

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45

Part I

**Counterfactuals, causality,
and mental representation**

PROOF ONLY

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45

PROOF ONLY

1 Counterfactual and causal explanation

From early theoretical views to new frontiers

David R. Mandel

In everyday and not-so-everyday life, we encounter situations that seem to demand an explanation of why something happened, how it happened, or how it could have been prevented. For example, following the 9/11 terrorist attacks, many people sought explanations for each of these questions. Explanations of *why it happened* have focused on Islamic fundamentalism and US hegemony in world politics. Explanations of *how it happened*, by contrast, have focused on the actions of the terrorists and their accomplices who were involved in instigating or directly carrying out the attacks. Differently still, explanations of *how the attacks might have been prevented* have focused on errors of judgment and ineffectual policies of US government agencies such as the CIA and the FBI that bore responsibility for preventing such attacks. As the example illustrates, explanations are “tuned” by the type of question they are meant to address. They are meant to be relevant, not “merely” true or probable (Hilton and Erb 1996).

That causal thinking plays a key role in the explanation process may seem obvious. After all, *how* and *why* questions *are* causal questions. As Kelley (1973: 107) put it, “Attribution theory is a theory about how people make causal explanations, about how they answer questions beginning with ‘why?’” Less obvious, perhaps, is the role that counterfactual thinking may play in that process. Yet, over the past two decades psychologists have proposed that counterfactual thinking does indeed play a key role. In this chapter, I examine some of the theoretical claims, critiques, and reconciliation attempts that have emerged from this literature. Although I draw on evidence from various pertinent sources, my focus in this chapter is on reasoning directed at explaining an effect in a specific case. Readers interested in the role of counterfactual reasoning about causal laws might consult Tetlock and Belkin (1996a).

Early theoretical views

Early claims regarding the effect of counterfactual thinking on cause explanation have focused on two interrelated routes of influence: (1) the selection of contrast cases used to define the effect to be explained and

12 *David R. Mandel*

(2) counterfactual conditional simulations used to test the plausibility of particular hypothesized causes (for reviews, see Spellman and Mandel 1999, 2003). I discuss these proposed routes in the subsections below.

Contrastive counterfactual thinking

The desire to explain is roughly proportional to the perceived discrepancy between expectancies and outcomes. When our expectancies are confirmed, there is little need for explanation. But, when expectancies are disconfirmed, they are likely to trigger spontaneous searches for causal explanations (Hastie 1984; Kanazawa 1992; Weiner 1985). The persistence of attention to disconfirmed expectancies is, by definition, then, a form of counterfactual thinking. Contrastive in nature, counterfactuals of this sort recapitulate expectancies that are juxtaposed against the reality of surprising outcomes. Although theorists tend to associate close counterfactuals with the term *almost* (Kahneman and Varey 1990; see also Chapter 8, Teigen) and counterfactual conditionals with the term *if only* and, more recently, *even if*, contrastive counterfactuals have not been provided with a natural-language marker. I propose that the term *rather than* might be appropriate in this regard. That is, contrastive counterfactuals often convey if not in form, then in gist, the following: “This unexpected event X occurred rather than Y, which I expected to occur instead.”

As many theorists have proposed, contrastive counterfactuals play an important role in causal selection by defining the nature of the *effect* to be explained (e.g., Einhorn and Hogarth 1986; Gorovitz 1965; Hesslow 1983; Hilton 1990; Mackie 1974). As Kahneman and Miller (1986) noted, in situations in which an outcome is viewed as normal in the circumstances – and hence expected – it is reasonable to answer the question “Why?” with the reply “Why not?” In such cases, no effect is meaningfully defined. The *why* question therefore presupposes a deviation between occurrence and what had been expected to occur by an explaine: “Why did *this* happen *rather than* what I *expected* would happen?” Theorists have proposed that, as a general rule, counterfactual thinking will recruit contrast cases that restore normality because people expect normal events to occur (Hart and Honoré 1985; Hilton and Slugoski 1986; Kahneman and Miller 1986). In the present context, normal not only means what was likely or is frequent, but also what is normative in the circumstances (McGill and Tenbrunsel 2000; Chapter 11, Catellani and Milesi).

These norm-restoring “downhill climbs” (Kahneman and Tversky 1982a) assist in defining a backgrounded set of factors – or *causal field* in Mackie’s (1974) terms – that are assumed to be common to both the factual case and the contrastive counterfactual set of cases and that are ruled out as causal candidates. Therefore, to the extent that contrastive counterfactual thinking plays a role in defining a causal field, it can be said to play an important role in determining what would *not* normally be deemed the cause. As we shall

1 see next, counterfactual thinking has also been ascribed a more positive role
2 in the process of causal explanation.

3 4 5 *Counterfactual conditional simulations*

6 While disconfirmed expectancies may be deemed counterfactual, not all
7 counterfactuals merely recapitulate disconfirmed expectancies. An important
8 function of counterfactual thinking, which Kahneman and Tversky (1982a)
9 brought to psychologists' research attention, is that it allows people to run
10 "if-then" simulations in working memory, which can allow us to explore
11 our own intuitions about how manipulations to aspects of a case might have
12 influenced how the case would have subsequently unfolded.

