An Informational Theory of Counterfactuals

Danilo Fraga Dantas (dfdantas@ucdavis.edu)

Final version: https://link.springer.com/article/10.1007

Abstract: Backtracking counterfactuals are problem cases for the standard, similarity based, theories of counterfactuals (e.g. Lewis, 1979). These theories usually need to employ extra-assumptions to deal with those cases (e.g. "the standard resolution of vagueness" in Lewis, 1979). Hiddleston (2005) proposes a causal theory of counterfactuals that, supposedly, deals well with backtracking. The main advantage of the causal theory is that it provides a unified account for backtracking and non-backtracking counterfactuals (no extra-assumption is needed). In this paper, I present a backtracking counterfactual that is a problem case for Hiddleston's account. Then I propose an informational theory of counterfactuals, which deals well with this problem case while mantaining the main advantage of Hiddleston's account (the unified account for backtracking and non-backtracking counterfactuals). In addition, the informational theory offers a general theory of backtracking that provides clues for the semantics and epistemology of counterfactuals. I propose that backtracking is reasonable when the (possibly non-actual) state of affairs expressed in the antecedent of a counterfactual transmits less information about an event in the past than the actual state of affairs. **Keywords:** Backtracking counterfactuals; Causal models; Information theory; Epistemology of modality.

Introduction

Backtracking counterfactuals admit counterfactual reasoning that claims that if things had been different at some time t_1, they would also have been different at some earlier time t_0 ('backtracking'). Consider the following situation, described by Jackson (1977): you see your friend Smith on the ledge of a twenty story building. You are afraid that he is going to jump, but (thankfully) Smith steps down and exits the building safely. You note that there was nothing underneath him besides the solid concrete of the sidewalk and conclude: "if Smith had jumped, he would have died". This is a non-backtracking counterfactual. Your friend Beth saw everything and disagrees with you. She argues that Smith is "a very rational person" and, had he jumped, there would have been a net underneath him to catch him safely. She concludes: "if Smith had jumped, he would have lived". This is a backtracking counterfactual. When it is reasonable to backtrack from a counterfactual is an issue for both the semantics and epistemology of counterfactuals.

Hiddleston (2005) proposes a causal theory of counterfactuals that, suposedly, deals well with backtracking. The main advantage of the causal theory is that it provides a unified account for backtracking and non-backtracking counterfactuals. In this paper, I present a backtracking counterfactual (case 3 in sec. 1) that is a problem case for Hiddleston's account (section 1). Then I propose an informational theory of counterfactuals, which deals well with this problem case while

mantaining the main advantage of Hiddleston's account (the unified account for backtracking and non-backtracking counterfactuals, section 2)[1]. In section 3, I propose a general theory of backtracking that provides clues for the semantics and epistemology of counterfactuals. I propose that backtracking is reasonable when the (possibly non-actual) state of affairs expressed in the antecedent of a counterfactual transmits less information about an event in the past than the actual state of affairs.

1 The causal theory of counterfactuals

Hiddleston (2005) develops a theory of counterfactuals using causal models of roughly the same sort as those used in Pearl (1988), Spirtes et al. (2000), and Glymour (1987). In evaluating a counterfactual of the type 'if ϕ had been, then ψ would have been' ($\phi \,\square\!\!\rightarrow\, \psi$), Hiddleston starts building a causal model of the actual situation ('actual model'). Then he introduces minimally altered models where ϕ is true ('counterfactual models'). The counterfactual models are built by introducing minimal causal breaks in the actual model and following the causal consequences of the breaks. If ψ is true in all counterfactual models, then the counterfactual $\phi \,\square\!\!\rightarrow\, \psi$ is true in the actual model.

A causal model M is a triple $\langle G, E, A\rangle$. The first element of a model, G, is a direct acyclic graph composed of nodes representing variables for events and edges representing causal relations[2]. Let the parents of a node X in M ($pa_M(X)$) be the set of nodes with edges into X. The second element of a model, E, is a set of equations of the form $(Y_1 = y_1 \wedge ... \wedge Y_n = y_n) \Rightarrow p(X = x) = z$, where the Y_i are all the members of $pa_M(X)$, the y_i are their respective values in the model (see A), z is the objective probability of X having the value x, and $\Rightarrow$ is a strict conditional. The set E must contain equations relating all possible values of all variables in G to all possible combinations of values for all of their parents. The third element of a model, A, is an assignment of values for all variables in G. The assignment A must be possible given the equations in E[3].

Hiddleston offers a notion of direct positive influence that is used to characterize the notions of a causal break and of a minimally altered model. Let M be an actual model with A containing actual values for all variables in G. Let $M', M'', ...$ be counterfactual models that differ from M only in having A', A'', etc with non-actual values for some variables in $G' = G, G'' = G$, etc (also, $E' = E, E'' = E$, etc). The notion of direct positive influence would therefore be:

Definition 1. Direct positive influence: Let M be a model in which $X = x$ is a parent of $Y = y$ and the other parents of $Y = y$ are in $\vec{Z} = \vec{z}$. Let M' be identical to M except that $X = x'$ in M'. Then $X = x$ has direct positive influence on $Y = y$ in M relative to M' iff $p(Y = y|X = x, \vec{Z} = \vec{z}) > p(Y = y|X = x', \vec{Z} = \vec{z})$.

Let the positive parents of $Y = y$ in M be the set $ppa_M(Y) = \{X : X = x$ has direct positive influence on $Y = y$ in $M\}$. Then the notion of a causal break is the following:

Definition 2. Causal break: A causal break in M' relative to M is a variable Y such that $A'(Y) \neq A(Y)$ and, for every $X \in ppa_M(Y), A'(X) = A(X)$.

