# Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning

**Conference Paper** · August 2019

DOI: 10.24963/ijcai.2019/876



**1 author:**

Ruth M.J. Byrne
Trinity College Dublin, University of Dublin, Ireland
**154** PUBLICATIONS **8,386** CITATIONS

# Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning

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

## Abstract

Counterfactuals about what could have happened are increasingly used in an array of Artificial Intelligence (AI) applications, and especially in explainable AI (XAI). Counterfactuals can aid the provision of interpretable models to make the decisions of inscrutable systems intelligible to developers and users. However, not all counterfactuals are equally helpful in assisting human comprehension. Discoveries about the nature of the counterfactuals that humans create are a helpful guide to maximize the effectiveness of counterfactual use in AI.

## 1 Introduction

Counterfactuals have become essential in many aspects of Artificial Intelligence. Their relevance to a variety of AI applications has been known for some time, ranging from subgoal construction to the identification of planning failures, from fault diagnosis to the determination of liability [Ginsberg, 1986; Halpern and Pearl, 2005]. Counterfactuals are used in AI in many diverse ways. One use is to supplement incomplete data. For example, information from the log data of recommender systems provides only partial feedback, such as the number of recommended articles a user downloaded. "Counterfactual risk minimization" algorithms aim to improve learning by estimating which articles the user would have downloaded if they had received a different set of recommendations [Swaminathan and Joachims, 2015]. Another use is in imperfect-information games such as poker. For example, "counterfactual regret minimization" ensures a deep learning algorithm adapts its strategy over successive iterations in self-play, by assigning values to reflect an opponent's hand using "deep counterfactual value networks" [Moravčík *et al*., 2017]. Counterfactuals are also used in generative adversarial networks (GANs) to improve training [Neal *et al*., 2018].

Perhaps the most striking use of counterfactuals in AI at present is in explainable AI. AI systems designed to guide complex tasks, which can range from decisions about creditworthiness to criminal sentencing, from social network information availability to automated vehicles, are often based on artificial neural networks (ANNs) trained on vast amounts of data which can produce behavior that appears unintelligible [Weld and Bansal, 2018]. To increase trust in such systems by human users, and accuracy in their training by designers, there is a need to enable AI systems to provide dynamic or *ad hoc* explanations of their decisions in ways that are intelligible to humans [Biran and Cotton, 2017]. An interpretable model allows a human user to mentally simulate some aspects of the system, and to understand the causes of its decision-making [Hoffman *et al*., 2018]. It enables the user to consider contrastive explanations and counterfactual analyses, such as why one decision was made instead of another, and to predict how a change to a feature will affect the system's output [Miller, 2019]. However, the number of counterfactuals that can be generated to explain any event is potentially limitless and it is a non-trivial problem to identify which counterfactuals best facilitate the construction of an explanatory model.

There is now a wealth of experimental data available from psychology and cognitive science about the capacities of human reasoners to comprehend and reason from counterfactuals, and about the sorts of counterfactuals they create. The use of counterfactuals in AI can benefit from incorporating insights from these discoveries. People use counterfactuals often in daily life and they create alternatives to reality guided by rational principles [Byrne, 2005]. This survey focuses on six discoveries about the ways in which people think counterfactually, relevant to XAI. First, people tend to create counterfactuals that add new information to what they know about the facts, rather than ones that delete information; the former aid creative problem solving whereas the latter aid logical reasoning. Second, people tend to create counterfactuals that imagine how the outcome could have been better, rather than worse. Counterfactuals about a better outcome aid the formation of intentions for the future whereas those about a worse outcome amplify positive emotions. Third, counterfactuals help people construct explanations by ensuring they identify cause-effect or reason-action relations between events. However, counterfactuals tend to focus on background enabling conditions that prevent a bad outcome whereas causal thoughts tend to focus on strong causes that co-vary with the outcome. Fourth, counterfactuals about how an outcome could have been different if an antecedent action had been different increase people's ascriptions of blame and fault to the action, whereas *semi-factuals* about how the outcome could have been the same "even if" the action had been different, reduce blame and fault ascriptions.

Fifth, people tend to imagine how an outcome could have been different by changing antecedents that are exceptional, controllable, recent, and action-based. They do not tend to focus on the most improbable events. Sixth, counterfactuals enable people to make inferences such as *modus tollens*, which they otherwise find difficult, because counterfactuals ensure that people simulate multiple possibilities.