13 Accordingly, the second proposed route of counterfactual influence on
14 causal explanation involves the idea that counterfactual "if-then" or "even
15 if-then" simulations can be used to identify – perhaps even verify – various
16 causal contingencies that may later be deemed "the cause" (e.g., Lipe 1991;
17 Mackie 1974; McGill and Klein 1993; Roese and Olson 1995a; Wells and
18 Gavanski 1989). These conditional representations may be regarded as
19 generic response forms to the counterfactual question "Would Y have hap-
20 pened if X had not?" For instance, according to this "counterfactual simula-
21 tion account," a student who learns that she failed an exam and who thinks
22 "If I had studied harder, I would have passed the exam" will be more likely
23 to view her lack of preparation as a cause (if not "the" cause) of her perform-
24 ance than a student in a comparable position who instead thinks "Even if I
25 had studied harder, I still would have failed" (McCloy and Byrne 2002).

26 According to Mackie (1974), the counterfactual test in which the cause is
27 negated is crucial because the very meaning of the expression "X caused Y"
28 is, first, that X and Y in fact happened and, second, that had X not hap-
29 pened, Y also would not have happened. This idea has been recurrent in psy-
30 chological literature linking counterfactual thinking and causal explanation.
31 For instance, Kahneman and Tversky (1982a: 202) proposed that "to test
32 whether event A caused event B, we may undo A in our mind, and observe
33 whether B still occurs in the simulation." More forcefully, Wells and Gavan-
34 ski (1989: 161) proposed that "an event *will* be judged as causal of an
35 outcome to the extent that mutations to that event would undo the
36 outcome" (*italics mine*).

37 Roese and Olson (1995a: 11) took the argument even further, claiming
38 that although "not all conditionals are causal . . . counterfactuals, by virtue
39 of the falsity of their antecedents, represent one class of conditional proposi-
40 tions that are *always* causal" (*italics mine*). These authors explain that "[t]he
41 reason for this is that with its assertion of a false antecedent, the counterfac-
42 tual sets up an inherent relation to a factual state of affairs" (1995a: 11).
43 Mental model theorists (Byrne and Tasso 1999; Thompson and Byrne 2002)
44 have similarly proposed that, whereas counterfactual conditionals automati-
45 cally recover factual models, thus establishing a salient contrast case, factual

14 David R. Mandel

conditionals do not automatically recover counterfactual models because people do not spontaneously represent false events. Thus, no contrast case would be evoked unless the implicit models were deliberately unpacked. These later accounts suggest not only that counterfactual thinking *may* influence causal explanation, but that counterfactual thinking *will* have a stronger influence on causal explanation than factual thinking.

Empirical and theoretical challenges

Although some evidence supporting the idea that causal judgments (and related attributions) are influenced by counterfactual thinking has accrued (e.g., see Branscombe *et al.* 1996; Roese and Olson 1997; Wells and Gavanski 1989), most of it is based on studies that have manipulated the mutability of antecedents or outcomes in a given case (but see Chapter 10, Dhimi *et al.*). Outcome mutability has been manipulated by constructing different versions of scenarios in which alternatives to a chosen option would have led either to the same outcome or to a better outcome (Wells and Gavanski 1989). Antecedent mutability has been manipulated, for instance, by varying the abnormality (Kahneman and Tversky 1982a) or controllability (Mandel and Lehman 1996) of antecedents in scenarios. The core assumption underlying such research has been that if the relevant manipulation influenced judgment, then it must have been mediated by counterfactual thinking.

Other research (e.g., Davis *et al.* 1995; N'gbala and Branscombe 1995), however, suggests that the effect of "mutability manipulations" on causal judgments may have had more to do with the particular hypothetical scenarios that had been used in previous research than with a robust effect of counterfactual thinking on causal judgment. For example, Mandel and Lehman (1996: Experiment 3) demonstrated that, when participants read about a hypothetical case that afforded the opportunity to make different counterfactual and causal selections, antecedent mutability manipulations influenced participants' counterfactual listings, but these same manipulations did not influence participants' causal judgments. Moreover, the antecedent that was perceived as most causal differed from that which was most frequently mutated as a way of undoing the outcome.

Trabasso and Bartolone (2003) pointed out that manipulations of antecedent normality in past studies are confounded with the extent to which such antecedents were themselves explained in the relevant scenarios. These authors independently manipulated level of explanation and normality in Kahneman and Tversky's (1982a) "Mr Jones" car-accident scenario and asked participants to rank the likely availability of four counterfactual "if only" statements that mutated either the route Jones took, the time he left work, his decision to brake at the yellow light, or the other vehicle charging into the intersection. Calling into question the relation between normality and counterfactual availability, Trabasso and Bartolone found that

Counterfactual and causal explanation 15

1 counterfactual rankings were influenced by level of explanation only.
2 Antecedent normality had no effect on participants' judgments of how likely
3 it would be that a given counterfactual statement would be generated.

4 Another set of studies (Mandel 2003b) examined the idea that counterfac-
5 tual thinking about what *could have been* has a stronger effect on attribution
6 than factual thinking about what *was*. For example, participants in Experi-
7 ment 2 first recalled an interpersonal event they had recently experienced
8 and were then instructed either to think counterfactually about something
9 they (or someone else) might have done that would have altered the outcome
10 or to think factually about something they (or someone else) did that con-
11 tributed to how the outcome actually occurred. Participants rated their level
12 of agreement with causality, preventability, controllability, and blame attri-
13 butions, each of which implicated the actor specified in the thinking manip-
14 ulation. Compared to participants who received no thinking directive,
15 participants in the factual and counterfactual conditions reported more
16 extreme attributions. However, mean agreement did not differ between the
17 factual and counterfactual conditions – a finding that was replicated in two
18 other experiments (cf. Mandel and Dhimi in press; Tetlock and Lebow
19 2001).

