Gettier Cases in Epistemic Logic

[penultimate draft; final version to appear in a symposium in *Inquiry*]

Timothy Williamson

University of Oxford

Abstract: The possibility of justified true belief without knowledge is normally motivated by informally classified examples. This paper shows that it can also be motivated more formally, by a natural class of epistemic models in which both knowledge and justified belief (in the relevant sense) are represented. The models involve a distinction between appearance and reality. Gettier cases arise because the agent's ignorance increases as the gap between appearance and reality widens. The models also exhibit an epistemic asymmetry between good and bad cases that sceptics seem to ignore or deny.

1. Thought experiments are a source of unmysterious philosophical knowledge. I have defended that statement elsewhere (Williamson 2007a). Nevertheless, when conclusions so reached are subsequently corroborated by other forms of argument, the latter need not be redundant. We are hardly good enough at philosophy to dispense with independent checks on our working. Moreover, less case-oriented methods of argument promise deeper theoretical insight into phenomena whose sheer occurrence has already been recognized.

Take Gettier cases. Since the publication of Gettier's paper (1963), there has been a consensus in epistemology that cases such as he presents are counterexamples to the view that justified true belief is equivalent to, or at least sufficient for, knowledge. The consensus survives, despite experiments that have been claimed to show that it depends on the ethnicity or gender of those who evaluate the cases (Weinberg, Nichols, and Stich 2001; Buckwalter and Stich 2011). More recent experiments suggest that the threat to the consensus may have been a false alarm, by calling into question the robust replicability of the results (Nagel 2012, 201X; Stich 201X dissents). Nevertheless, it would be reassuring to have some independent way of checking that in Gettier cases the subject does not know after all. Even granted that reflective verdicts on such cases do not vary significantly with ethnicity or gender, one can worry that those verdicts might still reflect some oddity of the human cognitive system for ascribing 'knowledge', rather than a genuine feature of the underlying epistemological kind to which the term refers (Weatherson 2003). An argument cast within a more general theoretical framework has a fair chance of bypassing any such putative oddity.

This paper uses the more general theoretical framework of epistemic and doxastic logic, in the tradition going back to Hintikka (1962). Models will be constructed in which justified true belief in the relevant sense can be compared with knowledge; cases of the former without the latter will be shown to arise under reasonable conditions.

We must take care about methodology. If we simply introduce three primitive operators for justification, belief, and knowledge respectively, and permit them to vary independently of each other, then modelling justified true belief without knowledge is trivial. By the same token, such models constitute no serious evidence for the genuine possibility of that combination. The point is instead to minimize the number of moving parts in the models, using reasonable idealizations and other well-motivated constraints, for if we can show that *all* models of the type so characterized contain cases of the required combination, that *is* significant for its genuine possibility. Relaxing the constraints will not make the combination less possible. What we seek from the model-building exercise is an independent test, not conclusive proof, of whether justified true belief implies knowledge. As in the natural sciences, the proper use of formal models involves an ineliminable informal element of good judgment. Readers who dislike the models in this paper can try to build better ones, to be judged by the same standard.

The models described below turn out to have other uses too. For example, they exhibit epistemic asymmetries of a sort that sceptical arguments have been accused of ignoring, and so provide independent evidence for an obstacle to scepticism.

2. For present purposes, an *epistemic model* is simply an ordered pair <W, R>, where W is any nonempty set and R is any binary relation (in extension) over W, that is, a set of ordered pairs of members of W.

Informally, we think of the members of W as *worlds*, or maximally specific states, although that informal interpretation is no part of the formal definition of what it is to be an epistemic model. The *propositions* in the model are just the subsets of W. Thus propositions are identified with sets of worlds. The variables *w*, *x*, … will be used to range over worlds, and the variables *p*, *q*, … to range over propositions. A proposition *p* is *true* in a world *w* if and only if *w* belongs to *p*; otherwise *p* is false in *w*. Consequently, a proposition *p entails* a proposition *q* if and only if *p* is a subset of *q*: *q* is true in every world in which *p* is true. Hence propositions are identical if and only if they are mutually entailing, true in the same worlds.

The model treats propositions as coarse-grained, without internal structure corresponding to the semantic structure of sentences that express them. Of course, such a treatment looks like an obstacle to understanding epistemic puzzles where agents seem to take different attitudes to the propositions expressed by 'Hesperus is bright' and 'Phosphorus is bright', or by '2 + 2 = 4' and 'There is no largest prime number' — although it is doubtful that even a finer-grained Russellian or Fregean treatment of propositions will solve all such puzzles by itself. In any case, even if the coarse-grained treatment is merely an idealization, it is a harmlessly simplifying one for our purposes. For given a model of justified true belief without knowledge for coarse-grained propositions, we can turn it into a model of justified true belief without knowledge for fine-grained propositions by associating in some uniform manner each coarse-grained

proposition in the model with one of the fine-grained propositions true at the same worlds. The distinction between coarse-grained and fine-grained propositions is not what does the work in the models below. Gettier cases and Frege cases raise very different problems.

