The predictive role of counterfactuals

Alfredo Di Tillio, Itzhak Gilboa, Larry Samuelson

The Predictive Role of Counterfactuals*

Alfredo Di Tillio†, Itzhak Gilboa‡, and Larry Samuelson§

February 1, 2011

Abstract

We suggest a model that describes how counterfactuals are constructed and justified. The model can describe how counterfactual beliefs are updated given the unfolding of actual history. It also allows us to examine the use of counterfactuals in prediction, and to show that a logically omniscient reasoner gains nothing from using counterfactuals for prediction.

Contents

*We thank Dov Samet and David Schmeidler for conversations that motivated and influenced this work. We are also grateful to Joe Altonji and Brian Hill for discussions, comments, and references. Itzhak Gilboa gratefully acknowledges ISF Grant 396/10 and ERC Grant 269754; Larry Samuelson gratefully acknowledges NSF grant SES-0850263.

†Universita Bocconi. alfredo.ditillio@unibocconi.it

‡Tel-Aviv University and HEC, Paris. tzachigilboa@gmail.com

§Yale University. larry.samuelson@yale.edu

The Predictive Role of Counterfactuals

1 Introduction

A counterfactual is a conditional statement whose antecedent is known to be false. For example, "If the Iranian hostages had been released before the US presidential election in 1980, President Carter would have been reelected" and "Even if the Iranian hostages had been released before the US presidential election in 1980, President Carter would not have been reelected" are both counterfactuals.

Counterfactuals are common in everyday parlance and they seem to play an important role in the way people reason about social, economic, and political phenomena. It is somewhat surprising that people can reason about counterfactuals and come up with arguments about them that are acceptable to others. After all, counterfactual statements cannot be empirically validated. And yet, some counterfactuals appear to be more reasonable than others. For example, one may ask what would have happened had the US government rescued Lehman Brothers in 2008, and find that some predictions make more sense than others. Moreover, the ranking of such predictions appears to be relevant to predicting the consequences of future bail-out decisions. What is the mechanism that allows such reasoning? How does counterfactual reasoning help rational agents form beliefs and make predictions?

In this note we extend a model of belief formation to encompass counterfactuals. In the context of this model, we can ask how counterfactuals can be used for prediction. Specifically, suppose that history provides compelling evidence for the truthfulness of a counterfactual belief. In this case, this belief can be added to the set of observations, as if it were actually experienced, and thereafter used for further predictions. Our central result is that a logically omniscient agent finds no benefit in using this type of counterfactual reasoning to make predictions.

The next subsection is devoted to a clearer definition of the problems we are concerned with. It is followed by a brief description of some fields of study that deal with counterfactuals, attempting to place this note in the context of the literature. We then proceed, in Section 2, to present our model, building on the unified model of induction presented in Gilboa, Samuelson, and Schmeidler (2010). Section 3 deals with prediction with the

aid of counterfactuals. The centerpiece of this section is the aforementioned impossibility result, showing that counterfactuals cannot add anything to a logically omniscient agent's predictive ability. Section 4 is devoted to a general discussion, and, in particular, to the extent to which counterfactuals can be valuable in enhancing the prediction of boundedly rational agents.

1.1 Modeling Counterfactuals

Our first task is to construct a model that will integrate counterfactuals with ordinary representations of non-counterfactual belief formation. Not all counterfactuals are created equal in this respect. It is helpful to consider three examples:

1. What would have happened had a person put her hand in the fire for several seconds?

2. What would have happened had President Bush, in September 2008, decided to save Lehman Brothers from bankruptcy?

3. What would have happened had the force of gravity not existed?[1]

All three counterfactual questions are similar in form and logical structure. But they differ in terms of our ability to reason about them. Question 1 is the simplest. Assuming that the person in question did not put her hand in the fire, we have no direct empirical evidence of what would have happened had she done so. But we have plenty of evidence regarding similar cases, as well as a rather good understanding of the underlying rules and mechanisms involved, so that we have no difficulty answering this question. This is the type of counterfactual question we routinely resolve when identifying the consequences of decisions in repeated, known contexts.

Question 3 is quite different. We have no empirical evidence from worlds remotely similar to ours without gravity. Everything we know in the natural sciences would have to be re-evaluated in order to answer this question. Question 3 is so difficult to reason about, that it does not pose a serious theoretical problem: the only reasonable answer is that we don't have any idea what would have happened in this case.

Question 2 represents an intermediate case. Like Question 3, we cannot claim to have a large database of similar situations whose outcomes were

[1] Questions 1 and 3 are classical examples given by Hume (1748).

actually observed. The financial crisis of 2008 is a unique event in history. The crisis of 1929 may have many similar features, but the two are not identical, and other crises take us yet further afield. Moreover, financial crises are global events that cannot be isolated and studied independently of each other: the very fact that the crisis of 1929 occurred before that of 2008 had an effect on the course of the latter. To complicate things further, it is doubtful that we have figured out the rules that govern the behavior of the world economy to the same extent that we have understood the laws that govern fire and its effect on the human body. Question 2 thus cannot be answered with scientific certainty as can Question 1. Nonetheless, Question 2 is not a matter of science-fiction speculation as is Question 3. We have some ways of reasoning about the effects of financial bail-out decisions, and often allude to counterfactual bail-out cases when debating current policy.

