewhat" to "very much so". They were also asked to what extent they *"ultimately concluded"* the man caused the elevator doors to open, and their further reflection tended to be only "somewhat" [Thorstad and Wolff, 2016].

Hence, when people construct a causal explanation, they seem to rely on two very different sorts of cognitive processes. They engage processes that are fast, intuitive, and automatic, perhaps based on heuristics to identify possible causes using simple cues, such as covariation, temporal contiguity, and so on. But they can also engage cognitive processes that are slower, deliberative, and controlled, perhaps based on examining underlying features, such as mechanisms. Such dual processes of fast and slow thinking underlie many sorts of decisions and inferences [Kahneman, 2011]. For example, people can very rapidly, within a time limit of just seconds, construct intuitive counterfactual explanations to try to justify a decision they otherwise consider unjustified [Tepe and Byrne, 2021]. Their subsequent more reflective counterfactual explanations tend to expand on these initial thoughts rather than develop alternative explanations.

An implication for XAI is that explanations provided by AI systems may match or mismatch users' immediate intuitions, or their more deliberative reflections, and hence be more or less successful as a result. For example, interpretability tools designed to assist data scientists to understand machine learning models can inadvertently encourage reliance on intuitive explanations [Kaur *et al.,* 2020]. Data scientists were asked to consider a machine learning adult income dataset, based on input features such as age, education, marital status, and so on, and output information about whether each person made a salary above a certain amount. They were provided with different sorts of interpretatbility tools which helped them visualize global and local explanations. But the

tools appeared to lead them to make quick decisions about the adequacy of the model based on superficial narrative evaluations, rather than reflective decisions based on critical evaluation [Kaur *et al.,* 2020]. Even more so for non-expert users, a user-centred XAI system may need to rule out intuitive explanations that are inaccurate, perhaps through interactive and iterative engagement between the user and the AI system. An important goal should be to ensure users are receptive to engaging in considered reasoning about the system, encouraged to generate potential counterexamples to putative conclusions, so they critically assess decisions and explanations.

2.3 Explanatory Stances

People expand their knowledge in real time when they understand an explanation [Keil, 2006]. They tend to adopt a variety of "explanatory stances" when they attempt to understand something, such as a *mechanical* stance, a *design* stance, or an *intentional* stance [Dennett, 1987; 1988]. Each of these stances provides a very different sort of explanation.

For example, from a physical or *mechanical* stance, a person may explain, say, how a bird flies, by describing the physical constitution of the system, that is, the mechanics of flying, considering thrust, propulsion, and so on. Mechanistic explanations can be of different sorts, including constitutive, that is, how a process works or what it is made up of, say, the role of the lightness of a bird's bones in flight; or etiological, that is, how something came to be or developed, say, the evolutionary developments of flight [Joo *et al.,* 2021]. Constitutive mechanisms explain the causal process of interacting parts, e.g., how a clock ticks; etiological mechanisms explain the chain of events that caused the outcome, e.g., how a tree grew its leaves [Joo *et al.,* 2021].

Alternatively, from a *design* stance, a person may explain how a bird flies by describing its elements and their functions, and assuming it will behave as it is designed to behave, for example, that a bird flies because wings are made for flying. Such design explanations, also known as teleological explanations, emerge early in children, and even adults can prefer them to at least some non-teleological, mechanistic ones [Joo *et al.,* 2021]. Scientifically questionable teleological explanations can be considered an "explanatory vice" [Lombrozo, 2016] and may contribute to intelligent design explanations that an agent made an item to work as it does [Keil, 2006].

From an *intentional* stance instead, a person may explain how a bird flies by describing the beliefs and desires a bird has about what it needs to do to be able to execute a move such as lift. Intentional stance explanations refer to mental states that have consequences for behavior [Dennett, 1987]. It is an explanation strategy that attributes beliefs and desires to systems, and predicts their behaviour based on what a system with those beliefs and desires would reasonably do. People often adopt an intentional stance to make sense of behaviour, not only of other people but also of other animals and artifacts such as computer programs [Dennett, 1988].

