Counterfactual reasoning (philosophical aspects)—quantitative

Alan Hájek

In N. J. Smelser & P. B. Baltes, *International Encyclopedia of the Social and Behavioral Sciences*, Elsevier, pp. 2872-2874. 2002.

This article should be read in conjunction with its companion, "Counterfactual Reasoning (Philosophical Aspects)—Qualitative". Here, after a general introduction and historical overview, we emphasize the role of counterfactual reasoning within the quantitative frameworks of probability theory, decision theory, and game theory.

1. Counterfactuals

Counterfactuals are a species of conditionals. They are propositions or sentences, expressed by or equivalent to subjunctive conditionals of the form 'if it were the case that *A*, then it would be the case that *B*', or 'if it had been the case that *A*, then it would have been the case that *B*'; *A* is called the *antecedent*, and *B* the *consequent.* Counterfactual reasoning typically involves the entertaining of hypothetical states of affairs: the antecedent is believed or presumed to be false, or contrary-to-the-fact, but its truth is imagined or supposed. Counterfactual reasoning is thus a form of *modal* reasoning, kindred to reasoning about necessity or possibility, and in contrast to reasoning about the way things actually are.

The philosophical study of conditionals goes back at least as far as the Stoics of ancient Greece, although their systems of logic apparently did not accord the counterfactual any emphasis. The rise in interest in counterfactuals has been a rather recent phenomenon, as it started to become clear to philosophers that counterfactuals are implicated in a host of other important concepts—laws of nature, confirmation, causation, scientific explanation, knowledge, perception, dispositions, free action, etc. The significance of counterfactuals has also become increasingly appreciated in the

social sciences—by psychologists (e.g., in understanding emotions such as regret), by historians (e.g., in studying the incremental contribution by some commodity to economic growth), in the law (e.g., in apportioning responsibility), etc. However, a widely agreed upon analysis of counterfactuals has proved elusive. The so-called *problem of counterfactuals*, raised notably by Chisholm and by Goodman in the 1940's, is that of providing adequate (in particular, non-circular) truth conditions for the counterfactual.

Let '*A* œ→ *B*' denote a counterfactual with antecedent *A* and consequent *B*. Fixing the truth value of *A* and *B* clearly does not fix the truth value of *A* œ→ *B*. For example, 'if I were to drop this fragile glass, it would break' is true, 'if I were to drop this fragile glass, it would turn into a hamster' is false, yet the antecedents ('I drop this fragile glass') and consequents ('the glass breaks', 'the glass turns into a hamster') are all false. This rules out as an adequate analysis the material conditional *A* £ *B* (defined to be equivalent to 'not-A or B'), which is automatically true when *A* is false. According to *metalinguistic* approaches, *A* œ→ *B* is true if and only if *A*, together with suitable further premises (on one version, stating appropriate laws of nature and initial conditions), implies *B*. Mackie takes counterfactuals to be themselves elliptical presentations of such arguments, and thus lacking truth values altogether. The advent of possible worlds semantics for modal logics in the 1960's gave new direction to the analysis of counterfactuals. Truth conditions for them are proposed in classic works by Stalnaker (1968) and Lewis (1973).

Stalnaker motivates his account by Ramsey's recipe for counterfactual reasoning ("Ramsey's test"): First, add the antecedent (hypothetically) to your stock of beliefs; second, make the minimal adjustments required in the rest of your beliefs to maintain consistency; finally consider whether or not you then believe the consequent. Stalnaker's truth conditions are intended to apply generally to conditionals of the form 'if *A*, then *B*', which Stalnaker denotes '*A* > *B*', pragmatics distinguishing between indicative conditionals and counterfactuals. He assumes that for each world w there is an ordering

of worlds according to their similarity, or 'closeness', to w. Call a world at which *A* is true an *A-world*; the proposition *A* is regarded as the set of *A*-worlds. His guiding idea is that

A > *B* is true at a world w

if and only if

B is true at the closest *A*-world to w.