We seek a model that can describe the generation of counterfactual beliefs, preferably in a way that is akin to the generation of non-counterfactual beliefs. As a test of reasonability, we would like the model to show how, with a reasonable choice of parameters, one gets (i) a more or less unique, deterministic answer to Question 1; (ii) complete ignorance when it comes to Question 3; and (iii) some intuitive though speculative reasoning about Question 2. Moreover, we would like the model to be able to describe how counterfactual beliefs, such as the answer to Question 2, might be updated as (factual) history unfolds, so that an agent might feel more or less confident about counterfactual statements given information that has been gathered while these statements were already known to be counter-factual.

We are then particularly interested in the way counterfactuals are used for prediction. There are situations in which counterfactuals appear to be useless, either because their consequents are too unclear, or because they do not add much to existing knowledge. For instance, Question 3 leaves too much room for speculation to be useful in prediction.[2] On the other hand, Question 1 is easy to answer, but such an answer is likely to simply confirm predictions that already follow from factual observations. By contrast, Question 2 appears to be more interesting: it seems to lie in the middle ground where it is sufficiently familiar so that something can be said about it, and yet sufficiently novel so that reasoning about it would teach us something new.

[2]Fortunately, Question 3 also does not seem to be relevant to any practical problem. We suspect that the fact that nothing can be said about it is related to its irrelevance.

1.2 Related Literature

Counterfactual reasoning comes in many forms, and it has been studied in different disciplines. The following is but a brief survey, highlighting the way our study merges with or differs from other approaches.

Philosophy. Starting with the work of Stalnaker (1968) and Lewis (1973), philosophers and logicians have studied the logic of counterfactuals, distinguished among types of counterfactuals, and considered their semantics. The formal model we present here employs a state space and is thus semantic in nature. In contrast to the philosophy literature, however, we do not insist on a syntactic model of counterfactuals. Our focus is on the process by which counterfactuals help form beliefs. Finally, we deal only with one type of counterfactual, namely, with the beliefs one has at the present about the evolution of history along paths not taken.

Decision theory. The point of departure for decision theory is a function that maps combinations of acts and states into consequences. This function is typically taken to be so obviously basic as to be passed over without further notice. In practice, much of the work in making a decision (or in offering expert advice in support of a decision) revolves around identifying these consequences. Doing so requires counterfactual reasoning. When crossing a road one engages in reasoning by conditional statements of the type, "If I cross the road in front of the car, I will get hit", as well as "If I wait to cross until after the car has passed, I will be safe" and so forth. Once the decision has been made, all but one of such statements will linger in the agent's mind as counterfactuals. As is common in the philosophical literature, we will often refer to such statements as "counterfactuals" even before the truth value of their antecedent has been determined.

Any model of reasoning about decision making is thus also a model of reasoning about counterfactuals. Indeed, it is here that we gain most of our intuition about counterfactuals. Our suspicion that answers to Question 1 (in Section 1.1) will be useful while those to Question 3 will be useless arises out of thinking about how these questions will help us make decisions. However, our interest in counterfactuals is not motivated by the observation that people often remember the reasoning behind their decisions. Instead, we are motivated by the observation that counterfactual beliefs are often revised as additional observations are gathered, even after the antecedent is known

to be false, and then used as inputs for subsequent belief formation. Thus, one might say, "Given my experience in the past 20 years, I believe that, had I chosen a different career, I would have been better off", and may utter this statement in the midst of advice about current choices. Put differently, one of the special features of counterfactuals is that our beliefs in them are continually updated and revised in light of new information, even after the conditional statement has been classified as *counter*-factual.

Psychology. Psychological studies suggest that counterfactual beliefs can have a significant impact on the way that actual outcomes are evaluated, and on the resulting affective reactions. In particular, the salience of alternative scenarios can play a role in the evaluation of actual ones. For example, Medvec, Maday, and Gilovich (1995) argue that Olympic bronze medal winners tend to be more pleased with their outcome than silver medal winners, because for the former the salient alternative consequence may be not to get a medal at all, whereas for the latter the salient alternative is often the gold medal. Specifically, the winner of a silver medal may engage in thinking along the lines of "Had I only done... I would have won the gold", whereas such counterfactual thoughts are less likely to burden the bronze medalist. Our focus is not on the emotional implications of counterfactual reasoning. Rather, we focus on the cognitive aspects, namely, how counterfactual reasoning is conducted and used.

Statistics. Statisticians often encounter a problem of missing data. Suppose there are multiple observations of variables $\{X_i^1, ..., X_i^m, Y_i\}$, but certain variables haven't been measured in certain observations. Restricting attention to observations for which all variables have been measured wastes some of the information in the data, while working with all of the information gives rise to a collection of missing-variables difficulties. In response, missing data are "filled in" using techniques such as kernel estimation, and are then used for further analysis.

Counterfactuals play a similar role in forming beliefs. Indeed, one may argue that, broadly construed, each problem can embed the other. A missing datum could be viewed as an answer to the question "what would we observe if we were to measure that which we didn't?", and so questions about missing data can be couched as counterfactuals. Conversely, any counterfactual could also be viewed as an observation that one would have liked to have but

doesn't. Specifically, we might observe the outcome of act a and wonder what would have resulted from other acts $b, c...$ Each such act could be viewed as another observation, with different X values and with an unobserved Y value.