Notably, each explanatory stance can be applied to explain the same device or action, but they have different consequences for understanding it. Each stance can lead to different kinds of insights, and to different kinds of erroneous inferences. The atypical application of a particular stance, say, a mechanical stance to explain an action more typically understood from an intentional stance, such as explaining travelers in a crowded airport as like pinballs careening around a pinball machine, may be interpreted analogically to yield new inferences [Keil, 2006].

People may tend to adopt multiple stances in their preferred explanations of an AI decision support system and its decisions, not unlike their tendencies in interacting with social robots [Clark and Fischer, 2023]. People are aware that a social robot is a machine, but interpret it as a *depiction* of a character, not unlike a ventriloquist dummy, and engage with it in pretense of interacting with the depicted character [Clark and Fischer, 2023]. Similarly, they may be aware that an AI decision support system is an algorithm but they may interpret its decisions as a *depiction* of those provided by a human, e.g., a bank loan assessor, or the organization the human represents, a bank. Hence, an intentional stance and a design stance may both be useful in different contexts for explaining how automated agents behave [Veit and Browning, 2023].

A potential implication for XAI is the necessity to consider when a particular stance is appropriate for explaining an AI system's decision. A question posed by a user as to why an AI system refused their loan may be a request for causal information of a mechanistic sort about how the AI system came to make that decision. Hence, it may require an explanation based on information about how the user's input features relate to a training data set, e.g., that loans have been refused for applicants of similar salary level, occupational status, credit history, as the user. Alternatively, it may be a request for functional information of a teleological sort about the purpose of the AI's decision. Accordingly, it may require instead information about the goal and consequences of the output, e.g., that decisions of this sort mitigate the risk of applicants defaulting from repayment. Each sort of explanation will lead users to develop a different understanding of aspects of an AI system, and impact their learning, satisfaction, and trust in it.

3 Causal Explanations

Causal explanations are central in the psychology of explanation [Einhorn and Hogart, 1986; Keil, 2006]. Causality is a complex and nuanced concept [Johnson-Laird and Khemlani, 2017]. I illustrate some of the rich discoveries of how people reason about causal explanations by considering each of the following tendencies in turn: enabling causes and inference suppression, causal discounting, and explanatory simplicity. I outline the potential implications of each one for XAI.

3.1 Enabling Causes and Inference Suppression

People distinguish between different sorts of causes. A *strong cause* refers to a one-to-one mapping between a single cause, say, a lightning storm, and a single effect, say, a forest fire. A strong cause is enough to bring about the outcome, and it is necessary in that the outcome will only occur when the cause occurs. Of course, in daily life, there are many causal relationships other than a one-to-one mapping between a single cause and a single outcome. An *enabling cause*, say, dry

leaves on the forest floor, is an additional background condition that must be met for the outcome to occur, the forest fire. An enabling condition is not enough by itself to bring about the outcome, but it is necessary.

The distinction between different causes is important because people make very different inferences from strong causes compared to enabling causes. They resist even the most obvious causal inferences when they know about background enabling conditions that also need to be met [Byrne, 1989]. For example, when participants are given a conditional about a strong cause, such as, "*if there was a lightning storm there was a forest fire"* and they are told *"there was a lightning storm"*, most of them make the simple *modus ponens* inference, *"there was a forest fire"*. But when they are also told about an enabling condition, *"if there was a lightning storm there was a forest fire, if there were dry leaves on the forest floor there was a forest fire"*, the inference, from *"there was a lightning storm"* to *"there was a forest fire"* is suppressed, that is, participants make far fewer of such inferences. They understand that there would also have to be dry leaves on the forest floor. Likewise, the *modus tollens* inference, from *"there was no forest fire"* to *"there was no lightning storm"* is also suppressed [Byrne, 1989].

The suppression of inferences when people think about enabling conditions is widespread in causal thinking. The presence of background enabling conditions affects how people make choices in various situations, ranging from which items to buy, to their legal decisions [Chandon and Janiszewski, 2008; De Neys *et al.,* 2003; Gazzo Castañeda and Knauff, 2016]. Inference suppression occurs even when participants are not explicitly told about enablers, but are prompted to retrieve them instead [Cummins *et al.,* 1991].

