# MODELS OF PHILOSOPHICAL THOUGHT EXPERIMENTATION

Jonathan Andy Tapsell

*A thesis submitted for the degree of*
*Master of Philosophy of*
*the Australian National University*

## STATEMENT

This thesis is solely the work of its author. No part of it has previously been submitted for any degree, or is currently being submitted for any other degree. To the best of my knowledge, any help received in preparing this thesis, and all sources used, have been duly acknowledged.

_____

Jonathan Andy Tapsell
4 September 2014

## ACKNOWLEDGEMENTS

I thank the chair of my supervisory panel, Daniel Nolan, for his advice and guidance. I also thank my partner, Leah Horsfall, for her love and support. This thesis would never have come into existence if it were not for them.

## ABSTRACT

The practice of thought experimentation plays a central role in contemporary philosophical methodology. Many philosophers rely on thought experimentation as their primary and even sole procedure for testing theories about the natures of properties and relations. This test procedure involves entertaining hypothetical cases in imaginative thought and then undergoing intuitions about the distribution of properties and relations in them. A theory's comporting with an intuition is treated as evidence in favour of it; but a clash is treated as evidence against the theory and may even be regarded as falsifying it.

The epistemic power of thought experimentation is mysterious. How can experiments carried out within the mind enable us to discover truths about the natures of properties and relations like knowledge, causation, personal identity, reference, meaning, consciousness, beauty, justice, morality, and free will? This epistemological challenge is urgent, but a model of philosophical thought experimentation would seem to be a necessary propaedeutic to any serious discussion of it. An adequate model would make the relevant test procedure explicit, thereby assisting in the identification of points of potential epistemic vulnerability.

In this monograph I advance the propaedeutical model-building work already done by Timothy Williamson, Anna-Sara Malmgren, and Jonathan Ichikawa and Benjamin Jarvis. Following the lead of these philosophers, I focus on a single Gettier-style thought experiment and the problem of identifying the real content of the Gettier intuition. My first contribution is to establish the inadequacy of all of the existing models. Each of them, I argue, fails to solve the content problem. It emerges from my discussion, however, that Ichikawa and Jarvis's truth in fiction approach holds out the prospect of a solution.

My second contribution is to develop and defend a new way of implementing the general idea behind the truth in fiction approach. The model I put forward does a better overall job of modelling Gettier-style thought experiments than any of the existing models. It has none of the defects which render those models inadequate and I

am unable to find any major defects peculiar to it. This should make us feel confident that my model is adequate. Moreover, since the Gettier-style thought experiment I focus on is paradigmatic, we should also feel confident that my model will generalise naturally to other philosophical thought experiments.

# TABLE OF CONTENTS

# 1. Introduction

The practice of thought experimentation plays a central role in contemporary philosophical methodology. A *thought experiment* is a special activity carried out within the imagination in order to test a theory or hypothesis. Experiments of this kind stand in marked contrast to *empirical experiments*, which involve testing theories by means of sensory observation of objects and events in the external physical world. Philosophers generally rely on thought experimentation as their primary and even sole test procedure; they rarely, if ever, get out of the armchair and go into the laboratory.[1] Such a heavy reliance on thought experimentation is perhaps one of the most conspicuous differences between the methodology of philosophy and the methodology of the natural sciences. There can be no doubt, of course, that thought experimentation has made valuable and sometimes crucial contributions to the development of modern science, especially physics. To appreciate the significance of its scientific contribution we need only reflect on the famous thought experiments devised by Galileo, Newton, and Einstein.[2] On the whole, however, thought experimentation is methodologically secondary to empirical experimentation in the sciences. The opposite goes for philosophy. Whereas the scientist pays most heed to the verdicts of the tribunal of perceptual experience, the philosopher chooses instead to bring his theories almost exclusively before the tribunal of the imagination. As

1 This statement might seem a bit anachronistic given the rise of so-called experimental philosophy in recent years. Experimental philosophy involves conducting surveys of laypeople and using the resultant empirical data to inform philosophical debates. However, despite the large amount of attention it has attracted, the proportion of philosophers who actually participate in this endeavour is, I think, very small. Furthermore, among those few philosophers who do engage in it, there are different views about the kind and degree of relevance of survey data to philosophical debates. One idea is that survey data constitutes evidence for or against theories of philosophically interesting properties and relations. Another is that it constitutes evidence for or against theories of how laypeople think about philosophically interesting properties and relations. And yet a further idea is that it constitutes evidence against the reliability of philosophical thought experimentation. The first of these ideas is operative in only *some* experimental philosophy research. See Alexander (2012) for a good introduction to the field of experimental philosophy. Knobe and Nichols (2008) collect some of the classic papers.

