Draft of paper to appear in Duncan Pritchard and Matthew Jope (eds.), *New Perspectives on Epistemic Closure*, London: Routledge.

Modal Epistemology and the Logic of Counterfactuals

Timothy Williamson

Modal epistemology, as understood here, is epistemology done using modal conditions to characterize epistemic states: for example, a modal connection between belief and truth to characterize actual knowledge. By contrast, the *epistemology of modality* is the epistemology of epistemic states whose content concerns modal matters: for example, knowledge that something is necessary or possible. Modal epistemology in the tradition descending from Fred Dretske (1970) and Robert Nozick (1981) has focussed on modal connections expressed by counterfactual conditions, such as *Sensitivity*: if the proposition weren't true, you wouldn't believe it (the simplest version). Sensitivity to a truth in that sense has been claimed necessary for knowing that truth.

Notoriously, such counterfactual conditions on knowledge seem to undermine plausible principles of *epistemic closure*: sometimes you know each premise of an argument, and believe the conclusion by competent deduction from the premises, yet you fail to *know* the conclusion, for although you are sensitive to each premise, you are not sensitive to the conclusion, so the alleged necessary condition for knowledge fails. Classic accounts of the semantics of counterfactual conditionals, in particular those of Robert Stalnaker (1968) and David Lewis (1973), vindicate such apparent counterexamples to epistemic closure, in a way intended by proponents of the sensitivity condition.

The contrapositive of the sensitivity condition has also been used to define a proposed modal condition on knowledge, *Counterfactual Safety*: if you believed the proposition, it would be true (again, the simplest version). Sensitivity and Counterfactual Safety are usually thought to work in quite different ways.

Section 1 discusses in more detail the relation between the logic and semantics of counterfactuals on one hand and counterexamples to epistemic closure on the other. Section 2 expounds a more natural approach to the logic and semantics counterfactual conditionals. Section 3 explains why the alleged counterexamples to epistemic closure are illusory on this alternative approach, and how the illusions arise. It also explains why Sensitivity and Counterfactual Safety are in principle equivalent. Section 4 formulates closure principles for those conditions validated by the alternative semantics. Section 5 concludes with a methodological moral.

1. *The standard picture*

One is sensitive to *A* just in case if *A* were not true, one would not believe *A*. More complex versions of sensitivity qualify 'believes' with 'by method M' or 'on basis B' or the like, and

even variation in the proposition believed. The argument of this chapter can be adjusted to allow for such qualifications; for simplicity, they will usually be omitted.

We formalize the operator 'one believes that …' as Bel, negation as ¬, and the counterfactual conditional with antecedent *A* and consequent *C* as $A > C$ ('If *A* were true, *C* would be true'). Thus 'one is sensitive to *A*' becomes $\neg A > \neg \mathrm{Bel}[A]$. For convenience, '*A*' takes sentence position in formulas but name position in their English paraphrases, where it acts as a (variable) name of the proposition expressed by the corresponding sentence.

Stalnaker and Lewis formulate their semantic accounts of counterfactual conditionals within the framework of possible worlds semantics. Sentences are true or false at (possible) worlds; an *A-world* is a world at which *A* is true. A counterfactual conditional $A > C$ is true at a world *w* if and only if the *A*-worlds closest to *w* are *C*-worlds. If there are no *A*-worlds (*A* is impossible), $A > C$ is true vacuously. Stalnaker stipulates that there is always a unique closest *A*-world; Lewis does not. Closeness is similarity on the relevant dimension; much can be said as to what that dimension is, but for present purposes we take it as vaguely understood. Thus, putting the pieces together, one is sensitive to *A* if and only if, at the closest worlds at which *A* is not true, one does not believe *A*.

According to the principle of *Strengthening the Antecedent*, if A^+ entails *A*, then $A > C$ entails $A^+ > C$. For example, since 'Exactly forty people were at the party' entails 'At least forty people were at the party', 'If at least forty people were at the party, the room was crowded' entails 'If exactly forty people were at the party, the room was crowded'. In the present setting, we can define entailment by stipulating that *D* entails *E* if and only if every *D*-world is an *E*-world. The Stalnaker-Lewis view predicts that Strengthening the Antecedent is invalid. For let *w*, *x*, and *y* be worlds such that *x* is closer than *y* to *w*. Let *A* be true just at *x* and *y*, A^+ true just at *y* (so A^+ entails *A*), and *C* true just at *x*. Thus the closest *A*-world to *w* is *x*, which is also a *C*-world, so $A > C$ is true at *w*. But the closest A^+-world to *w* is *y*, which is not a *C*-world, so $A^+ > C$ is not true at *w*. Thus $A > C$ does not entail $A^+ > C$, so Strengthening the Antecedent fails. This phenomenon leads to failures of epistemic closure, especially with respect to sceptical scenarios, just as Nozick and other proponents of a Sensitivity condition intended.

Let *H* be the proposition that I have hands. In the actual world I know *H* in the normal way. As a corollary, *H* is true and I believe *H*. I am also sensitive to *H*, for in the closest ¬*H*-worlds to the actual world, I lost my hands in a nasty accident, of which I am well aware, and so do not believe that I have hands, so $\neg H > \neg \mathrm{Bel}[H]$ is true at the actual world. Now let *BIV* be the proposition that I am a handless brain in a vat who believes that it has hands. Thus *H* entails ¬*BIV*. We may assume that in all relevant worlds in which I believe *H*, I also believe ¬*BIV* by competent deduction from *H*, for our present interest is not in failures of epistemic closure which merely reflect my deductive limitations. However, I am not sensitive to ¬*BIV*, for in the closest ¬¬*BIV*-worlds, in other words *BIV*-worlds, I believe *H*, and so also believe ¬*BIV*, so $BIV > \neg \mathrm{Bel}[\neg BIV]$ (omitting a double negation) is false at the actual world. Consequently, although I know *H*, and *H* entails ¬*BIV*, I do not know ¬*BIV*. This is so even though ¬*BIV* is true (since *H* is true), and I believe ¬*BIV* by having competently deduced it from *H*, which I already know. The non-closure of Sensitivity here is the main reason for the non-closure of knowledge, even though other details have to be checked too, since Sensitivity is not the only condition on knowledge, and we require competent deduction, not just entailment in itself.