Inference suppression, that is, the decreased frequency of inferences, arises because people envisage different possibilities for the different sorts of causes [Byrne *et al.,* 1999; Goldvarg and Johnson-Laird, 2001]. For a strong cause, they envisage the cause and effect, *a lightning storm and a forest fire*, and they also envisage the absence of the cause and absence of the effect, *no lightning storm and no forest fire*. In contrast for an enabling cause, they envisage these two possibilities, but they also simulate a third one, the cause and the absence of the effect, *a lightning storm and no forest fire*. The simple inference is suppressed because people construct a mental model that makes explicit a counterexample, *a lightning storm, but no dry leaves, and no forest fire*. Notwithstanding the importance of mechanism and related factors, the fundamental meaning of causality may depend on the mental simulation of such different possibilities [Johnson-Laird and Khemlani, 2017].

People often tend to focus on a strong cause in their communications with others because they believe that others will assume the presence of the relevant background enabling conditions [Hilton, 1996]. The difference between causes and enablers may be that causes are inconstant whereas enablers are constant; or causes violate the usual normal situation whereas enablers do not; or causes are abnormal whereas enablers are normal; or causes are conversationally relevant whereas enablers are not [Cheng and Novick, 1991; Einhorn and Hogarth, 1986; Hart and Honoré, 1985; Hilton, 1996]. In any case, people can readily distinguish causes and enablers [Frosch and Byrne, 2012; Goldvarg and Johnson-Laird, 2001]. They also can infer the normality of a cause, and in situations in which both a normal and abnormal cause lead to an outcome, they prefer an explanation based on the abnormal cause, at least when both causes are necessary [Kirfel *et al.,* 2022]. The difference between causes and enablers cannot be captured readily by probabilities since the conditional probabilities of an outcome given the cause or enabler can be high for either [cf. Oaksford and Chater, 2017; see Johnson-Laird and Khemlani, 2017].

An important implication for XAI is that when users are provided with a causal explanation for an AI system's decision, they may spontaneously notice or retrieve additional background enabling conditions. Such retrieved enablers may suppress inferences that users would otherwise be expected to make, that is, they will make fewer such inferences or even refrain from making them at all. If they retrieve even a single counterexample, it can suppress an inference, even though it could be an exception. For example, an AI system's decision, such as a prediction that a person is over the legal limit to drive, may result from a set of input features, some of which are strong causes of the output (e.g., number of units of alcohol drunk), and some of which are enabling conditions (e.g., an empty stomach) [Celar and Byrne, 2023; Warren*, et al.,* 2023]. Explanations for the AI decision that focus on a strong cause, *"if the person drank 6 units, they were over the legal limit to drive"* may not be as helpful to users as they may seem. Although the explanation invites the simple inference from "*the person drank 6 units"* to *"they were over the legal limit to drive*", the inference may be suppressed if users notice or retrieve enabling conditions that allow them to construct a counterexample, *the person drank 6 units, but they had a full stomach, and they were not over the legal limit to drive.* They may consider the explanation provided by the AI system to be misleading or incorrect since it did not acknowledge the enabling conditions. Inference suppression also occurs for counterfactuals, e.g., *"if the person had drunk 6 units, they would have been over the legal limit to drive"* [Espino and Byrne, 2020]. Explanations for AI decisions may need to consider the enabling conditions of which users are aware, and make clear in explanations whether the enabling conditions have been met. Automated explanations need to be not only machine robust [Ferrario and Loi, 2022; Virgolin and Fracaros, 2023], but also psychologically robust, interpreted by users in the way that they are intended.