Counterfactual tests of necessary causes

22 Another problem faced by counterfactual simulation accounts is that they
23 imply that causal reasoners assign greater weight to causes that are necessary
24 rather than sufficient to explain the relevant effect or yield the focal
25 outcome. Strictly speaking, X is a necessary cause of Y, if X is implied by Y
26 and, in contrapositive form, if $\neg X$ implies $\neg Y$ (the symbol \neg is read as “the
27 negation of”).¹ By contrast, X is a sufficient cause of Y, if X implies Y and,
28 in contrapositive form, if $\neg X$ is implied by $\neg Y$ (Cummins 1995; Fairley *et*
29 *al.* 1999). However, as Mackie (1974) explained, the concept of a causal field
30 requires a qualified interpretation of these relations, such that necessity and
31 sufficiency are interpreted as meaning necessary or sufficient *in the circum-*
32 *stances*, where the latter qualification may be interpreted as meaning “given
33 the presence of all the events in the case that are backgrounded.”

34 On either interpretation, counterfactual conditionals that proceed by
35 negating a hypothesized cause provide information relevant to the assess-
36 ment of whether that factor was necessary to bring about the effect. I regard
37 this as one of the principal limitations of counterfactual simulation accounts
38 because there is mounting evidence that what is meant by the term *cause* in
39 everyday discourse tends to reflect sufficiency in the circumstances rather
40 than necessity in the circumstances. Indeed, even Mackie (1974: 38), a key
41 proponent of the “necessary cause” view, wrote:

42
43
44 There is, however, something surprising in our suggestion that “X
45 caused Y” means, even mainly, that X was *necessary* in the circumstances

16 *David R. Mandel*

for *Y*. Would it not be at least as plausible to suggest that it means that *X* was sufficient in the circumstances for *Y*? . . . After all, it is tempting to paraphrase “*X* caused *Y*” with “*X* necessitated *Y*” or “*X* ensured *Y*,” and this would chime in with some of the thought behind the phrase “necessary connection”. But if “*X* necessitated *Y*” is taken literally, it says that *Y* was made necessary by *X* or became necessary in view of *X*, and this would mean that *X* was sufficient rather than necessary for *Y*.

One of the key problems with treating “*X* caused *Y*” as meaning, primarily, that “had *X* been absent, *Y* would not have occurred” is that the definition is too inclusive (Lombard 1990). As Hilton *et al.* (Chapter 3 below) put it, “Th[e] plethora of necessary conditions brings in its train the problem of causal selection, as normally we only mention one or two factors in a conversationally given explanation. . . .” Contrary to this reasonable proposal, according to the “negate-*X*, verify-*Y*” counterfactual criterion, oxygen would be the cause of all fires, and birth would be the cause of all deaths. Clearly, statements such as these violate our basic understanding of what causality means. Although we can all agree that birth is necessary for death, few would say that latter is brought about by the former. It is the quality of being instrumental in “bringing about,” even if not without the assistance of a background set of enabling conditions, which suggests that “*X* caused *Y*” means *X* was sufficient in the circumstances for *Y* to occur.

Since Mackie posed the preceding question, psychologists have conducted considerable research to empirically address it. Studies of naïve causal understanding indicate that people define causality primarily in terms of sufficiency. For example, Mandel and Lehman (1998: Experiment 1) asked participants to provide open-ended definitions of the words *cause* and *preventor*. They found that a majority of participants defined *cause* (71 percent) and *preventor* (76 percent) in terms of sufficiency (e.g., “if the cause is present, the effect will occur”). By contrast, only a minority defined these concepts in terms of necessity (22 percent and 10 percent for *cause* and *preventor*, respectively; e.g., “if the cause is absent, the effect won’t occur”). Although Mandel and Lehman (1998) did not report the cross-tabulated frequencies of response, a re-analysis of the dataset revealed an interesting result. Without exception, the minority of participants who provided a necessity definition also provided a sufficiency definition for the same term. That is, not a single participant in their study provided a definition of causality or preventability only in terms of necessity, whereas the majority provided definitions that focused exclusively on sufficiency.

Given the possibility for bias in coding of open-ended responses, Mandel (2003c: Experiment 2) attempted to replicate these findings by asking participants whether they thought the expression “*X* causes *Y*” means “When *X* happens, *Y* also will happen” (i.e., *X* is sufficient to cause *Y*) or “When *X* doesn’t happen, *Y* also won’t happen” (i.e., *X* is necessary to cause *Y*). Eighty-one percent of the sample interpreted the causal phrase in terms

Counterfactual and causal explanation 17

1 of sufficiency and a comparably high percentage (84.5 percent) thought that
2 other people would do so too.

3 Goldvarg and Johnson-Laird (2001) provide converging support for the
4 sufficiency view. Their research examined the types of mental models that
5 people view as being consistent with expressions like “X will cause Y.” In
6 Experiment 1, causal expressions were associated with the three possibilities
7 implied by the notion of a sufficient cause for roughly half of the sample
8 (i.e., X and Y, $\neg X$ and Y, and $\neg X$ and $\neg Y$).² The other half of the sample
9 indicated that the causal expressions were associated with the two possi-
10 bilities implied by the notion of a necessary and sufficient cause (i.e., X and
11 Y, and $\neg X$ and $\neg Y$). However, consistent with the cross-tabulation findings
12 just reported, no participant selected the set of possibilities implied by
13 the notion of a necessary cause alone (i.e., X and Y, X and $\neg Y$, and $\neg X$
14 and $\neg Y$).

