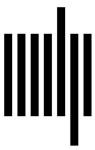
The Handbook of Rationality

Edited by: Markus Knauff, Wolfgang Spohn

Citation:

The Handbook of Rationality
Edited by: Markus Knauff, Wolfgang Spohn
DOI: 10.7551/mitpress/11252.001.0001
ISBN (electronic): 9780262366175
Publisher: The MIT Press
Published: 2021

Funding for the open access edition was provided by the MIT Libraries Open Monograph Fund.



The MIT Press

6.3 Conditional and Counterfactual Reasoning

Ruth M. J. Byrne and Orlando Espino

Summary

We consider a set of questions about conditionals and counterfactuals that have led to experimental discoveries able to distinguish between competing explanations of human reasoning. The first part of the chapter focuses on reasoning with factual conditionals, such as "If she made a good painting, she got a treat," and addresses two related questions: (1) Do people think about what is possible or what is probable when they understand a conditional? and (2) When do people infer a conditional relation? The second part of the chapter focuses on reasoning with counterfactual conditionals, such as "If she had done everything right, she would have passed her driving test," and also addresses two questions: (1) Do people make similar inferences from factual and counterfactual conditionals? and (2) Do people construct "embodied" mental representations of counterfactuals? The discoveries made from experimental investigations of these questions provide insights into how the mind accomplishes reasoning.

1. Conditionals and Counterfactuals

Conditional inferences are at the very center of everyday human reasoning. Even the youngest child who believes "If I make a good painting, I will get a treat," on hearing that her painting is good, will expect a treat. People make conditional inferences often in everyday life. Some seem immediate and obvious. Other conditional inferences require people to mull over them, to consider different possibilities. An office worker who knows "If I leave the office at 5 p.m., I will get home in time for dinner," on finding himself still at his desk as the clock strikes 5 p.m., may conclude that he will not get home on time—or, alternatively, he may begin to think about other routes to ensure he makes it. And people make inferences even more readily from counterfactual conditionals, about what once was possible but is so no longer. A learner driver who believes "If I had done everything right, I would have passed my driving test," reflecting on the fact that she did not pass, will conclude readily that she must have done something wrong. Inferences from counterfactuals occupy a special place at the center of conditional reasoning.

Conditionals and counterfactuals provide a unique window onto the human mind. They have been the subject of many intriguing and often conflicting analyses, especially in philosophy (e.g., Adams, 1975; Lewis, 1973; Stalnaker, 1968) and psychology (e.g., Evans & Over, 2004; Johnson-Laird & Byrne, 2002; Oaksford & Chater, 2007; Sloman & Lagnado, 2005). There have been many striking experimental discoveries about conditional inference (for a review, see Nickerson, 2015) and about counterfactual inference (for a review, see Byrne, 2016). Our goal in this chapter is to consider just a few of the crucial questions about them that have led to important findings since they first became the subject of intensive psychological study in the 1960s and 1970s. The discoveries provide important insights not only into how the mind accomplishes reasoning but also how the mind supports other sorts of thinking, including thinking that may seem quite remote from reasoning, such as creative imagination. People often create "if only . . ." thoughts about how the past could have turned out differently and "what if . . ." thoughts about how the future could be different (e.g., Roese & Epstude, 2017; Walsh & Byrne, 2004). Remarkably, the creation of such imagined alternatives to reality appears to depend on the same cognitive processes that support conditional and counterfactual reasoning (Byrne, 2005).

2. Conditional Reasoning

The frequency with which people make some conditional inferences is very high. Almost all participants in experiments, when they are told "If he wrote a good essay, he got a high mark" and "He wrote a good essay," make the modus ponens inference "He got a good mark" (see Nickerson, 2015). The frequency of other

conditional inferences varies. Usually about half to two thirds of participants, when they are told "He got a high mark," make the affirmation of the consequent inference "He wrote a good essay" (see Schroyens, Schaeken, & d'Ydewalle, 2001). Often about half of participants, when they hear "He did not get a high mark," make the modus tollens inference "He did not write a good essay" whereas the other half say that nothing follows. Similarly, when they hear "He did not write a good essay" about half to two thirds of participants make the denial of the antecedent inference, "He did not get a good mark" (see table 6.3.1).

How the human mind carries out reasoning is contended, and there is as yet no agreement on the cognitive processes that underlie conditional inference. An early set of theories proposed that people access a mental logic that consists of abstract inference rules, such as "If *p* then *q*; *p*; therefore *q*," and they construct a mental derivation of a conclusion, akin to a logical proof (e.g., Braine & O'Brien, 1998; Rips, 1994). On this sort of theory, people make modus ponens readily because it corresponds directly to an inference rule in their mental logic; they have difficulty with modus tollens because there is no rule in their mental logic corresponding to it, and they must instead construct a mental derivation using related rules. Early psychological studies of conditional reasoning, influenced by long-standing philosophical analyses of the validity of inferences, assumed propositional and predicate logics as the normative standard against which to compare human reasoning (e.g., Jeffrey, 1981). Another early set of theories proposed that the mind contains reasoning modules that consist of domain-sensitive inference rules, specialized for content such as obligation and permission, or social regulations of costs and benefits (e.g., Cheng & Holyoak, 1985; Cosmides, 1989). Domain-sensitive rule theories limited their explanations of conditional reasoning to performance in the Wason selection task (see Ragni, Kola, & Johnson-Laird, 2018). They proposed that people make inferences akin to modus ponens readily because the schemas for most domains contain an inference rule corresponding to it, for example, "If the action is to be taken, the precondition must be met"; they make inferences akin to modus tollens less often because few schemas contain an inference rule for it. These two sorts of theories—abstract and domain-sensitive inference rule theories—have few active champions in contemporary research on reasoning.

Nowadays, two other theories are hotly debated, one based on Bayesian probability and the other on mental models. The probabilist approach proposes that people make conditional inferences by relying on Bayesian probability calculations—different versions have been based on suppositions (e.g., Evans & Over, 2004), probabilistic logic (e.g., Oaksford & Chater, 2007), and causal Bayesian networks (e.g., Lucas & Kemp, 2015; Sloman & Lagnado, 2005). They propose that people make inferences from a conditional such as "If there are apples, there are oranges" by assessing conditional probability (the probability of an event occurring given that another event has already occurred). People compare the probability that there are oranges given there are apples, $P(B \mid A)$, to the probability that there are no oranges given there are apples, $P(\text{not-}B \mid A)$, and they do not think about the situations in which there are no apples (Evans, Handley, & Over, 2003). They make the modus ponens inference if their prior beliefs indicate $P(B \mid A)$ is greater than $P(\text{not-}B \mid A)$; they have difficulty with the modus tollens inference because they have not thought about the probability of situations in which there are no apples. On this view, probability logic is viewed as the normative standard against which to compare human reasoning (e.g., Cruz, Baratgin, Oaksford, & Over, 2015).