2 See Sorensen (1998) for general discussion of thought experimentation as it has been used in the work of these and other scientists.

James Brown and Yiftach Fehige (2011: 2) have remarked, "[p]hilosophy without thought experiments seems unthinkable. […] Philosophy, even more than the sciences, would be severely impoverished without thought experiments."

The procedure for testing theories in imaginative thought may be roughly characterised as follows. We first of all target a theory for evaluation. The target theory is taken to entail a *modal connection* between the specific properties or relations it is about, and thus to have implications for the distribution of those properties or relations in a range of possibilities, most of which are non-actual. We next consider a *hypothetical case* which we regard as falling within the modal scope of the target theory. This could be a case we think up ourselves, or it could be conveyed to us via a written description or some other medium. (Note that for a case to be hypothetical is simply for it to be entertained by the mind in imaginative thought. Although the hypothetical cases deployed in thought experimentation can, and usually do, fail to be realised in the actual world, they need not. Hypotheticality is grounded in imaginative mental activity, but since some of what we imagine could turn out to be really happening somewhere, hypotheticality does not entail non-actuality.) The final step of the procedure is to check the target theory's modal implications against our intuition about the distribution of the relevant properties or relations in the hypothetical case. In a *positive* thought experiment the target theory comports with our intuition and this is treated as at least some evidence in favour of it. We allow the target theory to survive another day. If, on the other hand, the target theory clashes with our intuition, this is treated as evidence against it. In both philosophy and science, but most particularly in philosophy, a *negative* thought experiment may even be regarded as enough to outright falsify the theory under evaluation.

The two thought experiments devised by Edmund Gettier (1963) as tests of the traditional theory of knowledge are paradigmatic examples. Knowledge, according to the traditional theory, is justified true belief; but Gettier claimed his thought experiments demonstrate otherwise, and the vast majority of philosophers who have carried them out agree with him. The traditional theory of knowledge, which arguably goes back as far as Plato, is nowadays widely acknowledged to have been refuted by Gettier's thought experiments.[3] Gettier is credited with having made one of

3 That, at any rate, is how things are portrayed in philosophical lore. We might find Williamson's comments on the status of Gettier's thought experiments reassuring here. "Sociologically," says

the greatest discoveries about the nature of knowledge in the history of epistemology, viz. that justification is not enough to make a true belief into knowledge. This discovery triggered an explosive research programme and continues to influence debates in epistemology right up to the present day.[4] Other paradigms of philosophical thought experimentation include such famous examples as Frank Jackson's (1982) neuroscientist experiment, Hillary Putnam's (1975) twin earth experiment, John Searle's (1980) Chinese room experiment, Saul Kripke's (1980) Schmidt/Gödel experiment, and Philippa Foot's (1967) and Judith Jarvis Thomson's (1976) trolley experiments—the list could go on. Like Gettier's thought experiments, these have all had enormous influence in their relevant sub-fields.

Given that thought experimentation is absolutely central to philosophical inquiry, it would be good to have a decent account of how it works. The question of the epistemic power of thought experimentation is often said to be especially urgent. It is natural to think that the principal or ultimate objective of much philosophical investigation is to discover truths about the natures of properties and relations in the world, not merely the words we use to talk about them or the concepts we use to think about them. Epistemology, for example, is naturally thought of as the branch of philosophy which investigates the nature of knowledge, not merely the word "knowledge" and its cognates, or the concept of knowledge. A similar point may be made about most other branches of philosophy. They are most naturally thought of as investigating the natures of things like causation, personal identity, reference, meaning, consciousness, beauty, justice, morality, and free will. But how, if at all, is it possible to discover truths about the natures of such properties and relations just