Despite the formal equivalence between "filling in" missing data and counterfactual reasoning, our focus here is quite different than that usually encountered in statistics. We are interested in cases in which an observation is counterfactual because the variables $\{X_i^1, \ldots, X_i^m\}$ did not occur, and therefore Y_i was not observed. It is rarely the case in statistics that the "outcome", namely the dependent variable Y, is being conjectured for values of the dependent variables that also do not appear in the data.

History. Counterfactuals are essential to the study of history. For example, consider the statement, "If General McClellan had pursued his advantage at Antietem, the American Civil War have ended a year earlier". Professionals as well as laypeople analyze history by comparing actual scenarios to counterfactual ones, pointing to possible causal relationships, which are, in turn, used to learn from historical events and to make predictions. Yet, the use and interpretation of such counterfactual statements is controversial (cf. Bunzl (2004)). The model we present below is most closely related to this use of counterfactuals, though it is clearly too theoretical to make a substantive contribution to the historical debate. Our model provides a theoretical framework for examining how people use counterfactuals, including (but not limited to) historical counterfactuals, to form beliefs and make predictions about likely outcomes in their current situation.

2 The Framework

2.1 The Unified Model

We adopt the unified model of induction of Gilboa, Samuelson, and Schmeidler (2010).[3] In each period, an agent makes predictions about the value of a variable y based on some observations x. She has a history of observations of past x and y values to rely on. We make no assumptions about

[3]We work with a special case of Gilboa, Samuelson, and Schmeidler's (2010) model that allows us to make the argument with a minimum of technical clutter. We present the model here, leaving most issues of motivation and interpretation to the original paper.

independence or conditional independence of the variables across periods, or any other assumption about the data generating process.

Let the set of periods be $\mathbb{T} \equiv \{0, 1, 2, \ldots, T\}$. At each period $t \in \mathbb{T}$ there is a *characteristic* $x_t \in X$ and an *outcome* $y_t \in Y$. The sets X and Y are finite and non-empty.[4] The set of all *states of the world* is

$$\Omega = \{\omega : \mathbb{T} \to X \times Y\}.$$

For a state ω and a period t, let $\omega(t) = (\omega_x(t), \omega_y(t))$ denote the element of $X \times Y$ appearing in period t. Let

$$h_t(\omega) = (\omega(0), \ldots, \omega(t-1), \omega_x(t))$$

denote the history of characteristics and outcomes in periods 0 through $t-1$, along with the period-t characteristic, given state ω.

For a history h_t, define

$$[h_t] = \{\omega \in \Omega \mid (\omega(0), \ldots, \omega(t-1), \omega_x(t)) = h_t\}.$$

Thus, $[h_t]$ is the event consisting of all states that are compatible with the history h_t. Similarly, for h_t and a subset of outcomes $Y' \subset Y$, we define the event

$$[h_t, Y'] = \{\omega \in [h_t] \mid \omega_y(t) \subset Y'\},$$

consisting of all states that are compatible with the history h_t and with the next outcome being in the set Y'.

In each period $t \in \mathbb{T}$, the agent observes a history h_t and makes predictions about the period-t outcome, $\omega_y(t) \in Y$. A *prediction* is a ranking of subsets in Y given h_t.

Predictions are made with the help of hypotheses. A *hypothesis* is an event $A \subset \Omega$. A hypothesis can represent a theory, an association rule, an analogy, or in general any reasoning aid one may employ in predicting y_t. Indeed, any such reasoning tool can be described extensively, by the set of states that are compatible with it. Let $\mathcal{A}$ denote the set of all hypotheses, and so $\mathcal{A} = 2^{\Omega}$.

The agent makes use of these hypotheses with the help of a model. Formally, a *model* is a function $\phi : \mathcal{A} \to \mathbb{R}_+$, where $\phi(A)$ is interpreted as the

[4] No conceptual problems arise in extending the analysis to infinite sets X, Y or $\mathbb{T}$, but we avoid a collection of technical complications by working with finite sets.

weight attached to hypothesis A for the purpose of prediction. The function ϕ is extended to subsets of hypotheses additively.

Given a history h_t, a hypothesis A that is disjoint from $[h_t]$ (i.e., a hypothesis that has been refuted by h_t) should not be taken into consideration in future predictions. Fixing a history h_t and a subset of outcomes $Y' \subset Y$, the set of hypotheses in $\mathcal{A}$ that have not been refuted by h_t and that predict the outcome will be in Y' is:

$$\mathcal{A}(h_t, Y') = \{A \in \mathcal{A} \mid \varnothing \neq A \cap [h_t] \subset [h_t, Y']\}. \tag{1}$$

Observe that the hypotheses in $\mathcal{A}(h_t, Y')$ are various events, many pairs of which may not be disjoint.