Nozick ingeniously used this pattern to explain the appeal of scepticism without surrendering to it: the sceptic rightly denies that we know that we are not in the sceptical scenario, but wrongly concludes that we fail to know ordinary truths incompatible with our being in the sceptical scenario, because the sceptic wrongly assumes epistemic closure.

Although Nozick mainly followed Lewis's semantics for counterfactual conditionals, in one respect he was unorthodox. Lewis was sympathetic to the Strong Centering assumption that each world is *the* closest to itself. Thus if *A* and *C* are both true at *w*, so is $A > C$, for the uniquely closest *A*-world to *w* is *w*, which is also a *C*-world. Nozick demurred. In addition to truth, belief, and sensitivity, his analysis of knowledge required a fourth condition: 'Not only is p true and S believes it, but if it were true he would believe it' (Nozick 1981: 176). He denied that the conjunction made the conditional redundant (as Strong Centering implies), insisting that a counterfactual is true only where the consequent is true at close (but not maximally close) worlds at which the antecedent is true. The symmetry between his third and fourth conditions, $\neg A > \neg \mathrm{Bel}[A]$ and $A > \mathrm{Bel}[A]$, was an attractive feature of his analysis, in Nozick's eyes: if the fourth condition were redundant, the symmetry would be an illusion.

Ernest Sosa has used something like the converse of Nozick's fourth condition, $\mathrm{Bel}[A] > A$, to formulate a version of another putative modal condition on knowledge, *Counterfactual Safety*: if one believed *A*, it would be true (Sosa 1999, 2003, 2007, 2009; for an approach to safety not employing counterfactual conditionals see Williamson 2000). Like Nozick, Sosa insists that the conjunction of the antecedent and consequent does not make the counterfactual conditional redundant.

The Sensitivity condition $\neg A > \neg \mathrm{Bel}[A]$ is just the contrapositive of the Counterfactual Safety condition $\mathrm{Bel}[A] > A$. Thus if counterfactual conditionals obeyed the rule of Contraposition ($A > C$ entails $\neg C > \neg A$), Sensitivity would be equivalent to Counterfactual Safety (since double negations cancel). But the Stalnaker-Lewis account predicts the failure of Contraposition. For let *w*, *x*, and *y* be worlds such that *x* is closer than *y* to *w*. Let *A* be true just at *x* and *y*, and *C* true at every world except *y*. Thus the closest *A*-world to *w* is *x*, which is also an *C*-world, so $A > C$ is true at *w*. But the closest $\neg C$-world to *w* is *y*, which is an *A*-world, so $\neg C > \neg A$ is not true at *w*. Thus Contraposition is invalid.

We can illustrate the failure of Contraposition on the Stalnaker-Lewis semantics with or without Strong Centering in terms of the sceptical scenario. The Counterfactual Safety conditional $\mathrm{Bel}[\neg BIV] > \neg BIV$ is true in the good case: at all worlds close to the actual world (including itself) in which I believe that I am not a handless brain in a vat who believes that it has hands, I am indeed not a handless brain in a vat who believes that it has hands, for I am not a brain in a vat in any close world. But the contrapositive of that Counterfactual Safety conditional is the Sensitivity conditional $BIV > \neg \mathrm{Bel}[\neg BIV]$ (omitting a double negation), which was seen above to be false on the semantics.

Conversely, on the Stalnaker-Lewis semantics, in some other cases the Sensitivity conditional is true while the Counterfactual Safety conditional is false. For example, let *A* be 'I am the Pope'. At every close world, I neither am the Pope nor believe that I am the Pope. Thus the Sensitivity conditional $\neg A > \neg \mathrm{Bel}[A]$ is actually true. However, in the nearest possible worlds in which I do believe that I am the Pope, I am merely deluded, and still not the Pope. Thus the counterfactual safety conditional $\mathrm{Bel}[A] > A$ is false on the Stalnaker-Lewis semantics (with or without Strong Centering).

Although the semantics does not make Counterfactual Safety equivalent to Sensitivity, it does make Counterfactual Safety, like Sensitivity, violate closure under entailment. For example, suppose that I am good at distinguishing between poodles and non-poodles, but poor at distinguishing between dogs and non-dogs: there are many wolves around, all of which I take to be dogs. Let *P* say that there is a poodle in front of me and *D* say that there is a dog in front of me. Thus *P* entails *D* (and I know the entailment). Plausibly, *P* satisfies counterfactual safety: Bel[*P*] > *P* is true. But, in the same circumstances, *D* may well violate counterfactual safety: Bel[*D*] > *D* may well be false. Thus Counterfactual Safety is not closed under entailment.

However, the non-closure of Counterfactual Safety differs from the non-closure of Sensitivity: the former, unlike the latter, depends on the gap between entailment in itself and competent deduction. For let us restrict attention to worlds in which (i) if I believe that there is a poodle in front of me, I also believe by competent deduction from that belief that there is a dog in front of me and (ii) if I believe that there is a dog in front of me, I do so by competent deduction from my belief that there is a poodle in front of me. If we pretend away all other worlds, Bel[*P*] and Bel[*D*] are true at exactly the same worlds and so have the same counterfactual implications; since *P* entails *D*, Bel[*P*] > *P* entails Bel[*D*] > *D*, for the Stalnaker-Lewis approach does at least make counterfactual implication closed under entailment in the consequent, for a fixed antecedent. Of course, this argument does not show that Counterfactual Safety is *really* closed under entailment. Rather, it suggests that we may be able to define some more restricted kind of belief-on-a-given-basis for which the appropriate analogue of Counterfactual Safety *is* closed under competent deduction. That suggestion is followed up in Section 3.