3.2 Causal Discounting

When people know about several alternative causes for an effect, they tend to discount some of them. For example, a person may infer that a wet lawn in the morning results from the proper working of their new overnight lawn sprinkler; but if they subsequently hear reports that it rained during the night, they may conclude instead that the wet lawn is a result of the

rain. The two causes are not mutually exclusive, it can rain while the sprinkler is working, but the presence of an alternative cause can lead people to resist inferring that the original cause led to the outcome. Many psychological experiments demonstrate that people tend to engage in such discounting, and a related "explaining away" of the likelihood of causes [see Khemlani and Oppenheimer, 2011, for a review]. Participants devalue one of the causes perhaps because the other "raises the bar" for decisions about its contribution [Laux *et al.,* 2010; see also Sloman and Lagnado, 2015].

People discount a cause based not only on the extent to which it covaries with the effect but also based on views about whether it is a believable cause [Fugelsang and Thompson, 2001]. When presented with multiple-cause scenarios, they appear to form theories of how the causes interact, inhibiting or enabling each other [Fugelsang and Thompson, 2001]. When a normal and abnormal cause lead to an outcome, they prefer an explanation based on the normal cause, when either cause is sufficient to bring about the outcome, unlike when both causes are necessary [Kirfel *et al.,* 2022].

Causal discounting is closely related to a second sort of suppression of inferences, this time from weak causes rather than enabling conditions. People make different inferences from a *many-to-one* weak causal relationship of several alternative causes leading to a single effect, compared to a *one-to-one* strong causal relationship of a single cause leading to a single effect. The *strong* causal relationship between a single cause, *it rained*, and a single effect, *the lawn is wet*, is contrasted in this case with a *weak* causal relationship, between several alternative causes, *it rained, the sprinkler was on,* and the effect, *the lawn was wet*. Each of the causes is sufficient to bring about the effect but neither one is necessary.

Participants make very different inferences from strong causes and weak causes [Rumain *et al.,* 1983]. When they are given a conditional about a strong cause, such as, *"if it rained the lawn was wet"* and they are told *"the lawn was wet"* many of them make the *affirmation of the consequent* inference, *"it rained"*. But when they are instead told about alternative causes, such as, *"if it rained the lawn was wet, if the sprinkler was on, the lawn was wet"*, the inference from *"the lawn was wet"* to *"it rained"* is suppressed, that is, people make far fewer of these inferences. Likewise, the *denial of the antecedent* inference from *"it did not rain"* to *"the lawn was not wet"* is also suppressed [Rumain *et al.,* 1983].

Causal discounting may arise because people envisage different possibilities for the strong *one-to-one* cause and effect relation, compared to the weak *many-to-one* alternative causes. For a strong cause, they envisage the cause and the effect, *rain and a wet lawn*, and the absence of the cause and absence of the effect, *no rain and no wet lawn*; but for a weak cause, they simulate these two possibilities and a third one, the absence of the cause but the presence of the effect, *no rain but a wet lawn*. The inferences are suppressed because people construct a mental model that makes explicit a counterexample, *no rain, but the sprinkler on, and a wet lawn* [Byrne, 1989]. Causal discounting is thus readily predicted based on the mental models of possibilities that people envisage [cf. Hall *et al.,* 2016].

An implication for XAI is that when human users are given an explanation for an AI system's decision that depends on a many-to-one mapping of several alternative causes for a single effect, they may be susceptible to such discounting effects. They may disregard the proposed causal explanation if they know of other alternative causes. Consider an AI decision support system that predicts that grass growth on a farm will be poor in the month ahead [Dai *et al.,* 2022]. The set of input features includes several alternative causes of poor grass growth, e.g., competition from weeds, poor soil. Explanations for the AI decision that focus on a single cause, *"if there had been less competition from weeds, grass growth would have been good"* may not be as helpful to users as they may at first seem. Although the explanation invites an inference from "*grass growth was poor",* to *"there was too much competition from weeds",* the inference may be suppressed if users notice or retrieve alternative causes that allow them to construct a counterexample, *grass growth was poor, but there was no competition from weeds, there was poor soil.* Once again, they may consider the explanation provided by the AI system to be misleading or incorrect since it did not acknowledge alternative causes. User-centred explanations for AI decisions may need to consider alternative causes of which users are aware, and make clear in explanations whether the alternative causes have been ruled out. Even though AI systems may appear to exhibit robustness and fidelity in the explanations provided to users, people may over-ride these explanations based on their own interpretations of the relevant causal relations.