Given a model $\phi : \mathcal{A} \to \mathbb{R}_+$, the total weight assigned to the hypotheses that are unrefuted by h_t and consistent with an outcome in Y' is thus given by

$$\phi(\mathcal{A}(h_t, Y')) = \sum_{A \in \phi(\mathcal{A}(h_t, Y'))} \phi(A).$$

The agent's prediction is then a ranking of the subsets of Y, with Y' considered more likely than Y'' if

$$\phi(\mathcal{A}(h_t, Y')) > \phi(\mathcal{A}(h_t, Y'')).$$

2.2 Counterfactual Beliefs

We now extend the unified model to capture counterfactual beliefs. Assume that history h_t has materialized, but the agent wonders what would happen at a different history, $h'_{t'}$. We focus on the case

$$[h_t] \cap [h'_{t'}] = \varnothing$$

in which, at h_t, $h'_{t'}$ is indeed counter-factual.[5]

If the agent were at $h'_{t'}$, she would simply apply (1) to identify the hypotheses consistent with $[h'_{t'}]$. But the agent is not actually at the history $h'_{t'}$: she has observed h_t, and should take this latter information into account. Hence, the agent should consider only those hypotheses that are compatible with h_t, namely, only those A's such that $A \cap [h_t] \neq \varnothing$. Therefore, the belief

[5] We do not distinguish in the formal model between the questions "what would happen if... were not the case" and "what would have happened if... had not been the case".

in outcomes $Y'' \subsetneq Y$ resulting from history $h'_{t'}$ *conditional on history* h_t is $\phi(\mathcal{A}(h'_{t'}, Y''|h_t))$, with

$$\mathcal{A}(h'_{t'}, Y''|h_t) = \left\{ A \in \mathcal{A} \;\middle|\; \begin{array}{c} A \cap [h_t], A \cap [h'_{t'}] \neq \varnothing \\ A \cap [h'_{t'}] \subset [h'_{t'}, Y''] \end{array} \right\}. \tag{2}$$

If it is the case that $[h_t] \cap [h'_{t'}] = \varnothing$ these beliefs will be referred to as *counterfactual*.[6] Observe that the hypotheses in $\mathcal{A}(h'_{t'}, Y''|h_t)$ are required to have a non-empty intersection with $[h_t]$ and with $[h'_{t'}]$ separately, but not with their intersection. Indeed, in the case of counterfactual conditional beliefs this intersection is empty.

Let us see how the definition given above captures intuitive reasoning in Questions 1-3 in the Introduction. Begin with Question 1, namely, what would happen to an agent who were to put her hand in the fire. The agent has not done so, and thus h_t specifies the choice to refrain from the dangerous act. However, when the agent (or at outside observer) contemplates a different history, $h'_{t'}$, in which the hand were indeed put in the fire, there are many hypotheses that suggest that the hand would burn. One such hypothesis is the general rule "objects put in the fire burn", which presumably received a positive ϕ value at the outset and has not be refuted since.[7] There are also many case-based hypotheses, each of which suggesting an analogy between the present case and a particular past case. Since in all past cases hands put in fires burned, each of these hypotheses suggests that this would be the outcome in the present case as well. In short, there is plenty of evidence about Question 1, captured in this framework both as general rules and as specific analogies, where practically all of them suggest the natural answer.

Consider now Question 3. What would have happened were gravity not to hold? There are many possible rules one can conjecture in this context, such as "without gravity no atoms would have existed" or "without gravity, only light atoms would have existed". However, in contrast to the rule "objects put in fire burn", none of these rules has been tested in the past, and they are all vacuously unrefuted. Thus, all of the conceivable rules remain with their original (and arbitrary) ϕ value, without the empirical mechanism allowing us to sift through the multitude of rules and find the unrefuted ones. Clearly,

[6] If $[h_t] \cap [h'_{t'}] \neq \varnothing$, then either h_t and $h'_{t'}$ are identical, or one is prefix of the other. If h_t is a prefix of $h'_{t'}$, then $\mathcal{A}(h'_{t'}, Y''|h_t) = \mathcal{A}(h'_{t'}, Y'')$, while the reverse inclusion gives $\mathcal{A}(h'_{t'}, Y''|h_t) = \mathcal{A}(h_t, Y'')$.

[7] As in Gilboa, Samuelson, and Schmeidler (2010), we do not deal here with probabilistic rules, though such an extension would obviously make the model more realistic.

in this question analogical reasoning will be of no help as well. The history we observed consists only of cases in which gravity held. In this sense, all these cases are dramatically different from the hypothetical case in which gravity does not hold. Thus, a reasonable analogical reasoning would suggest that there is no similarity between the past and hypothetical cases to be able to generate a meaningful belief.

Finally, we turn to the interesting case of Question 2. In September 2008 the US government decided not to bail out Lehman Brothers. At that point, the actual history h_t and the hypothetical one, in which the government decided otherwise, $h'_{t'}$, part forever: $[h_t] \cap [h'_{t'}] = \varnothing$. Yet, there are hypotheses A that are compatible with both, that is, that satisfy $A \cap [h_t], A \cap [h'_{t'}] \neq \varnothing$. One such hypothesis may be the rule "When the government bails out all large financial institutions confidence in the market is restored". Let us assume, for the sake of the argument, that such a rule is well-defined and holds in the observed history h_t. In this case, this rule will predict that, at $h'_{t'}$, confidence in the market will be restored. Alternatively, one may point to a rule that says "The government bails out a small number of institutions, and thereafter begins a crisis", predicting that a bail-out would not have averted the crisis. Along similar lines, one may also use analogical reasoning to generate the belief given $h'_{t'}$. For example, one case-based hypothesis holds that the problem of 2008 is similar to that of the previous year, and had the US government bailed out Lehman brothers, as it bailed out mortgage banks in 2007, the crisis would have been averted, as it was in 2007. Similarly, one might cite other cases in which a bailout did not avert a crisis.