By contrast, no such rehabilitation of closure is feasible for Sensitivity, on the Stalnaker-Lewis approach. For example, suppose that we restrict attention to worlds in which the analogues of (i) and (ii) hold: (i*) if I believe that I have hands, I also believe by competent deduction from that belief that I am not a handless brain in a vat who believes that it has hands and (ii*) if I believe that I am not a handless brain in a vat who believes that it has hands, I do so by competent deduction from my belief that I have hands. If we pretend away all other worlds, Bel[*H*] and Bel[¬*BIV*] are true at exactly the same worlds, as are their negations ¬Bel[*H*] and ¬Bel[¬*BIV*]. But that does not help, for they occur in the *consequents* of the Sensitivity conditionals. The corresponding antecedents are ¬*H* and *BIV*, which are not susceptible to such doctoring and still pick out different sets of worlds, even amongst those which satisfy both (i*) and (ii*). Thus, even under those restrictions, the Stalnaker-Lewis approach still predicts that Sensitivity is not closed under entailment.

2. *An alternative semantics for counterfactual conditionals*

Following Nozick, modal epistemology uses *counterfactual* conditionals, which are standardly contrasted with plain *indicative* conditionals. The difference is crucial. The counterfactual conditional ‘If there weren’t a poodle in front of me now, I wouldn’t now believe that there was a poodle in front of me’ is a sensible thing for an expert on poodles to say; by contrast, the indicative conditional ‘If there isn’t a poodle in front of me now, I don’t now believe that there is a poodle in front of me’ would be a weird thing for her to say. In some sense, the counterfactual permits one to develop scenarios in which some of what one

actually knows (for instance, that there *is* a poodle in front of her now) does not obtain, without thereby treating that actual knowledge as doubtful, while the indicative conditional does not. That allows one to evaluate the consequent of the counterfactual without the wrong kind of interference from one's background knowledge (for instance, that she *does* now believe that there is a poodle in front of her).

In English, the two kinds of conditional statement are distinguished syntactically, as in the indicative 'If X is the case, Y will be the case' versus the counterfactual 'If X was the case, Y would be the case'. Some other languages leave the distinction to be tacitly understood from the conversational context, but the explicit marking gives useful clues about the underlying structural difference. Oddly, the difference is usually described as one between the indicative and counterfactual *conditionals*, as though it were between two readings of 'if' itself, sometimes formalized as → and □→. But why locate the difference in a word in common between the two forms of sentence, when there are visible differences elsewhere, in the verb phrases? For example, in the antecedent, 'is the case' contrasts with 'was the case'; in the consequent, 'will be the case' contrasts with 'would be the case'. The simplest default hypothesis is that 'if' has exactly the same meaning in the two types of sentence, which combines with the distinct meanings of the different verb forms to give the different meanings of the two types of sentence within the standard framework of compositional semantics, which determines the meaning of a complex expression as a function of the meanings of its simpler constituents and the way in which they are put together.

The default hypothesis gains immediate support from the observation that 'would' has a life of its own, independent of 'if'. Syntactically, 'would' is of course the past tense of 'will'. For example, when two people part, someone might predict 'They will never meet again'. Later, after the prediction has turned out correct, one might end an account of their parting with 'They would never meet again'. But 'would' also has what linguists call a 'fake past' use, on which it ranges over relevant counterfactual possibilities instead of past times. For example, when you report 'Mary is not at home', I might comment 'She wouldn't be'. There is no suppressed 'if' clause. Nor am I talking about the past; rather, I am subsuming her absence under a modal generalization over a range of more or less close possibilities: it is no accident that she is not at home; perhaps she is always out for a walk at this time of day, or whatever. In such cases, 'would' functions as a contextually restricted local necessity operator. The natural hypothesis is that 'would' functions that way in counterfactual conditionals too. Thus 'If X was the case, Y would be the case' is the result of applying a contextually restricted local necessity operator, realized as a fake past tense, to 'If X is the case, Y will be the case' (the 'will' here may also be a fake future, as in 'Mary will be out walking now'). Consequently, 'If X was the case, Y would be the case' is true as uttered in a given context if and only if at every possibility relevant in that context, 'If X is the case, Y will be the case' is true.

To complete the compositional semantics of such modalized conditionals, we need an independent semantics of 'if' itself. The simplest hypothesis is that 'if' is just the truth-functional material conditional, so that the corresponding counterfactuals are just contextually restricted strict conditionals. Thus 'If X was the case, Y would be the case' is true as uttered in a given context if and only if at every possibility relevant to that context at which X is the case, Y is also the case. Where ⊃ is the material conditional ('if') and □ is the contextually restricted necessity operator, we can formalize the 'would' conditional with

antecedent *A* and consequent *C* as $\Box(A \supset C)$. Elsewhere, I have developed such an approach in detail, and shown that it has the resources to answer the salient kinds of objection likely to be brought against it (Williamson 2020, also for references to the literature on conditionals). I will not try to summarize all that argumentation here, but just mention a few features of special significance for its application to modal epistemology.[1]

On the non-modal side, the material interpretation of the English word 'if' faces numerous apparent counterexamples, which might be compelling were it not for the numerous apparent counterexamples to any rival interpretation. Fortunately, the evidence is not as anarchic as it seems. Our practice of using conditionals can be understood as built on a simple heuristic for assessing them (Williamson 2020: 19):

Suppositional Rule Take an attitude unconditionally to 'If *A*, *C*' just in case you take it conditionally to *C* on the supposition *A*.

For example, accept 'If *A*, *C*' unconditionally just in case you accept *C* conditionally on the supposition *A*; reject 'If *A*, *C*' unconditionally just in case you reject *C* conditionally on the supposition *A*. In probabilistic terms, your unconditional probability for 'If *A*, *C*' should equal your conditional probability for *C* on *A*. The Suppositional Rule is closely related to a test originally proposed by Frank Ramsey for assessing conditionals (Ramsey 1929: 143). There is considerable evidence that it is our primary means of assessing conditionals. Naturally enough, we also use it in assessing conditionals under further background suppositions, in common between the two sides of the Suppositional Rule. However, the Rule is not our *only* means of assessing conditionals. For example, one might accept 'If *A*, *C*' on the testimony of a trusted informant, without bothering to assess *C* on the supposition *A*.