---

Williamson (2007: 180), "the phenomenon is remarkable. Gettier had no previous publications and was unknown to most of the philosophical profession; he did not write as an establish authority. […] His three-page article turns on two imaginary examples. Yet his refutation of the justified true belief analysis was accepted almost overnight by the community of analytic epistemologists. His thought experiments were found intrinsically compelling." As is always to be expected in philosophy, though, not everyone has been persuaded. See Shope (1983: 26-33) for discussion of some of the early counter-arguments. Some prominent recent defenders of the traditional theory of knowledge include Hetherington (2001; 2011) and Weatherson (2003).

4 As BonJour (2002: 51) has pointed out, it is quite plausible "that Gettier's paper has given rise to a larger body of philosophical literature, consisting of replies, criticisms of replies, etc., in proportion to its size than any other piece of philosophical writing." Shope (1983) discusses the early history of the burgeoning research programme. Although interest in finding a solution to the Gettier problem has gradually and naturally waned over the sixty-year period since Gettier published his paper, attempts at solving it still steadily trickle out; furthermore, due to its astounding influence, the Gettier problem has recently become central to debates about the nature and scope of philosophical inquiry.

by carrying out experiments within one's imagination? How could thought experiments yield new knowledge of reality? The epistemological challenge is to explain how the practice of thought experimentation enables the mind to access the natures of philosophically interesting properties and relations. This is a very difficult challenge and it is not at all obvious how to go about addressing it. We have more than a rudimentary understanding of how our perceptual apparatus works. The power of empirical experimentation to yield knowledge of reality is far from incomprehensible. In comparison, the epistemic power of thought experimentation is apt to strike us as utterly mysterious. This mysteriousness threatens to do away with whatever epistemic legitimacy we may have attached to the use of thought experiments in philosophy. It could force us to conclude that, rather than playing with the denizens of his own imagination, a philosopher investigating, say, the nature of knowledge, would do better to use his eyes to observe epistemic agents in the real world around him.[5]

The epistemological challenge is indeed an urgent one for philosophers, but a *model* of philosophical thought experimentation would seem to be a necessary propaedeutic to any serious discussion of it. The reason is that an adequate model would assist in making the relevant test procedure explicit, and this would in turn assist in the identification of points of potential epistemic vulnerability. Several philosophers have already made a start on this project, most notably Timothy Williamson (2005; 2007), Jonathan Ichikawa and Benjamin Jarvis (2009), and Anna-Sara Malmgren (2011). Their efforts at modelling philosophical thought experimentation have contributed a great deal to our understanding of the practice, but there is much more work to be done. My aim in the present monograph is twofold: first, to establish the inadequacy of all of the existing models, and second, to develop and defend a new model of my own. The model I shall put forward not only has considerable advantages over its rivals but—as far as I am able to determine anyway—it neither partakes in any of their worst vices nor suffers from any major defects peculiar to itself.

I shall concentrate my efforts on developing an adequate model of a single Gettier-style thought experiment. This is the same approach to modelling as that taken

5 That might be the recommendation of proponents of naturalised epistemology, starting with Quine (1969). See Bishop and Trout (2005) and the series of works by Kornblith (1994; 1999; 2002; 2006) for some more recent proposals to naturalise epistemological investigation.

by Williamson and, following him, Ichikawa and Jarvis as well as Malmgren. Although such an approach to modelling could seem to lack a certain generality, it makes up for it, in my view, by firmly anchoring our discussion in reality. The use of a concrete example during the process of model-building does much to facilitate theoretical comprehension. In addition, it is more than probable that the majority of contemporary philosophers are already familiar with the nuts and bolts of Gettier-style thought experiments, having actually performed them on one or more previous occasions. This sort of antecedent familiarity with a subject matter also does much to facilitate theoretical comprehension when model-building. But the aforementioned lack of generality in my approach is more apparent than real anyway. After all, Gettier-style thought experiments are paradigms of the relevant phenomenon to be modelled. We therefore have excellent reason to think that the model we come up with should generalise to other philosophical thought experiments in a natural way. So, in light of these remarks, let us now proceed to construct a thought experiment in the style of Gettier's originals.