Thus, counterfactual beliefs are generated by considering hypotheses that are simultaneously consistent with the observed and with the counterfactual history. In Question 1, practically all such hypotheses point to the natural conclusion: were the hand put in fire, it would burn. In our notation, $\phi(\mathcal{A}(h'_{t'}, \{noburn\}|h_t)) = 0$ whereas $\phi(\mathcal{A}(h'_{t'}, \{burn\}|h_t)) > 0$.

In Question 3, there are no useful hypotheses to consult: no similar cases are known, and, relatedly, none of the conceivable rules one might imagine has been tested. Thus, the weight $\phi(\mathcal{A}(h'_{t'}, \{y\}|h_t))$ would reasonably be the same for any prediction y. (Indeed, it might be most reasonable to have a function ϕ for which this weight is zero.)

By contrast, in Question 2, there are hypotheses with positive weights that have been tested in the actual history ($A \cap [h_t] \neq \varnothing$) and that make predictions at the counterfactual history ($A \cap [h'_{t'}] \neq \varnothing$). Some of them suggest that a bail-out would have averted the crisis, some suggest the opposite.

The relative weight assigned to these classes of hypotheses would determine the counterfactual belief.

Observe that our model can also explain how the belief in a counterfactual conditional statement changes as new evidence is gathered, even after the statement's antecedent is known to be false. For example, assume that John is about to take an exam, and decides to study rather than party. Having observed his choice, we may not know how likely it is that he would have passed the exam, had he decided to party. But if we get the new piece of information that he failed the exam, we are more likely to believe that he would have failed, had he not studied. In our model, this would be reflected by the addition of a new observation to the factual history h_t, which rules out certain hypotheses and thereby changes the evaluation of the counterfactual at $h'_{t'}$.

2.3 Bayesian Counterfactuals

Gilboa, Samuelson, and Schmeidler (2010) define the set of *Bayesian hypotheses* to be

$$\mathcal{B} = \{\{\omega\} \,|\omega \in \Omega\} \subset \mathcal{A}.$$

Each of the Bayesian hypotheses fully specifies a single state of the world. A Bayesian agent will satisfy

$$\phi(\mathcal{A} \backslash \mathcal{B}) = 0,$$

that is,

$$\phi(A) = 0 \qquad \text{if} \qquad |A| > 1.$$

As discussed in Gilboa, Samuelson, and Schmeidler (2010), this reflects the Bayesian commitment not to leave any uncertainty unquantified. A Bayesian agent who expresses some credence in a hypothesis (event) A, should take a stance on how this event would occur, dividing all the weight of credence in A among its constituent states.

The following is immediate (cf. (2)) but worthy of note.

Observation 1 *If* $\phi(\mathcal{A} \backslash \mathcal{B}) = 0$ *then, whenever* $[h_t] \cap [h'_{t'}] = \varnothing$

$$\phi(\mathcal{A}(h'_{t'}, Y'' \,|\, h_t)) = 0$$

for all $Y'' \subset Y$.

Thus, a Bayesian agent has nothing to say about counterfactual questions. This result is obvious because a Bayesian agents assigns positive weight only to singletons, that is, to hypotheses of the type $A = \{\omega\}$, and no such hypothesis can simultaneously be consistent with both h_t and $h'_{t'}$. Hence, the history that has happened, h_t, rules out any hypothesis that could have helped one reason about the history that didn't happen, $h'_{t'}$. Intuitively, this is so because the Bayesian approach does not describe how beliefs are formed, by reasoning over various hypotheses. Rather, it presents only the bottom line, that is, the precise probability of each state. In the absence of the background reasoning, this approach provides no hint as to what could have resulted from an alternative history. Indeed, Bayesian accounts of counterfactuals either dismiss them as meaningless, or resort to additional constructions, such as lexicographic probabilities.

3 Counterfactual Predictions

We now ask how counterfactuals can help make predictions, essentially by adding information to the agent's database.

Imagine an agent has observed history h_t. In the absence of counterfactuals, she would make predictions by comparing weights of credence $\phi(\mathcal{A}(h_t, Y'))$, for various values of Y'. Now suppose she endeavors to supplement the information at her disposal by asking, counterfactually, what would have happened at history $h'_{t'}$, where $[h_t] \cap [h'_{t'}] = \varnothing$.

The agent first uses her counterfactual beliefs to associate a set of outcomes Y'' to the counterfactual history $h'_{t'}$. She then adds the counterfactual information $[h'_{t'}, Y'']$ to her data set. This counterfactual information may allow her to discard some hypotheses from consideration, thereby sharpening her predictions.

What set of outcomes Y'' should she associate with history $h'_{t'}$? To consider an extreme case, suppose that $\mathcal{A}(h'_{t'}, Y'' \,|\, h_t)$ is nonempty only for $Y'' = \{y_0\}$. Thus, the agent is certain that, had $h'_{t'}$ been the case, y_0 would have resulted. The counterfactual question posed by $h'_{t'}|h_t$ is then analogous to Question 1 in Section 1.1, with an obvious answer. In this case, she can add the hypothetical observation $[h'_{t'}, \{y_0\}]$ to her database, and continue to generate predictions based on the extended database, as if this observation had indeed been witnessed. This "extended database" cannot be described by a history, because no history can simultaneously describe the data in h_t

and in $h'_{t'}$ (recall that $[h_t] \cap [h'_{t'}] = \varnothing$). However, the agent can use both the actual history h_t and the hypothetical observation $[h'_{t'}, \{y_0\}]$ to rule out hypotheses and sharpen future prediction.