The sorts of counterfactuals that people create have been established in many different sorts of situations. Some experiments have required people to remember episodes from their own lives. Some require people to read stories about hypothetical events that happened to fictional protagonists concerning accidents or illnesses or a wide range of other sorts of content. Others require participants to engage in a task, such as solving puzzles. The experiments measure the types of counterfactuals people create, how quickly they read counterfactuals, the inferences they make, or they use eye-tracking to record where people look in a visual display when they hear a counterfactual, or brain imaging techniques such as event related potentials (ERP) or functional magnetic resonance imaging (fMRI) to examine areas of the brain activated during counterfactual comprehension and reasoning [for a review, see Byrne, 2016]. To illustrate the many diverse findings, the results will be described with reference throughout to the example of decisions made by an autonomous vehicle, to convey some of the relevance of the findings for XAI.

## 2 Counterfactual Structure

People tend to create counterfactuals about how things could have been different that add something new to what they already know about the situation, rather than ones that remove something from it. They tend to imagine how the outcome could have turned out better rather than how it could have turned out worse. These two tendencies are described in turn.

### 2.1 Additive and Subtractive Counterfactuals

People construct a model of the facts of a situation and they create a counterfactual by selecting for mutation an aspect of the situation that they have explicitly represented in their model. Suppose a human is attempting to understand the decision of an autonomous vehicle to swerve to avoid hitting a pedestrian and is considering why it did not brake instead. Suppose that the outcome of the decision was bad, the car hit a wall and its passenger sustained some minor injuries. An additive counterfactual adds some entirely new information to the simulation of reality, such as, "if the car had detected the pedestrian earlier and braked, the passenger would not have been injured". A subtractive counterfactual, in contrast, is constructed by deleting something already simulated about the factual reality, such as "if the car had not swerved and hit the wall, the passenger would not have been injured". Additive counterfactuals go beyond the information given to add new, extra information, whereas subtractive counterfactuals are restricted to modifications of the information given. People tend to create additive counterfactuals more than subtractive ones [Roese and Epstude, 2017].

The sorts of counterfactuals that people create have effects on their subsequent reasoning. Additive counterfactuals aid creative problem solving, whereas subtractive ones aid logical reasoning. For example, adults were asked to think about a bad thing that happened to them in the past and one group was instructed to generate an additive counterfactual, i.e., to complete a sentence stem "if I had… then the outcome would have been better/worse", and another group was instructed to generate a subtractive counterfactual, i.e., to complete a sentence stem "if I had not… then the outcome would have been better/worse". The additive counterfactual group performed better than the subtractive one on subsequent creative problem-solving tasks such as generating creative uses for an object [Markman *et al*., 2007]. The subtractive counterfactual group performed better than the additive one on subsequent logical reasoning tasks such as syllogistic inferences.

Since a person would tend to consider an additive counterfactual rather than a subtractive one, it may be effective in many situations for an AI agent in an explanatory exchange with a human user to do so. The provision of additive counterfactuals may also promote further creative problem-solving about other aspects of the system.

### 2.2 Better-Worlds and Worse-Worlds

People tend to imagine how things could have turned out differently most often after bad outcomes, such as accidents, illnesses or goal failures, although they also do so after exceptionally good outcomes, such as near misses or lucky wins. Most people tend to imagine how things could have been better rather than worse [Markman *et al*., 1993; Rim and Summerville, 2014]. For example, after a bad outcome such as an injury to a passenger, people tend to imagine how things could have been better (an upward comparison), e.g., "if the car had braked harder before it reached the wall, the passenger would have been uninjured" rather than how things could have been worse (a downward comparison), e.g., "if the car had swerved in the other direction into oncoming traffic, the passenger would have been killed". An AI agent can likewise provide an explanation by making a comparison to how things could have turned out better, or worse. But in doing so, it is noteworthy that upward and downward comparisons have very different consequences for human learning.