### 1.1. A Gettier-style thought experiment

The theory we are going to test is the traditional theory of knowledge, according to which knowledge is justified true belief. This is a theory about the *nature* of knowledge. As such, we should take it to entail a modal connection of the *metaphysical* kind between the relation of knowledge and the relation of justified true belief, and thus to have implications for the distribution of those epistemic relations in a very broad range of possibilities. Such broadness of modal scope is rather typical of philosophical theories. Many scientific theories, in contrast, have a much narrower modal scope; they only entail modal connections of the *nomological* kind between the properties or relations they are about. The modal scope of the traditional theory of knowledge encompasses not only the sphere of the nomologically possible, but the sphere of the metaphysically possible as well.[6] Put another way, our target theory

6 I am here assuming what I take to be the standard conception of modal reality, according to which nomological possibility is a proper subset of metaphysical possibility. It is compatible with this conception that some but not all scientific theories have metaphysical and not merely nomological modal scopes. The theory that water is composed of $H_2O$ molecules, for example, is a scientific theory if anything is; so if we are moved by the ideas of Kripke (1980) and Putnam (1975), we might consider its (implicit) modal scope to encompass both the nomological and the metaphysical spheres of possibility. Note that there are a few philosophers who have gone so far as to claim that these modal spheres are in fact identical. They endorse the doctrine that the true

entails that, as a matter of metaphysical necessity, one knows a proposition just in case one has a justified true belief in it. With the obvious symbolic correspondences, we may formally represent the entailed modal connection as follows:

(MC) $\Box\forall x\forall p\,(Kxp \leftrightarrow JTBxp)$

We should be wary of simply identifying the traditional theory of knowledge with (MC), for it is doubtful whether strict biconditionals per se should be interpreted as making claims about the natures of properties and relations. But the traditional theory of knowledge is true only if (MC) is true; contrapositively, it is false if (MC) is false. This entailment is what opens the door for us to test the traditional theory of knowledge by means of an experiment performed within the imagination. More generally, it is only because of the modal connections they entail that the tribunal of the imagination has any jurisdiction at all over philosophical and scientific theories.

The procedure for testing the traditional theory of knowledge in imaginative thought involves checking whether (MC) comports or clashes with our intuition about the distribution of the relations of knowledge and justified true belief in a hypothetical case. To commence our thought experiment, then, we may consider the following vignette, which for the sake of convenience I shall call *the Gettier case*:

> One day Smith is walking through the Australian countryside. He is an avid and experienced spotter of native wildlife who has the ability to identify animals such as kangaroos, wallabies, wombats, koalas, and so on, with a very high degree of reliability. The present conditions, furthermore, happen to be excellent for spotting native wildlife. Smith looks into a nearby paddock and sees what looks exactly like a mob of kangaroos standing next to a huge rock formation. "Ah," he thinks to himself, "there are kangaroos in this paddock." As it turns out, the objects Smith is looking at are not kangaroos at all. They are in fact sophisticated robots designed to perfectly mimic the appearance and behaviour of real kangaroos. Even so, Smith's belief in the proposition that there are

laws of physics, chemistry, biology, and so on, obtain as a matter of metaphysical necessity. See for example Bird (2007), Fales (1990), Shoemaker (1980; 1998), and Swoyer (1982). Of course, since I am assuming the standard conception of modality, I am by implication assuming these philosophers are mistaken. For critical discussion of their doctrine, see Sidelle (2002).

kangaroos in the paddock is actually true. A mob of real kangaroos is standing behind the huge rock formation, completely hidden from his view.[7]