All the epistemology in an epistemic model <W, R> is encoded in its second element, R. Informally, we think of R as the relation of *epistemic accessibility* for a given agent at a given time. In other words, a world *x* is accessible from a world *w* (*w*R*x*) if and only if whatever the agent then knows in *w* is true in *x*: for all the agent knows in *w*, she is in *x* (*x* is epistemically possible in *w*). For any proposition *p*, let K*p* be the proposition true in a world *w* if and only if *p* is true in every world accessible from *w*: K*p* = $\{w\colon \forall x\,(w\mathrm{R}x \Rightarrow x \in p)\}$. Informally, K*p* is identified with the proposition that the agent knows *p*. Thus the agent counts as knowing whatever is true in all the worlds that for all she knows she is in. This is in no deep sense an analysis of knowledge in terms of epistemic accessibility, but simply the imagined decoding of the epistemic information encoded in the relation R.

The treatment of K as a knowledge operator according to the definition above enforces a form of logical omniscience for knowledge beyond that already implicit in the coarse-grained treatment of propositions. Specifically, if the agent knows each premise of a valid deductive argument, then the agent ipso facto knows the conclusion. For since the argument is valid, the conjunction of the premises entails the conclusion, in other words, the intersection of the premises is a subset of the conclusion. Hence if the agent knows each premise in a world *w*, and a world *x* is accessible from *w*, then each premise is true in *x*, so the conclusion is true in *x*; thus the conclusion is true in every world accessible

from *w*, so she knows the conclusion in *w*. This logical omniscience too can be regarded as a harmlessly simplifying idealization of the models. If even perfect logicians are susceptible to Gettier cases, humans with more limited logical skills should not expect to be immune.

A convenient feature of epistemic models is that in each world *w* there is a strongest proposition R(*w*) known by the agent. It is simply the set of accessible worlds: R(*w*) = {*x*: *w*R*x*}. For, unpacking the definition of K, she knows R(*w*) in *w*, and for any proposition *p*, if she knows *p* in *w* then R(*w*) entails *p*. In terms of membership, R(*w*) is the smallest proposition the agent knows in *w*. More simply, R(*w*) is what she knows in *w*. It is the horizon of open epistemic possibility.

Knowledge is factive: whatever is known is true. That uncontentious principle requires the relation R to be reflexive, for if some world *w* were not accessible from itself, K*p* would be true but *p* false in *w*, where *p* is the proposition containing all worlds except *w*. Conversely, suppose that R is reflexive; let *q* be any proposition and *x* any world. Then K*q* is true in *x* only if *q* is true in every world accessible from *x*; since *x* is accessible from itself, *q* is true in *x*. Thus K is factive. All the models considered below have a reflexive accessibility relation.

3. To make progress, we must consider a more specific class of epistemic models. We do so by introducing a very simple form of the distinction between reality and appearance. Imagine the agent gaining perceptual knowledge of some environmental parameter (such as temperature) that takes values in a set E. For simplicity, imagine further that the parameter always appears to her in a maximally specific way, in the sense that exactly

one member of E appears to her to be the value of the parameter (for example, the temperature appears to be exactly 30 degrees Celsius). For present purposes it is unnecessary to build in the complications of unspecific appearances. There is no limit in principle to how far the apparent value of the parameter can diverge from its real value at a given time. As another idealization, we identify worlds with ordered pairs of members of E ($W = E^2$). Informally, the first member of an ordered pair represents the real value of the parameter; the second member represents its apparent value to the agent. Such impoverished worlds contain enough information to be adequate for present purposes; everything else can be ignored.

We treat present appearances to the agent as transparent to her: she knows all about them, in the sense that no world where the parameter has a different apparent value is epistemically possible for her. Formally, if $\langle e, f\rangle R\langle e^*, f^*\rangle$ then $f = f^*$. Although appearances are not in fact so epistemically privileged (Williamson 2000, pp. 93-109), treating them as such is yet another harmlessly simplifying concession for present purposes. If even agents to whom appearances are transparent suffer from Gettier cases, humans with more limited knowledge of their appearances should not expect to be immune.