Counterfactuals about how an outcome could have been *better* affect intentions for the future, unlike counterfactuals about how an outcome could have been worse. For example, when people carried out an anagram solving task, and then imagined how their performance could have been better, say, "I could have solved more anagrams if I had tried more combinations of letters", they formulated intentions to do so in the future, and their subsequent performance improved [Markman *et al*., 2008]. People create better-world counterfactuals, such as, "if the car had braked harder before it reached the wall, the passenger would have been uninjured", to work out how to prevent bad outcomes in the future. The counterfactual offers a roadmap to transition from the current situation, "the car swerved into a wall and the passenger was injured" to a different future one [Epstude and Roese, 2017]. It provides a blueprint for future intentions, e.g., "brake harder" [Smallman and McCulloch, 2012].

Hence, if an AI agent provides a better-world counterfactual comparison, a user may make inferences about different decisions the system could make in the future.

Counterfactuals about how things could have been better help people to prepare for the future, but they come at an affective cost - they amplify negative emotions such as regret or guilt [Kahneman and Tversky, 1982a]. A counterfactual comparison can even make an objectively better outcome appear worse. For example, people judged that Olympic competitors looked more unhappy at the moment they realized they had won silver - presumably they could imagine the better outcome of winning gold, compared to bronze medalists when they realized they had won bronze - presumably they could imagine the worse outcome of not winning a medal [Medvec *et al*., 1995].

In contrast, when people imagine how things could have turned out *worse,* such as "if the car had swerved into oncoming traffic, the passenger would have been killed", positive emotions such as relief or satisfaction are amplified. People tend to create counterfactuals about how things could have been worse when there are few opportunities for future preventative action, and they can deflect negative emotions [Beike *et al*., 2009]. Moreover, they sometimes opt to inhibit counterfactuals. For example, people decided not to be informed about the outcomes of unchosen options more often after large losses than small ones [Tykocinski and Steinberg, 2005]. The construction of counterfactuals about how things could have been worse, and the inhibition of counterfactuals, helps people to feel better, but at the cost of complacency - they do not benefit from learning from mistakes.

Although XAI may have the overall goal of improving trust in decisions made by AI systems, fragments of explanatory interactions can usefully be calibrated to specific sub-goals. An explanation of a decision intended to help the user understand the AI system and make inferences about its future performance could best rely on better-world counterfactuals; an explanation intended to ensure the user feels a decision was justified or excusable could rely on worse-world counterfactuals. Careful constraints on counterfactuals are required to provide interpretable models of the decisions of AI systems that people can trust and consider to be fair [see also Russell *et al*., 2017].

## 3 Counterfactual Relations

Counterfactuals enable people to explain past outcomes and predict future ones by helping them to identify the relations between events, such as cause-effect or reason-action relations. When people imagine how things could have been different, their counterfactuals also affect their judgments of blame and fault. These two tendencies are described next.

### 3.1 Counterfactuals and Causes

Counterfactual and causal inferences have long been considered two sides of the same coin [Hume, 1739/1978; Lewis, 1973]. Counterfactuals amplify causal judgments. For instance, when people know that an alternative course of action would have led to a different outcome, e.g., "if the car had swerved into the middle of the road instead, the passenger would not have been injured", their judgments of a causal relation between the antecedent, swerving into the wall, and the outcome, the passenger being injured, are amplified. But when they know that an alternative course of action would have led to the same outcome, e.g., "*even if* the car had swerved into the middle of the road, the passenger would have been injured", their judgments of the causal relation between the antecedent and the outcome are decreased [McCloy and Byrne, 2002]. Counterfactuals amplify the causal link between an action and its outcome; "even if" semi-factuals deny it, and can make the outcome appear inevitable.

Although counterfactuals amplify causal judgments, counterfactuals and causal explanations usually tend to refer to different sorts of causes. Events often have several causes, some of which preempt or supersede others [Kominsky *et al*., 2015]. People tend to construct causal explanations that refer to strong (necessary and sufficient) causes that co-vary with the outcome, such as "the passenger was injured because the car swerved into a wall", whereas they tend to create counterfactuals that refer to background enabling (necessary but not sufficient) conditions that could prevent the outcome, such as "if the car had braked harder before it reached the wall, the passenger wouldn't have been injured". In an experimental demonstration of this phenomenon, people read a story about a car accident that occurred when a drunk driver swerved into the protagonist as he was driving home on an unusual route. They identified the drunk driver as the cause of the accident, but they constructed counterfactuals about how the accident would have been avoided if the protagonist had driven home by his usual route [Mandel and Lehman, 1996]. Strikingly, when people think about an outcome, they spontaneously offer about twice as many causal explanations that describe the facts as they happened, rather than counterfactual thoughts that refer to an imagined alternative [McEleney and Byrne, 2006]. Accordingly, if an agent provides a counterfactual about an antecedent and an outcome, a human user will readily infer a causal relation between them. But an explanatory exchange that contains explicit causal explanations as well as counterfactual alternatives may be warranted given their different referents.