3.3 Simplicity of Explanations

People prefer simpler explanations, at least in terms of the number of causes for effects [Lombrozo, 2007]. They tend to judge themselves satisfied by explanations based on simple causal relationships over more complex ones, in that they prefer a *common cause* explanation in which one cause results in several effects, rather than an explanation in which multiple independent causes result in the different effects [Lombrozo, 2007]. For example, participants were told about three putative causes for two effects. The first was a *common cause* relation, of one cause leading to two effects, e.g., *"disease A causes symptom 1 and symptom 2".* The second was a *single-effect cause* that referred to a cause and one of the effects, and also ruled out any relation between the cause and the second effect, e.g., *"disease B causes symptom 1 and it does not cause symptom 2".* The third was the opposite, *"disease C causes symptom 2 and it does not cause symptom 1".* When participants were told of an instance in which the two effects were present, *"Treda has symptom 1 and symptom 2",* they tended to judge as the most satisfying explanation the simple *common-cause* explanation, *"she has disease A",* rather than the multiple *single-effect cause* explanation, *"she has disease B and disease C"* [Lombrozo, 2007].

People prefer such simple explanations to more complex ones, not only in deterministic situations where the cause

always leads to the effect, but even (albeit to a weaker extent) in stochastic situations where the cause sometimes leads to the effect [Johnson et *al.,* 2019]. Their preference for simple explanations varies for different domains, such as physics, biology, artifacts, and social domains, for example, it is strong for physical systems but not so for social causal systems [Johnson et *al.,* 2019]. Their preference for a simpler explanation as offered by a common cause persists until considerable evidence is amassed that the more complex explanation has greater probability [Lombrozo, 2007].

People also tend to spontaneously generate cognitively simpler explanations rather than more complex ones that require multiple possibilities to be simulated. For example, when participants wrote a diary entry to reflect on a fictional series of events, they spontaneously created twice as many causal explanations that focused on the facts, e.g., *"I haven't made any friends yet in this new town because I didn't go to my neighbor's party",* compared to counterfactual explanations, e.g., *"I would have made friends by now in this new town if I had gone to my neighbor's party".* The counterfactual requires the simulation of multiple possibilities including the imagined conjecture, and the facts [McEleney and Byrne, 2006; see also Dixon and Byrne, 2011].

Of course, the simplicity of an explanation depends not only on the simplicity of the relation between a cause and an effect, but also on the conceptual simplicity of the cause and the effect themselves. Although simplicity can be viewed as an "explanatory virtue" and a useful heuristic, it may have its limits in circumstances in which a simple explanation does not fit the data as well as a complex one [Johnson *et al.,* 2019; see also Lombrozo, 2016]. In some situations, people appear to prefer instead a *complex* explanation. For example, when asked to consider someone with disease A, and someone with disease B and disease C, and judge who is more likely to have symptom 1 and symptom 2, they favor the more complex explanation rather than the simpler *common cause* one [Johnson *et al.,* 2019]. Moreover, when they know only that symptom 1 occurred and not whether symptom 2 occurred, they prefer the narrow scope explanation, *"disease B causes symptom 1",* rather than the simpler *common cause* one [Khemlani *et al.,* 2011]. In some situations, people also appear to consider that the strength of a cause with multiple effects is diluted compared to a cause with a single effect [Stephan *et al.,* 2023]. Relatedly, for everyday explanations such as why China's population is rising despite their one-child policy, people prefer explanations that contain more causes, e.g., because ethnic minorities and rural communities are exempt, Chinese people are living longer and wealthy people can pay fines for violating the policy, rather than simpler explanations that refer to just one cause [Zemla *et al.,* 2017]. For such topics they seem to want it to be fully accounted for; they also seem to assume the causes provided are true even when they are not, e.g., Chinese people are not living longer [Zemla *et al.,* 2017].