15 Considerable support for the sufficiency view also comes from causal
16 induction studies, which have demonstrated that people assign greater
17 weight to the sufficiency-relevant cause-present cases than the necessity-rele-
18 vant cause-absent cases (e.g., Anderson and Sheu 1995; Cheng 1997; Kao
19 and Wasserman 1993; Mandel and Lehman 1998; McGill 1998; Schustack
20 and Sternberg 1981). In sum, the weight of available empirical evidence
21 favors the sufficiency interpretation of causality and, in the single-case
22 context, this assertion must be qualified by adopting the “in the circum-
23 stances” interpretation.

Theoretical reconciliations

Mandel and Lehman's (1996) prevention-focus account

24 To reconcile the idea that counterfactual conditionals do tell us something
25 of a causal nature with the substantial body of evidence indicating that
26 people place greater weight on sufficient rather than necessary conditions,
27 Mandel and Lehman (1996) proposed an alternative “prevention-focus”
28 account. This account posits that people often treat “negate-X” counterfac-
29 tual conditionals not as tests of the necessary generative cause of an effect
30 but as explanations of sufficient but foregone ways in which the effect might
31 have been prevented.

32 To illustrate the account, it will be instructive to map the representations
33 of factual and counterfactual events in a given case onto a 2×2 confusion
34 matrix, such as researchers have used to summarize contingency information
35 in multiple-case studies (e.g., Mandel and Lehman 1998). In the present
36 context, however, the cells of this matrix do not index the frequencies of the
37 relevant event conjunctions across a set of real cases but, rather, in line with
38 Kahneman and Tversky's (1982a) idea of the simulation heuristic, they
39 index the cognitive ease of representing ways in which the various possi-
40 bilities might have occurred in the focal case.
41
42
43
44
45

18 *David R. Mandel*

In Figure 1.1, *Y* represents an event to be explained (e.g., an unexpected car accident) and *X* represents a mutable antecedent event (e.g., the route the driver took). Further, let us assume that one is evaluating the hypothesis that “*X* caused *Y*.” Cell A represents the factual case in which *X* and *Y* co-occur. Cell B represents counterfactual cases in which *X* occurs but *Y* nevertheless does not. Cell C represents counterfactual cases in which the negation of *X* fails to undo *Y*. And, cell D represents counterfactual cases in which the negation of *X* successfully undoes *Y*.

Both cells A and D provide evidence that supports the focal hypothesis. By contrast, cells B and C provide evidence against the focal hypothesis, but the nature of the detraction differs in the two cases and is best thought of in terms of the contrast between competing sources of evidence. Specifically, the contrast of cause-present information in cells A and B permits a test of whether *X* is sufficient to cause *Y*, whereas the contrast of cause-absent information in cells C and D permits a test of whether *X* is necessary to cause *Y*.

As noted, theorists have focused on the counterfactual question “If *X* had not occurred, would *Y* still have occurred?” or the more restrictive probe “But for *X*, would *Y* still have occurred?” Answers to such questions will evoke representations that fall into either cells C or D (cf. Lipe 1991). If the negation of *X* does not undo *Y* (e.g., “Even if I had taken a different route, I probably still would have had an accident”), the C-cell representation will likely disconfirm the hypothesis that *X* was necessary in the circumstances for *Y*. Conversely, if the negation of *X* undoes *Y* (e.g., “If only I had taken a different route, I wouldn’t have had this accident”), then the D-cell representation will likely support the necessity claim.

The prevention-focus account shifts the emphasis on counterfactual conditionals in three key respects. First, emphasis is shifted from reasoning about generative causes to reasoning about inhibitory causes. Second, emphasis is shifted from assessing necessary conditions to assessing sufficient conditions. Third, there is an epistemological shift such that counterfactuals

	<i>Y</i>	$\neg Y$
<i>X</i>	A Confirmation by factual case	B Sufficiency violation by counterfactual case
$\neg X$	C Necessity violation by counterfactual case	D Confirmation by counterfactual case

Figure 1.1 Confusion matrix representing implications of factual and counterfactual cases for a hypothesis (“*X* caused *Y*”) about an actual generative cause.

Counterfactual and causal explanation 19

are intended to refer directly to hypotheses about counterfactual possibilities (what might have been) rather than hypotheses about factual possibilities (what was).

Figure 1.2 shows how the factual and counterfactual contingencies considered earlier change in terms of their implications when the focal hypothesis is reframed from “X caused Y” to “ $\neg X$ could have prevented Y.”³ In this case, the contrast of cells A and B provides information relevant to assessing whether $\neg X$ would have been necessary in the circumstances to prevent Y, whereas the contrast of cells C and D provides information relevant to assessing whether $\neg X$ would have been sufficient in the circumstances to prevent Y.

In support of the prevention-focus account, Mandel and Lehman (1996) demonstrated that participants who completed “If only . . .” sentence stems provided responses that were aligned more closely with other participants’ listings of how the effect might have been prevented than with participants’ listings of how the effect was caused. The intuition underlying the prevention-focus account is also captured by considering the types of explanations offered for the 9/11 attacks. As noted earlier, although the cause of the attacks was attributed to the terrorists and their accomplices, most counterfactual assessments have been directed at US prevention failures. For instance, one *New York Times* article began by stating “The director of the FBI, Robert S. Mueller III, acknowledged today for the first time that the attacks of Sept. 11 might have been preventable if officials in his agency had responded differently to all the pieces of information that were available” (Lewis 2002 § 1). Such counterfactuals are widespread and influence blame assignment, as the 9/11 commission in the United States is now proving (see, e.g., Johnston and Dwyer 2004), but one would be hard-pressed to find news stories claiming that US agencies *caused* 9/11.