### 3.2 Counterfactuals and Fault Assignment

People sometimes use counterfactuals to derogate actions, for example, in accident safety reports [Morris and Moore, 2000]. They justify poor performance by denying resources or control, e.g., "if there had been more time…". The tendency to use counterfactuals to excuse poor outcomes is especially prevalent when people imagine how things could have been different in the past, whereas when they create *prefactuals* about the future, "things could be different next time if…", they tend to focus on how they could better control the outcome [Ferrante *et al*., 2013].

People construct counterfactuals that imagine how a

decision could have been different, but they also evaluate whether it *should* have been different [Malle *et al*., 2014]. When a decision conforms to a social or moral norm, they tend not to mentally "undo" it in their counterfactual thoughts [McCloy and Byrne, 2000]. Counterfactuals are often used to determine legal culpability and they can amplify judgments of blame. For example, people read about two boys throwing bricks from an overpass bridge, and one boy's brick injured a driver whereas the other boy's brick fell to the side of the road. Their judgments were harsher of the boy whose brick injured someone, even though both carried out the same action with the same intentions and knowledge [Lench *et al*., 2015]. But when they created a counterfactual about how things could have been worse – the boy whose brick fell to the side of the road could have injured someone – they blamed him more for his actions [Parkinson and Byrne, 2017]. Similar effects of counterfactuals and semi-factuals occur for judgments about whether a person should have carried out a good action [Byrne and Timmons, 2018]. Hence, if an agent provides a counterfactual about a decision, a user may make inferences not only about causal relations but also about blame and fault based on moral norms. A human user's perception of a decision's quality may be affected by these moral inferences. Scaffolded construction of counterfactuals within a causal model may need to be informed by relevant social and moral norms for clarity.

# 4 Counterfactual Content

People show remarkable regularities in what they select to mutate in their representation of reality, so much so that there appear to be "fault-lines" - junctures that everyone identifies as pivotal points at which events could have followed a different path [Kahneman and Tversky, 1982a]. Some of these heuristics are described in the next sections.

## 4.1 Exceptions

People create a counterfactual by changing an exceptional event to be normal. Whatever is indicated as unusual in a situation tends to become a candidate for modification. For example, when people read that a car accident occurred when a man was driving home by a route he did not usually take, they tended to imagine that the outcome would have been different if he had taken his usual route; when they read instead that he was driving home by his usual route but at an earlier time than usual, they tended to imagine the outcome would have been different if he had driven home at his usual time [Kahneman and Tversky, 1982a]. The tendency to create a counterfactual by changing an exceptional event to be normal can be manipulated, however. People change an exceptional event to be exceptional in a different way, e.g., a different unusual route, rather than to be normal, if that would ensure a better outcome [Dixon and Byrne, 2011].

## 4.2 Controllability

People tend to create a counterfactual in which they change an event within a protagonist's control. For example, people read a story in which an individual took part in a game in which she could win a prize if she multiplied a sum in 30 seconds. She had to choose between two envelopes, A or B, one that contained an easy sum and one a difficult sum. She selected envelope A and it turned out to contain the difficult sum – multiply 67 x 86 - and she failed to complete it in 30 seconds. People tended to create counterfactuals that changed the event within her control, "if only she had chosen the other envelope…" [Girotto *et al*., 2007]. The tendency to create a counterfactual by changing a controllable event can also be manipulated. For example, people were invited to try the multiplication game themselves. They had to select envelope A or B which contained either an easy or a difficult sum, and they had to try to multiply it in 30 seconds. They all received the difficult sum and they all failed. They tended to create counterfactuals that changed events *outside* their control, "if only I had had more time", "if only I had had pen-and-paper" [Girotto *et al*., 2007]. Observers of the game also changed events outside the player's control [Pighin *et al*., 2011].