The foregoing text obviously describes a genuine metaphysical possibility, perhaps even a nomological possibility. Accordingly, the Gettier case must fall within the modal scope of the target theory. For the target theory implies that the relation of knowledge and the relation of justified true belief coincide in all metaphysical possibilities. If the target theory is true, then those epistemic relations must coincide in the Gettier case just as they do throughout the rest of the sphere of the metaphysically possible. This bears on the epistemology of what we are trying to do. If the Gettier case fell outside the target theory's modal scope, or if there were good reason for thinking it did, then at best our thought experiment would be of dubious evidential relevance to the truth or falsity of the target theory. At worst it would be wholly beside the point. By basing our thought experiment on an obvious metaphysical possibility, we forestall one potential epistemic threat. Note, furthermore, that the description of the Gettier case is neutral insofar as it makes no explicit mention of the relations of knowledge and justified true belief. It does, to be sure, explicitly specify that Smith has a justified true belief in the proposition that there are kangaroos in the paddock. But it does not specify that Smith knows, or fails to know, that proposition, nor that he has, or fails to have, a justified true belief in it. This also bears on the epistemology of our thought experiment. If we were to base it on a more partisan case description, our thought experiment would risk begging the question either for or against the target theory. Either way, its evidential relevance would be diminished, or even destroyed. So the neutrality of the Gettier case forestalls another potential epistemic threat to what we are trying to do.

With the Gettier case in mind, we may finalise our thought experiment by asking ourselves about the epistemic status of Smith's true belief. Intuitively, Smith has a justified true belief, but does not know, that there are kangaroos in the paddock. It will be convenient to call this intuition *the Gettier intuition*. The Gettier intuition clashes with (MC) and ipso facto with the target theory. The result of our thought experiment is therefore negative: the target theory's implications for the distribution of

7 This case is an Australianised variant of Chisholm's sheep-in-the-field case. See his (1989: 93).

the relations of knowledge and justified true belief in the Gettier case were not borne out in the tribunal of the imagination. The orthodox reaction to our thought experiment would be to acknowledge it as a falsification of the traditional theory of knowledge. We have, in other words, carried out what philosophical orthodoxy would consider to be a *successful* negative thought experiment. An adequate model of it must therefore elucidate our ostensibly legitimate transition from the Gettier intuition to the conclusion that the traditional theory of knowledge is wrong.

### 1.2. The content problem

Taken at face value the Gettier intuition is about a man called "Smith" and his epistemic relationship to the proposition that there are kangaroos in the paddock. As I reported it in the previous sub-section, the Gettier intuition appears to have the content that Smith justifiedly and truly believes that there are kangaroos in the paddock but does not know that there are kangaroos there. The apparent content of the Gettier intuition may be formally represented as follows, where the singular terms "s" and "k" are intended to stand for Smith and the proposition about kangaroos respectively:

(AC) $JTBsk \wedge \neg Ksk$

If (AC) is the real as well as the apparent content of the Gettier intuition, we should expect to find that our thought experiment leads us straightforwardly to the negation of the traditional theory of knowledge; for it is plain that (AC) would be a counter-instance to the left-to-right direction of the strict biconditional (MC). As a matter of fact, the transition we make to the negation of the traditional theory of knowledge *does* seem pretty straightforward. It could therefore be tempting to think that (AC) must be an integral component of any adequate model of our thought experiment. But the temptation should be resisted; whatever its initial attractions, (AC) is highly problematic.

The real content of the Gettier intuition is what we come to accept or believe when we imaginatively engage with the Gettier case and ask ourselves about Smith's epistemic situation in it. If the real content is simply (AC), then given the standard

framework of classical logic, the Gettier intuition would seem to commit us to the existence of both an object named "Smith" and an epistemic state of affairs in which this object has a justified true belief, but does not know, that there are kangaroos in the paddock. Formally, these commitments may be encapsulated by an existential claim:

(E) $\exists x\ (x{=}s \wedge JTBxk \wedge \neg Kxk)$

This existential claim is entailed by (AC) in classical logic. However, even if the idea is not completely absurd, it is most doubtful whether our acceptance of the Gettier intuition really imposes any such existential commitments upon us. We can accept the Gettier intuition and reject or withhold judgement on the existential commitments of its apparent content without thereby exhibiting any obvious irrationality. As a matter of fact, such a combination of attitudes would be perfectly natural and reasonable. Consequently, we should be extremely reluctant to treat the apparent content of the Gettier intuition as anything more than merely apparent. The real content of the Gettier intuition is *hidden*, and we must find out what it is if we wish to develop an adequate model of our thought experiment. That is what I shall call *the content problem*.[8]