The alert reader may have noticed that the material interpretation of 'if' does *not* fully validate the Suppositional Rule. For instance, on the material interpretation of 'if', 'NN is the Prime Minister' entails 'If NN has just died, NN is the Prime Minister'. Thus, if one is confident that NN is the Prime Minister, one should also be confident that if NN has just died, NN is the Prime Minister. But, of course, one's natural response to 'If NN has just died, NN is the Prime Minister' is to reject it, just as the Suppositional Rule predicts, because one rejects 'NN is the Prime Minister' on the supposition 'NN has just died'. Thus one might expect our reliance on the Suppositional Rule to *refute* the material interpretation of 'if', and support some other interpretation which *does* fully validate the Rule.

But things are not so simple. For the Suppositional Rule is actually *inconsistent*, and so is not fully valid on *any* interpretation of 'if'. A quick way to detect the inconsistency is by noting that the Rule implies the standard introduction and elimination rules for the conditional in systems of natural deduction, conditional proof and modus ponens. For, by the right-to-left direction of the Suppositional Rule, if *C* is conclusively accepted on the supposition *A* and background assumptions, then 'If *A*, *C*' should be conclusively accepted on those background assumptions, which is equivalent to the rule of conditional proof. Conversely, by the left-to-right direction of the Suppositional Rule, if 'If *A*, *C*' is conclusively accepted on background assumptions, then *C* should be conclusively accepted on the supposition *A* and those background assumptions, which is equivalent to the rule of modus ponens. Together, the rules of conditional proof and modus ponens for a conditional are well-known to force its equivalence to the material conditional. Thus the Suppositional Rule makes 'if' equivalent to the material conditional. But we saw previously that the

Suppositional Rule also has consequences inconsistent with the equivalence of 'if' to the material conditional. Thus the Suppositional Rule itself is inconsistent.

Naturally, one wonders how an inconsistent rule can be our primary way of assessing conditionals. Won't that make our whole practice of using 'if' collapse?

A useful analogy is with our practice of using the words 'true' and 'false'. Our primary means of assessing predications of 'true' and 'false' is by disquotation. We take the same attitude to '"Dubrovnik is in Croatia" is true' as to 'Dubrovnik is in Croatia', and the same attitude to '"Dubrovnik is in Croatia" is false' as to 'Dubrovnik is not in Croatia'. Such moves are essential to our practice of using 'true' and 'false'. Nevertheless, the disquotation rules are inconsistent, as the Liar and other semantic paradoxes show. Despite the inconsistency, our whole practice of using 'true' and 'false' does not collapse: it is a going concern. The inconsistencies are too marginal to do serious damage.

Philosophers who regard the disquotational rules as somehow *analytic* may be tempted to treat their validity as a precondition for 'true' and 'false' to be so much as meaningful. But that would be a mistake: the rules are invalid, because inconsistent, yet 'true' and 'false' are still meaningful words of English by any reasonable standard. Nor is any particular restriction built into the rules as we use them, to preserve consistency; the semantic paradoxes strike normal speakers of English like a bolt from the blue. Although research in philosophical logic may succeed in identifying more qualified rules for 'true' and 'false' which are valid without exception, that is a scientific *discovery*, not the articulation of something which normal speakers of English knew all along.

Similarly, despite the inconsistency of the Suppositional Rule, our whole practice of using 'if' does not collapse: it is a going concern. The inconsistency is not serious enough to undermine the practice, though it may be less marginal than the semantic paradoxes. Despite the Rule's invalidity, 'if' is still a meaningful word of English by any reasonable standard. The Rule is not somehow analytic. Nor is any particular restriction built into the Rule as we use it, to preserve consistency; the paradoxes of material implication strike normal speakers of English like a bolt from the blue. Although research in philosophical logic may succeed in identifying more qualified rules for 'if' which are valid without exception, that is a scientific *discovery*, not the articulation of something which normal speakers of English knew all along.

We do better to regard the Suppositional Rule as a *heuristic* for assessing conditionals in the psychologists' sense, a rule of thumb, normally quick and easy to apply, and reliable enough for practical purposes, but not perfectly reliable. Human cognition relies on many such heuristics. From the user's perspective, they need not wear their status as mere heuristics on their sleeve, especially when we have no better heuristic to check them by. That they are mere heuristics may be something else we have to discover, perhaps as a corollary from discovering that they are inconsistent.

In such cases, we cannot simply read the semantics of the word off the heuristic for its application. Instead, we must work out what semantics for the word makes the best sense of the practice, including all the relevant heuristics. For 'if' they include the primary heuristic, the Suppositional Rule, but also secondary heuristics, such as reliance on conditional testimony. There are clear theoretical reasons for taking the material interpretation of 'if' to make the best sense of our practice of using the word, in terms of both simplicity and charity (Williamson 2020: 89-121). Since those reasons are not available to normal speakers of

English except on complex theoretical reflection, the semantics of their own word 'if' is not transparent to them.

A consequence of this view is that, in cases where the Suppositional Rule gives the wrong result—for example, when they accept 'NN is the Prime Minister' but reject 'If NN has just died, NN is the Prime Minister'—they are typically in no position to recognize the error and correct it. They apply the Suppositional Rule correctly, and there is no special clue that it may not be giving the right result, so the case gives the illusion of being a counterexample to the material interpretation. Such predicaments are part of the unrecognized cost of relying on fallible heuristics.

Fortunately, heuristics on which we continually rely, such as the Suppositional Rule, also tend to have major benefits, otherwise in the long run they would hardly survive. In particular, the Suppositional Rule enables us to extract and apply information about various sorts of non-accidental connection encoded in our cognitive dispositions, by using imagination in reality-oriented ways to make offline assessments of some sentences on the supposition of others. By contrast, if we try to evaluate the material conditional in the disjunctive form $\neg A \vee C$ or the negated conjunctive form $\neg(A \wedge \neg C)$ by direct truth-functional calculation, we get stuck unless we know the truth-values of both constituent sentences. For instance, we can easily and correctly assess the conditional 'If he is sitting, he is not running' even when we know neither whether he is sitting nor whether he is running. In general, the Suppositional Rule does not make us overestimate the probabilities of conditionals, since the probability of the material conditional $A \supset C$ is always at least as high as the conditional probability of C on A. Of course, we may overestimate the probability of the material conditional if we already overestimate the conditional probability, but the Suppositional Rule is not to blame for that.