A mathematical temptation is to make the stronger biconditional stipulation that if $f = f^*$ then $\langle e, f\rangle R\langle e^*, f^*\rangle$ as well as vice versa. That temptation must be resisted, for it is tantamount to scepticism about the external world. It implies that the agent has no nontrivial knowledge at all of the real value of the parameter, for example of the external temperature. For let p be a non-trivial proposition about the real value. It is non-trivial in the sense that it is false in at least one world $\langle g, h\rangle$. It is about the real value in the sense

that its truth-value depends only on that value: *p* has the same truth-value in any two worlds with the same real value. Consider the agent in a world <*e*, *f*>. Since the real value is the same in <*g*, *h*> and <*g*, *f*>, and *p* is false in <*g*, *h*>, *p* is also false in <*g*, *f*>. But the envisaged converse stipulation implies that <*e*, *f*>R<*g*, *f*>, since the apparent value is the same in the two worlds. Thus *p* is false in a world epistemically accessible from <*e*, *f*>, so in <*e*, *f*> the agent does not know *p*: if the agent knows only appearances, for all she knows in <*e*, *f*> she is in <*g*, *f*>. The stronger, biconditional stipulation makes her perfectly knowledgeable about appearances but perfectly ignorant of the corresponding realities, even though we started with a scenario in which she was gaining ordinary perceptual knowledge of her environment. To avoid such crass scepticism, we must permit some worlds to be inaccessible from others with the same appearance.

A better picture is this. When the real value matches the apparent value, the agent knows *something* non-trivial about the real value. For example, when the temperature both is and appears to be 30 degrees, she knows that it is not zero. Some world with a different real value but the same apparent value is *not* epistemically possible for her. However, she still does not know *everything* about the real value, for her perceptual apparatus is not perfectly discriminating. For example, when the temperature is 30.0006 degrees, she does not know that it is not 30.0007 degrees. Some world with a different real value but the same apparent value *is* epistemically possible for her. The latter world is distinct from the one she is in, which is of course epistemically accessible from itself by factiveness. Formally, for any value *f* in E, the worlds accessible from the world <*f*, *f*> with identical real and apparent values are <*f*, *f*> itself (trivially, by reflexivity) and at least one other world:

#1 $\{\langle f, f\rangle\} \subset R(\langle f, f\rangle)$

To complete the picture, a natural postulate is that as the real value diverges more and more from the given apparent value, the agent knows less and less about the real value. To state the relation formally, we postulate a *metric* on E, in the mathematical sense of the term, to measure distance in the space of values. The metric is a function d from pairs of values in E to real numbers; d(*e*, *f*) measures the difference in value between *e* and *f*. The standard mathematical definition of a metric also requires d to have the following properties, for all *e*, *f*, *g* in E: $0 \leq d(e, f)$ (no distance is negative); $d(e, f) = 0$ if and only if $e = f$ (no value is as close to any other as any value is to itself); $d(e, f) = d(f, e)$ (distance is symmetric); $d(e, g) \leq d(e, f) + d(f, g)$ (the triangle inequality: metaphorically, the shortest journey in quality space from *e* to *g* is no longer than the journey from *e* to *f* followed by that from *f* to *g*).

We can now state the constraint on the accessibility relation R that ignorance grows with the distance of the real value from the apparent value:

#2 $d(e, f) \leq d(e^*, f)$ if and only if $R(\langle e, f\rangle) \subseteq R(\langle e^*, f\rangle)$

In other words, if the real value diverges no more from the given apparent value in the world $\langle e, f\rangle$ than it does in the world $\langle e^*, f\rangle$, then the agent knows at least as much in the former world as in the latter; every world she can exclude in the latter she can already exclude in the former. Conversely, the agent knows at least as much in one world as in

another only if the gap between appearance and reality is no wider in the former than in the latter.

The rationale for #2 is that any increase in the gap between appearance and reality has an epistemic cost for the agent: more knowledge is lost. However, #2 is not claimed to be an exceptionless law of epistemology. Rather, #2 is a natural idealization of a wide range of cases.

We check that our stipulations are reasonable by constructing a natural model that verifies them. Let E be the set of real numbers, so a world is an ordered pair of a real-valued real value and a real-valued apparent value. For real numbers e and f, let d(e, f) be the absolute distance $|e - f|$, so d is a metric. Let c be a positive real number; the epistemic structure of the model will not depend on the particular value of c. We define the relation R thus: $<e, f>R<e^*, f^*>$ if and only if $|e^* - f^*| \leq |e - f| + c$ and of course $f = f^*$. Thus the worlds accessible from a given world are those where the apparent value is exactly the same and the gap between it and the real value exceeds the gap in the given world by at most the constant c. Obviously R is reflexive. In any world, what the agent knows about the real value is that it falls within some margin for error of the apparent value; she does not know exactly what that margin is, since it varies across the accessible worlds.