[1] In the following, 'the causal theory' denotes Hiddleston's causal theory of counterfactuals and 'the informational theory' denotes the informational theory of counterfactuals proposed here. The theory in Shannon (1948) is referred as 'information theory'.

[2] A directed acyclic graph is a collection of nodes and directed edges in which the edges connect nodes such that it is impossible to start at a node n and follow a sequence of edges that loops back to n.

[3] In other words, A cannot assign a value $A(X)$ to a variable $X \in G$ if, according to E, $p(A(X)|A(pa(X))) = 0$.

In other words, a causal break occurs in a counterfactual model M' relative to an actual model M when a variable Y has a non-actual value in M' whereas all of its positive parents maintain their actual value in M'[4]. In this context, Hiddleston defines two sets:

$$Break(M', M) = \{Y : Y \text{ is a causal break in } M' \text{ relative to } M\}.$$
$$Intact(M', M) = \{Y : A'(Y) = A(Y) \text{ and for all } X \in ppa_M(Y), A'(X) = A(X)\}.$$

The next step is the characterization of the notion of a minimally altered model. Let a ϕ-model be a model in which ϕ is true, where ϕ is either atomic ($X = x$) or a complex (negation, conjunction, etc). $Break(M', M)$ is minimal among ϕ-models iff there is no ϕ-model M'' such that $Break(M'', M) \subset Break(M', M)$. $Intact(M', M)$ is maximal among ϕ-models iff there is no ϕ-model M'' such that $Intact(M', M) \subset Intact(M'', M)$[5]. Then the notion of a ϕ-minimal model is the following:

Definition 3. ϕ-Minimal model: M' is ϕ-minimal relative to M iff

(a) M' is a ϕ-model;

(b) for Z, the set of variables that are not descendants of ϕ, $Intact(M', M) \cap Z$ is maximal among ϕ-models[6]; and

(c) $Break(M', M)$ is minimal among ϕ-models.

If ϕ is true in M, then M is the ϕ-minimal model relative to M and $\{\}$ is the minimal $Break$. Finally, this is Hiddleston's causal theory:

Definition 4. Causal theory of counterfactuals: A counterfactual $\phi \mathrel{\square\!\!\rightarrow} \psi$ is true in a model M iff ψ is true in all ϕ-minimal models M'.

A counterfactual $\phi \mathrel{\square\!\!\rightarrow} \psi$ is true of a case C iff $\phi \mathrel{\square\!\!\rightarrow} \psi$ is true in M and M is an adequate model of C. M is an adequate model of C when (i) the properties represented in M are instantiated by the objects in C, (ii) the causal laws used in M are accurate (enough) for C, and (iii) M is complete enough to accurately represent the causal relations between the events of C that appear in M (Hiddleston, 2005, p. 648). In the following, I will not discuss the adequacy of models to cases, but only Hiddleston's theory for when a counterfactual is true in a model.

Case 1

> Suppose that the boss will randomly draw the name of an employee a, b, c, or d from a jar and write that name in a memo for a promotion. Suppose that a will randomly bet that the name of a specific colleague will appear in the memo. Suppose that c was drawn from the jar, c is written in the memo, a bet on d and a lost the bet.

The question is: 'if d were in the memo, then a would win the bet?' ($Memo = d \mathrel{\square\!\!\rightarrow} Win = 1$?). Hiddleston's answer is 'yes' and I think that this answer is correct. This answer is depicted in

[4]It follows that if Y has no parents, then any change in Y is a causal break.

[5]'Minimal' and 'maximal' are measured using set-inclusion rather than the number of breaks/intacts. In this context, two different $Break/Intact(M', M)$ and $Break/Intact(M'', M)$ may be both minimal/maximal.

[6]A child is a descendant; a child of a descendant is a descendant.

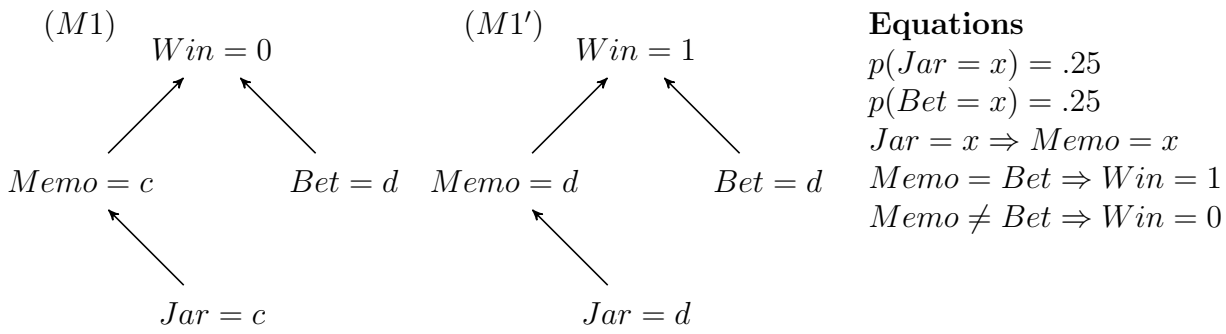


Figure 1: $M1$ is an actual model of case 1 and $M1'$ is the only $(Memo = d)$-minimal model allowed, where $x \in \{a, b, c, d\}$. $Break(M1', M1) = \{Jar\}$ and $Intact(M1', M1) \cap Z = \{Bet\}$.