In what follows, we will treat the material interpretation of 'if' as a working hypothesis, together with the defence of it just briefly sketched, including the Suppositional Rule as a heuristic. We must be correspondingly critical in our pre-theoretical assessments of conditionals, since they will usually rely on implicit applications of the rules.

For modal epistemology, our main interest is in conditionals modalized with 'would'. How do we assess them? If we simply reuse the Suppositional Rule as for plain indicative conditionals, 'would' becomes redundant, which it clearly is not: as examples above showed, it makes a difference. The fake past 'distances' the supposed situation in which the antecedent holds from actuality, and the contextual restriction on 'would' qualifies the supposed situation, not actuality. The Suppositional Rule can then be applied with respect to the supposed situation, as it were within the scope of the fake past. The overall way of assessing modalized conditionals feels quite like assessment by the Suppositional Rule, but it has a subtle counterfactual twist (Williamson 2020: 189-213 has details). We can conveniently describe it in terms like those for the original Suppositional Rule, but with 'counterfactually' inserted to mark the difference.

The contextual restriction in the semantics of 'would' has pragmatic effects too when we assess counterfactuals. For counterfactually supposing the antecedent modifies the context: it makes new possibilities relevant. Consider a context before the counterfactual with antecedent A and consequent C was entertained in which no A-worlds were relevant. Once we entertain the counterfactual and start assessing it in the usual way, we counterfactually suppose A, and assess C on that supposition, which tends to make A-worlds relevant. When A is possible, there are A-worlds, and at least some of them will become relevant in the

modified context. This effect is no part of the semantics, for 'would' applies equally well to unconditional sentences. It is just pragmatic, but it plays a major role when we confront examples.[2]

The alternative approach to the semantics of counterfactuals has significant implications for their logic. The next section focusses in particular on the repercussions for modal epistemology.

3. Applications to modal epistemology

The Stalnaker-Lewis approach to the semantics of counterfactuals invalidates the principle of Strengthening the Antecedent, as section 1 noted: sometimes, when *A* counterfactually implies *C*, and A^+ entails *A*, A^+ fails to counterfactually imply *C*. Nozick's modal epistemology exploits this phenomenon. Consider again (1)-(4):

(1) I do not have hands.

(2) I am a handless brain in a vat who believes that it has hands.

(3) If I did not have hands, I would not believe that I had hands.

(4) If I was a handless brain in a vat who believed that it had hands, I would not believe that I had hands.

In normal circumstances, on Nozick's view, although (3) is true, and (2) (the antecedent of (4)) entails (1) (the antecedent of (3)), (4) is not true. Indeed, the opposite counterfactual (5) to (4) is uncontroversially true, since its antecedent entails its consequent:

(5) If I was a handless brain in a vat who believed that it had hands, I would believe that I had hands.

Because (4) and (5) have the same possible antecedent and contradictory consequents, they seem to be inconsistent; therefore, since (5) is true, (4) is false. Thus Strengthening the Antecedent fails.

To complete the picture, we may further assume that in all the worlds at issue in which I believe (6) (which is contradictory to (1)), I competently deduce (7) (which is contradictory to (2)) and believe (7) on that basis:

(6) I have hands.

(7) I am not a handless brain in a vat who believes that it has hands.

Thus, given (5), (8) holds too:

(8) If I was a handless brain in a vat who believed that it had hands, I would believe that I was not a handless brain in a vat who believed that it had hands.

Compare (9):

(9) If I was a handless brain in a vat who believed that it had hands, I would not believe that I was not a handless brain in a vat who believed that it had hands.

Because (8) and (9) (like (4) and (5)) have the same possible antecedent and contradictory consequents, they seem to be inconsistent; therefore, since (8) is true, (9) is false. But (9) is in effect Nozick's Sensitivity condition for me to know that I am not a handless brain in a vat who believes that it has hands. Thus, on his view, the falsity of (9) excludes me from knowing that I am not a handless brain in a vat who believes that it has hands. By contrast, the Sensitivity condition for me to know that I have hands is (3), which is true in normal circumstances. Nozick's modal epistemology allows me to know that I have hands, but even if I competently deduce that I am not a handless brain in a vat who believes that it has hands and believe it on that basis, I cannot thereby know that I am not a handless brain in a vat who believes that it has hands.[3]

The invalidity of Strengthening the Antecedent for counterfactuals is central to Nozick's strategy, for if it were valid, the falsity of (4) would entail the falsity of (3): the insensitivity would extend to my ordinary belief that I have hands, and conceding the impossibility of knowing that one is not in a sceptical scenario would not enable one to retain the possibility of ordinary knowledge.

On the alternative approach of section 2 to the semantics of counterfactuals, Strengthening the Antecedent is *valid*. Formally, the counterfactual with antecedent A and consequent C is interpreted as the restricted strict conditional $\Box(A \supset C)$. Thus Strengthening the Antecedent becomes the principle that if A^+ entails A, then $\Box(A \supset C)$ entails $\Box(A^+ \supset C)$, where entailment is understood as *unrestricted* strict implication. If A^+ entails A, every A^+-world is an A-world. If, in addition, $\Box(A \supset C)$ is true, every contextually relevant A-world is a C-world. Consequently, every contextually relevant A^+-world is a C-world, so $\Box(A^+ \supset C)$ is true. Thus the alternative approach validates Strengthening the Antecedent.

In particular, Nozick's strategy fails because (3) entails (4), on the alternative semantics. However, that it validates Strengthening the Antecedent looks like a problem for the alternative semantics. For the natural pre-theoretic assessments of (3) and (4) are that (3) is true and (4) false.