Note that $R(<e, f>) = \{(<g, f>: |g - f| \leq |e - f| + c\}$. In particular, therefore, $R(<f, f>) = \{(<g, f>: |g - f| \leq c\}$. Thus $<f, f>R<f + c, f>$, so #1 holds. To check #2, first suppose that $|e - f| \leq |e^* - f|$. Then $R(<e, f>) = \{(<g, f>: |g - f| \leq |e - f| + c\} \subseteq \{(<g, f>: |g - f| \leq |e^* - f| + c\} = R(<e^*, f>)$. Conversely, suppose that $R(<e, f>) \subseteq R(<e^*, f>)$. Let $g = f + |e - f| + c$, so $|g - f| = |e - f| + c$. Hence

$<g, f> \in R(<e, f>)$, so $<g, f> \in R(<e^*, f>)$, so $|e - f| + c = |g - f| \leq |e^* - f| + c$, so $|e - f| \leq |e^* - f|$, as required. Thus the model verifies both constraints, #1 and #2.

In the rest of the paper, we explore some epistemic consequences of the constraints #1 and #2 themselves. For the sake of generality, we avoid reliance on the model just sketched. Nevertheless, the model may help the reader visualize the arguments below more clearly.[1]

4. One obvious consequence of the constraints is that whatever is epistemically possible in a world in which appearance matches reality is epistemically possible in any world with the same appearance. For #2 implies that since $d(f, f) = 0 \leq d(e, f)$ by the axioms for a metric, $R(<f, f>) \subseteq R(<e, f>)$. In particular, since the world in which appearance matches reality is accessible from itself, it is accessible from any world with the same appearance; $<f, f> \in R(<e, f>)$.

We can go further. By #1, $R(<f, f>)$ contains some world $<f^{\#}, f>$ other than $<f, f>$, so $f^{\#} \neq f$. Although appearance in fact matches reality perfectly, and the agent knows the appearance perfectly, she does not know the reality perfectly: for all she knows, it is $f^{\#}$ rather than f. Now suppose that $R(<f^{\#}, f>) \subseteq R(<f, f>)$. Then by #2, $d(f^{\#}, f) \leq d(f, f) = 0$, so $f^{\#} = f$ by the axioms for a metric, contrary to hypothesis. Thus $R(<f^{\#}, f>)$ contains some world $<f^{\#\#}, f>$ not in $R(<f, f>)$. Consequently, the accessibility relation is non-transitive. For $<f, f>R<f^{\#}, f>$ and $<f^{\#}, f>R<f^{\#\#}, f>$, but not $<f, f>R<f^{\#\#}, f>$. A natural interpretation of this result is that when appearance matches reality, for all the agent knows there is a small gap between them, and when there is a small gap between appearance and reality, for all she knows there is a larger gap, but when appearance matches reality, the agent

does know that there is not the larger gap between appearance and reality (scepticism is false). As always in epistemic logic, the non-transitivity of accessibility yields failures of the KK principle, also known as positive introspection and the 4 axiom of modal logic, that if one knows *p* then one knows that one knows *p* (for discussion see Williamson 2000, pp. 114-34). Specifically, in the world $\langle f, f \rangle$ the agent knows $R(\langle f, f \rangle)$, but she does not know that she knows $R(\langle f, f \rangle)$, because she does not know $R(\langle f, f \rangle)$ (even though it is true) in the accessible world $\langle f^{\#}, f \rangle$. There is no need to labour the failure of the KK principle here.

A more interesting consequence of the constraints is that accessibility is also non-symmetric, for $\langle f^{\#\#}, f \rangle R \langle f, f \rangle$ but not vice versa. A natural interpretation of this result is that there is an epistemic asymmetry between the good (anti-sceptical) scenario $\langle f, f \rangle$ where appearance matches reality and the bad (sceptical) scenario $\langle f^{\#\#}, f \rangle$ where appearance diverges widely from reality. In the bad case, for all the agent knows she is in the good case; by contrast, in the good case, the agent knows that she is not in the bad case. Sceptics often neglect the possibility of such epistemic asymmetries (for discussion see Williams 1978, pp. 310-13, Humberstone 1988, and Williamson 2000, pp. 164-8). As always in epistemic logic, the non-symmetry of accessibility yields failures of the 'Brouwerian' axiom B that if *p* is false then one knows that one does not know *p*. Specifically, in the bad case $\langle f^{\#\#}, f \rangle$ the proposition $R(\langle f, f \rangle)$ is false, but the agent does not know that she does not know $R(\langle f, f \rangle)$, because for all she knows she is in the good case $\langle f, f \rangle$, in which she does know $R(\langle f, f \rangle)$. The general failure of the B axiom for knowledge was pointed out by Hintikka (1962). It immediately yields failures of the 'negative introspection' principle (known as axiom 5 or E in modal logic) that if one does

not know *p* then one knows that one does not know *p*, for by factiveness if *p* is false then one does not know *p*. Although many researchers in epistemic logic continue to treat the accessibility relation for knowledge as symmetric and transitive as well as reflexive, and the assumptions are often harmless idealizations for their purposes, they are not harmless for purposes of serious epistemology.