At this point in our discussion it could be tempting to dismiss the content problem as an artefact of classical logic. The most natural alternative to classical logic in the present context is *free logic*. Free logic gets its name from the fact that it is free of existential assumptions with regard to terms.[9] In the framework of free logic, the quantifiers range over existent objects just as they do in the standard classical framework, but not all terms need refer to something in the domain of quantification. Terms which fail to refer to existent objects are *empty* and the formulae in which they occur are *empty-termed*. Free logic comes in three kinds, each of which differs from the other two in how it treats empty-termed *atomic* formulae. *Positive* free logic allows some empty-termed atomic formulae to be true. *Negative* free logic requires them to be false. And *neutral* free logic requires them to be truth-valueless.

8 I have borrowed this name for the problem of identifying the real content of the Gettier intuition from Malmgren (2011).

9 Lambert (2001) provides a concise introduction to the main ideas and motivations underlying free logic. See also Bencivenga (1986).

Since Smith is presumably not an existent object, the singular term "s" in (AC) should be taken to be an empty term, and the first conjunct of (AC) should be taken to be an empty-termed atomic formulae. It follows that (AC) must come out either false or truth-valueless in non-positive free logics. Since the Gettier intuition is presumably true, those logics are of no help vis-a-vis the content problem. If, however, we interpret (AC) and (E) in accordance with positive free logic, then not only could (AC) come out true, but it will fail to entail (E).[10] It may thus seem that the content problem vanishes. But actually our model-building endeavours have made no further headway. For the appeal to positive free logic does nothing but dissolve one manifestation of the content problem while giving rise to another. In positive free logic, we cannot validly infer (MC)'s falsity from (AC); we must first establish that Smith is an existent object.[11] But our thought experiment is successful without our having to do any such thing. We are simply not required to establish that Smith is an existent object in order to legitimately transition from the Gettier intuition to the negation

10 This logical point could do with elaboration. Let $\alpha$ be any term and let $\Phi(\alpha/x)$ be the formula which results from replacing all occurrences of the variable x in $\Phi$ with $\alpha$. The classical principle of existential generalisation may be stated as follows:

$\Phi(\alpha/x)$; so $\exists x\Phi$

But this principle is invalid in every kind of free logic. The free logic principle of existential generalisation may be stated using "E" as an existence predicate:

$E\alpha \wedge \Phi(\alpha/x)$; so $\exists x\Phi$

In positive free logic, (AC) does not entail (E) because there is no way to get Es (i.e. Smith's existence) from it. Interestingly, the entailment from (AC) to (E) does hold in negative free logic. This is owing to the fact that negative free logic validates the following:

$\Phi(\alpha/x) \rightarrow E\alpha$ if $\Phi(\alpha/x)$ is atomic

Since (AC) entails its own first conjunct and that conjunct is atomic, we can get Es from (AC) in negative free logic using conjunction elimination and then the above conditional. We would thus have everything needed to derive (E) using the free logic principle of existential generalisation.

11 It may be helpful to elaborate on this second logical point as well. Using the symbolism explained in the previous footnote, the classical principle of universal instantiation may be stated as follows:

$\forall x\Phi$; so $\Phi(\alpha/x)$

But it is invalid in every kind of free logic. The free logic principle of universal instantiation may be stated using E as an existence predicate:

$\forall x\Phi$; so $E\alpha \rightarrow \Phi(\alpha/x)$

So (MC) entails $(Es \wedge Ek) \rightarrow (Ksk \leftrightarrow JTBsk)$. But in positive free logic this conditional statement is logically compatible with (AC). Thus (AC) is not enough to get us to the negation of (MC). We also need to establish the conjunction $Es \wedge Ek$. In the main text I only mention the requirement to establish the first of these two conjuncts, i.e. the existence of Smith. Doing so is sufficient to convey the point I want to make there. But, strictly speaking, we would have establish the second conjunct as well, i.e. the existence of the proposition that there are kangaroos in the paddock. Note that, for the reasons discussed in the previous footnote, we can actually get from (AC) to the negation of (MC) in negative free logic. But the fact that (AC) entails the negation of (MC) in negative free logic does not make that kind of free logic suitable for dissolving the content problem. After all, in negative free logic, (AC) both entails (E) and comes out false.

of (MC) and thence to the conclusion that the traditional theory of knowledge is wrong. Indeed, as I have already emphasised, it would be perfectly natural and reasonable for us to reject or withhold judgement on the existence of Smith. So working within the framework of free logic would not enable us to avoid treating the apparent content of the Gettier intuition as merely apparent.