In contrast, the mental model theory proposes that the cognitive processes that underlie reasoning construct iconic simulations that correspond to the way the world would be if the assertion was true (e.g., Byrne & Johnson-Laird, 2009; Johnson-Laird & Byrne, 1991). People understand the conditional by envisaging a single possibility at the outset, "There are apples and there are oranges" (Johnson-Laird & Byrne, 2002). If need be, they can "flesh out" their models to think about other

Table 6.3.1
Four inferences from the conditional "If he wrote a good essay, he got a high mark" (If *A* then *B*)

Inference	Minor premise	Conclusion	Form
Modus ponens	He wrote a good essay	He got a high mark	$A \therefore B$
Affirm consequent	He got a high mark	He wrote a good essay	$B \therefore A$
Modus tollens	He did not get a high mark	He did not write a good essay	Not-$B \therefore$ not-A
Deny antecedent	He did not write a good essay	He did not get a high mark	Not-$A \therefore$ not-B

possibilities that are consistent with the conditional, such as "There are no apples and no oranges." The possibilities are conjunctive, that is, it is possible that there are apples and there are oranges, and it is possible that there are no apples and there are no oranges (e.g., Khemlani, Byrne, & Johnson-Laird, 2018). People interpret conditionals in many different ways—and so conditionals cannot be reduced to the truth-functional material implication of propositional logic (e.g., Johnson-Laird & Byrne, 2002). Instead, the normative standard against which to compare human reasoning is the semantic principle that an inference is valid if there are no counterexamples, that is, it is impossible for the premises to be true and the conclusion false (e.g., Johnson-Laird, Khemlani, & Goodwin, 2015). People make the modus ponens inference readily because it corresponds to the single possibility they have thought about at the outset; they find modus tollens more difficult because it requires them to flesh out their models to be more explicit.

Each of the theories predicts the difference between modus ponens and modus tollens, which they were designed to explain, but they make different predictions for other inferences (see Byrne & Johnson-Laird, 2009, for some examples). Difficulties in comparing the two sorts of theories arise because some probabilist accounts lack an algorithmic-level specification of the mental representations and cognitive processes that the human mind relies on to reason (see Oaksford, Over, & Cruz, 2019). Nonetheless, two crucial questions have led to discoveries that help to distinguish between the theories.

2.1 Do People Think about What Is Possible or What Is Probable?

According to the mental model theory, when people understand "if," they think about what is possible (Johnson-Laird & Byrne, 2002). They understand the conditional "If there are apples, there are oranges" by envisaging initially just a single possibility, in line with a principle of parsimony, because of working-memory constraints. Their models are small-scale dynamic simulations of the imagined possibility, but for convenience, they can be represented here in the following simple diagram:

```
apples      oranges
          . . .
```

The three dots in the diagram indicate that there may be other possibilities, and people can "flesh out" their models to think about some of them, for example,

```
apples      oranges
no apples   no oranges
```

Reasoners come to many different interpretations of "if," and some of them flesh out their initial representation to correspond to this *biconditional* interpretation, as if the conditional asserted "If and only if there are apples, then there are oranges." Others flesh out their models to include a further possibility:

```
apples      oranges
no apples   no oranges
no apples   oranges
```

which corresponds to a *conditional* interpretation of "if" (see table 6.3.2). In line with a principle of truth, people do not tend to think about the possibilities that are ruled out by an assertion. The conditional rules out as impossible one situation, the one in which there are apples and there are no oranges. The biconditional rules out as impossible two situations, the one in which there are apples and there are no oranges and another in which there are no apples and there are oranges. People do not tend to think about these impossibilities.

On the probabilist account, people understand "if" by assessing conditional probability (e.g., Evans & Over, 2004; Oaksford & Chater, 2007). To do so, they carry out a *Ramsey test* (after Ramsey, 1929/1990), in which they hypothetically suppose the "if" clause to be the case, make changes to their beliefs as needed, and assess, on this basis, the probability of the "then" clause. They compare the strength of their prior belief that there are apples and there are oranges to the strength of their prior belief that there are apples and there are no oranges (Evans et al., 2003). In other words, they think about the probability of a situation that is true according to the conditional and compare it to a situation the conditional rules out as false. To work out the likelihood

Table 6.3.2
The situations people think about for a conditional "If there are apples, then there are oranges" and a biconditional "If and only if there are apples, then there are oranges" according to two theories

	Mental models	Probability and prior beliefs
Conditional	apples and oranges no apples and no oranges no apples, and oranges	apples and oranges apples and no oranges
Biconditional	apples and oranges no apples and no oranges	apples and oranges apples and no oranges no apples, and oranges

that if there are apples, then there are oranges, people must think not only about what is possible if the conditional is true: there are apples and oranges; they must also think about what is not possible if the conditional is true: there are apples but no oranges. Moreover, they do not think about their beliefs concerning what happens when there are no apples, because they consider such instances irrelevant (Evans et al., 2003).

Recent empirical studies about judgments of truth and probability test predictions derived from these competing accounts (e.g., Byrne & Johnson-Laird, 2019; Cruz et al., 2015; Elqayam & Over, 2013; Goodwin, 2014; Hinterecker, Knauff, & Johnson-Laird, 2016; Pfeifer & Tulkki, 2017). I will illustrate just one of the discoveries. Consider the following short narrative:

> Carmen went shopping to the market. When she looked at the poster there, she saw it said, "If there are apples, there are oranges." When Carmen looked at the shelves, she saw that there were apples and there were oranges. Carmen checked her list of purchases.

When participants in experiments read a conditional such as "If there are apples, there are oranges" in this short narrative, they are able to read the subsequent conjunction "There were apples and there were oranges" very quickly, in just under a second and a half, on average (Espino, Santamaría, & Byrne, 2009). The conditional "primes" them to read the affirmative conjunction quickly. A biconditional such as "If and only if there are apples, there are oranges" also primes them to read the affirmative conjunction just as quickly compared to the conditional. Now consider instead the following narrative:

> Carmen went shopping to the market. When she looked at the poster there, she saw it said, "If there are apples, there are oranges." When Carmen looked at the shelves, she saw that there were no apples and there were oranges. Carmen checked her list of purchases.

Participants read the conjunction with the negated antecedent—"There were no apples and there were oranges"—more quickly when they were primed by a conditional compared to a biconditional (see figure 6.3.1). Why does the conditional prime participants to read the conjunction with the negative antecedent, "There were no apples and there were oranges," more quickly compared to the biconditional?

The mental model theory explains the experimental result readily: people mentally represent what is possible, not what is impossible. They read the conjunction "There were no apples and there were oranges" more quickly when it is primed by a conditional, because it is one of the true possibilities consistent with the conditional, whereas they take longer to read the conjunction when it is primed by the biconditional, because it is a false possibility ruled out by the biconditional (Espino et al., 2009).