We can also extract more positive consequences from the constraints. For example, we can show that the *only* world in which the agent knows R(<*f*, *f*>) is <*f*, *f*> itself. For suppose that the agent knows R(<*f*, *f*>) in a world <*e*, *g*>. By the structure of R, *g* = *f*. Thus KR(<*f*, *f*>) is true in <*e*, *f*>. Since R(<*e*, *f*>) is what she knows in <*e*, *f*>, R(<*e*, *f*>) $\subseteq$ R(<*f*, *f*>). Therefore, by #2, d(*e*, *f*) $\leq$ d(*f*, *f*) = 0, so *e* = *f*. Thus <*e*, *g*> = <*f*, *f*>, as required. Elsewhere, I have shown how such propositions known in only one world generate cases where the agent knows a truth *p* even though it is virtually certain on her own current evidence that she does not know *p*. Details are omitted here, because they involve the introduction of an apparatus of evidential probabilities.[2]

The consequences just drawn from our constraints have a strongly externalist feel. However, the constraints themselves were not motivated by externalism, but only by a preference for simplicity over complexity, and for anti-scepticism over scepticism. In particular, the agent was granted a knowledge of appearances as complete as internalists could desire. That makes the externalist upshot all the more significant.

5. To discuss Gettier cases, we must interpret belief and justification over the models. That is the task of this section.

In epistemic logic, the normal way to interpret belief is by adding another binary relation S between worlds as a further constituent of the model, to represent *doxastic accessibility* for the given agent at the given time. A world x is doxastically accessible from a world w (wSx) if and only if whatever the agent believes in w is true in x: x is doxastically possible in w. For any proposition p, let Bp be the proposition true in a world w if and only if p is true in every world doxastically accessible from w: Bp = $\{w\colon \forall x\,(w\mathrm{S}x \Rightarrow x \in p)\}$. Informally, B$p$ is identified with the proposition that the agent believes p. Thus the agent counts as believing whatever is true in all the worlds that for all she believes she is in. The treatment of B as a belief operator enforces a form of logical omniscience for belief beyond that already implicit in the coarse-grained treatment of propositions. Specifically, if the agent believes each premise of a valid deductive argument, then the agent ipso facto believes the conclusion, by the same reasoning as for K. As before, this logical omniscience too can be regarded as a harmlessly simplifying idealization of the models.

Since belief is not factive — some falsehoods are believed — the relation S is not required to be reflexive, unlike R. However, S may be required to be serial, in the sense that every world has S to at least one world. For if no world were doxastically accessible from a world w, then every proposition would count vacuously as believed in w, making the agent inconsistent. Although one can be irrational and have inconsistent beliefs, it is doubtful that one can have inconsistent justified beliefs.[3] Conversely, if from each world at least one world in the model is accessible, then neither Bp & B$\neg p$ nor B(p & $\neg p$) is true at any world: the agent never believes a contradiction, either in the sense of believing both of a pair of contradictories or in the sense of believing their conjunction.

A natural assumption is that if one knows *p*, one believes *p*. It is equivalent to the constraint that doxastic accessibility implies epistemic accessibility. For suppose that a world *x* is not epistemically accessible from a world *w*. Then in *w* the agent knows the proposition *p* containing all worlds except *x*. So if knowledge requires belief, in *w* the agent believes *p*, so *x* is not doxastically accessible from *w*. Conversely, if doxastic accessibility implies epistemic accessibility, then whatever is true in all epistemically accessible worlds is true in all doxastically accessible worlds, so knowledge requires belief. Let us accept that requirement.