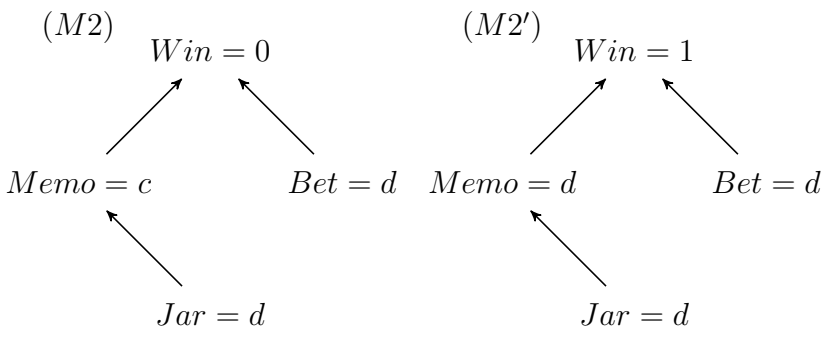
figure 1. $M1$ is an actual model of case 1 and $M1'$ is the only $(Memo = d)$-minimal model allowed. $Break(M1', M1) = \{Jar\}$ is the only minimal $Break$ possible because Jar needs to be in $Break$, otherwise $Memo = d$ would be impossible. $Intact(M1', M1) \cap Z = \{Bet\}$ is the only maximal $Intact$ possible because Bet and Jar are the only members of Z and Jar must be in $Break$. As a consequence, $M1'$ is the only $(Memo = d)$-minimal model allowed. $Win = 1$ is true in $M1'$. Then $Memo = d \boxdot\!\!\rightarrow Win = 1$ is true in $M1$. This answer involves backtracking from $Memo = d$ to $Jar = d$ (in $M1'$).

I think that this answer is correct for two reasons. The first reason, I think, is that the answer seems to be correct: since $Win = 1$ iff $Bet = Memo$, it seems to be true that if $Memo$ had the same value that Bet actually has, then Win would be 1 ($Memo = d \boxdot\!\!\rightarrow Win = 1$). The second reason is that there is a symmetry between $Memo = d \boxdot\!\!\rightarrow Win = 1$ and $Bet = c \boxdot\!\!\rightarrow Win = 1$: since $Win = 1$ iff $Bet = Memo$, asking whether Win would be 1 if $Memo$ had the same value that Bet actually has ($Memo = d \boxdot\!\!\rightarrow Win = 1$) and asking whether Win would be 1 if Bet had the same value that $Memo$ actually has ($Bet = c \boxdot\!\!\rightarrow Win = 1$) seems to be two different ways of asking the same question. Both $Memo = d \boxdot\!\!\rightarrow Win = 1$ and $Bet = c \boxdot\!\!\rightarrow Win = 1$ are true in Hiddleston's account and this is a good result.

Case 2

> Suppose that the boss will randomly draw the name of an employee a, b, c, or d from a jar and intends to write that name in a memo for a promotion. But the boss has dyslexia. If c is drawn, she will write c with probability .01 or d with probability .99. If d is drawn, she will write d with probability .01 and c with probability .99. Suppose that a will randomly bet that the name of a specific colleague will appear in the memo. Suppose that d was drawn from the jar, c is written in the memo, a bet on d and a lost the bet.

The question is: 'if d were in the memo, then a would win the bet?' ($Memo = d \boxdot\!\!\rightarrow Win = 1$?). Hiddleston's answer is 'yes' and I think that this answer is correct. This answer is depicted in figure 2. $M2$ is an actual model of case 2 and $M2'$ is the only $(Memo = d)$-minimal model allowed. $Break(M2', M2) = \{Memo\}$ is minimal because either $Memo$ or Jar need to be in $Break$, otherwise $M2'$ would not be a $(Memo = d)$-model. $Intact(M2', M2) \cap Z = \{Bet, Jar\}$ is the only maximal $Intact$ possible because Bet and Jar are the only members of Z



Equations

$p(Jar = x) = .25$
$p(Bet = x) = .25$
$Jar = y \Rightarrow Memo = y$
$Jar = c \Rightarrow p(Memo = c) = .01$
$Jar = c \Rightarrow p(Memo = d) = .99$
$Jar = d \Rightarrow p(Memo = d) = .01$
$Jar = d \Rightarrow p(Memo = c) = .99$
$Memo = Bet \Rightarrow Win = 1$
$Memo \neq Bet \Rightarrow Win = 0$

Figure 2: $M2$ is an actual model of case 2 and $M2'$ is the only $(Memo = d)$-minimal model allowed, where $x \in \{a, b, c, d\}$ and $y \in \{a, b\}$. $Break(M2', M2) = \{Memo\}$ and $Intact(M2', M2) \cap Z = \{Bet, Jar\}$.

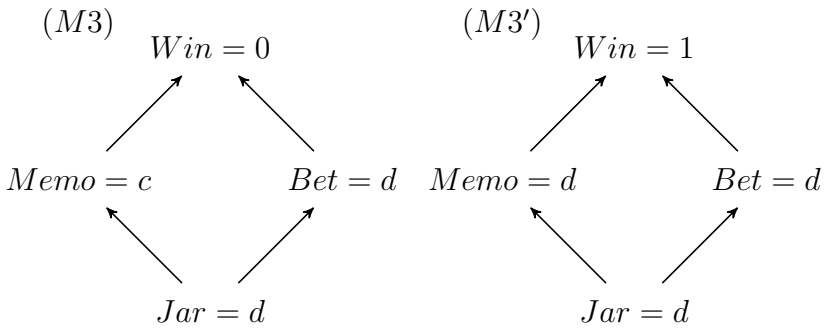
and both are in $Intact(M2', M2) \cap Z$. If Jar were in a $Break(M2'', M2)$, then Jar would not be in $Intact(M2'', M2) \cap Z$ and $Intact(M2'', M2) \cap Z$ would not be maximal. As a consequence, $M2'$ is the only $(Memo = d)$-minimal model allowed. $Win = 1$ is true in $M2'$. Then $Memo = d \; \square\!\!\rightarrow \; Win = 1$ is true in $M2$. This answer does not involve backtracking.