The experimental result is more challenging to explain on the probabilist account. People do not think about their beliefs concerning what happens when the "if" clause is not true (Evans et al., 2003). So the theory predicts no difference in reading times for "There are no apples and there are oranges" when primed by a conditional versus a biconditional. In fact, the suppositional version even predicts a difference in the opposite direction. A biconditional is understood as two related conditionals, "If there are apples, there are oranges, and if there are oranges, there are apples" (Evans et al., 2003; see also Handley, Evans, & Thompson, 2006). People must compare their belief that there are apples and there are oranges to their belief that there are apples and there are no oranges for the first conditional, and in addition, they must compare their belief that there are oranges and there are apples to their belief that there are oranges and there are no apples for the second conditional. Thus, for a biconditional, they think about three prior beliefs, rather than just two (see table 6.3.2). The theory must predict that people will read the conjunction "There are no apples and there are oranges" *more* quickly when it is primed by the biconditional than the conditional, the opposite of the experimental result (Espino et al., 2009).

We can conclude that when people understand a conditional, their initial thoughts are about what is possible, not what is impossible. We can also conclude that when they understand a conditional, their initial thoughts are not about what is probable (since that would require them to have thought about what is impossible as well as what is possible).

2.2 When Do People Infer a Conditional Relation?

Many studies examine how people infer a causal relationship from the observation of objects and events (for reviews, see Johnson-Laird & Khemlani, 2017; Oaksford & Chater, 2017). But people can infer a conditional relationship even in non-causal situations. Suppose you know "There are roses in the garden or there are lilies or both." Can you infer "If there are no roses there are lilies"? About half of participants in experiments make the inference (Espino & Byrne, 2013). It occurs even when the assertions contain negated clauses, for example, when people are told "There are no daisies in the garden or there are no buttercups or both," they infer "If there are daisies, there are no buttercups" (see figure 6.3.2).

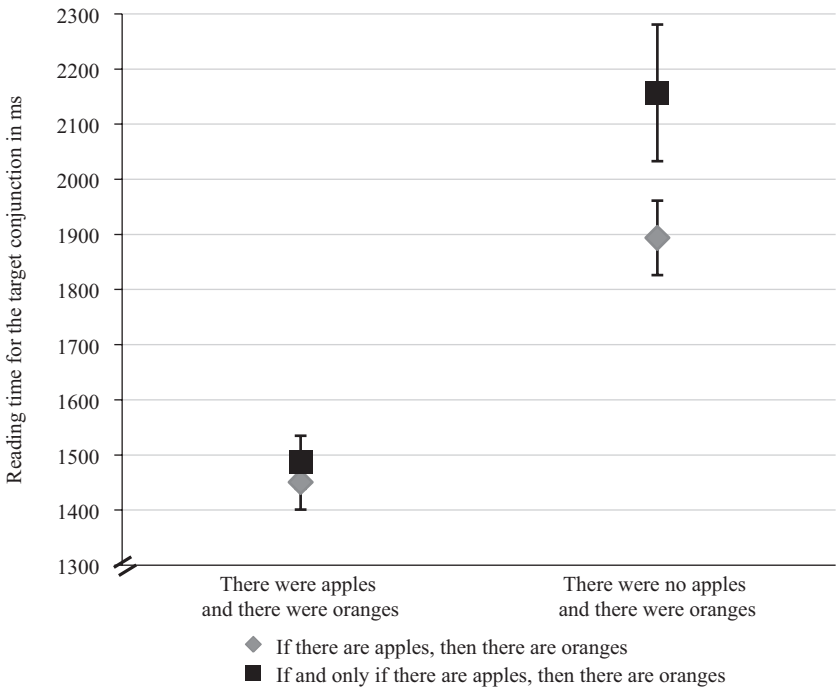


Figure 6.3.1
Reading times for conjunctions primed by conditionals or biconditionals (adapted from Espino et al., 2009, experiment 2). Error bars are standard error of the mean.

The discovery that when people hear about disjunctive alternatives, they make an inference of a conditional relation between them, is intriguing. Why would they do so?

According to the mental model theory, the inference is a reasonable one to make (Johnson-Laird & Byrne, 2002; see also Ormerod & Richardson, 2003). The disjunction "There are roses or lilies or both" is consistent with three possibilities:

roses	no lilies
no roses	lilies
roses	lilies

And the conditional "If there are no roses there are lilies" is consistent with the same three possibilities:

no roses	lilies
roses	no lilies
roses	lilies

Because there are no possibilities that the two assertions do not have in common, the inference from the disjunction to the conditional is valid—there are no counterexamples to it. Of course, the inference is difficult, requiring the comparison of multiple possibilities for each assertion, which is likely to exceed most people's working-memory capacity (see also Murray & Byrne, 2005). Unsurprisingly, only about half of participants make the inference (see figure 6.3.2). Instead, many of them make a revealing error: they choose the conclusion that matches their initial model of the conditional (see Espino & Byrne, 2013).

Crucially, according to the mental model theory, the inference between the disjunction and the conditional is valid in either direction, from "or" to "if" and from "if" to "or." For example, suppose you know instead "If there are no tulips in the garden, there are daffodils." Can you infer "There are tulips or daffodils or both"? About one third of participants make the inference. Tellingly, in most cases, participants make the inference from "or" to "if" more often than the one from "if" to "or," as figure 6.3.2 shows.

The difference is revealing because it challenges the predictions of probabilist theories (Oberauer, Geiger, & Fischer, 2011; Over, Evans, & Elqayam, 2010). According to probabilist theories, the inference from "or" to "if"—the one that people make most often—is not valid; only the one from "if" to "or" is valid. A probabilistically valid (p-valid) inference is one in which the likelihood of the conclusion is not lower than the likelihood of the premise. The inference from "or" to "if" is not p-valid because the likelihood of the conditional can be *lower* than the likelihood of the disjunction. The equation for the general relation between the probability of a disjunction "*A* or *B*," $P(A \text{ or } B)$, and the probability of a conditional "If not-*A*, *B*," $P(B \mid \text{not-}A)$, is as follows (as given in Over et al., 2010, p. 142):

$$P(A \text{ or } B) = P(A) + P(B \mid \text{not-}A) - P(A)P(B \mid \text{not-}A)$$

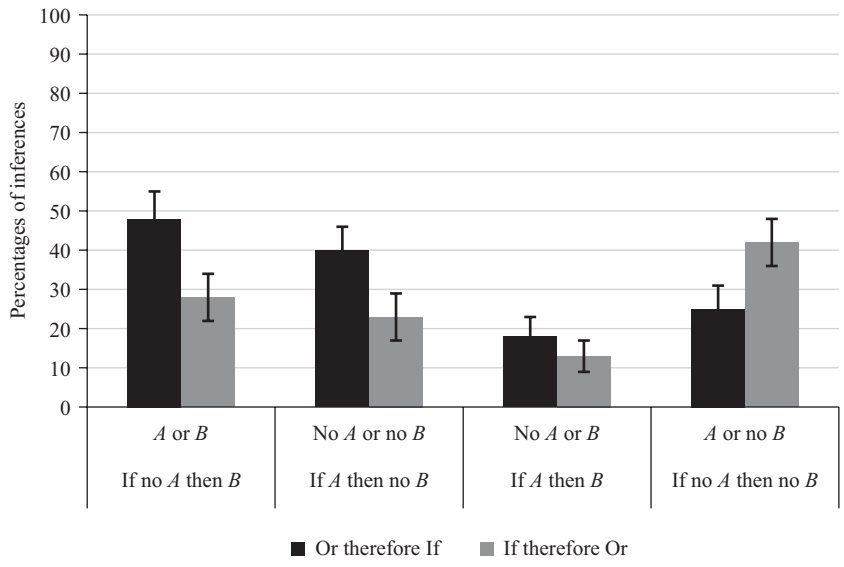


Figure 6.3.2
Percentages of inferences from a disjunction to a conditional and from a conditional to a disjunction (adapted from Espino & Byrne, 2013, experiment 1b). Error bars are standard error of the mean.