An implication for XAI is that in at least some situations, users may prefer simple explanations for an AI system's decision, such as a *common cause* explanation of a single cause that has several different effects, rather than a more complex explanation of multiple different causes for different effects. Many automated explanations in XAI offer an explanation based on a change to a single input, e.g., *"if you had earned $10,000 more, your loan application would have been approved".* In some situations, users may benefit from explanations that elaborate instead on a common cause that has several effects; in other situations, they may benefit from explanations that provide enough information to appear to fully account for an outcome. Such explanations may be more effective in improving satisfaction and trust, even if they impose additional constraints on the provision of automated explanations in XAI.

4 Conclusions

Given the widespread use of AI decision support systems, there is a recognised need for the development of automated explanations of their decisions to ensure fairness and transparency. I suggest that central to that endeavour is the requirement that the automated explanations provided by an AI system must make sense to human users. Knowledge of key discoveries in cognitive science about human explanatory inference is crucial for XAI to ensure that explanations generated by AI systems can interface directly with tendencies people exhibit in explanation construction and causal reasoning.

A limitation of a survey of this nature is that it can highlight only a limited number of psychological phenomena, from among the rich and diverse research on human explanatory reasoning. For example, the choice to focus on causal explanations reflects the vast research on them, but necessarily leads to omission of other sorts of explanations such as graphical ones. Similarly, a focus on individual cognitive tendencies omits consideration of the impact of social and cultural factors. It is also vital to acknowledge the real concern that AI systems themselves may introduce biases into decisions and explanations, and current explanation strategies in XAI may not fully address them. Although factors such as algorithmic tractability, computational accessibility, context of use, and so on, place core demands on the development of explanations for AI decisions, it is crucial to recognise that XAI automated explanations can only be robust and effective if human users understand them appropriately.

Acknowledgments

I am grateful to Mark Keane, Greta Warren, Lenart Celar, and Xinyue Dai for helpful discussions on explanations in XAI.

References

[Adadi and Berrada, 2018] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box. *IEEE access*, *6*, 52138-52160.

[Alter *et al.,* 2010] Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest. *Journal of Personality and Social Psychology*, *99* (3), 436-451.

[Buçinca *et al.,* 2020] Buçinca Z., Lin, P. Gajos, K.Z., & Glassman, E.L. (2020). Proxy tasks and subjective

measures can be misleading in evaluating explainable AI systems. *IUI '20*, 454–464.

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

[Byrne, 2005] Byrne, R.M.J. (2005). *The Rational Imagination*. Cambridge, MA: MIT press.

[Byrne, 2019] Byrne, R.M.J. (2019). Counterfactuals in explainable Artificial Intelligence (XAI). In *IJCAI-19,* 6276-6282.

[Byrne *et al.,* 1999] Byrne, R.M.J., Espino, O., & Santamaria, C. (1999). Counterexamples and the suppression of inferences. *Journal of Memory and Language*, *40*(3), 347-373.

[Cadario *et al.,* 2021] Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour*, *5*(12),1636-1642.

[Celar and Byrne 2023] Celar, L., & Byrne, R.M.J. (2023). How people reason with counterfactual and causal explanations for Artificial Intelligence decisions in familiar and unfamiliar domains. *Memory & Cognition*, 1-16.

[Chandon and Janiszewski, 2008]. Chandon, E., & Janiszewski, C. (2008). The influence of causal conditional reasoning on the acceptance of product claims. *Journal of Consumer Research*, *35*(6),1003–1011.

[Cheng and Novick, 1991]. Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, *40*(1-2), 83-120.

[Chromik *et al.,* 2021] Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021). I think I get your point, AI! In *IUI'21*: *Proceedings of 26th Intelligent User Interfaces,* 307-317.

[Clark and Fischer, 2023] Clark, H.H., & Fischer, K. (2023). Social robots as depictions of social agents. *Behavioral and Brain Sciences*, *46*, e21.

[Collaris *et al.,* 2018] Collaris, D., Vink, L. M., & van Wijk, J.J. (2018). Instance-level explanations for fraud detection. *arXiv:1806.07129*.

[Cummins *et al.,* 1991] Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition*, *19* (3), 274–282.