Lewis rejects Stalnaker's assumption that, for any given w and *A*, there is exactly one closest *A*-world to w. His truth conditions take as primitive the notion of comparative closeness of worlds:

$A \mathrel{\text{œ}}\!\!\rightarrow B$ is true at w

if and only if

some *AB*-world is closer to w than any $A\overline{B}$-world.

For both Stalnaker and Lewis, the counterfactual is vacuously true if the antecedent is impossible.

2. Probabilities of counterfactuals

One important line of research considers counterfactuals within the quantitative framework of probability theory. In particular, there has been much interest in the connection between *probabilities of conditionals* (of which counterfactuals are an important class) and *conditional probabilities*. Let P be your subjective probability function, so that $P(A \mathrel{\text{œ}}\!\!\rightarrow B)$ is your degree of belief in $A \mathrel{\text{œ}}\!\!\rightarrow B$. There are various reasons (linguistic intuitions, considerations stemming from Ramsey's test, etc.) for thinking that counterfactual reasoning is, or should be, subject to the following constraint:

(*) $P(A \mathrel{\text{œ}}\!\!\rightarrow B) = P(B|A)$ for all *A*, *B* in the domain of P such that $P(A) > 0$.

Stalnaker proposed a version of this identity, which can be shown to vindicate his truth conditions above, and to vitiate Lewis'. In response, Lewis (most famously in 1976)

proved a number of 'triviality results': under certain assumptions, the only classes of probability functions that can sustain (*) consist solely of members that take at most four values, and that are thus trivial. Since then, a number of authors have strengthened these results—see Eells and Skyrms (1994) for a collection of relevant articles.

Nonetheless, an identity somewhat in the spirit of (*) *does* hold for the Stalnaker conditional. Start with probability function P. Among other things, P assigns probabilities to individual worlds. Define a new probability function P_A as follows: for each world w, P_A shifts the probability that P assigns to w to the closest *A*-world to w. P_A is said to be derived from P by *imaging on A*. Some authors model counterfactual reasoning by imaging (eschewing conditional probabilities as in (*)). Lewis (1976) shows that there is an intimate connection between the Stalnaker conditional and imaging:

$$P(A > B) = P_A(B) \quad \text{(for all P, } A \text{ and } B\text{)}$$

The exploration of tenable variants of (*) is likely to continue to be a lively area of research.

3. Counterfactual reasoning in decision theory

The distinction between probabilities of conditionals and conditional probabilities is important for *decision theory*, a normative theory of how one's beliefs and desires in tandem determine what one should do. Such a theory typically combines an agent's utility function *u* and probability function P to give a figure of merit for each possible action, called the *expectation*, or *desirability* of that action. Let S_1, S_2, ..., S_n be a partition of possible states of the world. *Evidential decision theory*, as presented by Jeffrey (1966), measures the choiceworthiness of action A by a conditional probability-weighted average of the utilities of its possible outcomes (the conjunctions $A \,\&\, S_i$):

$$V(A) = \sum_i u(A \,\&\, S_i)P(S_i|A)$$

It thus prescribes that an agent act so as to maximize V. Various authors believe that evidential decision theory gives incorrect verdicts in Newcomb problem cases, in which an action is thought to be evidence for the obtaining of a given state, without in any way causing that state (see *Newcomb's problem*). This has led to the development of rival versions of decision theory, sharing the name *causal decision theory*, that replace the 'evidential' conditional probability weights with weights that are supposed to capture the agent's degrees of belief about the causal efficacy of the action in bringing about each possible state of affairs. Gibbard and Harper (1981), following Stalnaker, use probabilities of counterfactuals as the weights:

$$U(A) = \sum_i u(A \,\&\, S_i)P(A \mathrel{\square\!\!\rightarrow} S_i)$$

They then advocate the maximization of U. Proponents of causal decision theory believe that it agrees with evidential decision theory in unproblematic cases, and that it delivers intuitively correct answers in Newcomb problem cases.

