Presidential Address, *Proceedings of the Aristotelian Society* **105** (2005): 1-23

Armchair Philosophy, Metaphysical Modality and Counterfactual Thinking

Timothy Williamson

ABSTRACT A striking feature of the traditional armchair method of philosophy is the use of imaginary examples: for instance, of Gettier cases as counterexamples to the justified true belief analysis of knowledge. The use of such examples is often thought to involve some sort of a priori rational intuition, which crude rationalists regard as a virtue and crude empiricists as a vice. It is argued here that, on the contrary, what is involved is simply an application of our general cognitive capacity to handle counterfactual conditionals, which is not exclusively a priori and is not usefully conceived as a form of rational intuition. It is explained how questions of metaphysical possibility and necessity are equivalent to questions about counterfactuals, and the epistemology of the former (in particular, the role of conceiving or imagining) is a special case of the epistemology of the latter. A non-imaginary Gettier case is presented in order to show how little difference it makes.

I

If anything can be pursued in an armchair, philosophy can. Its traditional method is thinking, without observation or experiment. If the pursuit is conceived as social, rather than solely individual, then speaking must be added to thinking, and several armchairs are needed, but that still leaves philosophy looking methodologically very far from the natural sciences. Loosely speaking, their method is a posteriori, philosophy's a priori. What should we make of this difference?

Crude rationalists regard philosophy's a priori method as a virtue. According to them, it makes philosophical results specially reliable, because immune from perceptual error. *Crude empiricists* regard philosophy's a priori method as a vice. According to them, it makes philosophical results specially unreliable, because immune from perceptual correction.

Few contemporary philosophers have the nerve to be crude rationalists. Given the apparent absence of a substantial body of agreed results in philosophy, crude rationalism is not easy to maintain. Many contemporary philosophers have some sympathy for crude empiricism, particularly when it goes under the more acceptable name of 'naturalism'. However, that sympathy sometimes has little effect on their philosophical practice: they philosophize in the grand old manner, merely adding naturalism to their list of a priori commitments.

Perhaps the 'linguistic turn' in twentieth century philosophy can be understood as a response to naturalism. Holding fixed the a priori nature of philosophy, one asks what philosophy could be good for. Since confinement to an armchair does not deprive one of one's linguistic competence, a tempting answer is that the proper business of philosophy is with language. The more individualistic version of this idea is that the subject matter of philosophy is thought. In the twenty-first century, such answers no longer seem satisfying. Taken at face value, much of contemporary philosophy is not about words or concepts. For instance, many analytic metaphysicians want to understand the nature of time itself, not just our concept of time or what we mean by the word ‘time’; they suspect that the latter may somehow embody misconceptions about time. Attempts to reinterpret their interest as really conceptual or semantic do not withstand critical scrutiny, revealing more about the preconceptions of the interpreters than they do about the subjects of interpretation.[1]

The issues to be discussed here concern the method of philosophy more directly than its subject matter. We shall question how deeply different the traditional method of philosophy is from more empirical methods of investigation. Crude rationalists and crude empiricists share the assumption that the a priori method of philosophy is profoundly unlike the a posteriori methods of the natural sciences: it is no mere difference between distinct applications of the same underlying methods. In particular, both rationalists and empiricists focus on philosophers’ appeals to *intuition*. A paradigm of such appeals is in the use of imaginary counterexamples supposedly to refute philosophical analyses or theories. The analysis or theory entails a classification of the example; the philosopher has an armchair intuition to the opposite effect, and declares the analysis or theory refuted. The technique is used both by philosophers who take themselves to be assessing an analysis of thought or language and by philosophers who take themselves to be assessing a theory about a reality independent of mind and language.[2]

The literature on philosophical intuition takes as a paradigm case Edmund Gettier's refutation of the traditional analysis of knowledge as justified true belief (1963). Almost overnight, the vast majority of epistemologists throughout the analytic community rejected what for more than two thousand years, since Plato, had been the standard analysis of the central epistemological concept, in response to a couple of imaginary examples in a three page article by someone most of them had never heard of: sociologically, a striking phenomenon.[3] The example is analysed in detail below, with the aim of shedding some light on the nature of traditional philosophical method.

Of course, philosophy is not simply the activity of producing imaginary counterexamples. An analysis or theory may be confirmed when its verdicts about cases seem to be vindicated by intuition. There is also the activity of thinking up the analysis or theory in the first place. Indeed, much philosophizing may not fit anywhere into this simple schema of constructing analyses or theories and then testing them against examples. A fully satisfying account of traditional philosophical method would require careful study of a far more diverse variety of cases than can be considered here. Nevertheless, we can make a start by assessing a paradigm of one very recognizable sort of philosophical activity.

II

Gettier's cases exploit the elementary logical point that some logical consequences of falsehoods are truths. They also exploit the elementary epistemological point that deduction is a way of transmitting justification from the premises to the conclusion of an argument. Moreover, they assume that the truth component of the justified true belief analysis is not redundant: some justified beliefs are false; advocates of the analysis usually grant that assumption. Gettier imagines a subject with a justified belief in a falsehood *q*, who deduces a truth *p* from *q* and believes *p* on that basis. Consequently, the subject has a justified true belief in *p*. Nevertheless, the subject does not know *p*, because a belief that depends in that way on an error does not constitute knowledge, no matter how justified and true it is.

For present purposes, we can ignore most of the structure specific to Gettier cases, and concentrate on the logical structure that they share with most other imaginary counterexamples to philosophical analyses. In particular, we may assume that the verdict that the subject lacks knowledge in the particular case has epistemic priority over the general diagnosis that a true belief essentially derived from a false one does not constitute knowledge. Thus the primary direction of justification here is from the particular verdict to the general principle (by inference to the best explanation) rather than from the general principle to the particular verdict (by universal instantiation). Gettier's own focus is on the particular verdicts, and that is how the counterexamples have usually been conceived as working. At any rate, Gettier's examples *can* be used in that way, and methodologically it is best to start with the simplest case, in which the particular verdict has priority. The discussion can easily be generalized to very many imaginary counterexamples that have been deployed against philosophical analyses in just that way.