## 4.3 Actions

People modify actions rather than failures to act when they create a counterfactual. For example, people read about two individuals: one has shares in company A, thinks about switching to Company B, and decides to do so, and she loses $1,000; the other has shares in Company B, thinks about switching to Company A, but decides to stay where he is, and also loses $1,000. People created counterfactuals focused on the person who acted, "if only she hadn't switched…" rather than the person who failed to act [Kahneman and Tversky, 1982b]. The tendency to create a counterfactual by changing an action rather than an inaction can be manipulated too. It is reversed when people take a long-term perspective on events. For example, people judged that the individual who acted would feel worse in the short term, but the individual who did not act would feel worse in the long term [Gilovich and Medvec, 1995], although only when the counterfactual outcome was unknown [Byrne and McEleney, 2000].

## 4.4 Recent Events

People create a counterfactual in which they change the most recent event in a temporal sequence of independent events. The tendency is particularly evocative for counterfactuals about historical events, or about sports, e.g., "if only the striker had won the last penalty shot…". People were asked to imagine a game in which two people toss a coin, and if they toss the same face coin, they will both win $1,000. The first individual tossed heads, the second tossed tails, and so they both lost. People tended to imagine that the outcome could have been different if the second person had tossed heads, that is, they focused on the most recent event [Miller and Gunasegaram, 1990]. The tendency to create a counterfactual by changing the most recent event can also be manipulated. It does not occur when the context provides an alternative [Walsh and Byrne, 2004]. One example is that

people read that the first player tossed heads but there was a technical hitch and the game was restarted, and this time the first player tossed tails and the second tossed heads. They imagined a counterfactual in which the first player had tossed heads, i.e., they focused on the earlier event rather than the more recent one. The tendency is also reversed when the sequence of events is causally connected: people mentally undo the first cause in a causal sequence [Segura *et al*., 2002].

### 4.5 Probability

Counterfactual thoughts tend to be rooted in reality – people rarely imagine fantastical alternatives, such as that a passenger would not have been injured if he had been made of steel. They construct plausible counterfactuals [De Brigard *et al*., 2013] and can distinguish between those that have a high likelihood and those that do not [Petrocelli *et al*., 2011]. A counterfactual's probability may be determined by beliefs [Over *et al*., 2007], or by the possibilities that people envisage [Khemlani *et al*., 2018]. But, the counterfactuals people create do not appear to be based on likelihood. For example, when people think about a car accident in which two cars crashed into each other at a crossroads, the most improbable event is that the cars were in exactly the same place at exactly the same time. Yet no one imagines an alternative to this highly unlikely event [Kahneman and Tversky, 1982a].

Hence, people tend to create counterfactuals that focus on certain aspects of reality, as they have represented it, more than others. Each of these content effects has implications for XAI [Miller *et al.*, 2017]. Selective navigation through the natural "fault-lines" can ensure that an agent provides analyses that resonate with those that people produce.

## 5 Counterfactual Inferences

People make inferences very readily from counterfactuals. From a counterfactual conditional in the subjunctive mood such as, "if the car had continued forwards, the pedestrian would have been killed", people tend to judge that someone who asserted the counterfactual meant to imply "the car did not continue forwards" and "the pedestrian was not killed" [Thompson and Byrne, 2002]. When they read a story that contained such a counterfactual, they subsequently read a conjunction such as, "the car did not continue forwards and the pedestrian was not killed" far more rapidly than when the story contained a factual conditional in the indicative mood, such as "if the car continued forwards, the pedestrian was killed" [Santamaria *et al*., 2005]. When they understand a counterfactual, they recover the presumed or known facts.

Moreover, people read the conjunction, "the car continued forwards and the pedestrian was killed" equally rapidly whether they have been primed by the counterfactual or the factual conditional [Santamaria *et al*., 2005]. The finding indicates that they understand a counterfactual by envisaging a mental model of the counterfactual conjecture - the imagined alternative to reality, as well as a model of the presumed factual reality [Espino and Byrne, 2018]. They simulate two possibilities, one imagined and one real. Accordingly, counterfactuals activate brain regions associated with conflict detection [Ferguson *et al*., 2008; Van Hoeck *et al*., 2013].