In epistemic models of the sort introduced above, we can define the doxastic accessibility relation S well enough for present purposes without adding it as a separate component. To motivate the definition, let us make the simplifying but natural assumption that what the agent believes, unlike what she knows, depends only on the apparent value of the relevant parameter: for any values *e*, *e**, and *f*, exactly the same worlds are doxastically accessible from <*e*, *f*> as from <*e**, *f*>. Although an agent's beliefs could be sensitive to factors other than appearance, we focus on agents who do not base their beliefs on such factors. Since the agent has the same beliefs in a world <*e*, *f*> as in the world <*f*, *f*>, all that remains is to determine what the agent believes in a world where appearance matches reality. Since knowledge requires belief, whatever she knows in <*f*, *f*> she believes in <*f*, *f*>. Conversely, there is no relevant reason to attribute to her beliefs that fail to constitute knowledge in <*f*, *f*>, the good case where appearances match reality perfectly. Thus we may assume that whatever she believes in <*f*, *f*> she knows in <*f*, *f*>. Consequently, she believes *p* in <*e*, *f*> if and only if she knows *p* in <*f*, *f*>. Thus doxastic accessibility is definable in terms of epistemic accessibility in our models: <*e*,

f>S<*e**, *f**> if and only if <*f*, *f*>R<*e**, *f**> (so *f* = *f**). What the agent believes in a case is what she knows in the corresponding good case. As with #1 and #2, this is not claimed to be an exceptionless law of epistemology; rather, it is a natural idealization of a wide range of cases.

One consequence of our definition is that S is serial, for <*e*, *f*>S<*f*, *f*> since <*f*, *f*>R<*f*, *f*>. Thus the agent's beliefs are always consistent. Moreover, a perfect match between appearance and reality is always doxastically possible.

For knowledge to imply belief, doxastic accessibility must imply epistemic accessibility. Our definition of the former in terms of the latter yields that result, given our constraints. For suppose that <*e*, *f*>S<*e**, *f**>. By the definition of S, <*f*, *f*>R<*e**, *f**>, so *f* = *f**. But by #2 R(<*f*, *f*>) $\subseteq$ R(<*e*, *f*>). Since <*e**, *f*> $\in$R(<*f*, *f*>), <*e**, *f**> = <*e**, *f*> $\in$R(<*e*, *f*>), so <*e*, *f*>R<*e**, *f**>, as required. That is some evidence that our definition of doxastic accessibility is a reasonable one.

The agent in our models is omniscient about her own current beliefs, in the sense that if she believes *p* then she knows that she believes *p*, and if she does not believe *p* then she knows that she does not believe *p*. For what she believes supervenes in the model on appearances to her, and she is omniscient about those appearances. More precisely, the definition of S ensures for all *e*, *e**, and *f* in E, B*p* is true in <*e*, *f*> if and only if it is true in <*e**, *f*>. Consequently, if B*p* is true in <*e*, *f*> then it is true in every world epistemically accessible from <*e*, *f*>, so KB*p* is true in <*e*, *f*>. Likewise, if B*p* is false in <*e*, *f*> then it is false in every world epistemically accessible from <*e*, *f*>, so K¬B*p* is true in <*e*, *f*>.

The agent does *not* satisfy the principle that if she believes *p* then she also believes that she knows *p* (it is an axiom of the system in Stalnaker 2006).[4] For example, in $\langle f, f\rangle$, she believes, and indeed knows, $R(\langle f, f\rangle)$, but we noted in section 4 that she fails to know $R(\langle f, f\rangle)$ in the other worlds epistemically accessible from $\langle f, f\rangle$; since those worlds are also doxastically accessible from $\langle f, f\rangle$, she neither knows nor believes in $\langle f, f\rangle$ that she knows $R(\langle f, f\rangle)$. This is an admirable form of epistemic modesty on her part. In the model, she cannot know that she knows $R(\langle f, f\rangle)$; she avoids believing that she knows $R(\langle f, f\rangle)$ because in the model such a belief cannot constitute knowledge.

6. We have modelled an internally rational agent whose beliefs depend only on appearances. Whatever she believes in a world, she knows in some world with the same appearance. In any world in such a model, all her beliefs are justified in an internalist sense. In some worlds, she has justified false beliefs. We saw in section 4 that the world $\langle f^{\#\#}, f\rangle$ is epistemically inaccessible from the world $\langle f, f\rangle$: in the good case, the agent has non-trivial knowledge of the real value of the parameter. By the definition of S, $\langle f^{\#\#}, f\rangle$ is doxastically inaccessible from itself. Thus in $\langle f^{\#\#}, f\rangle$ the agent has a justified false belief that she is not $\langle f^{\#\#}, f\rangle$. In an internalist sense, she is justified in believing the gap between appearance and reality to be narrower than it really is. Similarly, internalists often say that a brain in a vat is justified in believing that it is not a brain in a vat.