That challenge neglects the contextualist aspect of the semantics. In an ordinary context, no worlds in which I am a handless brain in a vat who believes that it has hands are relevant; thus (2) is false in all relevant worlds. The relevant worlds are much closer to actuality: in those in which I lack hands, I am straightforwardly aware of lacking hands, and do not believe that I have hands, so (3) is true. But (4) is also true in the ordinary context, vacuously, because in every relevant world the antecedent (2) of the embedded material conditional is false, so the material conditional itself is true. But once we start entertaining (4), and assessing it in the usual suppositional way, we change the context, by making relevant some worlds in which I am a handless brain in a vat who believes that it has hands. Those worlds falsify the antecedents and verify the consequents of the material conditionals embedded in (3) and (4). Consequently, both (3) and (4) are false in the new context. Thus the apparent failure of the entailment from (3) to (4) was a mere artefact of the unnoticed shift from the ordinary context in which we considered (3) to the special epistemological

context created by considering (4). In any fixed content in which (3) is true, (4) is also true. The entailment holds.

The contextualist account gains support from what happens when we reconsider (3) immediately *after* considering (4) (this is an example of what are known in the literature as 'reverse Sobel sequences'). Now (3) can easily look false, because we can still easily count as relevant the newly added worlds in which I am a handless brain in a vat who believes that it has hands, and they falsify the material conditional embedded in (3). Which worlds count depends on the extrinsic context, not on the intrinsic semantics.[4]

How does this critique of Nozick's modal epistemology compare with longstanding contextualist objections to his approach? It has in common with them the complaint that Nozick's apparent cases of non-closure are artefacts of unnoticed context-shifting, and the suggestion that a suitably qualified principle of epistemic closure will hold without exception in each fixed context. However, the conclusion usually drawn is that Nozick's counterfactual analysis of knowledge is incorrect, because it endorses those cases of non-closure (compare Luper 1984, BonJour 1987, DeRose 1995, Roush 2005). On the present account, by contrast, what generates the non-closure is *not* the counterfactual analysis of knowledge but rather Nozick's application of it, which ignores the contextual shiftiness of his counterfactuals themselves. In principle, an epistemologist could retain the letter of Nozick's analysis, but reinterpret it as already an implicitly contextualist account of 'know', fully consistent (for all that Nozick has shown) with a suitably qualified principle of epistemic closure holding without exception in each fixed context.

For myself, I endorse neither Nozick's analysis nor contextualism about 'know' (Williamson 2000, 2005). Nevertheless, counterfactual conditions are of epistemological interest in their own right, and we need to understand how they really work, once contextual effects are properly controlled for. Section 4 will show how to formulate appropriate closure principle for suitable variants of Sensitivity and Counterfactual Safety, without prejudice to their relation to knowledge.

Another respect in which traditional modal epistemology relies on the Stalnaker-Lewis semantics concerns the distinction between Counterfactual Safety and Sensitivity, as section 1 noted. Simple versions of those conditions are often formalized as contrapositives: for Counterfactual Safety, Bel[A] > A, for Sensitivity, $\neg A > \neg$Bel[A]. But the approach of section 2 makes contrapositive counterfactuals logically equivalent. For $A > C$ and $\neg C > \neg A$ are analysed as $\Box(A \supset C)$ and $\Box(\neg C \supset \neg A)$ respectively, which are logically equivalent, because $A \supset C$ and $\neg C \supset \neg A$ are truth-functionally equivalent and $\Box$ is a normal modal operator. Thus Contraposition is valid.

As with Strengthening the Antecedent, the validity of Contraposition seems to pose a problem for the alternative semantics, since it faces apparent counterexamples. We can illustrate them from modal epistemology. Here is a Counterfactual Safety condition:

(10) If I believed that there were other people, there would be other people.

A plausible pre-theoretic assessment of (10) is that it is true: one believes that there are other people because one has met some. The contrapositive (11) of (10) gives a Sensitivity condition:

(11) If there were no other people, I would not believe that there were other people.

A plausible pre-theoretic assessment of (11) is that it is false: were I the only person left after some global disaster, I would desperately believe that there must be at least a few other people left alive somewhere.

As before, the challenge neglects the contextualist aspect of the semantics. In an ordinary context, no worlds in which I am the only person left are relevant, so (10) is true. Then (11) is also true in the ordinary context, vacuously. Hence (10) and (11) have the same truth-value in the ordinary context. But once we start entertaining (11), and assessing it in the usual suppositional way, we change the context, by making relevant some worlds in which I am the only person left. If, in those worlds, I desperately believe that there must be at least a few other people left alive somewhere, (11) is false, but so too is (10). Hence (10) and (11) also have the same truth-value in the new context. Thus the apparent failure of the equivalence between (10) and (11) was a mere artefact of the unnoticed shift from the ordinary context in which we considered (10) to the dystopian-minded context created by considering (11). In any fixed content, (10) and (11) have the same truth-value. The equivalence holds.

Again, the contextualist account gains support from what happens when we reconsider (10) immediately *after* considering (11). Now (10) can easily look false, because we can still easily count as relevant the newly added worlds in which I am the only person left. Which worlds count depends on the extrinsic context, not on the intrinsic semantics.

On reflection, like apparent counterexamples to Strengthening the Antecedent, apparent counterexamples to Contraposition emerge as mere artefacts of context-shifting; they support the contextualist account of counterfactuals on which both principles are valid with respect to any fixed context. Thus, in their present forms, Sensitivity and Counterfactual Safety are logically equivalent. Any failure of equivalence is pragmatic: considering them naturally creates different contexts, owing to their different antecedents, so Sensitivity in its natural context is not equivalent to Counterfactual Safety in *its* natural context. Nevertheless, in any one context, Sensitivity is equivalent to Counterfactual Safety. Which contexts a philosopher wishes to propose them for is another matter.

4. An epistemic closure principle for a counterfactual condition

Section 3 explained how an apparent case of non-closure for Sensitivity dissolves into a mere artefact of context-shifting. Assuming the alternative semantics for counterfactuals, this section formulates and validates closure principles for both Sensitivity and Counterfactual Safety; since they are equivalent, the same principle does for both. We do not assume any particular relation between either Sensitivity or Counterfactual Safety and knowledge.