One consequence of constructing multiple mental models during the comprehension of counterfactuals is that people make many more inferences from counterfactual conditionals than from factual ones. Given the counterfactual, and information about the facts, such as "the pedestrian was not killed," most people readily conclude "the car did not continue forwards". They tend to make this *modus tollens* inference about twice as often from a counterfactual conditional compared to a factual one. When they are told instead that, in fact, "the car continued forwards," they readily make the *modus ponens* inference, "the pedestrian was killed", and they do so as often from the counterfactual as from the factual conditional [Byrne and Tasso, 1999]. The observation that people make inferences readily from counterfactuals, including inferences that they are otherwise reluctant to make from factual conditionals, confirms their potential usefulness in XAI.

Nonetheless, it is worth noting that there are considerable individual differences in the comprehension of counterfactuals. For example, when people heard a counterfactual, eye-tracking showed their eyes moved immediately in a matter of just a few hundred milliseconds to fixate on an image corresponding to the presumed facts, and they alternated their gaze between this image and one corresponding to the conjecture, consistent with the idea that they envisaged two possibilities. But almost half of participants fixated on just the image corresponding to the conjecture, suggesting they constructed an interpretation corresponding to a single possibility [Orenes *et al*., 2019]. The result suggests that it may be prudent for an AI agent to probe that a human user has mentally simulated the provided counterfactual in an elaborated manner.

## 6 Conclusions

The inclusion of counterfactuals in interpretable models of complex AI systems can pay dividends in many ways. However, to maximize their effectiveness, it will be useful for XAI to incorporate information from psychological experiments about the way people create and comprehend counterfactuals, for counterfactuals of different structure and content, and with various relations. XAI can benefit from including the rich knowledge in cognitive science about the cognitive capacities of human reasoners. Enabling an agent to simulate the same sorts of alternatives to reality as a human may also go some way towards creating imaginative agents.

## Acknowledgments

Thanks to Mark Keane for comments on an earlier version.

## References

[Beike *et al*., 2009] Denise R. Beike, Keith D. Markman, and Figen Karadogan. What we regret most are lost opportunities: A theory of regret intensity. *Personality and Social Psychology Bulletin*, 35(3): 385-397, 2009.

[Biran and Cotton, 2017] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A

survey. In *IJCAI-17 Workshop on Explainable AI (XAI)*, 8: 8-13, 2017.

[Byrne, 2005] Ruth M.J. Byrne. *The Rational Imagination: How People Create Alternatives to Reality*. MIT Press, Cambridge, Massachusetts, 2005.

[Byrne, 2016] Ruth M.J. Byrne. Counterfactual thought. *Annual Review of Psychology,* 67: 135-157, 2016.

[Byrne and McEleney, 2000] Ruth M.J. Byrne and Alice McEleney. Counterfactual thinking about actions and failures to act. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 26(5): 1318-31, 2000.

[Byrne and Tasso, 1999] Ruth M.J. Byrne and Alessandra Tasso. Deductive reasoning with factual, possible, and counterfactual conditionals. *Memory & Cognition*, 27(4): 726-40, 1999.

[Byrne and Timmons, 2018] Ruth M.J. Byrne and Shane Timmons. Moral hindsight for good actions and the effects of imagined alternatives to reality. *Cognition*, 178: 82-91, 2018.

[De Brigard *et al*., 2013] Felipe De Brigard, Karl K. Szpunar, and Daniel L. Schacter. Coming to grips with the past: Effect of repeated simulation on the perceived plausibility of episodic counterfactual thoughts. *Psychological Science,* 24(7): 1329-1334, 2013.

[Dixon and Byrne, 2011] James Dixon and Ruth M.J. Byrne. Counterfactual thinking about exceptional actions. *Memory & Cognition*, 39(7): 1317-31, 2011.

[Espino and Byrne, 2018] Orlando Espino and Ruth M.J. Byrne. Thinking about the opposite of what is said: Counterfactual conditionals and symbolic or alternate simulations of negation. *Cognitive Science,* 42: 2459-501, 2018.

[Ferguson *et al*., 2008] Heather J. Ferguson, Anthony J. Sanford, and Hartmut Leuthold. Eye-movements and ERPs reveal the time course of processing negation and remitting counterfactual worlds. *Brain Research,* 1236: 113-25, 2008.

[Ferrante *et al*., 2013] Donatella Ferrante, Vittorio Girotto, Marta Straga, and Clare Walsh. Improving the past and the future: A temporal asymmetry in hypothetical thinking. *Journal of Experimental Psychology: General,* 142(1): 23-27, 2013.