[Dai *et al.,* 2022] Dai, X., Keane, M. T., Shalloo, L., Ruelle, E., & Byrne, R.M.J. (2022). Counterfactual explanations for prediction and diagnosis in XAI. In *AIES'22*, 215-226.

[De Neys *et al.,* 2003] De Neys, W., Schaeken, W., & D'Ydewalle, G. (2003). Inference suppression and semantic memory retrieval. *Memory & Cognition*, *31* (4), 581-595.

[Dennett, 1987] Dennett, D.C. (1987). *The Intentional Stance*. Cambridge MA: MIT press.

[Dennett, 1988] Dennett, D.C. (1988). Précis of The Intentional Stance. *Behavioral and Brain Sciences*, *11* (3), 495-505.

[Dixon and Byrne, 2011]. Dixon, J. E., & Byrne, R. M. J. (2011). "If only" counterfactual thoughts about exceptional actions. *Memory & Cognition*, *39*, 1317-1331.

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

[Einhorn and Hogarth, 1986] Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99* (1), 3-19.

[Fernbach *et al.,* 2013] Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*, *24* (6), 939-946.

[Ferrario and Loi, 2022] Ferrario, A., & Loi, M. (2022). The robustness of counterfactual explanations over time. *IEEE Access*, *10*, 82736-82750.

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

[Fugelsang and Thompson, 2001] Fugelsang, J. A., & Thompson, V. A. (2001). Belief-based and covariation-based cues affect causal discounting. *Canadian Journal of Experimental Psychology*, *55*(1),70-76.

[Fugelsang and Thompson, 2003] Fugelsang, J. A., & Thompson, V. A. (2003). A dual-process model of belief and evidence interactions in causal reasoning. *Memory & Cognition*, *31* (5), 800-815.

[Gazzo Castañeda and Knauff, 2016] Gazzo Castañeda, L. E., & Knauff, M. (2016). Defeasible reasoning with legal conditionals. *Memory & Cognition*, *44* (3), 499–517.

[Goldvarg and Johnson-Laird, 2001] Goldvarg, E., & Johnson-Laird, P.N. (2001). Naive causality. *Cognitive Science*, *25* (4), 565-610.

[Guidotti *et al.,* 2018] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *CSUR*, *51* (5), 1-42.

[Hall *et al.,* 2016] Hall, S., Ali, N., Chater, N., & Oaksford, M. (2016). Discounting and augmentation in causal conditional reasoning. *PloS one*, *11*(12), e0167741.

[Hart and Honoré, 1985] Hart, H.L.A., & Honoré, T. (1985). *Causation in the Law*. Oxford: Oxford University Press.

[Hilton, 1996]Hilton, D.J. (1996). Mental models and causal explanation. *Thinking & Reasoning*, *2* (4), 273-308.

[Hoffman *et al.,* 2018] Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arxiv.org/abs/1812.04608*

[Johnson *et al.,* 2019] Johnson, S. G., Valenti, J. J., & Keil, F. C. (2019). Simplicity and complexity preferences in causal explanation. *Cognitive psychology*, *113*, 101222.

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

Waldman (Ed). *Oxford Handbook of Causal Reasoning.* (pp. 169-188). Oxford: Oxford University Press.

[Joo *et al.,* 2021] Joo, S., Yousif, S. R., & Keil, F. (2021). What is a 'mechanism'? In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. 1609-1613.

[Kahneman, 2011] Kahneman, D. (2011). *Thinking, Fast and Slow*. NY: Farrar, Straus and Giroux.

[Karimi *et al.,* 2020] Karimi, A.H., Barthe, G., Schölkopf, B. & Valera, I. (2020). A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv*:*2010.04050*.

[Kaur *et al.,* 2020] Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability. In *Proceedings of Human Factors in Computing Systems*, 1-14.

[Keane *et al.,* 2021] Keane, M. T., Kenny, E. M., Delaney, E. and Smyth, B. (2021). If only we had better counterfactual explanations. In *IJCAI-21,* 4466-4474.

[Keil, 2006] Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology, 57*, 227-254.