Let us read the traditional analysis as the claim that justified true belief is necessary and sufficient for knowledge: necessarily, for any subject *x* and proposition *p*, *x* knows *p* if and only if *x* has a justified true belief in *p*.[4] In symbols:

(1) $\Box \forall x \forall p\, (\mathrm{K}(x,p) \equiv \mathrm{JTB}(x,p))$

Consider a particular Gettier case (henceforth 'the Gettier case'), neutrally described without prejudice to the traditional analysis. For instance, the description does not specify that *x* does not know *p*. It fills in the schematic account above with picturesque details such as one find in Gettier's paper.[5] The objection to (1) depends on two claims. First, the case is possible: someone could stand in that relation to a proposition. Using 'GC' for this neutral description, we symbolize the claim that a subject *x* could have stood as described to a proposition *p* thus:

(2) $\Diamond \exists x \exists p\, \mathrm{GC}(x,p)$

The second claim is that if the Gettier case had occurred, then the subject would have had a justified true belief in *p* without knowing *p*. More precisely, if someone had stood as described to a proposition, then whoever stood as described to a proposition would have had justified true belief without knowledge in respect of that proposition. In symbols, where '$r \mathbin{\Box\!\!\rightarrow} s$' means that if *r* had been the case then *s* would have been the case (the counterfactual or subjunctive conditional):[6]

(3) $\exists x \exists p\ \mathrm{GC}(x,p) \;\Box\!\!\rightarrow\; \forall x \forall p\ (\mathrm{GC}(x,p) \supset (\mathrm{JTB}(x,p)\ \&\ \neg\mathrm{K}(x,p)))$

Given (2) and (3), it follows that someone could have had justified true belief without knowledge in respect of some proposition:

(4) $\Diamond \exists x \exists p\ (\mathrm{JTB}(x,p)\ \&\ \neg\mathrm{K}(x,p))$

But (4) is clearly inconsistent with the original analysis (1), in particular, with its right-to-left direction. Justified true belief is not sufficient for knowledge.

Several features of the account deserve comment.

The Gettier case constitutes an objection to the claim (1) that justified true belief is necessary and sufficient for knowledge, not merely to the stronger claim that knowledge is identical with justified true belief or to the even stronger claim that ‘knowledge’ is synonymous with ‘justified true belief’ or that the concept *knowledge* is identical with the concept *justified true belief*; (1) does not make any of those stronger claims. However, the modal element in (1) is crucial. Since the Gettier case is imaginary, it constitutes no objection to the non-modal claim that in fact every case of knowledge is a case of justified true belief and *vice versa*. Gettier can claim that his case is possible, as in (2), not that it is actual. Since the ‘possibly’ qualification is essential in (2), the ‘necessarily’ qualification is essential in (1) if the objection is to stand. This point about the use of imaginary counterexamples in refuting philosophical analyses generalizes far beyond Gettier cases.

One might wonder why the complication of the counterfactual conditional is essential in (3). Would it not be simpler to use the claim of strict implication, that necessarily anyone who stands as described in the Gettier case to any proposition has justified true belief without knowledge in respect of that proposition?

(3*) $\Box \forall x \forall p\, (\mathrm{GC}(x,p) \supset (\mathrm{JTB}(x,p)\ \&\ \neg \mathrm{K}(x,p)))$

For (2) and (3*) entail (4) at least as easily as (2) and (3) do; moreover, (3*) entails (3). However, (3*) faces the difficulty that we have much weaker grounds for it than for (3). Examples are almost never described in complete detail; a mass of background must be taken for granted; it cannot all be explicitly stipulated. Many of the missing details are irrelevant to whatever philosophical matters are in play, but not all. This applies not just to highly schematic descriptions of examples, such as the formal sketch at the beginning of this section, but even to the much richer stories such as Gettier's that philosophers like to tell. For example, a subject with sufficiently awkward background beliefs who derives *p* from *q* would thereby lose justification for *q* rather than gaining it for *p*, even in cases like those Gettier described. Without specifically addressing the question, we simply envisage the case differently from that. Nor do we worry about whether our verdicts would hold even if mad scientists were interfering with the subject's brain processes in various ways; those possibilities do not normally occur to us when we assess Gettier cases. Any humanly compiled list of such interfering factors is likely to be incomplete. We envisage Gettier's descriptions as realized in ways that minimize departures from actuality in respects about which nothing is explicitly stipulated. In effect, we assess what *would* hold if the description were realized, not what is a strictly necessary consequence of its being realized. What we should endorse is the counterfactual conditional (3), not the strict implication (3*). The latter is quite probably false. Again, this point about the use of imaginary counterexamples in refuting philosophical analyses extends far beyond Gettier cases. It also generalizes to their use in refuting philosophical claims of necessity that lack the form of an analysis, such as one-way strict implications.

By what principles can (4) be inferred from (2) and (3)? Let r symbolize the non-modal proposition that the Gettier case is realized ($\exists x \exists p$ GC(x,p)), s the non-modal proposition that whenever the Gettier case is realized the subject has justified true belief without knowledge in respect of the proposition ($\forall x \forall p$ (GC(x,p) $\supset$ (JTB(x,p) & ¬K(x,p)), and t the non-modal proposition that someone has justified true belief without knowledge in respect of some proposition ($\exists x \exists p$ (JTB(x,p) & ¬K(x,p))).[7] Thus (2) is $\Diamond r$, (3) is $r \Box\!\!\rightarrow s$ and (4) is $\Diamond t$. The key logical relation between r, s and t is that r and s together entail t (by elementary syllogistic logic). One can easily check the validity of the inference from (2) and (3) to (4) if one assumes a simple possible worlds semantics for $\Diamond$ and $\Box\!\!\rightarrow$, giving conditions for the truth of sentences at any world w:

> $\Diamond r$ is true at w if and only if r is true at some world.

> $r \Box\!\!\rightarrow s$ is true at w if and only if either r is true at no worlds or, of worlds at which r is true, s is true at some more similar (in the relevant respect) to w than is any at which s is false.

(Lewis 1986). An argument is valid in the relevant sense if and only if (in any model) at any world at which the premises are true the conclusion is also true. Now suppose that (2) and (3) are true at w. By (2), r is true at some world, so by (3) s is true at some world at which r is true; since r and s entail t, the latter is also true at some world; thus (4) is true at w.