In a strongly externalist sense, a belief is *fully* justified only if it constitutes knowledge, and is therefore true. Although one may have a cast-iron excuse for a false belief, that excuse does not amount to a justification. Similarly, consider a competent and non-negligent surgeon who causes the death of a patient because a lab technician with a

grievance switched the labels on two bottles: he has a cast-iron *excuse* for killing the patient, but it does not amount to a *justification* for killing the patient (Williamson 2007b). Trivially, in that externalist sense, there are no cases of justified true belief without knowledge; indeed, there are no cases of justified belief without knowledge. The interest of Gettier cases concerns a more internalist sense of 'justified', closer to excusability than to genuine justification. Gettier himself seems to have had such a reading in mind; he emphasizes that someone can be justified in believing a false proposition (Gettier 1963). After all, a sense of 'justification' explained in terms of 'knowledge' would not help the project of analysing knowledge in terms of justification and other factors. The internalist sense of 'justified' in which the beliefs of the agent in the model are justified even when false is much more relevant to the debate over Gettier cases.

Every model of the kind we have described also contains cases of justified true belief without knowledge: Gettier cases. For recall the scenario derived in section 4 from our constraints: a world $<f^{\#\#}, f>$ is epistemically accessible from a world $<f^{\#}, f>$, which is in turn epistemically accessible from $<f, f>$, although $<f^{\#\#}, f>$ is not itself epistemically accessible from $<f, f>$. The proposition $R(<f, f>)$ is true in $<f^{\#}, f>$. Moreover, the agent believes $R(<f, f>)$ in $<f^{\#}, f>$, for by the definition of S what she believes in $<f^{\#}, f>$ is what she knows in $<f, f>$, and she knows $R(<f, f>)$ in $<f, f>$. We have just seen that in the relevant models the agent's beliefs count as justified in the relevant sense. Thus in $<f^{\#}, f>$ the agent has a justified true belief in $R(<f, f>)$. However, in $<f^{\#}, f>$ she does not *know* $R(<f, f>$, for $R(<f, f>)$ is false in the epistemically accessible world $<f^{\#\#}, f>$. Thus in

$<f^{\#}, f>$ with respect to R($<f, f>$), the agent has a justified true belief that is not knowledge. Our constraints predict Gettier cases in epistemic logic.

The Gettier cases just described are not of the traditional type, in which the agent derives her justified true belief from a justified false belief in something else. Our Gettier cases arise in $<f^{\#}, f>$, a world in which the agent lacks false beliefs: for if she believes p in $<f^{\#}, f>$ then she knows p in $<f, f>$, so p is true in $<f^{\#}, f>$ because the latter is epistemically accessible from $<f, f>$. The cases above are more like 'fake barn' Gettier cases (Goldman 1976), in which the agent may lack relevant false beliefs but still the circumstances are not favourable enough for knowledge of the given truth. In our case, the unfavourable circumstance is the gap between appearance and reality, even though the agent does not doxastically endorse the appearance specifically enough to produce falsity.

Some but not all of our models also contain Gettier cases of a more traditional type. For example, in the real-valued models described in section 3, the agent in the world $<2c, 0>$ has a justified belief in the false proposition R($<0, 0>$), which is true just in worlds of the form $<e, 0>$ where $-c \leq e \leq c$. Since her beliefs are closed under entailment, in $<2c, 0>$ she also believes the 'disjunctive' proposition R($<0, 0>$) $\vee$ $\{<2c, 0>\}$. Her belief in the disjunction is justified in $<2c, 0>$ because she has a justified belief in its left disjunct, although no belief in its right disjunct. Moreover, in $<2c, 0>$ the disjunction is true because its right disjunct is true, although its left disjunct is false. But in $<2c, 0>$ she does not know the disjunction, for it is false in the world $<3c, 0>$, which is epistemically (though not doxastically) accessible from $<2c, 0>$. That is much closer to Gettier's original cases.[5]

With respect to Gettier cases, the informal judgments of epistemologists and a more formal model-building methodology in epistemic logic converge. One can arrive at the same conclusion either way, and each method lends support to the other. The Gettier effect is robust.[6]

Notes

1 There are also less natural models of all the constraints. Here is one. Let E = {0, 1, 2}. Let R hold between all pairs in W with just these exceptions: <0, 0> does not have R to <1,0>; <1, 1> does not have R to <2, 1>; <2, 2> does not have R to <0, 2>. The metric d is simple: d(w, x) = 1 if $w \neq x$. Since the model is symmetric between members of E, we need only consider worlds of the form <e, 0>. Then #1 holds because <0, 0>R<2, 0>; #2 holds since both sides of #2 are false when e* = 0 and $e \neq 0$ and both sides of #2 are true otherwise. Although such models are hard to interpret, for present purposes there is no need to exclude them, since they do no harm.