[Kenny *et al.,* 2021] Kenny, E. M., Ford, C., Quinn, M., & Keane, M.T. (2021). Explaining black-box classifiers using post-hoc explanations-by-example. *Artificial Intelligence*, *294*, 103459.

[Khemlani and Oppenheimer, 2011] Khemlani, S. S., & Oppenheimer, D.M. (2011). When one model casts doubt on another. *Psychological Bulletin*, *137*(2), 195-210.

[Khemlani *et al.,* 2011] Khemlani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2011). Harry Potter and the sorcerer's scope. *Memory & Cognition*, *39*, 527-535.]

[Kirfel *et al.,* 2022] Kirfel, L., Icard, T., & Gerstenberg, T. (2022). Inference from explanation. *Journal of Experimental Psychology: General*, *151*(7), 1481-1501.

[Laux *et al.,* 2010] Laux, J.P., Goedert, K.M., & Markman, A.B. (2010). Causal discounting in the presence of a stronger cue is due to bias. *Psychonomic Bulletin & Review*, *17*, 213-218.

[Lombrozo, 2007] Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55* (3), 232-257.

[Lombrozo, 2016] Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*(10), 748-759.

[Mandel and Lehman, 1996] 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.

[McEleney and Byrne, 2006] McEleney, A., & Byrne, R.M.J. (2006). Spontaneous counterfactual thoughts and causal explanations. *Thinking & Reasoning*, *12*(2), 235-255.

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

[Oaksford and Chater, 2017] Oaksford, M., & Chater, N. (2017). Causal models and conditional reasoning. In M. Waldman (Ed.), *Oxford Handbook of Causal Reasoning* (pp. 327–346). Oxford: Oxford University Press.

[Orenes *et al.,* 2022] Orenes, I., Espino, O., & Byrne, R. M.J. (2022). Similarities and differences in understanding negative and affirmative counterfactuals and causal assertions. *Quarterly Journal of Experimental Psychology*, *75*(4), 633-651.

[Pearl, 2009]. Pearl, J. (2009). *Causality*. Cambridge: Cambridge University Press.

[Rozenblit and Keil, 2002] Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26* (5), 521-562.

[Rumain *et al.,* 1983] Rumain, B., Connell, J., & Braine, M. D. (1983). Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults. *Developmental Psychology*,*19*(4),471-481.

[Sloman and Lagnado, 2015] Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, *66*, 223-247.

[Sokol and Flach, 2020] Sokol, K., & Flach, P. (2020). Explainability fact sheets. In *Proceedings of Fairness, Accountability, and Transparency*, 56-67.

[Stephan *et al.,* 2023] Stephan, S., Engelmann, N., & Waldmann, M. R. (2023). The perceived dilution of causal strength. *Cognitive Psychology*, *140*, 101540.

[Tepe and Byrne, 2022] Tepe, B., & Byrne, R.M.J. (2022). Cognitive processes in imaginative moral shifts. *Memory & Cognition*, *50*(5), 1103-1123.

[Thorstad and Wolff, 2016] Thorstad, R., & Wolff, P. (2016). What causal illusions might tell us about the identification of causes. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 991-996.

[Veit and Browning, 2023] Veit, W., & Browning, H. (2023). Social robots and the intentional stance. *Behavioral and Brain Sciences*, *46*, e47.

[Virgolin and Fracaros, 2023] Virgolin, M., & Fracaros, S. (2023). On the robustness of sparse counterfactual explanations to adverse perturbations. *Artificial Intelligence*, *316*, 103840.

[Verschueren *et al.,* 2005] Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2005). A dual-process specification of causal conditional reasoning. *Thinking & Reasoning*, *11* (3), 239-278.

[Warren *et al.,* 2023] Warren, G., Byrne, R.M.J., & Keane, M.T. (2023). Categorical and continuous features in counterfactual explanations of AI systems. In *IUI'23*,171-187.

[Zemla *et al.,* 2017] Zemla, J. C., Sloman, S., Bechlivanidis, C., & Lagnado, D. A. (2017). Evaluating everyday explanations. *Psychonomic Bulletin & Review*, *24*, 1488-1500.