More generally, assume that the conditional beliefs $\phi(\mathcal{A}(h'_{t'}, Y'' \,|\, h_t))$ are positive only for a subset of outcomes $Y_0 \subset Y$ and subsets thereof, i.e.,

$$\phi(\mathcal{A}(h'_{t'}, Y_0|h_t)) > 0 \tag{3}$$

$$\phi(\mathcal{A}(h'_{t'}, Y''|h_t)) > 0 \;\; \Rightarrow \;\; Y'' \subset Y_0, \tag{4}$$

so that the agent is absolutely sure that, had $h'_{t'}$ materialized, the outcome would have been in Y_0. Thus, no other subset of Y competes with outcomes in Y_0 for the title "the set of outcomes that could have resulted had $h'_{t'}$ been the case". We are then dealing with a counterfactual analogous to question 2 in Section 1.1) (with the previous paragraph dealing with the special case in which $Y_0 = \{y_0\}$). In this case the agent adds to the database the hypothetical observation that $h'_{t'}$ results in an outcome in Y_0.

Now the agent uses the information that history h_t has occurred, and the counterfactual information that history $h'_{t'}$ would have resulted in an outcome from Y_0, to winnow the set of hypotheses to be used in prediction. In particular, the hypotheses used the the agent include:

- All hypotheses that are consistent with h_t but not with $h'_{t'}$. Indeed, since $h'_{t'}$ did not materialize, it cannot make a claim, as it were, to rule out hypotheses that are consistent with observations.

- All hypotheses that are consistent with each of h_t and $h'_{t'}$, provided that they are consistent with the counterfactual prediction Y_0 (satisfying (3)–(4)).

In other words, define the new set of hypotheses relevant for evaluating the set of outcomes Y' at history h_t, given counterfactual information $[h'_{t'}]$, to be

$$\mathcal{A}(h_t, Y'|h'_{t'}, Y_0) = \left\{ A \in \mathcal{A} \;\middle|\; \begin{array}{c} \varnothing \neq A \cap [h_t] \subset [h_t, Y'] \\ A \cap [h'_{t'}] \subset [h'_{t'}, Y_0] \end{array} \right\}. \tag{5}$$

The agent then uses ϕ to rank the sets $\mathcal{A}(h_t, Y'|h'_{t'}, Y_0)$, for various values of Y', and then to make predictions.[8]

[8] We have added the result of a single counterfactual consideration to the reasoner's database. Adding multiple counterfactuals is a straightforward elaboration.

Our model allows us to consider agents who are not Bayesian, but are nonetheless rational. This is important, as Observation 1 ensures that there is no point in talking about counterfactual predictions made by Bayesians. Indeed, we view the model as incorporating the two essential hallmarks of rationality: the consideration of all states of the world, capturing beliefs by a comprehensive, a priori model ϕ containing all the information available to the agent, and the drawing of subsequent inferences by deleting falsified hypotheses. An agent who is rational in this sense need not be Bayesian, which is to say that the agent need not consider only singleton hypotheses. In this case, counterfactuals are potentially valuable in making predictions.

Our result is that counterfactual reasoning adds nothing to prediction:

Proposition 1 *Assume that* $[h_t] \cap [h'_{t'}] = \varnothing$ *and that* Y_0 *satisfies (3)–(4). Then, for every* $Y' \subset Y$*,*

$$\phi(\mathcal{A}(h_t, Y')) = \phi(\mathcal{A}(h_t, Y'|h'_{t'}, Y_0)). \tag{6}$$

Predictions made without the counterfactual information (governed by $\phi(\mathcal{A}(h_t, Y'))$ *thus match those made with the counterfactual information (governed by* $\phi(\mathcal{A}(h_t, Y'|h'_{t'}, Y_0))$*.*

Thus, the counterfactual information has no effect on prediction. The (immediate) proof of this result consists in observing that, for Y_0 to include all possible predictions at $h'_{t'}$, it has to be the case that, among the hypotheses consistent with h_t, the only ones that have a positive ϕ value are those that are anyway in $\mathcal{A}(h_t, Y'|h'_{t'}, Y_0)$.[9]

This result has a flavor of a "cut-elimination" theorem (Gentzen, 1934-5):[10] it basically says that, if a certain claim can be established with certainty, and thereby be used for the proof of further claims, then one may also skip the explicit statement of the claim, and use the same propositions that could be used to prove it to directly deduce whatever could follow from the unstated claim. Clearly, the models are different, as the cut-elimination theorem deals with formal proofs, explicitly modeling propositions and logical steps, whereas our model is semantic, and deals only with states of the

[9]Formally, it is obvious that $\mathcal{A}(h_t, Y'|h'_{t'}, Y_0) \subset \mathcal{A}(h_t, Y')$, since the first condition in the definition of $\mathcal{A}(h_t, Y'|h'_{t'}, Y_0)$ is precisely the definition of $\mathcal{A}(h_t, Y')$. Suppose the hypothesis A is in $\mathcal{A}(h_t, Y')$ but not in $\mathcal{A}(h_t, Y'|h'_{t'}, Y_0) \subset \mathcal{A}(h_t, Y')$. Then, from (5), it must be that $A \cap [h'_{t'}]$ is not a subset of $[h'_{t'}, Y_0]$. But then, from (3)–(4), it must be that $\phi(A) = 0$.