For simplicity, we start with the case of single-premise closure for Counterfactual Safety. The formula Bel[C / A] means ‘One believes C on the basis of believing A’. Thus Bel[C / A] entails both Bel[C] and Bel[A]. The formula SafeBel[A] abbreviates the Counterfactual Safety condition $\Box$(Bel[A] $\supset$ A), which can be paraphrased as ‘If one believed A, A would be true’, or more briefly as ‘A is safe to believe’. Similarly, the formula SafeBel[C / A] abbreviates the qualified Counterfactual Safety condition $\Box$(Bel[C / A] $\supset$ C), which can be paraphrased as ‘If one believed C on the basis of believing A, C would be true’, or more briefly as ‘C is safe to believe given A’.

We can now formulate a single-premise closure principle for Counterfactual Safety:

SPC-CS If *A* entails *C*, then SafeBel[*A*] entails SafeBel[*C* / *A*].

Thus if *A* is safe to believe, then whatever *A* entails is safe to believe given *A*. In other words, if *A* entails *C*, and *A* would be true if one believed *A*, then *C* would be true if one believed *C* on the basis of believing *A*. Naturally, SafeBel[*A*] and SafeBel[*C* / *A*] must be understood with respect to the same context.

We can establish **SPC-CS** thus. Assume SafeBell[*A*] and that *A* entails *C*. Suppose that in a relevant world *w* one believes *C* on the basis of believing *A*. Thus in *w* one believes *A*. Given SafeBel[*A*], *A* is true in every relevant world in which one believes *A*, and so in particular *A* is true in *w*. But *A* entails *C*, so *C* is true in *w* too. Thus *C* is true in every relevant world in which one believes *C* on the basis of believing *A*, which is SafeBel[*C* / *A*]. QED.

The premise SafeBel[*A*] is essential to the validity of **SPC-CS**, even given that *A* entails *C*. Without it, one might believe *C* on the basis of believing *A* at a relevant world *w* even though both *A* and *C* were false at *w*.

We can easily generalize **SPC-CS** to a multi-premise closure principle for Counterfactual Safety. The formula Bel[$C / A_1, \ldots, A_n$] means 'One believes *C* on the basis of believing $A_1, \ldots, A_n$'. Thus Bel[$C / A_1, \ldots, A_n$] entails both Bel[*C*] and Bel[A_1], …, Bel[A_n]. Correspondingly, SafeBel[$C / A_1, \ldots, A_n$] abbreviates the qualified Counterfactual Safety condition $\Box$(Bel[$C / A_1, \ldots, A_n$] $\supset$ *C*], which can be paraphrased as 'If one believed *C* on the basis of believing $A_1, \ldots, A_n$, *C* would be true', or more briefly as '*C* is safe to believe given $A_1, \ldots, A_n$'.

We can now formulate the multi-premise closure principle for Counterfactual Safety, of which **SPC-CS** is simply the special case for $n = 1$:

MPC-CS If $A_1, \ldots, A_n$ jointly entail *C*, then SafeBel[A_1], …, SafeBel[A_n] jointly entail SafeBel[$C / A_1, \ldots, A_n$].

Thus if each of $A_1, \ldots, A_n$ is safe to believe, then whatever they jointly entail is safe to believe given them. In other words, if $A_1, \ldots, A_n$ jointly entail *C*, and for each *i* A_i would be true if one believed A_i, then *C* would be true if one believed *C* on the basis of believing $A_1, \ldots, A_n$. Naturally, SafeBel[A_1], …, SafeBel[A_n], and SafeBel[$C / A_1, \ldots, A_n$] in **MPC-CS** must all be understood with respect to the same context.

The proof of **MPC-CS** is omitted, since it is an obvious generalization of the proof of **SPC-CS**. The premises SafeBel[A_1], …, SafeBel[A_n] are essential to the validity of **MPC-CS**, just as the premise SafeBel[*A*] is essential to the validity of **SPC-CS**, even given the corresponding entailment, and for a similar reason.

Of course, since Counterfactual Safety is equivalent to Sensitivity with respect to a given context, a corresponding version of multi-premise closure holds for Sensitivity too. It says that if $A_1, \ldots, A_n$ jointly entail *C*, and for each *i* one would not believe A_i if A_i was not true, then one would not believe *C* on the basis of believing $A_1, \ldots, A_n$ if *C* was not true (for any fixed context). The corresponding single-premise closure principle for Sensitivity is just the special case where $n = 1$.

The validity of such multi-premise closure principles for natural versions of both Sensitivity and Counterfactual Safety suggests that treating either condition as necessary for

knowledge would not jeopardize a corresponding multi-premise closure principle for knowledge too. If they are objectionable conditions on knowledge, it must be for some other reason.

Of course, Dretske and Nozick regarded the failure of natural closure principles for knowledge as the *right* result. If the proposed conditions on knowledge do not deliver that result, once the counterfactual conditionals in them are properly understood, an epistemologist in their spirit might still stipulate that those counterfactual conditionals are to be assigned the meanings Dretske or Nozick wanted, perhaps those which some version of the Stalnaker-Lewis approach assigns them, even if those meanings diverge from their actual meanings in English. Subtler means to the same end would be by stipulating that any occurrence of 'would' in the analysis of knowledge is to range over just those worlds which Dretske or Nozick would have expected it to range over. However, such manoeuvres make vivid the unnaturalness involved in resisting an appropriate closure principle for knowledge.

5. *Methodological moral*

In this paper, we have seen how treacherous counterfactual conditionals can be as servants of philosophical theorizing. We already had grounds to suspect them in that role, given the 'conditional fallacy': they may fail to serve their purpose because realizing the antecedent could have unintended side-effects and alter the very thing we intended to probe for (Shope 1978). The conditional fallacy does not depend on counterfactual conditionals' contextual shiftiness, though the two may interact: the side-effects may themselves depend on context. The contextual shiftiness introduces another parameter which the theorist needs to keep under control. The content of a theoretical claim made with a counterfactual conditional varies with the context in which the claim is made or interpreted. Of course, the theorist can in principle stipulate a canonical context which the counterfactual is to be interpreted as if uttered in, but it may be hard in practice to filter out unconscious effects from the different context in which the discussion is actually taking place. It might be easier and safer to make the intended sub-domain of worlds explicit, rather than relying on our shifty implicit understanding of the counterfactual. That is in effect to dispense with the counterfactual conditional itself.