More complex accounts make possibility and necessity contingent on which world obtains; they qualify ‘some world’ in the semantics for $\Diamond$ by ‘possible relative to w’. One can still demonstrate the validity of the inference from (2) and (3) to (4), on the plausible assumption that any world more similar (in the relevant respect) to a given world w than is some world possible relative to w is itself possible relative to w (that assumption follows from the idea that the impossibility of one world at another is a radical degree of dissimilarity between them). For suppose that (2) and (3) are true at w. By (2), r is true at some world x that is possible relative to w. If s is also true at x then t is true at x, so (4) is true at w. Otherwise, s is false at x, so by (3) some worlds at which r and s are both true, and therefore t is true, are more similar to w than is x; by the assumption, those worlds are themselves possible relative to w, and again (4) is true at w.

However, let us be cautious and not assume that $\Diamond$ and $\Box\!\!\rightarrow$ should be given any such possible worlds semantics. Instead, we can state the required principles without mentioning possible worlds. Here ‘$\models$’ symbolizes a notion of validity (truth-preservation in all models) that is not assumed to advert to possible worlds:

REFLEXIVITY $\models A \Box\!\!\rightarrow A$

CLOSURE If $\mathrm{B}_1, \ldots, \mathrm{B}_n \models \mathrm{C}$

then $\mathrm{A} \Box\!\!\rightarrow \mathrm{B}_1, \ldots, \mathrm{A} \Box\!\!\rightarrow \mathrm{B}_n \models \mathrm{A} \Box\!\!\rightarrow \mathrm{C}$

POSSIBILITY $\Diamond\mathrm{A}, \mathrm{A} \Box\!\!\rightarrow \mathrm{B} \models \Diamond\mathrm{B}$

These rules yield the easy derivation:

(i) $r, s \models t$ Non-modal logic

(ii) $r \Box\!\!\rightarrow r, r \Box\!\!\rightarrow s \models r \Box\!\!\rightarrow t$ (i), CLOSURE

(iii) $\models r \Box\!\!\rightarrow r$ REFLEXIVITY

(iv) $r \Box\!\!\rightarrow s \models r \Box\!\!\rightarrow t$ (ii), (iii)

(v) $\Diamond r, r \Box\!\!\rightarrow t \models \Diamond t$ POSSIBILITY

(vi) $\Diamond r, r \Box\!\!\rightarrow s \models \Diamond t$ (iv), (v)

Thus (2), (3) $\models$ (4), as desired.

The three principles are pretheoretically plausible. If something were the case, it would indeed be the case (REFLEXIVITY; compare Lewis 1986: 132, axiom schema (3)). If various things would be the case if A were the case, and it follows from their being the case that C would be the case, then C would be the case if A were the case (CLOSURE; compare Lewis 1986: 132, the rule of Deduction within Conditionals).[8] If A could be the case, and C would be the case if A were the case, then C could itself be the case (POSSIBILITY).

The correctness of POSSIBILITY does depend on reading ◊ in terms of non-epistemic possibility, given that the counterfactual conditional □→ is itself read non-epistemically. For if A is ‘Hesperus is not Phosphorus’ and B is ‘Hesperus is not Hesperus’, then the counterfactual ‘If Hesperus had not been Phosphorus, Hesperus would not have been Hesperus’ is vacuously true, but A may be epistemically possible while B is not: someone may be in a position to rule out B without being in a position to rule out A. The truth of counterfactuals is indeed not determined by what the speaker knows: ‘If I had pressed the button, the light would have gone on’ may be false as I utter it even if all my evidence suggests otherwise. POSSIBILITY may even hold on more than one non-epistemic reading of ◊: for example, both as metaphysical possibility and as physical possibility.

The preceding points about the logical structure underlying the use of imaginary counterexamples in refuting necessity claims also generalize far beyond Gettier cases.

Is this reliance on modal reasoning what makes philosophical method so distinctive? Two modal premises about the Gettier case were needed: the possibility claim (2) and the counterfactual conditional claim (3). Only (3) corresponds to the ‘intuition’ that the subject in the Gettier lacks knowledge (but has justification), for (2) says only that the Gettier case as neutrally described is possible. Since standard Gettier cases are not far-out science fictions but mundane practical and physical possibilities, (2) is normally uncontested; it is not where the philosophical action is. That takes place around (3).

Asserting counterfactuals is not distinctive of a priori methodology. They are often asserted on a posteriori grounds. My wife sees a fox run past outside and say to me

‘If you had looked out of the window just then, you would have seen a fox run past’. Obviously, such a counterfactual is also highly contingent; it would have been false if the fox had not run past just then, as might very easily have happened. We often use counterfactuals in our practical dealings with the world. I observe that if that rock had fallen five seconds later it would have hit me, infer that the path I am on is dangerous, and resolve to avoid it in future. Many counterfactuals are closely linked to causal statements, even if it is over-optimistic to expect necessary and sufficient conditions for causal claims in counterfactual terms, or necessary and sufficient conditions for counterfactual claims in causal terms. ‘If the lightning strike had not occurred, the forest fire would not have occurred’ is just as a posteriori and contingent as ‘The lightning strike caused the forest fire’. However, this causal element is necessary for neither a posteriority nor contingency. ‘If Dr Johnson had kicked George Berkeley, he would have kicked a bishop’ is a posteriori and contingent, but the relation between its antecedent and consequent is not causal: the kicking of George Berkeley would have *been* the kicking of a bishop. Again, famously, counterfactuals follow from a posteriori claims of natural law (perhaps *ceteris paribus*): if it is a natural law that salt dissolves in water, then if this quantity of salt had been put in water, it would have dissolved.