[10]We thank Brian Hill for this observation.

world and the events that do or do not include them. Yet, the similarity in the logic of the results suggests that Proposition 1 may be significantly generalized to different models of inference.

4 Discussion

4.1 Why do Counterfactuals Exist?

Proposition 1 suggests that counterfactuals are of no use in making predictions, and hence for making better decisions. At the same time, we find counterfactual reasoning everywhere. Why do counterfactuals exist? We can suggest three reasons.

Lingering decisions. Section 1.2 noted that counterfactuals are an essential part of connecting acts to consequences, and hence in making decisions. The counterfactuals we encounter may simply be recollections of this prediction process, associated with past decisions. Before the agent knew whether h_t or $h'_{t'}$ would materialize, it was not only perfectly legitimate but necessary for her to engage in predicting the consequences of each possible history. Moreover, if the distinction between h_t and $h'_{t'}$ depends on the agent's own actions, then it would behoove her to think how each history would evolve (at least if she has any hope to qualify as rational). Thus, the agent would have engaged in predicting outcomes of both h_t and $h'_{t'}$, using various hypotheses. Once h_t is known to be the case, hypotheses consistent with both histories may well still be vivid in the agent's mind, generating counterfactual beliefs. According to this view, counterfactual beliefs are of no use; they are simply left-overs from previous reasoning, and they might just as well fade away from memory and make room for more useful speculations.

New information. We assumed that counterfactual outcomes are "added" to the database of observations only when they are a logical implication of the agent's underlying model. However, one might exploit additional information to incorporate counterfactual observations even if they are not logical implications of the model ϕ. For example, as mentioned above, statisticians sometimes fill in missing data by kernel estimation. This practice relies on certain additional assumptions about the nature of the process generating the data. In other words, the agent who uses ϕ for her predictions may resort

to another model, $\hat{\phi}$, in order to reason about counterfactuals. The additional assumptions incorporated in the model $\hat{\phi}$ may not be justified, strictly speaking, but when data are scarce, such a practice may result in better predictions than more conservative approaches. In fact, our results suggest that such a practice may be useful precisely because it relies on additional assumptions.

It is, however, not clear that adding such "new information" is always rational. Casual observations suggest that people may support their political opinions with counterfactual predictions that match them. It is possible that they first reasoned about these counterfactuals and then deduced the necessary political implications from them. But it is also possible that some of these counterfactuals were filled in in a way that fits one's pre-determined political views. Our analysis suggests that the addition of new information to a database should be handled with care.

Bounded rationality. We presented a model of logically omniscient agents. While logical omniscience is a weaker rationality assumption than the standard assumptions of Bayesian decision theory, it is still a restrictive and often unrealistic assumption. Our agent must be able to conceive of all hypotheses at the outset of the reasoning process and capture all of the information she has about these hypotheses in the function ϕ. Nothing can surprise such an agent, and nothing can give her cause to change her model ϕ as a result of new observations. Given the vast number of hypotheses, this level of computational ability is hardly realistic, and it accordingly makes sense to consider agents who are imperfect in their cognitive abilities. For such an agent, a certain conjecture may come to mind only after a counterfactual prediction Y_0 at $h'_{t'}$ is explicitly made, and only then can the agent fill in some parts of the model ϕ. According to this account, counterfactual predictions are a step in the reasoning process, a preparation of the database in the hope that it would bring to mind new regularities.

In this bounded-rationality view, discussions about counterfactuals are essentially discussions about the appropriate specification of ϕ. An agent may well test a particular possibility for ϕ by examining its implications for counterfactual histories, leading to revisions of ϕ in some cases and enhanced confidence in others. The function ϕ lies at the heart of the prediction model, so that counterfactuals here are not only useful but perhaps vitally important to successful prediction. In a sense, this view of counterfactuals takes us back

to Savage (1954), who viewed the critical part of a learning process as the massaging of beliefs that goes into the formation of a prior belief, followed by the technically trivial process of Bayesian updating. The counterpart of this massaging in our model would be the formation of the function ϕ. Whereas in most models of rational agents this function simply springs into life, as if from divine inspiration, in practice it must come from somewhere, and counterfactuals may play a role in its creation.

4.2 Extension: Probabilistic Counterfactuals

The counterfactual predictions we discuss above are deterministic. It appears natural to extend the model to quantitative counterfactuals. In particular, if the credence weights $\phi(\mathcal{A}(h'_{t'}, Y'|h_t))$ happen to generate an additive measure (on sets of outcomes Y'), they can be normalized to obtain a probability on Y, generating probabilistic counterfactuals along the lines of "Had $h'_{t'}$ been the case, the result would have been $y \in Y$ with probability $p(y|h'_{t'}, h_t)$".