It appears that although upward counterfactual thinking is, broadly speaking, causally informative and that it does play a key role in the explanation process, the focus of counterfactual explanations differs from

	Y	$\neg Y$
X	A Confirmation by factual case	B Necessity violation by counterfactual case
$\neg X$	C Sufficiency violation by counterfactual case	D Confirmation by counterfactual case

Figure 1.2 Confusion matrix representing implications of factual and counterfactual cases for a hypothesis (“ $\neg X$ could have prevented Y”) about a foregone inhibitory cause.

20 *David R. Mandel*

that of direct causal explanations. In particular, as Mandel and Lehman (1996) observed, counterfactual listings are more likely than generative-cause listings to focus on controllable behaviors (also see Girotto *et al.* 1991; Mandel 2003a; McCloy and Byrne 2000). The emphasis on personal control over the outcome and on “human error” (Morris *et al.* 1999) suggests that counterfactual explanations tend to conform to notions described by some manipulability theories of causality. As Collingwood (1940: 296) put it, “the question ‘What is the cause of an event y ?’ means in this case ‘How can we produce or prevent y at will?’” Collingwood (1940: 307) even proposed that “for the mere spectator (meaning someone who cannot produce or prevent any of the conditions of an event) there are no causes.” Although this definition of causality is too restrictive, it captures fairly well the sense in which people offer counterfactual explanations for past events.

Spellman’s (1997) probability-updating account

Another reconciliatory account was offered by Spellman (1997; Spellman and Kincannon 2001; Chapter 2, Spellman *et al.*), who proposed a probability-updating account of causal selection akin to a stepwise multiple regression analysis (cf. Hilton 1988). According to SPA (for Spellman’s probability-updating account), a reasoner first identifies the factual outcome to be explained (Y), as well as a set of causal candidates (X_1, X_2, \dots, X_n) from the chain of events in the relevant case. Next, the reasoner assesses the post-hoc probability of Y prior to X_1 and then reconditionalizes the probability on X_1 , observing the change in probability accounted for by X_1 . This process is repeated for each candidate in the set, each time assessing the change in probability from the last step. For instance, the causal efficacy of X_2 would be assessed by assessing $P(Y|X_1 \wedge X_2)$ and subtracting from it the prior assessment, $P(Y|X_1)$. Causal selection is determined by identifying the candidate that accounts for the largest change in outcome probability.

SPA is essentially a probabilistic model of causal selection adapted for single-case judgments, but it also posits that counterfactual thinking can influence causal explanation when the resultant counterfactuals affect relevant subjective probabilities. In particular, SPA predicts that if it is easy to imagine counterfactual alternatives to the actual outcome, then $P(Y)$ will be lowered, correspondingly increasing the chance that X_1 will be selected as the cause. More precisely, if Y can occur k ways from a set of n alternatives, then holding k constant, the number of counterfactuals alternatives to Y should be a positive monotonic function of n . Although this idea is plausible, at present there is a lack of direct evidence demonstrating that subjective probability mediates the effect of counterfactual thinking on causal explanation. Indeed, it has yet to be demonstrated that the factor that yields the largest change in subjective probability will be deemed the cause. In the next section, I report on a recent test of the latter prediction (see also Chapter 3, Hilton *et al.*).

Reflections and new frontiers

It is indisputable that counterfactual, causal, and for that matter, covariational statements *can* convey similar meanings, thereby indicating some degree of conceptual overlap. For instance, if a child becomes ill after eating a poisonous mushroom, one observer might say “He got sick *because* he ate the mushroom,” while another observer might say “Yes, *if only* he hadn’t eaten that mushroom, he wouldn’t have been sick,” while yet another observer might say “*Given that* he ate the mushroom, the *chances* of him getting sick were almost certain.” Although each observer uses a different form of expression, the reasoning in each case seems closely aligned. For this reason, many theories of causal explanation have posited a central role of counterfactual and/or covariational reasoning. Surely, counterfactual simulations can be used to probe the plausibility of causal explanations, sometimes with surprising effects (Tetlock and Belkin 1996b), but it is also true that causal knowledge influences the counterfactuals that people generate. Rather than arguing that one form of thinking precedes the other, it may be more sensible to examine the features of cognition that account for similarities and differences between causal and counterfactual explanations.

Reflections on similarities: is the mind a determinist or an indeterminist?

On the surface, one similarity between people’s causal and counterfactual explanations is that both reveal, as Kahneman (1995: 383) put it, that “the mind is not a determinist.” In hindsight, people tend to perceive outcomes and potential causes as having been more or less mutable. I say “on the surface,” however, because in other respects people don’t appear to be indeterminists either. To illustrate this, consider a simple thought experiment. First, imagine a chain of events culminating in a particular outcome. Now, replay the scenario keeping everything *exactly* the same right up to the point before the final outcome and then ask yourself whether the outcome would be identical or even slightly different. Assuming that the antecedents were *exactly* the same (and notwithstanding any knowledge one might have about quantum electrodynamics), Concurring with Tetlock and Henik (Chapter 12) who assert that historical observers tend to err in the direction overly deterministic thinking, I suspect most would conclude that the outcome would also be exactly the same *in every possible replay*.