Counterfactuals play a significant role in the explanation of our evidence for many accepted empirical claims. For example, we might support the claim ‘There are no kangaroos on this island’ with the counterfactual conditional ‘If there were any, we’d have seen some of them by now’. The counterfactual is what we need. In such cases, our information does not enable us to replace it by anything more rigorous. Obviously, ‘There are kangaroos on this island’ does not logically or conceptually entail ‘We have seen some kangaroos on this island'. The corresponding indicative conditional 'If there are kangaroos on this island, we have seen some of them by now' is of course hopeless here: we are sure that we have not seen any kangaroos on the island, and must rather say 'If there are kangaroos on this island, we have not seen any of them yet'. Similarly, given our current evidence, it is very improbable that we have seen some kangaroos on this island, even conditional on the hypothesis that there are kangaroos on this island. In any case, we are in no position to calculate such probabilities without making arbitrary assumptions.[9] It is plausible that counterfactuals play such a role in explanations of evidential status even in scientifically more significant cases. Consequently, the use of counterfactuals such as (3) shows nothing very distinctive about philosophy.

Someone might reply that the distinctively philosophical element consists in the special kind of grounds on which philosophically interesting counterfactuals such as (3) are asserted. But if the thought is that in philosophical examples the grounds provide an analytic or at least necessary connection of the antecedent to the consequent, we have already seen it to be mistaken: our grounds for (3) provide much weaker support for the quite probably false strict implication (3*). We have not deduced (3) from some much stronger claim.

It would be equally incorrect to suppose that (3) is special because anyone who appears to deny that the subject in the Gettier case has a justified true belief without knowledge must be exhibiting linguistic incompetence; they fail to understand the word 'know' or some other relevant expression. A few philosophers do deny that the subject in the Gettier case has a justified true belief in the relevant sense without knowledge. Those philosophers exhibit theoretical deviance, perhaps bad epistemological judgement, but not linguistic incompetence. Some are native speakers of English; other native speakers of English do not classify them as incompetent at English. By any reasonable criterion, they understand the word 'know' and possess the concept *know*.[10]

The modal element in Gettier cases is not even essential. For Gettier cases have actually occurred. Just to make sure of that, I recently brought one about when giving a visiting lecture on intuitions. At the start of the lecture, I told the audience that I had got the idea for it while on holiday in Algeria. They had no reason to distrust me on such matters. They believed that I had been to Algeria; their belief was justified in the relevant sense. I also made sure that they inferred that I had been to North Africa, and believed on that basis that I had been to North Africa. That belief was justified too. In fact, I had never been to Algeria, although I had been to North Africa. So my audience formed a justified true belief that I had been to North Africa: but they did not know that I had been to North Africa. Thus (1) was directly falsified by a particular instance of $\mathrm{JTB}(x,p)$ & $\neg\mathrm{K}(x,p)$, without any need of modal thinking beyond the uncontentious point that knowledge is necessarily coextensive with justified true belief only if knowledge *is* coextensive with justified true belief ($\Box A \models A$). My judgement that the audience had justified true belief without knowledge is not based on intuition in any sense that would distinguish it from other judgements. It is just an ordinary a posteriori judgement, philosophical only in the use I make of it. In particular, I did not first describe the situation to myself in neutral terms, then intuit a priori that any case with those features is a case of justified true belief without knowledge, and finally conclude that the situation was one of justified true belief without knowledge. After all, not every judgement can be inferential in that way, on pain of an infinite regress. I no more need a priori intuition to judge whether the audience knew that I had been to North Africa than I need it to judge whether an audience is getting bored, or whether Newton knew that the planets move in elliptical orbits. I am simply applying the concept of knowledge to empirically encountered cases.

A real-life Gettier case makes almost no difference to the epistemic status of the result that justified true belief is insufficient for knowledge. Those who already accepted Gettier's fictional cases did not need real-life examples to convince them. Those who objected to Gettier's cases will doubtless make similar objections to the real-life case, for their objections were not to the role of modality. Equally, real-life cases would have done just as well as fictional ones in the original article. Concerns about errors of perception, memory or testimony would not have significantly weakened their force. Crude rationalists will explain that point by saying that all we really need of the examples is their possibility. But when there is no special doubt about our empirical information, we do not first retreat from it to the modal premises (2) and (3) before we deploy the counterexample to (1); we deploy our empirical information directly against (1). Perhaps a sense can be made out in which the argument against (1) from (2) and (3) counts as 'a priori' while the argument against (1) from the real-life case (described as actual, not merely as possible) counts as 'a posteriori': but what is more significant is how little the difference matters, at least here.

We have a general cognitive ability to handle counterfactual conditionals. When we have some conception of the circumstances in which *r* is true, and some conception of the circumstances in which *s* is true, we also have some conception of the circumstances in which the counterfactual $r \mathrel{\Box\!\!\rightarrow} s$ is true. Sometimes we can reason from *r* to *s*, or from *r* to ¬*s*, perhaps using as auxiliary premises background beliefs that in some sense to be specified do not conflict with *r*. But for many cases 'reasoning' seems to imply a more formal or conceptually articulated process than we actually employ. Even concerning many counterexamples to philosophical analyses, our verdicts do not seem to be based on reasoning in any useful sense. Perhaps our ability to assess the truth-values of counterfactuals involves some capacity to simulate mentally the truth of the antecedent and to determine the truth-value of the consequent under that simulation, although just what that would involve is frustratingly obscure. Surely we need a better epistemology of counterfactuals than we currently have. But we have no good reason to expect that the evaluation of 'philosophical' counterfactuals such as (3) uses radically different cognitive capacities from the evaluation of ordinary 'unphilosophical' counterfactuals. We can evaluate (3) without leaving the armchair; we can also evaluate many 'unphilosophical' counterfactuals without leaving the armchair. After all, we had plenty of experience before we sat down. Very often, the background knowledge needed to evaluate a counterfactual consists not of specific items of information acquired on specific occasions but of a more general sense of how things go, honed over long experience. Such a sense is typically not presented to the subject in usably verbal form; one says things like 'She would do that because she is that sort of person'. Of course, underlying the inarticulate sense of how things go must be some conformation of the brain, but the latter does not constitute a theory from which the subject can infer the counterfactual or its negation.

Some evaluations are more sensitive than others to the specific course of the subject's prior experience. The evaluation of (3) is comparatively insensitive (given the experience needed to understand (3)), although the evidence suggests that it is not totally insensitive.[10] Both crude rationalists and crude empiricists in effect assume that evaluations of counterfactuals can somehow be divided according to their sensitivity or otherwise to experience into two drastically contrasted classes. That assumption is neither intrinsically plausible nor adequately supported by evidence.