Hence, the inference from "or" to "if" is not p-valid (see also Oberauer et al., 2011). In contrast, the inference from "if" to "or" *is* p-valid because the likelihood of the disjunction cannot be *smaller* than the likelihood of the conditional. The relation is shown by the expansion of the probability of the conditional (as provided in Oberauer et al., 2011, p. 97), given by $P(B \mid \text{not-}A)$:

$$P(A) \times P(B \mid \text{not-}A) + P(\text{not-}A) \times P(B \mid \text{not-}A)$$

and the probability of the disjunction:

$$P(A) + P(\text{not-}A) \times P(B \mid \text{not-}A)$$

People judge the probability of a conditional premise to be higher than, or the same as, the probability of a disjunctive conclusion for "if" to "or" inferences, and they judge the probability of a conditional conclusion to be higher than, or the same as, the probability of a disjunctive premise for "or" to "if" inferences (Cruz et al., 2015; see also Gilio & Over, 2012). But nonetheless, as figure 6.3.2 shows, they make the inference from "or" to "if" more often than the inference from "if" to "or," even though the former is p-invalid and the latter is p-valid. The experimental result is the opposite of what would be expected if people understood conditionals by calculating likelihood, and it runs counter to the predictions of probabilist theories.

We can conclude that people are willing to infer a conditional relation from alternative possibilities. They infer a conditional from a disjunction because they think about the possibilities that are consistent with each assertion and appreciate that they are the same. Although the inference is valid in both directions in that there are no counterexample possibilities, people may be more inclined to make the inference from the disjunction to the conditional because the disjunction appears to assert categorically that something is the case, in contrast to the conditional, which appears to assert a hypothetical relation.

These two key discoveries shed light on the nature of the mental algorithms that underlie conditional inference. Further light is shed by discoveries about inferences from counterfactual conditionals.

3. Counterfactual Reasoning

A counterfactual such as "If she had done everything right, she would have passed her driving test" can seem to mean something very different from its factual counterpart, "If she did everything right, she passed her driving test." The counterfactual seems to convey that its antecedent is not the case—"She did not do everything right"—and it seems to convey that its consequent is not the case either—"She did not pass her driving test." Hence, assessing the truth of a counterfactual poses unique logical problems—since every counterfactual has a false antecedent, every counterfactual must be true on a traditional truth-functional account of conditionals, according to which a conditional is true if its antecedent is false or its consequent is true (see Nickerson, 2015). The development of possible-worlds semantics in modal logics led to several alternative analyses of counterfactuals (e.g., Lewis, 1973; Stalnaker, 1968). Their study in the psychology of reasoning has also had a significant impact in the past 20 years since they have come under increasing scrutiny in experiments. Two research

questions, and the discoveries they led to, illustrate the novel contribution of the study of counterfactuals to understanding the reasoning mind.

3.1 Do People Make Similar Inferences from Factual and Counterfactual Conditionals?

From a conditional in the indicative mood—which is often called a *factual* conditional—such as, "If Paolo was in Venice, then Marco was in Padua," when people are told, "Marco was not in Padua," only about half of them make the modus tollens inference, "Paolo was not in Venice," and the remainder tend to say that nothing follows. Strikingly, inferences from a conditional in the subjunctive mood—which is often called a *counterfactual* conditional—such as, "If Paolo had been in Venice, then Marco would have been in Padua," show a different pattern. When people know "Marco was not in Padua," most of them readily make the modus tollens inference "Paolo was not in Venice." The otherwise complex inference is made easily from the counterfactual (Byrne & Tasso 1999), as figure 6.3.3 shows.

Moreover, people continue to make the modus ponens inference, from "Paolo was in Venice" to "Marco was in Padua," as readily from the counterfactual as from the factual conditional. Similar effects occur for the *denial of the antecedent* inference, which is made more often from the counterfactual than the factual conditional, and the *affirmation of the consequent* inference, which is made equally from both sorts of conditionals.

This "counterfactual inference effect" is robust and has been studied extensively (for a review, see Byrne, 2017). It occurs not only for locational content but also for causal and definitional counterfactuals (Thompson & Byrne, 2002) and various everyday scenarios (Frosch & Byrne, 2012). It has been examined for inducements such as promises and threats (Egan & Byrne, 2012) and deontic content such as obligations (Quelhas & Byrne, 2003). Why are otherwise difficult inferences easy from a counterfactual?

The discovery helps distinguish alternative explanations of human reasoning. According to the mental model theory, people understand a counterfactual such as "If there had been apples there would have been oranges" by thinking about two possibilities from the outset. They mentally represent the conjecture mentioned in the conditional, "There are apples and there are oranges," and