I propose that, although the mind is a determinist at heart, it behaves like an indeterminist, mutating Xs and Ys left and right. That the mind is a determinist at heart is revealed, for instance, by people’s proclivity for causal explanations and their dissatisfaction with purely statistical ones. Indeed, a key heuristic strategy for judging likelihood is representativeness, which comprises not only similarity assessments but assessments of causality too (Kahneman and Tversky 1972). Moreover, statistical information such as

22 *David R. Mandel*

base rates (Tversky and Kahneman 1982b) and conditional probabilities (Tversky and Kahneman 1982a) are given greater weight if they are interpreted in causal terms that directly shed light on how the relevant outcomes or criteria may have been determined. Deterministic thinking is also revealed by the examples of finding “too much meaning” discussed by Galinsky *et al.* (Chapter 7).

What keeps the mind from experiencing considerable dissonance, I suspect, is its bounded sense of rationality and its pragmatism. Unlike a true philosophical determinist, the mind is rarely concerned with heady questions about whether the initial conditions of the universe necessitated all that has followed. It simply is not equipped to handle such complexities. First, unlike Laplace’s demon, it cannot store knowledge about everything. Second, what the mind thinks it knows must be manipulated within the confines of working memory. Hence, the mind focuses on a piece of a larger puzzle that wins out, at least momentarily, in a contest of pragmatic relevance. The piece that it focuses on will likely be organized into a scenario that has story-like properties including goals, means, and outcomes (Read 1987).

Accordingly, when one searches for either a causal or counterfactual explanation, the antecedents in the scenario that serve as candidates will only be partially constrained. For instance, social perceivers often explain their own and other actors’ behaviors in terms of choices, intentions, and so on without accounting for (or apparently feeling the need to account for) the determinants of these ostensibly indeterminate explanatory constructs. If the causal determinants of those antecedents are located outside of the focal scenario or are otherwise inaccessible, attention to the constraints on causes – namely, the *meta-causes* – will be highly restricted. The steepness of attentional decline can foster the appearance that the mind is an indeterminist. However, within its field of attention, the mind constructs simplified explanations of *how* or *why*, which conform well to a deterministic stance. In essence, the appearance of indeterminism in causal and counterfactual reasoning is a consequence of focalism, akin to that found in other areas of research (e.g., Schkade and Kahneman 1998; Wilson *et al.* 2000)

New frontiers: judgment dissociation theory

In this subsection, I describe an account called *judgment dissociation theory* (JDT, Mandel 2003c), which attempts to explain how counterfactual, causal, and covariational (i.e., conditional-probability) reasoning differ from each other. Consistent with past functionalist proposals (Roeser and Olson 1997; Taylor and Pham 1996; Weiner 1985), JDT posits that causal, counterfactual, and covariational reasoning can aid in prediction, control, and explanation across many situational contexts. Nevertheless, JDT draws functional distinctions between these types of reasoning and predicts dissociations in the content of causal, counterfactual, and covariational judgments under specified conditions.

Counterfactual and causal explanation 23

1 A key proposal of JDT is that causal, counterfactual, and covariational
2 judgments differ in terms of how “the outcome” of a scenario is conceptual-
3 ized. The theory proposes that causal selection is guided by the *actuality*
4 *principle* – namely, causal reasoners tend to focus on antecedents that played
5 a critical role in how the *actual* outcome of a case was generated.
6 Antecedents that “merely” explain how an outcome similar to the actual
7 one, which might have been probable or even inevitable up to a point in the
8 case, will likely be rejected as “the” cause. That is, JDT posits that causal
9 selections will be perceived as informative to the extent that they elucidate
10 causal processes that *in fact* generated the outcome.

11 In this regard, JDT draws conceptual connections to accounts that posit
12 the importance of perceived causal mechanisms (Ahn and Kalish 2000;
13 Michotte 1963), force dynamics (Wolff and Song 2003), and perceived
14 changes of propensity that can serve as indicators of causal processes (Kahne-
15 man and Varey 1990). As the philosopher W.C. Salmon (1984: 170) put it,
16 “causal processes are the means by which causal influence is propagated, and
17 changes in processes are produced by causal interactions.” According to
18 JDT, then, causal explanations will focus on causal interactions that produce
19 salient changes in causal processes whose propagated force is perceived as
20 having culminated in the actual outcome.

21 By contrast, JDT proposes that counterfactual and covariational assess-
22 ments are guided by the *substitution principle* – namely, they tend to focus on
23 *ad hoc* categories of outcome in which the actual outcome represents the cat-
24 egorical norm. *Ad hoc* categories (e.g., “ways to prevent Mr Jones from having
25 been injured on this day”) are constructed in response to short-term goals
26 that, once satisfied, allow the category to be disbanded (Barsalou 1983,
27 1991). In JDT, the goals that generate such categories are associated in pre-
28 dictable ways with different types of reasoning. Specifically, category exem-
29 plars in counterfactual assessments are constrained by the goal of *finding ways*
30 *to undo the outcome or something like it*, whereas category exemplars in covari-
31 ational assessments are constrained by the goal of *finding ways that would*
32 *increase the probability of the outcome or something like it*.