The point is not that no distinction at all can be drawn between the a priori and the a posteriori, or that nothing falls on the a priori side, but that such a distinction lacks the significance with which it is often credited. It does not cut at the joints. Although, if we like, we can classify the evaluation of (3) as a priori, we should not suppose it to involve fundamentally different cognitive capacities from evaluations of counterfactuals that we classify as a posteriori. Consequently, we should not suppose it to raise fundamentally new questions of reliability. A rough analogy: in a logical system, we can distinguish arguments with no premises from arguments with at least one premise, but that is no reason to claim that a special issue arises about the validity of arguments with no premises, dissimilar from any issue about the validity of arguments with at least one premise.

If a priori intuition is understood as a distinctive cognitive capacity or pathology, it is not required for using imaginary counterexamples against philosophical theories or analyses. We have our ordinary capacities for making judgements about what we encounter, and a further capacity to evaluate counterfactuals by running those capacities ‘offline’; that is already enough for philosophy to get going, without any need of a kickstart from a special faculty of intuition.

III

Of the two premises in the argument underlying a Gettier case, the discussion so far has concentrated on the counterfactual claim (3), that the subject would have a justified true belief without knowledge. The other premise (2), that the case as neutrally described could occur, seems too banal to be the focus of methodological debate. With other philosophical counterexamples, however, the question of possibility is far more urgent. For instance, if someone challenges holism about concepts with an example in which a subject has the concept *red* and no other concepts whatsoever, it is not at all obvious that the possibility of the case should be granted. Even examples that the philosophical community finds convincing may involve conditions that we have no chance of realizing, by contrast with Gettier cases, which we can easily bring about. For instance, many philosophers would regard it as metaphysically contingent that the Big Bang occurred, on the basis of unavoidably imaginary courses of history in which it does not occur. A ban on such examples would severely cramp our philosophical style. Does the reliance on claims of metaphysical possibility involve a distinctively philosophical use of intuition or imagination?

We should not assume too quickly that our capacity to handle counterfactuals does not already involve a capacity to handle claims of metaphysical possibility. For metaphysical possibility and metaphysical necessity have been defined in terms of counterfactuals. Stalnaker (1968) defined necessity and possibility by equating $\Box A$ with $\neg A \mathrel{\Box\!\!\rightarrow} A$ and $\Diamond A$ with $\neg(A \mathrel{\Box\!\!\rightarrow} \neg A)$ (in present notation), thus preserving the duality by which $\Box A$ is equivalent to $\neg\Diamond\neg A$ and $\Diamond A$ to $\neg\Box\neg A$. Informally, ‘A is necessary’ becomes ‘If A were not the case, A would be the case’; ‘A is possible’ becomes ‘It is not the case that if A were the case, A would not be the case’.

Although the equivalences for $\Box$ and $\Diamond$ look unnatural at first sight, they have a simple rationale. It turns on the case of vacuous truth. The need for true counterfactuals with impossible antecedents is already clear from REFLEXIVITY, since otherwise $A \Box\!\!\rightarrow A$ would fail when A was impossible. Let us make two generally accepted assumptions about the distinction between vacuous and non-vacuous truth: (a) $B \Box\!\!\rightarrow C$ is vacuously true if and only if B is impossible (this could be regarded as a definition of ‘vacuously’ for counterfactuals); (b) when $B \Box\!\!\rightarrow C$ is non-vacuously true, B and C are logically compatible (see also Lewis 1986: 24-6).[12] We can now argue for Stalnaker’s equivalences without assuming any possible worlds semantics. If $\Box A$ is true, then $\neg A$ is impossible, so by (a) $\neg A \Box\!\!\rightarrow A$ is vacuously true; conversely, if $\neg A \Box\!\!\rightarrow A$ is true, then by (b) it is vacuously true, so by (a) $\neg A$ is impossible, so $\Box A$ is true. Similarly, if $\Diamond A$ is true, then A is not impossible, so by (a) $A \Box\!\!\rightarrow \neg A$ is not vacously true, and by (b) not non-vacuously true, so $\neg(A \Box\!\!\rightarrow \neg A)$ is true; if $\Diamond A$ is not true, then A is impossible, so by (a) $A \Box\!\!\rightarrow \neg A$ is vacuously true, whence $\neg(A \Box\!\!\rightarrow \neg A)$ is not true. Thus the metaphysical modalities are tantamount to special cases of the counterfactual conditional, so understanding the counterfactual conditional implies understanding equivalents of the metaphysical modalities (given understanding of negation).[13]

To appreciate the equivalence of $\Box A$ and $\neg A \Box\!\!\rightarrow A$, one can reflect that A is a counterfactual consequence of its own negation if and only if it is a counterfactual consequence of everything. We may therefore consider this alternative definition:[14]

DEF(□) $\Box A =_{\text{def}} \forall p\, (p \mathrel{\Box\!\!\rightarrow} A)$

A is necessary if and only if whatever were the case, A would still be the case (see also Lewis 1986: 23). That is a natural way of explaining informally what 'metaphysically necessary' means, even if it falls short of strict synonymy. The dual definition for possibility is this:

DEF(◊) $\Diamond A =_{\text{def}} \exists p\, \neg(p \mathrel{\Box\!\!\rightarrow} \neg A)$

A is possible if and only if A is not such that it would fail in every eventuality. DEF(□) and DEF(◊) have several nice consequences for logic:

(I) Given Def(◊), POSSIBILITY (the principle that possibility transmits across the counterfactual conditional) is derivable from independently plausible principles of counterfactual reasoning (see Appendix). We need an auxiliary principle corresponding to the case of vacuous truth, that something is a counterfactual consequence of its own negation only if it is a counterfactual consequence of everything (Lewis 1986: 132, axiom schema 4):

VACUITY $\neg A \mathrel{\Box\!\!\rightarrow} A \models C \mathrel{\Box\!\!\rightarrow} A$

We also need the further principle that a difference between logically equivalent antecedents makes no difference to the logical consequences of the conditional (compare the slightly stronger axiom schema (a7) in Stalnaker 1968 and the rule of Interchange of Logical Equivalents in Lewis 1986: 132); if A and A* are logically equivalent, what would be the case if A were the case is just what would be the case if A* were the case:

EQUIVALENCE If $A \Dashv\vDash A^{*}$ then $A \mathrel{\Box\!\!\rightarrow} B \Dashv\vDash A^{*} \mathrel{\Box\!\!\rightarrow} B$

Conversely, VACUITY is derivable from POSSIBILITY and the other principles. VACUITY has the extra advantage of making DEF(□) and DEF(◊) equivalent to Stalnaker's definitions. That is immediate for DEF(□); for DEF(◊) a trivial use of EQUIVALENCE is needed.