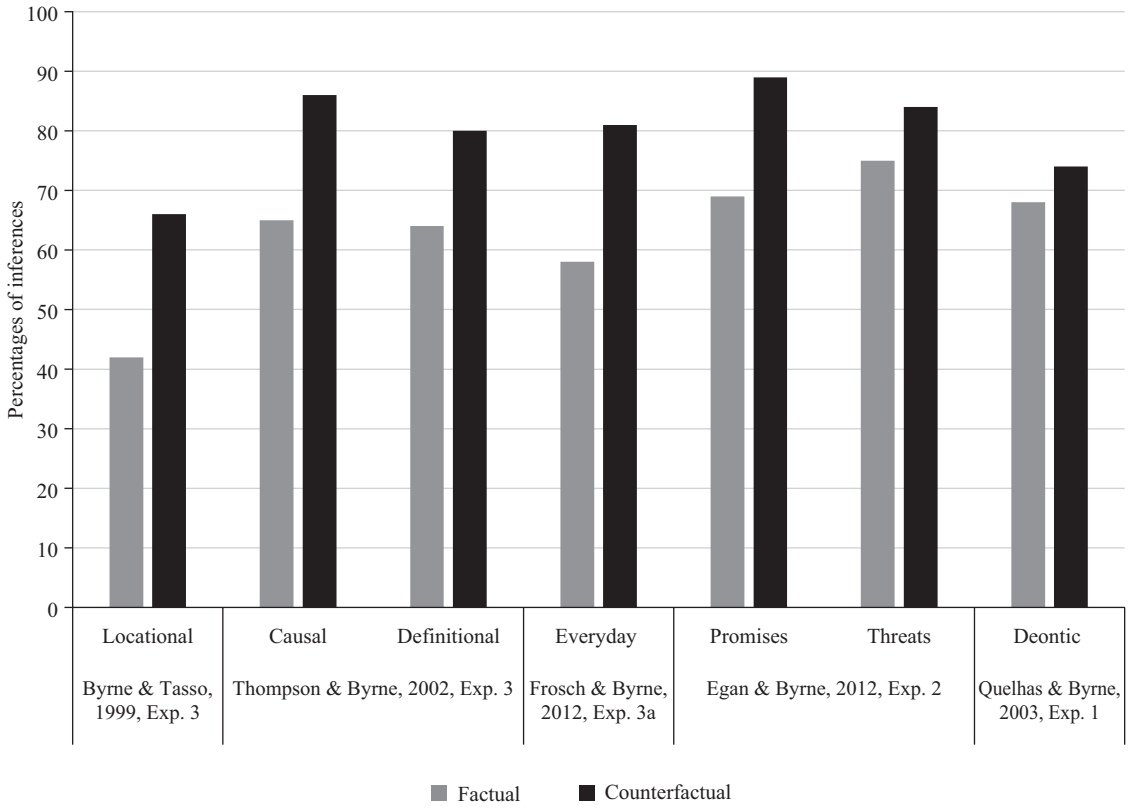


Figure 6.3.3
The percentages of modus tollens inferences made from factual and counterfactual conditionals for different contents (adapted from Byrne & Tasso, 1999; Egan & Byrne, 2012; Frosch & Byrne, 2012; Quelhas & Byrne, 2003; Thompson & Byrne, 2002).

they also mentally represent the presupposed or known facts, "There are no apples and there are no oranges." They keep track of the epistemic status of their models, as corresponding to a counterfactual possibility, or the facts (e.g., Johnson-Laird & Byrne, 1991):

Counterfactual:	apples	oranges
Facts:	no apples	no oranges

. . .

The theory explains a number of curious findings about counterfactuals. For example, when people hear the counterfactual, they mistakenly remember that they were told there were no apples and no oranges (Fillenbaum, 1974). They judge that someone who utters the counterfactual means to imply there were no apples and no oranges (Thompson & Byrne, 2002). Strikingly, they read the conjunction "There were no apples and there were no oranges" more quickly when they are primed by the counterfactual compared to the factual conditional, but they read the conjunction "There were apples and there were oranges" equally quickly when they are primed by either conditional, as figure 6.3.4 shows (Santamaría, Espino, & Byrne, 2005; see also de Vega, Urrutia, & Riffo, 2007). They look at images corresponding to "no apples and no oranges" more often for the counterfactual than the factual conditional, as shown by eye-tracking studies (Orenes, García-Madruga, Gómez-Veiga, Espino, & Byrne, 2019; see also Ferguson & Sanford, 2008; Nieuwland & Martin, 2012). The dual meaning of counterfactuals has been corroborated by studies that rely on measures of brain activity such as event-related potential (ERP) measures and functional magnetic resonance imaging (fMRI) measures (e.g., Ferguson, Sanford, & Leuthold, 2008; Kulakova, Aichhorn, Schurz, Kronbichler, & Perner, 2013; Van Hoeck et al., 2013).

The proposal that people think of two possibilities from the outset for the counterfactual predicts the finding that they readily make the modus tollens inference (Byrne & Tasso, 1999). When they hear, "There are no oranges," they do not have to flesh out their models, as they do for a factual conditional. They have already represented a possibility in which there are no oranges. They can eliminate the model corresponding to the counterfactual conjecture and conclude, "There are no apples." Similarly, they make the modus ponens inference as readily from the counterfactual as the factual conditional because when they hear, "There are apples," they can update their models to eliminate the model corresponding to the presupposed facts and note instead that the facts are that there are apples and oranges. They can then conclude, "There are oranges."

Of course, the fully fleshed out models for the counterfactual are the same as those for the factual conditional, even though the epistemic status of the models

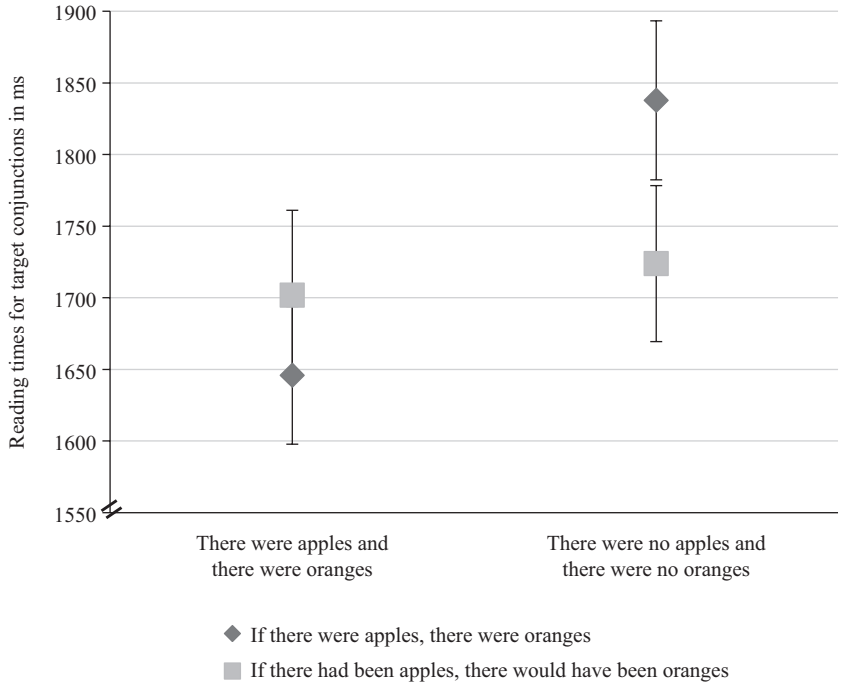


Figure 6.3.4
Reading times for conjunctions primed by factual or counterfactual conditionals (adapted from Santamaría et al., 2005, experiment 1). Error bars are standard error of the mean.

is different, for example, on a conditional interpretation of "if":

Counterfactual:	apples	oranges
Facts:	no apples	no oranges
Counterfactual:	no apples	oranges

Even though counterfactuals can seem to mean something very different from factual conditionals, their semantics runs in parallel (e.g., Byrne & Johnson-Laird, 2019). A counterfactual such as "If Oswald hadn't shot Kennedy, then someone else would have" seems debatable, whereas the corresponding factual one, "If Oswald didn't shoot Kennedy, then someone else did," seems true (Adams, 1970). But the comparison is misleading (Byrne & Johnson-Laird, 2019). The parallel to Adams's counterfactual should be "If Oswald hasn't shot Kennedy, then someone else will," which is also debatable. And the counterfactual parallel to his factual conditional, which assumes Kennedy has been shot, should be, "If Oswald hadn't been the person who shot Kennedy, then someone else must have been," which indeed seems true (see Byrne & Johnson-Laird, 2019).