I think that this answer is correct for two reasons. The first reason, I think, is that the answer seems to be correct: since $Win = 1$ iff $Bet = Memo$, it seems to be true that if $Memo$ had the same value that Bet actually has, then Win would be 1 ($Memo = d \; \square\!\!\rightarrow \; Win = 1$). The second reason is that there is a symmetry between $Memo = d \; \square\!\!\rightarrow \; Win = 1$ and $Bet = c \; \square\!\!\rightarrow \; Win = 1$: since $Win = 1$ iff $Bet = Memo$, asking whether Win would be 1 if $Memo$ had the same value that Bet actually has ($Memo = d \; \square\!\!\rightarrow \; Win = 1$) and asking whether Win would be 1 if Bet had the same value that $Memo$ actually has ($Bet = c \; \square\!\!\rightarrow \; Win = 1$) seems to be two different ways of asking the same question. Both $Memo = d \; \square\!\!\rightarrow \; Win = 1$ and $Bet = c \; \square\!\!\rightarrow \; Win = 1$ are true in Hiddleston's account and this is a good result.

Case 3

> Suppose that the boss will randomly draw the name of an employee a, b, c, or d from a jar and intends to write that name in a memo for a promotion. But the boss has dyslexia. If c is drawn, she will write c with probability .01 or d with probability .99. If d is drawn, she will write d with probability .01 or c with probability .99. Suppose that a has an infallible method to know which name was drawn from the jar. She will bet that this name is in the memo. Suppose that d was drawn from the jar, c is written in the memo, a bet on d and a lost the bet.

The question is: 'if d were in the memo, then a would win the bet?' ($Memo = d \; \square\!\!\rightarrow \; Win = 1$?). Hiddleston's answer is 'yes' and I think that this answer is *incorrect*. This answer is depicted in figure 3. $M3$ is an actual model of case 3 and $M3'$ is the only $(Memo = d)$-minimal model allowed. $Break(M3', M3) = \{Memo\}$ is minimal because either $Memo$ or Jar need to be in $Break$, otherwise $M3'$ would not be a $(Memo = d)$-model. $Intact(M3', M3) \cap Z = \{Bet, Jar\}$ is the only maximal $Intact$ possible because Bet and Jar are the only members of Z and both are in $Intact(M3', M3) \cap Z$. If Jar were in a $Break(M3'', M3)$, then Jar would not be



Equations

$p(Jar = x) = .25$
$Jar = x \Rightarrow Bet = x$
$Jar = y \Rightarrow Memo = y$
$Jar = c \Rightarrow p(Memo = c) = .01$
$Jar = c \Rightarrow p(Memo = d) = .99$
$Jar = d \Rightarrow p(Memo = d) = .01$
$Jar = d \Rightarrow p(Memo = c) = .99$
$Memo = Bet \Rightarrow Win = 1$
$Memo \neq Bet \Rightarrow Win = 0$

Figure 3: $M3$ is an actual model of case 3 and $M3'$ is the only $(Memo = d)$-minimal model allowed, where $x \in \{a, b, c, d\}$ and $y \in \{a, b\}$. $Break(M3', M3) = \{Memo\}$ and $Intact(M3', M3) \cap Z = \{Bet, Jar\}$.

in $Intact(M3'', M3) \cap Z$ and $Intact(M3'', M3) \cap Z$ would not be maximal. As a consequence, $M3'$ is the only $(Memo = d)$-minimal model allowed. $Win = 1$ is true in $M3'$. Then $Memo = d \boxrightarrow Win = 1$ is true in $M3$. This answer does not involve backtracking.

I think that this answer is incorrect for two reasons. The first reason, I think, is that this counterfactual seems to admit backtracking: if $Memo = d$, then, most probably, $Jar = c$, which entails that $Bet = c$ and $Win = 0$. I think that this is good reasoning because $Memo = d$ is a good reason (as good as any non-conclusive reason) for $Jar \neq d$ and for $Jar = c$ ($p(Jar = d|Memo = d) = .01 < p(Jar \neq d|Memo = d) = p(Jar = c|Memo = d) = .99$). Hiddleston defends a similar line of reasoning in discussing his example #4:

> *Example #4:* Alice offers Ben a bet on a coin toss, but this time Alice can influence its outcome. Ben bets tails. Hoping to win, Alice flips the coin so that it has a higher chance (.8, say) of landing heads. It does land heads. She says to Ben, 'If you had bet heads, then you would have won'. That seems false.

The second reason has to do with the symmetry between $Memo = d \boxrightarrow Win = 1$ and $Bet = c \boxrightarrow Win = 1$. In Hiddleston's account, $Memo = d \boxrightarrow Win = 1$ is true, but $Bet = c \boxrightarrow Win = 1$ is false[7]. However, since $Win = 1$ iff $Bet = Memo$, asking whether Win would be 1 if $Memo$ had the same value that Bet actually has ($Memo = d \boxrightarrow Win = 1$) and asking whether Win would be 1 if Bet had the same value that $Memo$ actually has ($Bet = c \boxrightarrow Win = 1$) seems to be two different ways of asking the same question. This is a bad result.