(II) Given Def(□), the validity of CLOSURE for the counterfactual conditional immediately implies the validity of a corresponding closure principle for necessity:

CLOSURE(□) If $B_1, \ldots, B_n \vDash C$

then $\Box B_1, \ldots, \Box B_n \vDash \Box C$

The logical consequences of necessary truths are themselves necessary truths. This is the distinctive principle that makes the logic of □ *normal* in the technical sense of modal logic (given that the background propositional logic is classical).[15]

(III) Still given Def(□), the principle that necessity implies truth follows from the generally accepted rule of modus ponens for the counterfactual conditional (see Lewis 1986: 132, axiom schema (6)):[16]

MODUS PONENS $A \mathrel{\Box\!\!\to} B, A \models B$

For by DEF(□), $\Box B \models (B \supset B) \mathrel{\Box\!\!\to} B$; since $\models B \supset B$, by MODUS PONENS $\Box B \models B$.

Thus, starting with the counterfactual conditional, we can build a promising theory of metaphysical necessity and possibility. The significance of this point does not depend on the assumption that our ordinary concept of the counterfactual conditional is really more primitive than our ordinary concepts of metaphysical necessity and possibility, or that DEF(□) and DEF(◊) provide strict synonyms. If we replace '$=_{\mathrm{def}}$' in them by mutual entailment, they are still strong enough for counterfactual thinking to address questions of metaphysical necessity and possibility. The capacity for thought about the metaphysical modalities cannot be isolated from the capacity for ordinary thought about the natural world, or excised without loss to the latter, for it is implicit in the latter.

Counterfactuals are often regarded as desperately vague and context-sensitive. Would DEF(□) and DEF(◊) transmit all that vagueness and context-sensitivity to □ and ◊, interpreted as metaphysical modalities? Not automatically. For instance, within a Lewis-Stalnaker framework, different readings or sharpenings of □→ might differ on the similarity ordering of worlds while still agreeing on what worlds there are, so that the differences would cancel out in the right-hand sides of DEF(□) and DEF(◊). If not all the differences do cancel out, perhaps the residual vagueness and context-sensitivity are genuine features of □ and ◊, interpreted as metaphysical modalities.

Discussions of the epistemology of modality often focus on imaginability or conceivability as a test of possibility: a notoriously unreliable test, unless imaginability is defined in terms of possibility.[17] Such discussions typically ignore the role of the imagination in evaluating counterfactuals. In doing so, the preceding considerations suggest, they omit the appropriate context for understanding the relation between metaphysical modality and the imagination. When we work out what would have happened if such-and-such had been the case, we frequently cannot do it without imagining such-and-such to be the case and letting things run.

Obviously, the use of the imagination in evaluating counterfactuals would generally be useless if it were not disciplined by background knowledge. Would she have left him if he had admitted the affair? You imagine him admitting the affair and her subsequent state of mind, in a way richly informed by your sense of what he and she are like. Although you could imagine them having quite different beliefs and desires, that would not help answer the question. Would the stone have reached the bottom of the slope if it had been dislodged? Although you could imagine the slope much smoother than it actually is, you do not. Even when the imaginative exercise is disciplined by one's background knowledge, one can easily misjudge the truth-value of the counterfactual, because one's background knowledge is inevitably incomplete and often over-estimated.

The imagination is a standard means for running our cognitive capacities 'offline' in evaluating counterfactuals. The process is manifestly fallible and practically indispensable. The same may well apply to the special cases tantamount to the evaluation of claims of metaphysical possibility. Of course, when one evaluates $\neg A \mathrel{\Box\!\!\rightarrow} A$ as true, something unusual has happened, for the antecedent is judged incapable of discriminating some possibilities from others. Nevertheless, provision for that special case must be built into our general means for processing counterfactuals. Even if we had a sieve to filter out counterfactuals with impossible antecedents, that would already involve making exactly the modal discriminations at issue. To have the counterfactual conditional and negation is to have the capacity to make modal distinctions, even if one is not very good at applying those distinctions in practice. At the very least, we cannot expect an adequate account of the role of the imagination in the epistemology of modality if we lack an adequate account of its role in the epistemology of counterfactuals.

In some loose sense, we may well have a special cognitive faculty or module dedicated to evaluating counterfactuals. It would have significant practical utility. If we wanted, we could call it ‘intuition’, although it would not in general be a priori. What seems quite unlikely is that we have a special cognitive faculty or module dedicated just to evaluating counterfactuals whose antecedents are incompatible with their consequents: the case is too special. Yet that is the crucial case for the metaphysical modalities. It is far more likely that the general cognitive capacities that enable us to evaluate counterfactuals whose antecedents are compatible with their consequents also enable us to evaluate counterfactuals whose antecedents are incompatible with their consequents, and therefore the metaphysical modalities. Indeed, the undecidability of first-order logical consequence implies that there is no general effective test for the incompatibility of the consequent of a counterfactual with its antecedent.