4. Counterfactual reasoning in game theory

Counterfactual reasoning can play a crucial role in the rational deliberation of two or more agents strategically playing against each other, their pay-offs depending on what they all do—the province of *game theory*. Paralleling the distinction we found in decision theory, Stalnaker (1996) distinguishes two sorts of counterfactual reasoning, *causal* and *epistemic* (the latter could also be called *evidential).* In the former, agents reason about what would happen if they or others were to act in ways that they know or believe to be non-actual. In the latter, they consider how their own or others' beliefs would change if they were to learn things that they expect not to learn—for example, contrary to what is standardly taken to be common knowledge (and thus regarded as certain) by all the players, an irrational choice is hypothetically made. Indeed, such (epistemic) counterfactual reasoning leads naturally to some important refinements of Nash equilibrium, such as Selten's subgame perfect and ('trembling hand') perfect

equilibrium. [NOTE TO EDITORS: CROSS-REFERENCE TO ENTRY ON GAME THEORY HERE?]

Various models have been proposed of how an agent's beliefs are disposed to change in the face of unexpected actions by others, and the consequences of such dispositions for rational action—see, e.g., Bicchieri (1988) (for extensive form games), and Stalnaker (1996) (for normal form games). Closely related is Hyun Song Shin's (1989) theory of equilibrium in normal-form games. He gives a counterfactual-based account of rationality—a rational agent never acts in such a way that he would have done better had he acted differently—and proves that his account is equivalent, under certain assumptions, to those that employ game theory's standard solution concepts of Nash equilibrium and correlated equilibrium. Skyrms (1998) integrates the themes that we have discussed in this and the previous two sections. He gives a probabilistic theory of subjunctive conditionals that finds its inspiration in (*), with 'assertability value' replacing the 'P' on the left-hand side, and expectation of conditional chance replacing the right-hand side. He then uses this theory to give a unified treatment of counterfactual reasoning in individual decision theory, and in normal form and extensive form games.

Finally, in what might be taken as an indicator of a future trend in this field, more and more sophisticated probabilistic models are being developed for the representation of conditional beliefs in games. The economists Battigalli and Siniscalchi (1998), for example, construct a space of infinite hierarchies of conditional probability systems in which one can represent any statement about the players' dispositions to hold (arbitrarily high-order) beliefs regarding the other players.

Bibliography

Battigalli, Pierpaolo, and Marciano Siniscalchi (1998): "Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games", EUI Working Paper ECO No. 98/29, Badia Fiesolana, San Domenico.

Bicchieri, C. (1988): "Strategic Behavior and Counterfactuals", *Synthese* 76, 135-69.

Eells, Ellery and Brian Skyrms (eds.) (1994): *Probability and Conditionals*, Cambridge University Press.

Gibbard, Allan and William L. Harper (1981): "Counterfactuals and Two Kinds of Expected Utility", in *Ifs*, eds. W.L. Harper, R. Stalnaker, G. Pearce, Reidel.

Jeffrey, R. C. (1966): *The Logic of Decision*, Chicago University Press.

Lewis, D. K. (1973): *Counterfactuals*, Blackwell.

Lewis, David (1976): "Probabilities of Conditionals and Conditional Probabilities", *Philosophical Review* 85, 297-315

Shin, H.S. (1989): "Two Notions of Ratifiability and Equilibrium in Games", in M. Bacharach and S. Hurley (eds.), *Foundations of Decision Theory*, Blackwell.

Skyrms, Brian (1998): "Subjunctive Conditionals and Revealed Preference", *Philosophy of Science* 65, 545-574.

Stalnaker, R. (1968): "A Theory of Conditionals", in N. Rescher (ed.), *Studies in Logical Theory*, Blackwell.

Stalnaker, R. (1996): "Knowledge, Belief and Counterfactual Reasoning in Games", *Economics and Philosophy* 12, 133-163.