One probabilist approach proposes that people understand a counterfactual by assessing its likelihood, which at the present time is the same as the probability of the factual conditional at a previous time (Over et al., 2007). People think only about one situation, in line with a singularity principle (Evans, 2007, p. 74). They combine their prior belief in the conditional probability of the counterfactual, that is, the likelihood of the counterfactual consequent "There would have been oranges" given the counterfactual antecedent "if there had been apples," with their prior beliefs in each of the implied facts, "There were no apples" and "There were no oranges," to arrive at a single assessment of their belief in the counterfactual (Over et al., 2007). Hence, it is challenging for this suppositional theory to explain the findings that people make modus tollens more often from counterfactuals than factual conditionals and that they appear to have ready access not only to the counterfactual conjecture but also to the presupposed facts.

Another probabilist theory is based on causal Bayes nets (e.g., Pearl, 2013). People construct causal models, for example,

in which nodes represent causes and effects, arrows indicate causal direction, and a conditional probability table at each node gives the probability that a node is present or absent (e.g., Lucas & Kemp, 2015; Sloman & Lagnado, 2005). Changes to the causal model are made by counterfactual "interventions" on a node (e.g., imagining that *B* is absent deletes links into *B*). A counterfactual's probability relies on the "do" intervention operator, $P(c \mid b.\ do(\neg b))$, that is, the conditional probability of *c*, given that *b* was observed, but counterfactually removed (e.g., Dehghani, Iliev, & Kaufmann, 2012; Lagnado, Gerstenberg, & Zultan, 2013; Meder, Hagmayer, & Waldmann, 2009). The account predicts the *opposite* of the counterfactual inference effect, namely, that people *refrain* from modus tollens when counterfactuals describe the prevention of a cause (e.g., Sloman & Lagnado, 2005). For example, participants tend to make the modus tollens inference for a factual question, "Suppose *A* was observed to not be moving, would *B* still be moving?" but not for a counterfactual one, "Suppose *A* were prevented from moving, would *B* still be moving?" (Sloman & Lagnado, 2005). A challenge for the causal model theory is that the counterfactual inference effect is observed for conditionals that convey not only causal relations but also a wide variety of other sorts of conditional relation, as figure 6.3.3 shows.

Of course, not all conditionals in the subjunctive mood are counterfactuals (e.g., Dudman, 1988). So-called Anderson conditionals (e.g., "If he had taken arsenic, he would have exactly these symptoms"; Adams, 1975) make use of the subjunctive mood, but their antecedents may be indicatives in disguise (e.g., "If he in fact took arsenic, he would have exactly these symptoms"). And other counterfactuals, such as counterpossibles (counterfactuals with impossible antecedents), for example, "If Hobbes had squared the circle, sick children in the mountains of South America at the time would have cared" (see Berto & Jago, 2018; Williamson, 2018), have received scant psychological attention.

3.2 Do People Construct "Embodied" Mental Representations of Counterfactuals?

There is abundant evidence that when people understand a counterfactual such as "If the flowers had been roses, the trees would have been orange trees," they think about two possibilities, the conjecture, "There were roses and orange trees," and its opposite, the presupposed facts. But how are the presupposed facts mentally represented? They could be represented explicitly through the use of negation, "There are no roses and there are no orange trees," or they could be represented by alternates, "There are poppies and there are apple trees." Which sorts of mental representations do people rely on?

The idea that the presupposed facts are represented as alternates is derived from the theory that the meaning of concepts, such as "roses," is *embodied*, in a modality-specific,

experientially based representation grounded in characteristics of sensory or motor processes (e.g., Barsalou, Simmons, Barbey, & Wilson, 2003). The embodied proposal extends even to abstract concepts such as negation (e.g., Glenberg, Robertson, Jansen, & Johnson-Glenberg, 1999). In an embodied system, a negation such as "There is not a rose" is represented by an alternate, such as "There is a poppy," and it may require several steps, for example, first representing "There is a rose" and then inhibiting the representation (e.g., Kaup, Lüdtke, & Zwaan, 2006; Mayo, Schul, & Burnstein, 2004).

In an experimental test of the embodied idea, participants were told that in a garden, the flowers are roses or poppies and the trees are orange trees or apple trees, and they then read a counterfactual, "If the flowers had been roses, the trees would have been orange trees." When they were told, "The trees were not orange trees," most of them tended to infer, "The flowers were poppies" (Espino & Byrne, 2018). They rarely said, "The flowers were not roses." In other words, they tended to infer what *is* the case, from information about what is *not* the case. This "inference-to-alternates effect" is a robust strategy that occurs in many different situations.

One situation in which the inference-to-alternates effect does *not* occur is revealing about the source of the tendency. Participants were told that the flowers are roses or poppies *or lilies* and the trees are orange trees or apple trees, that is, the context for flowers was a multiple one rather than a binary one. Now when they read a counterfactual "If the flowers had been roses, the trees would have been orange trees," and "The trees were not orange trees," many of them said, "The flowers were not roses," as figure 6.3.5 shows.

Why do people exhibit an inference-to-alternates effect in a binary context but not in a multiple context? On the embodied view, the presupposed facts for the counterfactual are represented as alternates, in both sorts of context. It explains the observation of an inference-to-alternates effect in the binary context, but it makes the wrong prediction for the multiple context. When people are told, "The trees are not orange trees," they should conclude, "There are poppies or lilies." But they do not do so.

According to the mental model theory, the mental representation of negation is as iconic as possible, but it can include symbols, such as a propositional-like tag "no" or some other annotation to capture negation (e.g., Johnson-Laird & Byrne, 2002; Orenes, Beltrán, & Santamaría, 2014). In a binary context, people simulate the presupposed facts by thinking about alternates:

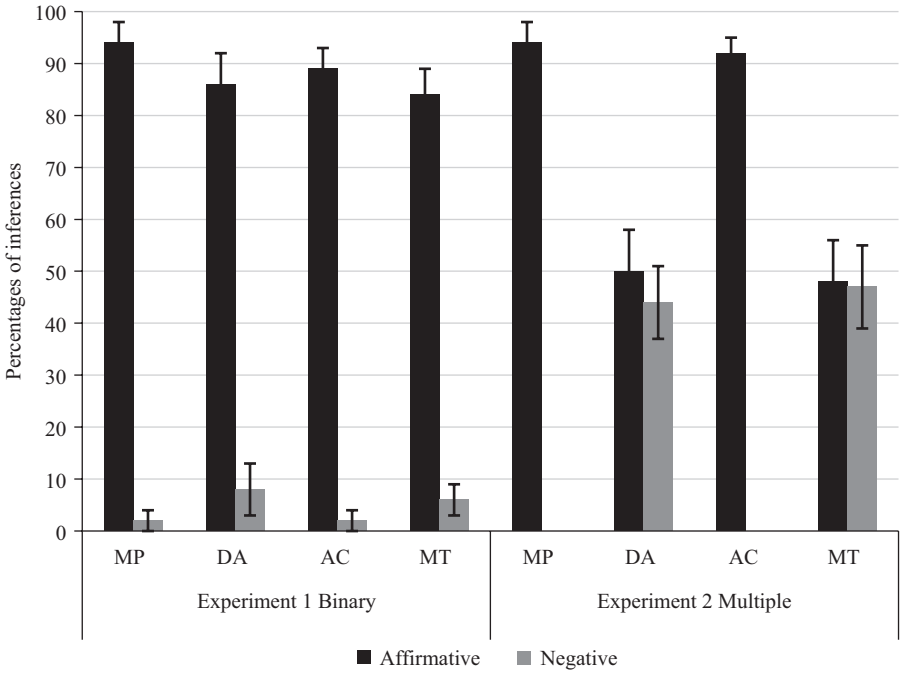


Figure 6.3.5
Percentages of modus ponens (MP), denial of the antecedent (DA), affirmation of the consequent (AC), and modus tollens (MT) inferences from counterfactuals in a binary or multiple context phrased as affirmative, e.g., "The flowers were poppies," or as negative, e.g., "The flowers were not roses" (adapted from Espino & Byrne, 2018). Error bars are standard error of the mean.

Counterfactual:	roses	oranges
Facts:	poppies	apples
	. . .	

When they are told, "There are no oranges," they eliminate the first model and conclude, based on the second model, "There are poppies." Hence, the inference-to-alternates effect is observed. But in a multiple context, the number of alternates they would have to consider is likely to exceed working memory:

Counterfactual:	roses	oranges
Facts:	poppies	apples
	lilies	apples
	. . .	

Instead they can switch to models that are annotated with propositional-like tags or symbols:

Counterfactual:	roses	oranges
Facts:	no roses	no oranges
	. . .	

Hence, when they are told, "There are no oranges," they conclude, based on the second model, "There are no roses," that is, the inference-to-alternates effect is eliminated. We can conclude that how people represent the presupposed facts depends on the context. People tend to mentally simulate the presupposed facts by considering alternates but they can switch strategy to a symbolic annotation when the possibilities exceed working-memory constraints.

These two key discoveries about inferences from counterfactuals shed further light on the nature of the mental algorithms that underlie conditional inference.

4. Conclusion

Conditional inferences are at the very center of everyday human reasoning, and counterfactual inferences occupy a special place there. Conditionals and counterfactuals provide a window onto the nature of mental simulations and algorithms in the human mind. In this chapter, we have considered some important discoveries that help distinguish explanations of conditional and counterfactual reasoning. When people understand, and reason about, a *factual* conditional, they think about what the conditional indicates is possible, not what it rules out as impossible. When they hear about alternative possibilities, they infer a conditional relation between them. The findings indicate that people think about what is possible, not what is probable. When they understand, and reason about, a *counterfactual* conditional, they are able to make complex inferences very readily. They make inferences about what *is* the case, even when they are told what is *not* the case. The findings indicate that people think about dual possibilities for a counterfactual—not only the counterfactual conjecture but also the presupposed facts—and they construct iconic mental simulations that can include symbolic annotations. These discoveries about conditional and counterfactual reasoning provide important insights into how the mind accomplishes reasoning by imagining possibilities of different sorts.

Acknowledgments

We thank our colleagues, Phil Johnson-Laird and Mark Keane, for helpful discussions related to this chapter.

References

Adams, E. W. (1970). Subjunctive and indicative conditionals. *Foundations of Language*, *6*, 89–94.

Adams, E. W. (1975). *The logic of conditionals*. Dordrecht, Netherlands: Reidel.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Science*, *7*, 84–91.

Berto, F., & Jago, M. (2018). Impossible worlds. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/archives/fall2018/entries/impossible-worlds/

Braine, M. D. S., & O'Brien, D. P. (Eds.). (1998). *Mental logic*. Mahwah, NJ: Erlbaum.

Byrne, R. M. J. (2005). *The rational imagination: How people create alternatives to reality*. Cambridge, MA: MIT Press.

Byrne, R. M. J. (2016). Counterfactual thought. *Annual Review of Psychology*, *67*, 135–157.

Byrne, R. M. J. (2017). Counterfactual thinking: From logic to morality. *Current Directions in Psychological Science*, *26*(4), 314–322.

Byrne, R. M. J., & Johnson-Laird, P. N. (2009). 'If' and the problems of conditional reasoning. *Trends in Cognitive Sciences*, *13*, 282–287.

Byrne, R. M. J., & Johnson-Laird, P. N. (2020). *If* and *or*: Real and counterfactual possibilities in their truth and probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(4), 760–780.

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

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, *17*(4), 391–416.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*(3), 187–276.

Cruz, N., Baratgin, J., Oaksford, M., & Over, D. E. (2015). Bayesian reasoning with ifs and ands and ors. *Frontiers in Psychology*, *6*, 192.

Dehghani, M., Iliev, R., & Kaufmann, S. (2012). Causal explanation and fact mutability in counterfactual reasoning. *Mind & Language*, *27*(1), 55–85.

de Vega, M., Urrutia, M., & Riffo, B. (2007). Cancelling updating in the comprehension of counterfactuals embedded in narratives. *Memory & Cognition*, *35*, 1410–1421.

Dudman, V. H. (1988). Indicative and subjunctive. *Analysis*, *48*, 113–122.

Egan, S. M., & Byrne, R. M. J. (2012). Inferences from counterfactual threats and promises. *Experimental Psychology*, *59*(4), 227–235.

Elqayam, S., & Over, D. E. (2013). New paradigm psychology of reasoning: An introduction to the special issue edited by Elqayam, Bonnefon, and Over. *Thinking & Reasoning*, *19*, 249–265.

Espino, O., & Byrne, R. M. J. (2013). The compatibility heuristic in non-categorical hypothetical reasoning: Inferences between conditionals and disjunctions. *Cognitive Psychology*, *67*(3), 98–129.

Espino, O., & Byrne, R. M. J. (2018). Thinking about the opposite of what is said: Counterfactual conditionals and symbolic or alternate simulations of negation. *Cognitive Science*, *42*(8), 2459–2501.

Espino, O., Santamaría, C., & Byrne, R. M. J. (2009). People think about what is true for conditionals, not what is false: Only true possibilities prime the comprehension of "if." *Quarterly Journal of Experimental Psychology*, *62*, 1072–1078.

Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and Judgement*. Hove, England: Psychology Press.

Evans, J. St. B. T., Handley, S. J., & Over, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(2), 321–335.

Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford, England: Oxford University Press.

Ferguson, H. J., & Sanford, A. J. (2008). Anomalies in real and counterfactual worlds: An eye-movement investigation. *Journal of Memory and Language*, *58*, 609–626.

Ferguson, H. J., Sanford, A. J., & Leuthold, H. (2008). Eye-movements and ERPs reveal the time course of processing negation and remitting counterfactual worlds. *Brain Research*, *1236*, 113–125.

Fillenbaum, S. (1974). Information amplified: Memory for counterfactual conditionals. *Journal of Experimental Psychology*, *102*(1), 44–49.