Our investigation of the possibility premise in the refutation of a philosophical theory by an imaginary counterexample has supported the results of the investigation of the other premise. Neither case involves *sui generis* philosophical thinking, a special faculty of rational intuition or the illusion of such. They simply involve particular applications of general cognitive capacities − most notably, the capacity to process counterfactuals − that are widely used throughout our cognitive engagement with the spatio-temporal world. A reasonable speculation is that their applications in philosophy have roughly the same degree of reliability as their applications elsewhere. Both crude rationalism and crude empiricism distort the epistemology of philosophy by treating it as far more distinctive than it really is. They forget how many things can be done in an armchair, including significant parts of natural science (doing thought experiments, making predictions, constructing explanations, …). That is not to say that philosophy *is* a natural science, for it also has much in common with mathematics. Armchair thinking dominates its method to a degree which the word 'natural' would merely obfuscate. But the epistemology of that thinking is fundamentally the same as the epistemology of thinking in any other position.[18]

Appendix

Derivation of POSSIBILITY from VACUITY given DEF(◊), CLOSURE and EQUIVALENCE:

(i)	$\neg\Diamond B \vDash A \mathrel{\Box\!\!\rightarrow} \neg B$	DEF(◊)
(ii)	$B, \neg B \vDash \neg A$	Non-modal logic
(iii)	$A \mathrel{\Box\!\!\rightarrow} B, A \mathrel{\Box\!\!\rightarrow} \neg B \vDash A \mathrel{\Box\!\!\rightarrow} \neg A$	(ii), CLOSURE
(iv)	$A \mathrel{\Box\!\!\rightarrow} B, \neg\Diamond B \vDash A \mathrel{\Box\!\!\rightarrow} \neg A$	(i), (iii)
(v)	$A \dashv\vDash \neg\neg A$	Non-modal logic
(vi)	$A \mathrel{\Box\!\!\rightarrow} \neg A \vDash \neg\neg A \mathrel{\Box\!\!\rightarrow} \neg A$	(v), EQUIVALENCE
(vii)	$A \mathrel{\Box\!\!\rightarrow} B, \neg\Diamond B \vDash \neg\neg A \mathrel{\Box\!\!\rightarrow} \neg A$	(iv), (vi)
(viii)	$\neg\neg A \mathrel{\Box\!\!\rightarrow} \neg A \vDash p \mathrel{\Box\!\!\rightarrow} \neg A$	VACUITY (*p* not in A)
(ix)	$\neg\neg A \mathrel{\Box\!\!\rightarrow} \neg A \vDash \neg\Diamond A$	(viii), DEF(◊)
(x)	$A \mathrel{\Box\!\!\rightarrow} B, \neg\Diamond B \vDash \neg\Diamond A$	(vii), (ix)
(xi)	$A \mathrel{\Box\!\!\rightarrow} B, \Diamond A \vDash \Diamond B$	(x)

Derivation of VACUITY from POSSIBILITY given DEF(◊), CLOSURE and REFLEXIVITY:

(i)	$\models \neg(A \,\&\, \neg A)$	Non-modal logic
(ii)	$\models p \mathrel{\Box\!\!\rightarrow} \neg(A \,\&\, \neg A)$	(i), CLOSURE (*p* not in A)
(iii)	$\models \neg\Diamond(A \,\&\, \neg A)$	(ii), DEF(◊)
(iv)	$\neg A \mathrel{\Box\!\!\rightarrow} (A \,\&\, \neg A), \Diamond\neg A \models \Diamond(A \,\&\, \neg A)$	POSSIBILITY
(v)	$\neg A \mathrel{\Box\!\!\rightarrow} (A \,\&\, \neg A) \models \neg\Diamond\neg A$	(iii), (iv)
(vi)	$\neg\Diamond\neg A \models C \mathrel{\Box\!\!\rightarrow} \neg\neg A$	Def(◊)
(vii)	$\neg\neg A \models A$	Non-modal logic
(viii)	$C \mathrel{\Box\!\!\rightarrow} \neg\neg A \models C \mathrel{\Box\!\!\rightarrow} A$	(vii), CLOSURE
(ix)	$\neg A \mathrel{\Box\!\!\rightarrow} (A \,\&\, \neg A) \models C \mathrel{\Box\!\!\rightarrow} A$	(v), (vi), (viii)
(x)	$A, \neg A \models A \,\&\, \neg A$	Non-modal logic
(xi)	$\neg A \mathrel{\Box\!\!\rightarrow} A, \neg A \mathrel{\Box\!\!\rightarrow} \neg A \models \neg A \mathrel{\Box\!\!\rightarrow} (A \,\&\, \neg A)$	(x), CLOSURE
(xii)	$\models \neg A \mathrel{\Box\!\!\rightarrow} \neg A$	REFLEXIVITY
(xiii)	$\neg A \mathrel{\Box\!\!\rightarrow} A \models \neg A \mathrel{\Box\!\!\rightarrow} (A \,\&\, \neg A)$	(xi), (xii)
(xiv)	$\neg A \mathrel{\Box\!\!\rightarrow} A \models C \mathrel{\Box\!\!\rightarrow} A$	(ix), (xiii)

Notes

<1> For discussion of the linguistic turn see Williamson 2004b.

<2> For a more extensive discussion of ‘intuitions’ in philosophy and further references see Williamson 2004a.

<3> For details of the reaction to Gettier 1963 see Shope 1983. See Weatherson 2003 for an interesting recent discussion of the role of intuition in the epistemology of Gettier cases along lines very different from those here. For the state of recent discussion of philosophical intuition see the papers in DePaul and Ramsey 1998.

<4> The assumption that propositions are the objects of knowledge is convenient, but not essential to the argument.

<5> The use of fictional names like ‘Jones’ in constructing imaginary examples may introduce a special element of pretence: that the sentences in which they occur express propositions. Usually, however, modal claims such as (2) below with no corresponding element of pretence are easily extracted from the discourse. The fictional names function as picturesque substitutes for bound variables such as ‘x’ and ‘p’.

<6> The operator □→ should be read as taking the widest possible scope.

<7> The description ‘non-modal’ is restricted to the structure that is overt in the given formulas; nothing is implied either way as to whether knowledge, justification or belief is modal.