2 The informational theory of counterfactuals

The causal theory fails in case 3 because, in this case, there is a relevant positive influence between $Memo$ and Bet that is different from Hiddleston's direct positive influence. However, while $Memo$ and Bet are not positive parents of each other, they are dependent nevertheless. Two variables X and Y are independent ($X \perp Y$) iff their joint probability equals the product of their probabilities

[7]If $Bet = c$, then Jar must be c. Then there are two $(Bet = c)$-minimal models: one in which $Memo = c$ and another in which $Memo = d$ ($Break$ and $Intact$ are equal in both models because the value of Jar has changed). The value of Win is 1 in the first model, but 0 in the second. Then $Bet = c \boxrightarrow Win = 1$ is false in $M3$.

(i.e. $X \perp Y$ iff $p(X = x, Y = y) = p(X = x)p(Y = y)$). In case 3, $p(Memo = c, Bet = d) = .247 \neq p(Memo = c)p(Bet = d) = .062$. Then $Memo \not\perp Bet$. This is what generates the problem for Hiddleston's account. In order to evaluate case 3 correctly, a theory needs to take into account all relations of dependence. In this section, I propose a theory of counterfactuals based on information theory which takes into account the dependence relation.

Information theory (Shannon, 1948) provides a measure of how much information is associated with a given state of affairs. The amount of information associated with a proposition $X = x$ ($i(X = x)$) is calculated as follows[8]:

$$i(X = x) = -log_2(p(X = x)).$$

Pointwise mutual information (PMI) is a measure of the mutual information between two propositions, which accounts for dependence. The PMI between two propositions $X = x$ and $Y = y$ ($pmi(X = x; Y = y)$) is calculated as follows:

$$\begin{aligned} pmi(X = x; Y = y) &= i(X = x) + i(Y = y) - i(X = x, Y = y) \\ &= -log_2(p(X = x)) - log_2(p(Y = y)) + log_2(p(X = x, Y = y)). \end{aligned}$$

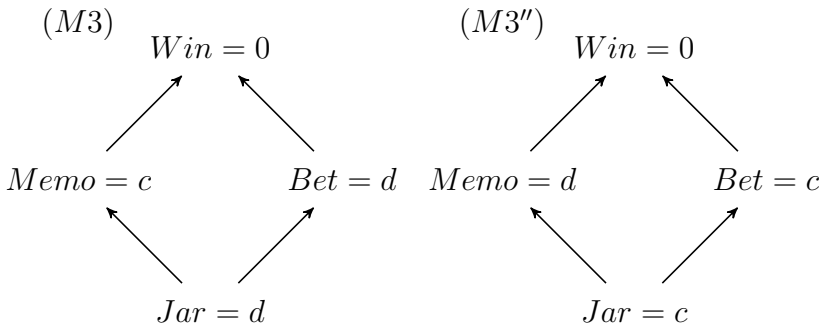
The PMI between $X = x$ and $Y = y$ is such that $-\infty \leq pmi(X = x; Y = y) \leq min(i(X = x), i(Y = y))$. If $pmi(X = x; Y = y) < 0$, then $X = x$ and $Y = y$ are negatively correlated. If $pmi(X = x; Y = y) = 0$, then $X = x$ and $Y = y$ are independent. If $pmi(X = x; Y = y) > 0$, then $X = x$ and $Y = y$ are positively correlated. The PMI between two propositions $X = x$ and $Y = y$ conditional to some propositions $\vec{Z} = \vec{z}$ ($pmi(X = x; Y = y|\vec{Z} = \vec{z})$) is calculated as follows:

$$\begin{aligned} pmi(X = x; Y = y|\vec{Z} = \vec{z}) = &- log_2(p(X = x|\vec{Z} = \vec{z})) - log_2(p(Y = y|\vec{Z} = \vec{z})) \\ &+ log_2(p(X = x, Y = y|\vec{Z} = \vec{z})). \end{aligned}$$

In this context, the informational theory is as follows:

Definition 5. Informational theory of counterfactuals: Let M be an actual model with the actual values $X = x$ and $Y = y$. Consider a counterfactual $X = x' \,\square\!\!\rightarrow Y = y'$. The evaluation of a counterfactual is as follows:

1. Check whether the antecedent $X = x'$ is true in M (whether $x' = x$);
   (a) If yes, check whether $Y = y'$ is also true in M (whether $y' = y$). If yes, return 'true'; otherwise, return 'false'.
   (b) If not, do the following:
      i. For every $Z = z$ in M ($Z \neq X$), check whether $pmi(X = x'; Z = z|\vec{W} = \vec{w}) \geq pmi(X = x; Z = z|\vec{W} = \vec{w})$, where $\vec{W} = \vec{w}$ are the parents of $Z = z$ that are in $Holdfix$. If yes, add Z to the set $Holdfix$. If not, do nothing.
      ii. Build all $(X = x')$-models that are consistent with the equations in M and in which all variables in $Holdfix$ maintain their actual values.
      iii. If $Y = y'$ is true in all of these models, return 'true'; return 'false' otherwise.



Equations

$p(Jar = x) = .25$
$Jar = x \Rightarrow Bet = x$
$Jar = y \Rightarrow Memo = y$
$Jar = c \Rightarrow p(Memo = c) = .01$
$Jar = c \Rightarrow p(Memo = d) = .99$
$Jar = d \Rightarrow p(Memo = d) = .01$
$Jar = d \Rightarrow p(Memo = c) = .99$
$Memo = Bet \Rightarrow Win = 1$
$Memo \neq Bet \Rightarrow Win = 0$

Figure 4: $M3$ is an actual model of case 3. $M3''$ is the only $(Memo = d)$-counterfactual model allowed according to the informational theory ($Holdfix = \{Win\}$).