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

Gilio, A., & Over, D. (2012). The psychology of inferring conditionals from disjunctions: A probabilistic study. *Journal of Mathematical Psychology*, *56*(2), 118–131.

Glenberg, A. M., Robertson, D. A., Jansen, J. L., & Johnson-Glenberg, M. C. (1999). Not propositions. *Cognitive Systems Research*, *1*(1), 19–33.

Goodwin, G. P. (2014). Is the basic conditional probabilistic? *Journal of Experimental Psychology: General*, *143*, 1214–1241.

Handley, S., Evans, J. St. B. T., & Thompson, V. (2006). The negated conditional: A litmus test for the suppositional conditional? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 559–569.

Hinterecker, T., Knauff, M., & Johnson-Laird, P. N. (2016). Modality, probability, and mental models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10), 1606–1620.

Jeffrey, R. C. (1981). *Formal logic*. New York, NY: McGraw-Hill.

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

Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, *109*, 646–678.

Johnson-Laird, P. N., & Khemlani, S. S. (2017). Mental models and causation. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 169–188). Oxford, England: Oxford University Press.

Johnson-Laird, P. N., Khemlani, S. S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, *19*, 201–214.

Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, *38*(7), 1033–1050.

Khemlani, S. S, Byrne, R. M. J., & Johnson-Laird, P. N. (2018). Facts and possibilities: A model theory of sentential reasoning. *Cognitive Science*, *42*(6), 1887–1924.

Kulakova, E., Aichhorn, M., Schurz, M., Kronbichler, M., & Perner, J. (2013). Processing counterfactual and hypothetical conditionals: An fMRI investigation. *NeuroImage*, *72*, 265–271.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *37*(6), 1036–1073.

Lewis, D. (1973). *Counterfactuals*. Oxford, England: Blackwell.

Lucas, C. G., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review*, *122*, 700–734.

Mayo, R., Schul, Y., & Burnstein, E. (2004). "I am not guilty" vs "I am innocent": Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology*, *40*(4), 433–449.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. *Memory & Cognition*, *37*(3), 249–264.

Murray, M. A., & Byrne, R. M. J. (2005). Attention and working memory in insight problem solving. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (pp. 1571–1576). Mahwah, NJ: Erlbaum.

Nickerson, R. (2015). *Conditional reasoning*. Oxford, England: Oxford University Press.

Nieuwland, M. S., & Martin, A. E. (2012). If the real world were irrelevant, so to speak: The role of propositional truth-value in counterfactual sentence comprehension. *Cognition*, *122*, 102–109.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford, England: Oxford University Press.

Oaksford, M., & Chater, N. (2017). Causal models and conditional reasoning. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 327–346). Oxford, England: Oxford University Press.

Oaksford, M., Over, D., & Cruz, N. (2019). Paradigms, possibilities, and probabilities: Comment on Hinterecker, Knauff, and Johnson-Laird (2016). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(2), 288–297.

Oberauer, K., Geiger, S., & Fischer, K. (2011). Conditionals and disjunctions. In K. Manktelow, D. Over, & S. Elqayam (Eds.), *The science of reason: A Festschrift for Jonathan St B. T. Evans* (pp. 93–118). Hove, England: Psychology Press.

Orenes, I., Beltrán, D., & Santamaría, C. (2014). How negation is understood: Evidence from the visual world paradigm. *Journal of Memory and Language*, *74*, 36–45.

Orenes, I., García-Madruga, J. A., Gómez-Veiga, I., Espino, O., & Byrne, R. M. J. (2019). The comprehension of counterfactual conditionals: Evidence from eye-tracking in the visual world paradigm. *Frontiers in Psychology*, *10*, 1172.

Ormerod, T., & Richardson, J. (2003). On the generation and evaluation of inferences from single premises. *Memory & Cognition*, *31*(3), 467–478.

Over, D. E., Evans, J. St. B. T., & Elqayam, S. (2010). Conditionals and non-constructive reasoning. In M. Oaksford & N. Chater (Eds.), *Cognition and conditionals* (pp. 135–151). Oxford, England: Oxford University Press.

Over, D. E., Hadjichristidis, C., Evans, J. St. B. T., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, *54*, 62–97.

Pearl, J. (2013). Structural counterfactuals: A brief introduction. *Cognitive Science*, *37*, 977–985.

Pfeifer, N., & Tulkki, L. (2017). Conditionals, counterfactuals, and rational reasoning: An experimental study on basic principles. *Minds and Machines*, *27*, 119–165.

Quelhas, A. C., & Byrne, R. M. J. (2003). Reasoning with deontic and counterfactual conditionals. *Thinking & Reasoning*, *9*, 43–66.

Ragni, M., Kola, I., & Johnson-Laird, P. N. (2018). On selecting evidence to test hypotheses: A theory of selection tasks. *Psychological Bulletin*, *144*(8), 779–796.

Ramsey, F. P. (1990). General propositions and causality (original manuscript 1929). In D. H. Mellor (Ed.), *Philosophical papers* (pp. 145–163). London, England: Humanities Press.

Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.

Roese, N. J., & Epstude, K. (2017). The functional theory of counterfactual thinking: New evidence, new challenges, new insights. *Advances in Experimental Social Psychology*, *56*, 1–79.

Santamaría, C., Espino, O. & Byrne, R. M. J. (2005). Counterfactual and semifactual conditionals prime alternative possibilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1149–1154.

Schroyens, W. J., Schaeken, W., & d'Ydewalle, G. (2001). The processing of negations in conditional reasoning: A meta-analytic case study in mental model and/or mental logic theory. *Thinking & Reasoning*, *7*(2), 121–172.

Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science*, *29*, 5–39.

Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (pp. 98–112). Oxford, England: Blackwell.

Thompson, V., & Byrne, R. M. J. (2002). Reasoning counterfactually: Making inferences about things that didn't happen. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1154–1170.

Van Hoeck, N., Ma, N., Ampe, L., Baetens, K., Vandekerckhove, M., & Van Overwalle, F. (2013). Counterfactual thinking: An fMRI study on changing the past for a better future. *Social Cognitive Affective Neuroscience*, *8*, 556–564.

Walsh, C. R., & Byrne, R. M. J. (2004). Counterfactual thinking: The temporal order effect. *Memory & Cognition*, *32*, 369–378.

Williamson, T. (2018). Counterpossibles. *Topoi*, *37*, 357–368.

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Knauff, Markus, editor. | Spohn, Wolfgang, editor.
Title: The handbook of rationality / edited by Markus Knauff and Wolfgang Spohn.
Description: Cambridge : The MIT Press, 2021. | Includes bibliographical references and index.
Identifiers: LCCN 2020048455 | ISBN 9780262045070 (hardcover)
Subjects: LCSH: Reasoning (Psychology) | Reason. | Cognitive psychology. | Logic. | Philosophy of mind.
Classification: LCC BF442 .H36 2021 | DDC 153.4/3—dc23
LC record available at https://lccn.loc.gov/2020048455