<8> CLOSURE is not quite as straightforward as it looks, for if the inference from ‘A’ to ‘Actually A’ is counted as valid, then CLOSURE (with REFLEXIVITY) would imply that ‘If it had rained, it would have actually rained’ is valid; but that sentence is false on the rigid reading of ‘actually’ if it did not actually rain. Thus if the language contains operators such as ‘actually’, validity must be understood in a way that invalidates the inference from ‘A’ to ‘Actually A’, as truth-preservation with respect to all worlds, not just with respect to the actual world. The same understanding is required for the ordinary principle of Necessitation in modal logic (if $\models C$ then $\models \Box C$), to block the derivation of ‘Necessarily, if it rains then it actually rains’. Necessitation corresponds to the special case of CLOSURE with no premises ($n = 0$): if $\models C$ then $\models A \mathrel{\Box\!\!\rightarrow} C$. Lewis states his rule only for $n \geq 1$, but the case for $n = 0$ is easily derived from the case for $n = 0$ and REFLEXIVITY.

<9> The point in the text is connected to the well-known problem of old evidence that has already been priced into current probabilities. The claim is not that the problem refutes all abstract probabilistic analyses; for one account and further references see Williamson 2000: 220-221. The point is rather that in practice we need counterfactual conditionals to handle some concrete evidential situations.

<10> See Williamson 2003 for more on this theme.

<11> See Weinberg, Stich and Nichols 2001. On the present view, the cultural variations for which they find evidence are not very disturbing for philosophy. Most intellectual disciplines rely at various points on claims over which there is cultural variation. Eye-witnesses often disagree in their descriptions of recent events, but it would be foolish to react by dismissing all eye-witness reports. It would not be at all surprising if the reliability of a person's application of concepts such as *knowledge* and *justification* to particular cases can be improved by training (compare the training of lawyers in the careful application of very general concepts to specific cases). It is a plausible empirical hypothesis that similar cultural variations arise in the application of epistemic concepts to empirically encountered cases. The natural sciences rely on such applications: for example, in judging that one theory is better confirmed than another or that an experimental result is not to be trusted because the apparatus was not known to be working correctly. In particular, Weinberg, Stich and Nichols rely on such judgements in drawing conclusions from their data.

<12> The warning in n. 8 also applies to the notion of logical incompatibility here.

<13> Counterfactuals are sometimes claimed to lack truth-values (Edgington 2003), but that view makes little sense of their embedded occurrence in readily intelligible and evaluable sentences such as 'Everyone who would have gained if the measure had been

passed voted for it' and 'Most of the area that would have been flooded if the dam had burst was forested'. See Bennett 2003: 252-6 for further discussion and references.

<14> The occurrence of '*p*' in sentence position in DEF(□) requires that it also be assigned sentence position in (1)-(4), as is quite natural; 'S knows that' takes a sentential complement. The quantification into sentence position need not be understood substitutionally. In purely modal contexts it can be modelled as quantification over all sets of possible worlds, even if not all of them are intensions of sentences that form the substitution class, although this modelling presumably fails for hyperintensional contexts such as 'K(*x*,*p*)'. A more faithful semantics for it might use non-substitutional quantification into sentence position in the meta-language. Such subtleties are inessential for present purposes.

<15> The warning in n. 8 also applies to both EQUIVALENCE and CLOSURE(□).

<16> One can accept a counterfactual when rationally unwilling to apply modus ponens to it, in the sense that on learning its antecedent one would reject the counterfactual rather than accept its consequent. For example, I accept 'If Oswald had not shot Kennedy, Kennedy would not have been shot', but if I come to accept 'Oswald did not shoot Kennedy' I will not conclude 'Kennedy was not shot'. But that is no threat to the validity of modus ponens. In circumstances in which both 'If Oswald had not shot Kennedy, Kennedy would not have been shot' and 'Oswald did not shoot Kennedy' are true, so is 'Kennedy was not shot'.

<17> For recent examples see the essays in Gendler and Hawthorne 2002.

<18> I thank the Centre for Advanced Study at the Norwegian Academy of Science and Letters in Oslo for its hospitality during the writing of this paper. The real-life Gettier case in the text was realized and later discussed in a lecture at the University of Bilkent in Ankara. I thank audiences there, at a conference on Modalism and Mentalism in Contemporary Epistemology hosted by Aarhus University in Copenhagen and at classes in Oxford for helpful comments on much earlier versions of this material. I also thank Dorothy Edgington, Pascal Engel, Olav Gjelsvik and Scott Sturgeon for useful suggestions. The paper develops a theme adumbrated in Williamson 2004a, from which it inherits older debts; the two papers complement each other.

Bibliography

Bennett, J. 2003. *A Philosophical Guide to Conditionals*. Oxford: Clarendon Press.

DePaul, M., and Ramsey, W. (eds.) 1998. *Rethinking Intuition: The Psychology of Intuition and its Role in Philosophical Inquiry*. Lanham, Maryland: Rowman and Littlefield.

Edgington, D. 2003. 'Counterfactuals and the benefit of hindsight', in P. Dowe and P. Noordhof (eds.), *Causation and Counterfactuals*. London: Routledge.

Gendler, T. Szabó and J. Hawthorne (eds.) 2002. *Conceivability and Possibility*. Oxford: Clarendon Press.

Gettier, E. 1963. 'Is justified true belief knowledge?', *Analysis* **23**: 121-3.

Lewis, D. 1986. *Counterfactuals*, revised edn. Cambridge, Mass.: Harvard University Press.

Shope, R. 1983. *The Analysis of Knowing: A Decade of Research*. Princeton: Princeton University Press.

Stalnaker, R. 1968. 'A theory of conditionals', in *American Philosophical Quarterly Monographs* **2** (*Studies in Logical Theory*): 98-112.

Weatherson, B. 2003. 'What good are counterexamples?', *Philosophical Studies* **115**: 1-31.

Weinberg, J., Stich, S., and Nichols, S. 2001. 'Normativity and epistemic intuitions', *Philosophical Topics* **29**: 429-60.

Williamson, T. 2000. *Knowledge and its Limits.* Oxford: Oxford University Press.

Williamson, T. 2003. 'Understanding and inference', *Aristotelian Society* sup. vol. **77**:

249-93.

Williamson, T. 2004a. ‘Philosophical “intuitions” and skepticism about judgement’, *Dialectica* **58** (2004): 109-153.

Williamson, T. 2004b. ‘Past the linguistic turn?’, in B. Leiter (ed.), *The Future for Philosophy*, Oxford: Oxford University Press.