2 The models in Williamson 2011 differ from those here in having a symmetric accessibility relation. The more complex models introduced in Williamson 201X are more similar to those in the present paper; in particular, they have a non-symmetric accessibility relation. However, they do not fully satisfy #2 with respect to their natural metric, since they allow d(e, f) to continue growing after R(<e, f>) has attained a maximum (the models are finite to simplify the probability distribution). They satisfy a good enough approximation to #2 for KR(<f, f>) to be {<f, f>}. Incidentally, those models were originally devised for quite different purposes, with no thought of Gettier cases. They simply turned out to exhibit Gettier-like behaviour.

3 On some views, Kripke's puzzling Pierre may simultaneously have both a justified belief that London is pretty, which he expresses by sincerely asserting 'Londres est jolie', and a justified belief that London is not pretty, which he expresses by sincerely asserting 'London is not pretty' (Kripke 1979). The coarse-grained contents of those beliefs are mutually inconsistent. As already explained, the complexities of such Frege cases have been ignored for purposes of this paper.

4 The principle B$p \rightarrow$ BKp yields the principle B$p \rightarrow$ BK…Kp for any number of iterations of K by repeated substitutions for p, and so is stronger than first appears.

5 By contrast, the model in footnote 1 lacks traditional Gettier cases. For suppose that in <e, 0> the agent believes a false proposition p that entails a true proposition q. By the definition of S, she knows p in <0, 0> since she believes p in <e, 0>. The worlds epistemically accessible from <0, 0> are itself and <2, 0>. Thus p is true in <0, 0> and <2, 0>. Since p is false in <e, 0>, e = 1. Hence q is true in <1, 0>. Since p entails q, q is true in <0, 0> and <2, 0>. Thus q is true in all worlds epistemically accessible from <e, 0>, so in <e, 0> the agent knows q. Thus <e, 0> is no Gettier case for q. By the symmetry of the model, the argument extends to all worlds.

6 This paper is based on a talk at the 2012 workshop in Svolvær on formal epistemology; thanks to the participants for lively discussion, and to the Centre for the Study of Mind in Nature in Oslo for its support of the event.

References

Buckwalter, Wesley, and Stich, Stephen 2011: 'Gender and the philosophy club', *The Philosophers' Magazine*, 52, pp. 60-65.

Dougherty, Trent (ed.) 2011: *Evidentialism and its Discontents*. Oxford: Oxford University Press.

Gettier, Edmund 1963: 'Is justified true belief knowledge?', *Analysis*, 23, pp. 121-123.

Goldman, Alvin 1976: 'Discrimination and perceptual knowledge', *The Journal of Philosophy*, 73, pp. 771-791.

Hintikka, Jaakko 1962: *Knowledge and Belief*. Ithaca, N.Y.: Cornell University Press.

Humberstone, Lloyd 1988: 'Some epistemic capacities', *Dialectica*, 42, pp. 183-200.

Kripke, Saul 1979: 'A puzzle about belief', in Margalit 1979, pp. 239-83.

Margalit, Avishai (ed.) 1979: *Meaning and Use*. Dordrecht: Reidel.

Nagel, Jennifer 2012: 'Intuitions and experiments: a defense of the case method in epistemology', *Philosophy and Phenomenological Research*, forthcoming.

Nagel, Jennifer 201X: 'Defending the evidential value of epistemic intuitions: a reply to Stich', *Philosophy and Phenomenological Research*, forthcoming.

Stalnaker, Robert. 2006. 'On logics of knowledge and belief', *Philosophical Studies*, 128, pp. 169-199.

Stich, Stephen 201X: 'Do different groups have different epistemic intuitions? A reply to Jennifer Nagel', *Philosophy and Phenomenological Research*, forthcoming.

Timmons, Mark, Greco, John, and Mele, Alfred (eds.) 2007: *Rationality and the Good: Critical Essays on the Ethics and Epistemology of Robert Audi*. Oxford: Oxford

University Press.

Weatherson, Brian 2003: ‘What good are counterexamples?’, *Philosophical Studies*, 115, pp. 1-31.

Weinberg, Jonathan, Nichols, Shaun, and Stich, Stephen 2001: ‘Normativity and epistemic intuitions’, *Philosophical Topics*, 29, pp. 429-460.

Williams, Bernard 1978: *Descartes: The Project of Pure Enquiry*. London: Penguin.

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

Williamson, Timothy 2007a: *The Philosophy of Philosophy*. Oxford: Blackwell.

Williamson, Timothy 2007b: ‘On being justified in one’s head’, in Timmons, Greco, and Mele 2007, pp. 106-122.

Williamson, Timothy 2011: ‘Improbable knowing’, in Dougherty 2011, pp. 147-164.

Williamson, Timothy 201X: ‘Very improbable knowing’, *Erkenntnis*, forthcoming.