Case 3 again

The question was: 'if d were in the memo, then a would win the bet?' ($Memo = d \boxdot\!\!\rightarrow Win = 1$?). The answer from the informational theory is 'no' and I think that this answer is correct. This answer is depicted in figure 4. $M3$ is an actual model of case 3 and $M3''$ is the only $(Memo = d)$-counterfactual model allowed. Jar is not in $Holdfix$ because $pmi(Memo = d; Jar = d) = -4.634 < pmi(Memo = c; Jar = d) = 1.985$. Bet is not in $Holdfix$ because $pmi(Memo = d; Bet = d) = -4.634 < pmi(Memo = c; Bet = d) = 1.985$. Win is in $Holdfix$ because $pmi(Memo = d; Win = 0) = pmi(Memo = c; Win = 0) = .971$. Then $Holdfix = \{Win\}$. $M3''$ is the only $(Memo = d)$-counterfactual model compatible with the equations in $M3$ and in which all variables in $Holdfix$ maintain their actual values. $Win = 1$ is false in $M3''$. Then $Memo = d \boxdot\!\!\rightarrow Win = 1$ is false in $M3$. This answer involves backtracking from $Memo = d$ to $Jar = c$.

I think that this answer is correct for two reasons. The first reason, I think, is that this counterfactual seems to admit backtracking: if $Memo = d$, then, most probably, $Jar = c$, which entails that $Bet = c$ and $Win = 0$. I think that this is good reasoning because $Memo = d$ is a good reason (as good as any non-conclusive reason) for $Jar \neq d$ and for $Jar = c$ ($p(Jar = d|Memo = d) = .01 < p(Jar \neq d|Memo = d) = p(Jar = c|Memo = d) = .99$). This backtracking reasoning is represented in model $M3''$. The second reason is that there is a symmetry between $Memo = d \boxdot\!\!\rightarrow Win = 1$ and $Bet = c \boxdot\!\!\rightarrow Win = 1$: since $Win = 1$ iff $Bet = Memo$, asking whether Win would be 1 if $Memo$ had the same value that Bet actually has ($Memo = d \boxdot\!\!\rightarrow Win = 1$) and asking whether Win would be 1 if Bet had the same value that $Memo$ actually has ($Bet = c \boxdot\!\!\rightarrow Win = 1$) seems to be two different ways of asking the same question. From the perspective of informational theory, $Memo = d \boxdot\!\!\rightarrow Win = 1$ and $Bet = c \boxdot\!\!\rightarrow Win = 1$ are both false[9]. This is a good result.

[8] All the definitions in this section may be easily adapted to complex propositions (negations, conjunctions, etc).

[9] Jar is not in $Holdfix$ because $pmi(Bet = c; Jar = d) \approx -\infty < pmi(Bet = d; Jar = d) = 2$. $Memo$ is not in $Holdfix$ because $pmi(Bet = c; Memo = c) = -4.634 < pmi(Bet = d; Memo = c) = 1.985$. Win is in $Holdfix$ because $pmi(Bet = c; Win = 0) = pmi(Bet = d; Win = 0) = .971$. Then $M3''$ is the only $(Memo = d)$-model compatible with the equations of $M3$ and with actual values for all variables in $Holdfix$. $Win = 1$ is false in $M3''$. Therefore $Bet = c \boxdot\!\!\rightarrow Win = 1$ is false in $M3$.

The other cases

The causal and the informational theories evaluate a counterfactual $\phi \boxdot\!\!\rightarrow \psi$ following roughly the same steps: first, check whether there is a change in the actual value of ϕ; if yes, execute a procedure to check whether the uncertainty regarding that value propagates to other variables; finally, build all allowed ϕ-models given the equations in the actual model. The counterfactual $\phi \boxdot\!\!\rightarrow \psi$ is true iff ψ is true in all such models. Propagation may be forward (from ancestor to descendant) or backward (from descendant to ancestor) and may occur via deterministic or nondeterministic relations[10].

Almost all propagation in the causal theory is forward. The reason is that the only causal influence considered in Hiddleston's account is positive parenthood, which is forward. The cases of forward propagation in the causal theory are exactly the cases of forward propagation in the informational theory. The cases of forward propagation in the causal theory are cases in which $X = x$ is a positive parent of $Y = y$. If $X = x$ is a positive parent of $Y = y$, then $p(Y = y|X = x, \vec{Z} = \vec{z}) > p(Y = y|X = x', \vec{Z} = \vec{z})$ (def. 1). The following equivalences hold:

$$
\begin{aligned}
p(Y = y|X = x, \vec{Z} = \vec{z}) &> p(Y = y|X = x', \vec{Z} = \vec{z}) \Leftrightarrow \\
&> p(Y = y, X = x', \vec{Z} = \vec{z})/p(X = x', \vec{Z} = \vec{z}) \Leftrightarrow \\
&> [p(Y = y, X = x'|\vec{Z} = \vec{z})p(\vec{Z} = \vec{z})]/[p(X = x'|\vec{Z} = \vec{z})p(\vec{Z} = \vec{z})] \Leftrightarrow \\
&> -log_2(p(X = x'|\vec{Z} = \vec{z})) + log_2(p(X = x', Y = y|\vec{Z} = \vec{z})) \Leftrightarrow \\
&> -log_2(p(X = x'|\vec{Z} = \vec{z})) - log_2(p(Y = y|\vec{Z} = \vec{z})) \\
&\quad + log_2(p(X = x', Y = y|\vec{Z} = \vec{z})) \Leftrightarrow \\
pmi(X = x; Y = y|\vec{Z} = \vec{z}) &> pmi(X = x'; Y = y|\vec{Z} = \vec{z}) \Leftrightarrow \\
\neg(pmi(X = x'; Y = y|\vec{Z} = \vec{z}) &\geq pmi(X = x; Y = y|\vec{Z} = \vec{z})).
\end{aligned}
$$

Then $p(Y = y|X = x, \vec{Z} = \vec{z}) > p(Y = y|X = x', \vec{Z} = \vec{z})$ iff $\neg(pmi(X = x'; Y = y|\vec{Z} = \vec{z}) \geq pmi(X = x; Y = y|\vec{Z} = \vec{z})$. In the informational theory, Y is not in $Holdfix$ iff $\neg(pmi(X = x'; Y = y|\vec{Z} = \vec{z}) \geq pmi(X = x; Y = y|\vec{Z} = \vec{z}))$. These are the cases of forward propagation in the informational theory (def. 5). As a consequence, the cases of forward propagation in the causal theory are exactly the cases of forward propagation in the informational theory.