33 Thus, whereas selected counterfactual conditionals are sensitive to factors
34 that are sufficient in the circumstances to prevent an entire category of
35 outcome, conditional-probability assessments are sensitive to contingencies
36 that raise the probability of any exemplar. More generally, the substitution
37 principle coheres with Kahneman and Tversky’s (1982c: 149) idea that “it is
38 frequently appropriate in conversation to extend the definition of an event X
39 to ‘X or something like it.’” However, the substitution principle generalizes
40 this idea by proposing that a similar extension from X to “X or something
41 like it” aptly characterizes the mental representations that people invoke in
42 counterfactual and covariational reasoning.

43 Mandel (2003c) tested JDT’s predictions by constructing cases with two
44 sequential sufficient conditions for a given type of outcome. For example, in
45 Experiment 2, the protagonist (John) works at one company division for a

24 *David R. Mandel*

year, after which he can keep the same job or switch to another division. He decides to switch and, soon afterwards, he is relocated. Unbeknownst to John, he is exposed to asbestos fibers at the new location and he develops a terminal case of lung cancer. One day, on a routine medical visit to the hospital, a nurse gives John the wrong medication. He immediately suffers cardiac arrest and, although there is a resuscitation attempt, he dies moments after receiving the drug. After reading the scenario, participants listed factors that caused John's death and ways of undoing John's death in counterbalanced order. They also rated the importance of each listing from 0 to 10. Participants then estimated the probability of John dying prematurely conditional on the points at which he: (1) was first hired (2) switched jobs, (3) was moved to the new location with asbestos, (4) spent thirty years in the building with asbestos, and (5) was given the wrong medication.

According to JDT, this scenario will produce a full dissociation in the content of causal, counterfactual, and conditional-probability assessments. Specifically, causal selections will focus primarily on the drug error because that factor played a critical role in how John actually died. By contrast, participants will judge the move to the new location and the ensuing exposure to asbestos as the joint factor that accounted for the greatest increase in the probability of his death – even though he actually died as a consequence of the drug error. Hence, contrary to the prediction of SPA, JDT predicts that participants will judge the drug error as having done little to increase the probability of John's death even though they will select it as the primary cause. Finally, participants' counterfactuals will focus primarily on John's decision to switch jobs because this factor alone is sufficient in the circumstances to undo the actual death *and* the inevitable death by cancer. Note that this prediction is at odds with counterfactual simulation accounts that predict strong overlap in counterfactual and causal content (Wells and Gavanski 1989).

As JDT predicts, participants were more likely to list the drug error as the cause (95 percent) than either the cancer (66 percent) or the decision to switch jobs (38 percent). By contrast, the most frequent counterfactual listing focused on the decision to switch jobs (79 percent), although a sizable percentage (71 percent) mentioned the nurse's error, probably because of its highly controllable nature. These findings do not support counterfactual simulation accounts. Indeed, contrary to that view (Roese and Olson 1997), the mean correlation in rated importance across antecedent targets was not significantly greater when the counterfactual tasks preceded rather than followed the causal tasks.

Analyses of participants' probability judgments also supported JDT. Participants judged that the move to the new location and the ensuing asbestos exposure, taken together, accounted for a 45 percent increase in the probability of John's premature death, whereas the decision to switch jobs and the drug error accounted for only 11 percent and 3 percent increases, respectively. These findings are inconsistent with SPA, which predicts on

Counterfactual and causal explanation 25

1 the basis of these findings that participants would be most likely to select
2 the relocation and subsequent cancer-causing exposure to asbestos as the
3 cause. Moreover, disconfirming the fundamental test of SPA called for
4 earlier, the correlations between probability-change scores and causal-
5 importance ratings were invariably nonsignificant across the targets.

6 In summary, the findings of this experiment and others reported in
7 Mandel (2003c) support the idea that, under certain boundary conditions,
8 causal, counterfactual, and covariational assessments will diverge in focus.
9 These dissociations are particularly likely in scenarios with multiple suffi-
10 cient causes, as in the experiment just described. In such cases, SPA and
11 JDT predict different forms of asymmetric causal discounting. SPA predicts
12 that the earliest sufficient condition will be deemed the cause because of its
13 effect on probability updating, whereas JDT predicts that the last sufficient
14 condition is likely to be deemed the cause because of its probable role in the
15 process by which the effect was generated. In this regard, there is strong
16 agreement between JDT and Mackie's analysis. Mackie (1974: 44–6) con-
17 sidered causal scenarios that included cases much like the one just described.
18 He concluded that, in such cases, "What we accept as causing each result,
19 though not necessary in the circumstances for the result described in some
20 broad way, was necessary in the circumstances for the result *as it came about*"
21 (1974: 46, italics in original).

Conclusion

22
23
24
25 Our understanding of the relationship between counterfactual and causal
26 thinking has evolved from the idea that counterfactuals provide input into
27 the causal explanation process to the idea that counterfactuals provide a dis-
28 tinct form of explanation that shares similarities with causal explanation but
29 that also has important differences. Recent developments indicate several
30 avenues for future research. Although research has shown that causal reason-
31 ing influences categorization (Rehder and Hastie 2001), the effect of catego-
32 rization on causal and counterfactual reasoning is just starting to be
33 examined (Mandel 2003c).