Probabilistic counterfactuals of this nature can also be used to enrich the database by hypothetical observations. Rather than claiming that one knows what would have been the outcome had $h'_{t'}$ occurred, one may admit that uncertainty about this outcome remains, and quantify this uncertainty using counterfactuals. Further, one may use the probability over the missing data to enhance future prediction. However, under reasonable assumptions, a result analogous to Proposition 1 would hold. For instance, if the agent makes predictions by taking the expected prediction given the various hypothetical observations, she will make the same probabilistic predictions as if she skipped the counterfactual reasoning step.

4.3 A Possible Application: Extensive Form Games

Consider an extensive form game with a choice of a strategy for each of the n players. Assume for simplicity that these are pure strategies, so that it is obvious when a deviation is encountered.[11] Should a rational player follow her prescribed strategy? This would depend on her beliefs about what the other players would do, should she indeed follow it, but also what they would

[11] When one considers mixed (or behavioral) strategies, one should also consider some statistical tests of the implied distributions in order to make sure that the selection of strategies constitutes a non-vacuous theory.

do if she were to deviate from her strategy. How would they reason about the game in face of a deviation?

For concreteness, assume that player I is supposed to play a at the first node of the game. This is part of an n-tuple of strategies whose induced play path is implicitly or explicitly assumed to be common belief among the players.[12] Player I might reason, "I should play a, because this move promises a certain payoff; if, by contrast, I were to play b, I would get ..."—namely, planning to play a, the player has to have beliefs about what would happen if she were to change her mind, at the last minute as it were, and play b instead.

This problem is related, formally and conceptually, to the question of counterfactuals. Since player I intends to play a, she expects this to be part of the unfolding history, and she knows that so do the others. However, she can still consider the alternative b, which would bring the play of the game to a node that is inconsistent with the "theory" provided by the n-tuple of strategies. Differently viewed, we might ask the player, after she played a, why she chose to do so. To provide a rational answer, the player should reason about what would have happened had she chosen to do otherwise. The answer to this counterfactual question is, presumably, precisely what the player had believed would have happened had she chosen b, before she actually made up her mind.

Our model suggests a way to derive counterfactual beliefs from the same mechanism that generates regular beliefs. For example, consider the backward induction solution in a perfect information game without ties. Assume that for each k there is a hypothesis A_k "All players play the backward induction solution in the last k stages of the game". These hypotheses may have positive ϕ values based on past plays of different games, perhaps with different players. Suppose that this ϕ is shared by all players.[13] For simplicity, assume also that these are the only hypotheses with positive ϕ values. At the beginning, all players believe the backward induction solution will be followed. Should a deviation occur, say, k stages from the end of the game, hypotheses A_l will be refuted for all $l \geq k$. But the deviation would leave $A_{k-1}, ..., A_1$ unrefuted. If the player uses these hypotheses for the counterfactual prediction, she would find that the backward induction solution

[12] See Aumann (1995), Samet (1996), Stalnaker (1996), Battigalli and Siniscalchi (1999).

[13] Such a model only involves beliefs about other players' behavior. To capture higher-order beliefs one has to augment the state space and introduce additional structure to model the hierarchy of beliefs.

would remain the only possible outcome of her deviation. Hence she would reason that she has nothing to benefit from such a deviation, and would not refute A_k. Note that other specifications of ϕ might not yield the backward induction solution. Importantly, the same method of reasoning that leads to the belief in the equilibrium path is also used for generating off-equilibrium, counterfactual beliefs, with the model providing a tool for expressing and evaluating these beliefs.

5 References

Aumann, Robert J. (1995), "Backward induction and common knowledge of rationality," *Games and Economic Behavior* 8, 6-19.

Battigalli, Pierpaolo and Marciano Siniscalchi (1999), "Hierarchies of conditional beliefs and interactive epistemology in dynamic games," *Journal of Economic Theory* 88, 188-230.

Bunzl, Martin (2004), "Counterfactual History: A User's Guide," *The American Historical Review* 109, 845–858.

Gentzen, Gerhard (1934-1935), "Untersuchungen Uber das logische Schliessen," *Mathematische Zeitschrift* 39, 405–431.

Gilboa, Itzhak, Larry Samuelson, and David Schmeidler (2010), "Dynamics of Induction in a Unified Model", mimeo.

Hume, David (1748). *An Enquiry Concerning Human Understanding* (Oxford: Clarendon Press).

Lewis, David (1973). *Counterfactuals* (Oxford: Blackwell Publishers).

Medvec, Victoria, Scott Madey and Thomas Gilovich (1995), "When Less is More: Counterfactual Thinking and Satisfaction Among Olympic Medalists," *Journal of Personality and Social Psychology* 69, 603–610.

Samet, Dov (1996), "Hypothetical knowledge and games with perfect information," *Games and Economic Behavior* 17, 230-251.

Savage, Leonard J. (1954). *The Foundation of Statistic*s (New York: John Wiley and Sons; Second Edition 1972, Dover).

Stalnaker, Robert (1968), "A Theory of Counterfactuals," in Nicholas Rescher, ed., *Studies in Logical Theory: American Philosophical Quarterly, Monograph 2* (Oxford: Blackwell Publishers), 98-112.

Stalnaker, Robert (1996), "Knowledge, belief and counterfactual reasoning in games," *Economics and Philosophy* 12, 133-163.