There is some backward propagation in the causal theory. This backward propagation, however, only occurs through deterministic relations. The reason is that, in Hiddleston's account, backward propagation does not happen directly because of any causal influence considered in the theory (e.g. positive parenthood), but rather because of the procedure for selecting the allowed counterfactual models given the equations in the actual model[11]. Since this procedure is the same in both the causal and informational theories , the cases of deterministic backward propagation in the causal theory are exactly the cases of deterministic backward propagation in the informational theory. Then the causal and the informational theories agree in all cases of forward propagation on uncertainty and in all cases of deterministic backward propagation.

The two theories may disagree in cases of backward propagation through nondeterministic relations. These cases occur in the informational theory, but do not occur in the causal theory.

[10]The ancestors of a node X are the nodes Y such that X is a descendant of Y.

[11]If the relation between parent and child is deterministic, a change in the child may force a change in the parent because, otherwise, the resulting model may not be allowed given the equations in the actual model.

For example, the informational (but not the causal) theory considers backward propagation from $Memo = d$ to $Jar = c$ in case 2. The consequence of this disagreement is that the causal and the informational theories may return different results about backtracking. For example, the informational (but not the causal) theory advises backtracking from $Memo = d$ to $Jar = c$ in case 2. The divergence regarding backward propagation, however, does not necessarily generate different results. For example, in case 2, the theories agree that the counterfactual $Memo = d \mathrel{\square\!\!\rightarrow} Win = 1$ is true.

The theories return different results for two groups of cases. The first group is composed of counterfactuals that are manifestly backtracking (the antecedent of the counterfactual is a descendant of the consequent, e.g. $Memo = d \mathrel{\square\!\!\rightarrow} Jar = d$ in case 2.), but it seems reasonable to backtrack when evaluating a manifestly backtracking counterfactual. Since the theories may disagree in cases of nondeterministic backward propagation, they may also disagree in cases of forward-backward and backward-forward propagation when the backward propagation is nondeterministic[12]. These cases may occur in situations of nondeterministic siblings (this is the second group of cases)[13]. In these cases, the theories may return different results even for non-manifestly backtracking counterfactuals. These are the least consensual cases. At least in some of these cases, I think that the informational theory is correct. For example, I think that the informational theory is correct in case 3, which is a case of backward-forward propagation where $Memo$ and Bet are nondeterministic siblings.

It is difficult to assess the general situation of these two theories regarding cases involving nondeterministic siblings because these cases are not common in the literature. For example, there is not a single case of nondeterministic siblings among the examples discussed in Hiddleston (2005). Examples 1−7 do not present siblings. In examples 4 and 7, $Coin/Dice$ and Win share a common ancestor Bet, but they are not siblings because $Coin/Dice$ is a parent of Win (figure 5). There is a case that Hiddleston discusses outside the main examples that presents siblings (figure 5). In this case, $Flash$ and $Bang$ are siblings, but the relations between $Explosion$ and $Flash$ and between $Explosion$ and $Bang$ are deterministic. This feature alone makes the causal theory deal correctly with this case, but this would not happen if some relations were not deterministic (as in case 3).

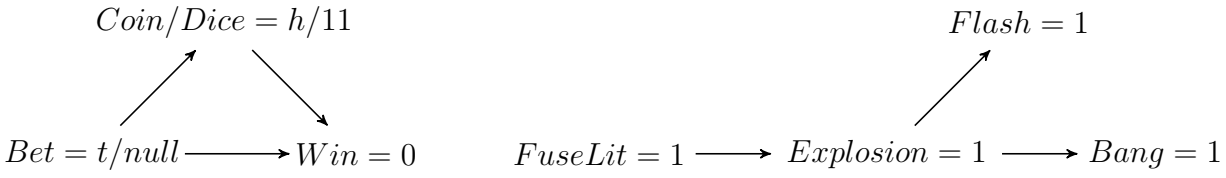


Figure 5: Left, examples 4 and 7 (Hiddleston, 2005). Right, a case discussed in Hiddleston (2005).

[12]Backward-forward propagation is propagation from a node to an ancestor and then to another descendant of the ancestor. Forward-backward propagation is propagation from a node to a descendant and then to another ancestor of the descendant.

[13]A common ancestor of X and Y is an ancestor of both X and Y. X and Y are siblings iff they share a common ancestor, but are not ancestors of each other.