34 As JDT predicts, the explanation of an effect depends on how the effect is
35 conceptualized. Whereas causal explanations favor an instance-based view of
36 effects, counterfactual explanations favor a category-based view. But what
37 types of features influence the perceived similarity between actual and once-
38 inevitable outcomes? JDT posits that perceived functional similarity of con-
39 sequences is one factor, but *what* people treat as similar, as well as possible
40 asymmetries in similarity assessments, remains to be explored. For example,
41 a counterfactual explanation might require undoing an inevitable premature
42 death that was precluded by an unrelated severe injury. But would an
43 inevitable injury precluded by an unrelated premature death also require
44 undoing? And, if so, how serious would the injury have to be?

45 There are also many intriguing questions concerning the validity and

26 *David R. Mandel*

normative status of counterfactual explanations. As Dawes (1996) noted, counterfactuals may be accurate yet vacuous if they are not supported by reliable statistical evidence. In hindsight, it is easy to imagine ways that one could have made an outcome turn out better, and the availability of such thoughts can obscure the fact that, in foresight, success by the counterfactual course of action may have been even less probable (Miller and Turnbull 1990; Sherman and McConnell 1995). Research by Goldinger *et al.* (2003) has also shown that, compared to people with relatively high working memory capacity, those with lower capacity were more likely to blame victims when their actions were highly mutable, thus raising interesting questions about whether people can make appropriate inferential corrections once counterfactuals lacking in validity have been considered (see also Koehler 1991).

Dawes' warning echoes Kahneman and Tversky's (1982c) earlier point that forecasts and decisions based on "inside view" scenarios of a case are poor substitutes for assessments of statistical information based on the "outside view" of many cases. Counterfactual simulations are prime examples of inside-view thinking and, therefore, they are poor methods for rigorously assessing support for hypothesized causes. For instance, counterfactual simulations would seldom provide accurate frequency estimates in the four cells of the confusion matrices shown in Figures 1.1 or 1.2, although they could assist in determining whether it is *possible* that a given cell has a non-zero value. Such thought experiments, in fact, have proved to be very important in science. Error will arise, however, when the ease of generating such examples is mistaken for the frequency or likelihood of such examples.

Note, however, that the role of counterfactual thinking in possibility probing may be far greater than that previously assigned to it. As Figure 1.1 revealed, not only can people inquire about the status of *Y* had *X* been negated (drawing attention to cells C or D depending on the simulation result), they can also probe whether it would have been possible for *Y* not to have happened even if *X* had occurred (drawing attention to cells A and B). If people do in fact assign the term *cause* to conditions that are perceived as sufficient in the circumstances to bring about the actual outcome, then "affirm-*X*, verify-*Y*" probes may be even more important in the causal explanation process than the oft-discussed "negate-*X*, verify-*Y*" probes. For instance, B-cell counterfactuals could signal the insufficiency of a causal candidate by draw attention to the importance of enabling conditions that might not have been present or the possibility of disabling conditions might have intervened.

Another normative issue concerns the co-tenability of counterfactual explanations. Counterfactual thinkers often simulate the consequences of a "minimal rewrite" (Tetlock and Belkin 1996b) by changing *X* and nothing else, but as Jervis (1996) points out, in complex systems it may be impossible to change just one thing. Counterfactual simulations may thus resemble

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45

Counterfactual and causal explanation 27

1 tightly controlled experiments with high internal validity, but they may
 2 also be externally invalid, reflecting the perception that causes are separable
 3 in cases in which main effects are the exception and higher-order interac-
 4 tions are the rule.

5 Despite all their real and possible shortcomings, counterfactuals and the
 6 cognitive processes that produce them are indispensable, and may be essen-
 7 tial for mental health (Hooker *et al.* 2000). The ability to think counterfac-
 8 tually represents a truly remarkable evolutionary achievement, which
 9 probably much more than we have yet acknowledged, defines what it is to
 10 be human. Humans not only relive the past in the present, they also relive a
 11 multitude of pasts that never existed, and they use these simulated realities
 12 to map out possible futures whose truth values remain indeterminate. As
 13 Tulving (2004) argued in a keynote address to the American Psychological
 14 Society, this ability for forward “time travel,” which he calls proscopic
 15 chronesthesia, is a necessary condition for the inception and evolution of all
 16 human cultures. This is an important part of the reason why the study of
 17 counterfactual thinking continues to intrigue. If progress over the last
 18 decade of counterfactual thinking research is a reasonable indicator of what
 19 is to come, then the next decade of research examining the determinants and
 20 consequences of how people explain *what was*, *what might have been*, and *what*
 21 *may be* will surely be revealing. I recommend that we treat this as our best-
 22 guess hypothesis and proceed *as if!*

Notes

23 Preparation of this chapter was facilitated by research grant 249537-02 from the
 24 Natural Sciences and Engineering Research Council of Canada.

- 25 1 The nature of the conditional implications defining necessity and sufficiency is of
 26 considerable debate and beyond the scope of this chapter. Elsewhere (Mandel
 27 2004), I argue that these concepts and their associated conditional implications
 28 are not interpreted as the material conditional implication in everyday discourse
 29 but rather in terms of a modified Ramsey test.
 - 30 2 Goldvarg and Johnson-Laird (2001) define true possibilities in terms of the
 31 material conditional implication (cf. Evans *et al.* 2003; Mandel 2004).
 - 32 3 This framing effect may be restated as the following truth-functional question: Is
 33 $\neg X$ to be interpreted as “X is false” or as “ $\neg X$ is true?” Whereas counterfactual-
 34 simulation accounts assume the former interpretation, the prevention-focus
 35 account assumes the latter interpretation.
- 36
37
38
39
40
41
42
43
44
45