[Gilovich and Medvec, 1995] Thomas Gilovich, and Victoria H. Medvec. The experience of regret: what, when, and why. *Psychological Review,* 102: 379-95, 1995.

[Ginsberg, 1986] Mathew L. Ginsberg. Counterfactuals. *Artificial intelligence*, 30(1): 35-79, 1986.

[Girotto *et al*., 2007] Vittorio Girotto, Dontella Ferrante, Stefania Pighin, and Michel Gonzalez. Postdecisional counterfactual thinking by actors and readers. *Psychological Science,* 18: 510-15, 2007.

[Halpern and Pearl, 2005] Joseph Halpern and Judea Pearl. Causes and explanations: A structural-model approach. Part 1: causes. *British Journal for the Philosophy of Science*, 56(4): 843-87, 2005.

[Hoffman *et al*., 2018] Robert Hoffman, Tim Miller, Shane T. Mueller, Gary Klein, and William J. Clancey. Explaining Explanation, Part 4: A Deep Dive on Deep Nets. *IEEE Intelligent Systems*, 33(3): 87-95, 2018.

[Hume, 1739/1978] David Hume. *A Treatise of Human Nature.* Oxford University Press, Oxford, 1978. [Original 1739].

[Kahneman and Tversky, 1982a] Daniel Kahneman and Amos Tversky. The simulation heuristic. In Daniel Kahneman, Paul Slovic, and Amos Tversky (Eds). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York, pages 201-8, 1982a.

[Kahneman and Tversky, 1982b] Daniel Kahneman and Amos Tversky. The psychology of preferences. *Scientific American*, 246(1): 160-73, 1982b.

[Khemlani *et al*., 2018] Sangheet Khemlani, Ruth M.J. Byrne, and Philip N. Johnson-Laird. Facts and possibilities: A model-based theory of sentential reasoning. *Cognitive Science,* 42(6): 1887-924, 2018.

[Kominsky *et al*., 2015] Jonathan F. Kominsky, Jonathan Phillips, Tobias Gerstenberg, David Lagnado, and Joshua Knobe. Causal superseding. *Cognition,* 137: 196-209, 2015.

[Lench *et al*., 2015] Heather C. Lench, Darren Domsky, Rachel Smallman, and Kathleen E. Darbor. Beliefs in moral luck: When and why blame hinges on luck. *British Journal of Psychology,* 106(2): 272-87, 2015.

[Lewis, 1973] David Lewis. Causation. *Journal of Philosophy*, 70: 556-67, 1973.

[Malle *et al*., 2014] Bertram F. Malle, Steve Guglielmo, and Andrew E. Monroe. A theory of blame. *Psychological Inquiry*, 25(2): 147-86, 2014.

[Mandel and Lehman, 1996] David R. Mandel and Darrin R. Lehman. Counterfactual thinking and ascriptions of cause and preventability. *Journal of Personality & Social Psychology,* 71: 450-63, 1996.

[Markman *et al*., 1993] Keith D. Markman, Igor Gavanski, Steven Sherman, and Matthew McMullen. The mental simulation of better and worse possible worlds. *Journal of Experimental Social Psychology*, 29(1): 87-109, 1993.

[Markman *et al*., 2007] Keith D. Markman, Matthew J. Lindberg, Laura J. Kray, and Adam D. Galinsky. Implications of counterfactual structure for creative generation and analytical problem solving. *Personality & Social Psychology Bulletin*, 33(3): 312-24, 2007.

[Markman *et al*., 2008] Keith D. Markman, Matthew N. McMullen, and Ronald A. Elizaga. Counterfactual thinking, persistence, and performance: A test of the reflection and evaluation model. *Journal of Experimental Social Psychology*, 44(2): 421-28, 2008.

[McCloy and Byrne, 2000] Rachel McCloy and Ruth M.J. Byrne. Counterfactual thinking about controllable actions. *Memory & Cognition*, 28: 1071-78, 2000.

[McCloy and Byrne, 2002] Rachel McCloy and Ruth M.J. Byrne. Semifactual "even if" thinking. *Thinking & Reasoning*, 8: 41-67, 2002.

[McEleney and Byrne, 2006] Alice McEleney and Ruth M.J. Byrne. Spontaneous causal and counterfactual thoughts. *Thinking and Reasoning,* 12: 235-55, 2006.