3 General theory of backtracking

Backtracking counterfactuals admit counterfactual reasoning that claims that if things had been different at some time t_1, they would also have been different at some earlier time t_0. When it is reasonable to backtrack is an issue for both the semantics and epistemology of counterfactuals. My initial idea is that backtracking is reasonable when the (possibly non-actual) state of affairs expressed in the antecedent of a counterfactual transmits less information about an event in the past than the actual state of affairs. The informational theory states that, in evaluating a counterfactual $X = x' \,\square\!\!\rightarrow Y = y$, we should consider a change in the value of some variable Z when the (possibly non-actual) value $X = x'$ transmits less information about the actual value of Z than the actual value $X = x$ (possibly, $x = x'$) − given the actual values for $\vec{W}$, which are the parents of Z that are not subject to change (def. 5). From that definition, it follows that:

Definition 6. General theory of backtracking: A counterfactual $X = x' \,\square\!\!\rightarrow Y = y$ admits backtracking when the (possibly) non-actual value $X = x'$ transmits less information about the actual value of Z (an ancestor X) than the actual value $X = x$ (possibly $x = x'$) − given the actual values of $\vec{W}$, which are the parents of Z that are not subject to change, in other words, when it is the case that $pmi(X = x'; Z = z|\vec{W} = \vec{w}) < pmi(X = x; Z = z|\vec{W} = \vec{w})$.

The first thing to note is that the general theory of backtracking may express my initial idea because the relation of parenthood represents the causal relation and, consequently, if Z is an ancestor of X, then Z is prior to X in time[14].

In the following, I will defend that the informational theory and the general theory of backtracking are reasonable. An intuitive reading of $pmi(X = x; Y = y)$ is 'the amount of information that $X = x$ transmits about $Y = y$'. This interpretation is used in Dretske (1981, p. 15-16, my emphasis):

> We are now asking about the informational value of situation r, but we are not asking about $I(r)$. We are asking how much of $I(r)$ is information received from or *about* s. I shall use the symbol $I_s(r)$ to designate this new quantity. The r in parentheses indicates that we are asking about the amount of information associated with r, but the subscript s is meant to signify that we are asking about that portion of $I(r)$ that is information received from s. ...$I_s(r)$ is a measure of the information in situation r *about* situation s.

Then the general theory of backtracking and the informational theory state that we should consider a change in the value of a variable in a counterfactual situation when the state of affairs expressed in the antecedent of the counterfactual transmits less information about the actual value of the variable than the corresponding actual state of affairs.

In some situations, almost all information about a variable is shared with another variable. For example, in case 3, the variable $Memo$ transmits almost all existing information about Jar. The amount of information associated with $Memo$ or Jar is 2 bits ($i(Memo) = 2, i(Jar) = 2$)[15]. The amount of information shared between $Memo$ and Jar is 1.96 bits ($i(Memo; Jar) = 1.96$)[16].

[14]Supposing that a cause is prior to its effect.

[15]The amount of information in a variable X ($i(X)$) is the expected value of the amount of information in each value $X = x$: $i(X) = \sum_x i(X = x)p(X = x)$.

[16]The mutual information between two variables X and Y ($mi(X; Y)$) is the expected value of the pointwise mutual information between each pair of value $X = x$ and $Y = y$: $mi(X; Y) = \sum_{x,y} pmi(X = x; Y = y)p(X = x, Y = y)$.

Then $Memo$ transmits almost all existing information about Jar. Therefore, it is reasonable to form beliefs about Jar from the observation of $Memo$. In a counterfactual situation in which a non-actual antecedent $X = x'$ is true, the only source of information about the value of some other variable Z is the value of X itself (and the causal structure of the model). Then loosing information about the value of Z from X is, in fact, loosing information about which value Z holds. Loosing information about which value Z holds is loosing information about whether Z maintains its actual value. Loosing information about whether Z maintains its actual value is loosing justification for the belief that Z maintains its actual value. In this situation, it seems to be reasonable to consider a change in the value of Z. Given this, I think that the informational theory and the general theory of backtracking are reasonable.

The main advantage of the causal theory is to provide a unified account for both backtracking and non-backtracking counterfactuals. The fact that the general theory of backtracking follows as a special case of the informational theory shows that this theory also provides a unified account for both non-backtracking and backtracking counterfactuals. When we should backtrack is a special case of when we should consider changing the value of a variable in general.

4 Conclusions

In this paper, I have presented a backtracking counterfactual that is a problem case for Hiddleston's theory and proposed an informational theory that deals with this problem case while maintaining Hiddleston's correct results for the other problem cases. The main advantage of Hiddleston's theory was the elimination of the asymetry between the treatment of backtracking and non-backtracking counterfactuals. The informational theory maintains this advantage and eliminates the last asymmetry of Hiddleston's theory: the qualitatively different treatment for deterministic and indeterministic relations.

There are counterfactual theories of causality (Lewis, 1973) and causal theories of counterfactuals (Hiddleston, 2005). There are counterfactual theories of information (Cohen and Meskin, 2006) and informational theories of counterfactuals. These three notions seem to be related. In my opinion, information is the primitive notion from which causality and counterfactuality are derivable. However, this is an issue for further research.

Acknowledgements

I would like to thank Prof. Adam Sennet for the class on counterfactuals for which this paper was my final paper and for his comments on the paper and Phillip Villani for his careful reading.

References

Cohen, J. & Meskin, A. (2006). An objective counterfactual theory of information. *Australasian Journal of Philosophy*, *84* (3), 333–352.

Collier, J. D. (1999). Causation is the transfer of information. In H. Sankey (Ed.), *Causation and laws of nature* (pp. 215–245). Kluwer Academic Publishers.

Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA: The MIT Press.

Glymour, C. (1987). *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling*. Academic Press.

Hiddleston, E. (2005). A causal theory of counterfactuals. *Noûs*, *39(4)*, 632–657.

Jackson, F. (1977). A causal theory of counterfactuals. *Australasian Journal of Philosophy*, *55*(1), 3–21.

Lewis, D. (1973). Causation. *Journal of Philosophy*, *70*(17), 556–567.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, *13*(4), 455–476.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Pearl, J. (2000). *Causality. models, reasoning, and inference*. Cambridge: Cambridge University Press.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd). Cambridge: MIT Press.