Might a contextualist in epistemology *welcome* the contextual shiftiness of counterfactual conditionals, using them in theoretical formulations to inject the desired shiftiness? That possibility was briefly raised in section 3. The trouble is that although 'would' is shifty, its shifts may not stay within the required lines. Epistemological contextualists typically envisage 'know' and other terms of epistemic appraisal as varying on quite specific dimensions, such as a scale of epistemic standards. By contrast, 'would' varies with contextual relevance, which in turn depends on whatever happens to be salient to participants in the conversation. It may well include factors which by most standards count as non-epistemic. Such uncontrolled contextual shiftiness is unlikely to favour theoretical insight.

Counterfactual conditionals are treacherous indeed. In our key theoretical formulations, we may do better to stick to plain material 'if' and more overtly restricted modal operators, keeping the action on the table rather than under it.

Notes

1 For other semantic accounts of counterfactual conditionals as strict conditionals see von Fintel 2001 and Gillies 2007. However, they rely on a framework of dynamic semantics and do not provide a compositional semantics derived from independent semantic accounts of 'if' and 'would'.

2 von Fintel 2001 and Gillies 2007 explain such effects non-pragmatically by building them into a dynamic semantics. However, such an account is unduly inflexible; in reality, the effects are not mandatory (Williamson 2020: 228). Moreover, it is more explanatory to derive such effects from more general considerations, as here, rather than building them into the semantics by hand.

3 The underlying point still holds once Nozick's further complications, such as relativization to methods, have been factored in.

4 For early resistance to Nozick's approach on related lines see Wright 1983. Sarah Moss (2012) has responded on behalf of the Stalnaker-Lewis approach to appeals to reverse Sobel sequences; see Williamson 2020: 223-228 for a reply.

5 See Dretske 2013a and 2013b and Hawthorne 2013 for a debate on the merits of closure for knowledge. Williamson 2009 discusses further complexities for versions of Safety which vary the proposition at issue.

6 Thanks to participants for discussion of a predecessor of this paper at the 2019 Edinburgh workshop on epistemic closure.

References

Bonjour, Laurence. 1987: 'Nozick, externalism, and skepticism', in Luper 1987: 297-313.
DeRose, Keith. 1995: 'Solving the sceptical problem', *Philosophical Review*, 104: 1-52.
Dretske, Fred. 1970: 'Epistemic operators', *Journal of Philosophy*, 67: 1007-1023.
Dretske, Fred. 2013a: 'The case against closure', in Steup et al. 2013: 27-40.
Dretske, Fred. 2013b: 'Reply to Hawthorne;, in Steup et al. 2013: 56-59.
von Fintel, Kai. 2001: 'Counterfactuals in a dynamic context', in Michael Kenstowicz (ed.), *Ken Hale: A Life in Language*, 123-152. Cambridge, Mass.: MIT Press.
Gillies, Anthony. 2007: 'Counterfactual scorekeeping', *Linguistics and Philosophy*, 30: 329 -360.
Hawthorne, John. 2013: 'The case for closure', in Steup et al. 2013: 40-56.
Lewis, David. 1973: *Counterfactuals*. Oxford: Blackwell.
Luper, Steven. 1984: 'The epistemic predicament: knowledge, Nozickian tracking, and skepticism', *Australasian Journal of Philosophy*, 62: 26-50.
Luper, Steven (ed.). 1987: *The Possibility of Knowledge: Nozick and his Critics*. Totowa, NJ: Rowman and Littlefield.
Luper, Steven (ed.). 2003: *The Skeptics*. Farnham: Ashgate Publishing.
Moss, Sarah. 2012: 'On the pragmatics of counterfactuals', *Noûs*, 46: 561-586.
Nozick, Robert. 1981: *Philosophical Explanations*. Oxford: Clarendon Press.
Ramsey, Frank. 1929: 'General propositions and causality', reprinted in Hugh Mellor (ed.), *Foundations: Essays in Philosophy, Logic, Mathematics and Economics*: 133-151. London: Routledge and Kegan Paul.
Roush, Sherrilyn. 2005: *Tracking Truth: Knowledge, Evidence and Science*. Oxford: Oxford University Press.
Shope, Robert. 1978: 'The conditional fallacy in modern philosophy', *Journal of Philosophy*, 75: 397-413.
Sosa, Ernest. 1999: 'How to defeat opposition to Moore', *Philosophical Perspectives*, 13: 141-152.
Sosa, Ernest. 2003: 'Neither contextualism nor skepticism', in Luper 2003: 165-182.
Sosa, Ernest. 2007: *A Virtue Epistemology: Apt Belief and Reflective Knowledge, volume I*. Oxford: Oxford University Press.
Sosa, Ernest. 2009: *Reflective Knowledge: Apt Belief and Reflective Knowledge, volume II*. Oxford: Oxford University Press.
Stalnaker, Robert. 1968: 'A theory of conditionals', *American Philosophical Quarterly Monographs*, 2: 98-112.
Steup, Matthias, John Turri, and Ernest Sosa (eds.) 2013: *Contemporary Debates in Epistemology*, 2nd. ed. Oxford: Wiley-Blackwell.
Williamson, Timothy. 2000: *Knowledge and its Limits*. Oxford: Oxford University Press.
Williamson, Timothy. 2005: 'Contextualism, subject-sensitive invariantism and knowledge of knowledge', *Philosophical Quarterly*, 55: 213-235.
Williamson, Timothy. 2009: 'Probability and danger' *The Amherst Lecture in Philosophy*, 4: 1-35. http://www.amherstlecture.org/williamson2009/
Williamson, Timothy. 2020: *Suppose and Tell: The Semantics and Heuristics of Conditionals*. Oxford: Oxford University Press.
Wright, Crispin. 1983: 'Keeping track of Nozick', *Analysis*, 43: 134-140.