[Medvec *et al*., 1995] Victoria H. Medvec, Scott F. Madey, and Thomas Gilovich. When less is more: counterfactual thinking and satisfaction among Olympic medalists. *Journal of Personality & Social Psychology,* 69: 603-10, 1995.

[Miller and Gunasegaram, 1990]. Dale Miller and Saku Gunasegaram. Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of Personality & Social Psychology,* 59(6): 1111-18, 1990.

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

[Miller *et al*., 2017] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, 2017.

[Moravčík *et al*., 2017] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, *356*(6337): 508-13, 2017.

[Morris and Moore, 2000] Michael W. Morris, and Paul C. Moore. The lessons we (don't) learn: Counterfactual thinking and organizational accountability after a close call. *Administrative Science Quarterly,* 45: 737-65, 2000.

[Neal *et al*., 2018] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 613-28, 2018.

[Orenes, *et al*., 2019] Isabel Orenes, Juan A. García-Madruga, Isabel Gómez-Veiga, Orlando Espino, and Ruth M.J. Byrne. The comprehension of counterfactual conditionals: Evidence from eye-tracking in the visual world paradigm. *Frontiers in Psychology*, 10: 1172, 2019.

[Over *et al*., 2007] David E. Over, Constantinos Hadjichristidis, Jonathan St. B.T. Evans, Simon J. Handley, and Steven A. Sloman. The probability of causal conditionals. *Cognitive Psychology,* 54(1): 62-97, 2007.

[Parkinson and Byrne, 2017] Mary Parkinson and Ruth M.J. Byrne. Counterfactual and semi-factual thoughts in moral judgements about failed attempts to harm. *Thinking & Reasoning,* 23(4): 409-48, 2017.

[Petrocelli *et al*., 2011] John V. Petrocelli, Elise J. Percy, Steven J. Sherman, and Zakary L. Tormala. Counterfactual potency. *Journal of Personality & Social Psychology,* 100: 30-46, 2011.

[Pighin *et al*., 2011] Stefania Pighin, Ruth M.J. Byrne, Donatella Ferrante, Michel Gonzalez, and Vittorio Girotto. Counterfactual thoughts about experienced, observed, and narrated events. *Thinking & Reasoning*, 17: 197-211, 2011.

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

[Roese and Epstude, 2017] Neal J. Roese and Kai Epstude. The functional theory of counterfactual thinking. In *Advances in Experimental Social Psychology*, Vol. 56, pages 1-79. Academic Press. 2017.

[Russell *et al*., 2017] Chris Russell, Matt J. Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. *Advances in Neural Information Processing Systems*, 6414-23, 2017.

[Segura *et al*., 2002] Susana Segura, Pablo Fernandez-Berrocal, and Ruth M.J. Byrne. Temporal and causal order effects in counterfactual thinking. *Quarterly Journal of Experimental Psychology,* 55: 1295-305, 2002.

[Smallman and McCulloch, 2012] Rachel Smallman and Kathleen McCulloch. Learning from yesterday's mistakes to fix tomorrow's problems. *European Journal of Social Psychology,* 42 (3): 383-90, 2012.

[Swaminathan and Joachims, 2015] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814-823, 2015.

[Thompson and Byrne, 2002] Valerie Thompson and Ruth M.J. Byrne. Reasoning counterfactually: Making inferences about things that didn't happen. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 28: 1154-70, 2002.

[Tykocinski and Steinberg 2005] Orit E. Tykocinski and Noa Steinberg. Coping with disappointing outcomes: retroactive pessimism and motivated inhibition of counterfactuals. *Journal of Experimental Social Psychology,* 41(5): 551-58, 2005.

[Van Hoeck *et al*., 2013] Nicole Van Hoeck, Ning Ma, Lisa Ampe, Kris Baetens, Marie Vandekerckhove, and Frank Van Overwalle. Counterfactual thinking: an fMRI study on changing the past for a better future. *Social Cognitive & Affective Neuroscience,* 8: 556–64, 2013.

[Walsh and Byrne, 2004] Clare R. Walsh and Ruth M.J. Byrne. Counterfactual thinking: The temporal order effect. *Memory & Cognition*. 32: 369-78, 2004.

[Weld and Bansal, 2018] Daniel S. Weld and Gagan Bansal. Intelligible artificial intelligence. *ArXiv e-prints,* 2018.