Another alternative is to divest the quantifier ∃ in (E) of the existential import standardly imputed to it in classical logic (and free logic). The entailment from (AC) to (E) could then be preserved without imposing any unpalatable existential commitments upon us. One approach of this kind involves *neutral quantification*. In logics which deploy neutral quantification, the quantifiers range over both *existent* and *non-existent* objects.[12] The domain of quantification can include you, me, the Colosseum, and my mug, as well as Pegasus, Yahweh, the golden mountain, and the round square. Inferences to what "there is" are not the same as inferences to existence. From the statement that my mug is blue, we can validly infer that "there is" something which is blue, but we cannot validly infer the existence of something which is blue. From the statement that the round square is round and square, we can validly infer that "there is" something which is round and square, but we cannot validly infer the existence of something which is round and square. Accordingly, in the framework of neutral quantification, (AC) entails (E), but (E) only affirms that "there is" something identical to Smith which justifiedly and truly believes, but fails to know, the proposition that there are kangaroos in the paddock. It is silent as to whether this alleged entity or thing called "Smith" really exists. The content problem may thus seem to vanish. Once again, however, we have done nothing but dissolve one manifestation of the content problem while giving rise to another. The quantifier ∀ in (MC) is not neutral; it must be stronger. For suppose otherwise. Then falsifying the traditional theory of knowledge would be as easy as coming up with a description explicitly attributing a non-knowledge justified true belief to some "non-existent object". An example, perhaps, is the man who lives on the golden mountain, draws pictures of the round square, and has a non-knowledge justified true belief. But even if "there is" such a man, the traditional theory of knowledge obviously cannot be falsified so easily. Whatever kind of quantification is involved in (MC), it is

12 Routley (1980) discusses and defends neutral quantification at length. See Lewis (1990) for criticism.

stronger than neutral quantification.[13] If, therefore, we confine ourselves to interpreting (AC) and (E) in accordance with the framework of neutral quantification, it is hard to see how (AC) could have any logical bearing on the truth or falsity of (MC) and the traditional theory of knowledge. Other approaches to divesting the quantifier ∃ in (E) of existential import lead to pretty much the same result.[14] So the real content of the Gettier intuition is still hidden and the content problem is still a genuine problem.

A further point it is important for me to emphasise here is that the content problem is a problem about the content of the the Gettier intuition, not its metaphysics. The question of the metaphysics of intuition has to do with the status of intuitions in the true theory of the mind. Broadly speaking, there are two rival metaphysical conceptions: *sui generisism* and *reductionism*. According to the sui generisist conception, intuitions are irreducible propositional attitudes.[15] They are standardly held by its proponents to be conscious representational states (or episodes) with a special kind of phenomenal character. This alleged phenomenology is variously described as being forceful, assertive, and even revelatory; roughly, it may be said to consist in something akin to a more or less compelling feeling of ascertaining that things are really as they are represented to be.[16] Furthermore, although on the sui generis conception intuitions are non-doxastic states, they are held to causally generate doxastic states like beliefs whenever their veridicality is not in doubt. The reductionist alternatives, on the other hand, deny that intuitions comprise

13 More generally, whatever kind of quantification is involved in the modal connections entailed by our philosophical theories of the natures of properties and relations, it must be stronger than neutral quantification. Otherwise, all such theories would be absurdly easy to falsify. Consider, for example, the theory that the nature of triangleness is having three straight sides and three angles. This theory entails a modal connection between the property of triangleness and the properties of having three straight sides and three angles. But the triangle which lacks three straight sides is triangular and lacks three straight sides. So "there is" something which is triangular and lacks three straight sides. So the foregoing theory of triangleness is false (assuming the quantifier ∀ in the entailed modal connection is neutral). But obviously that theory cannot be falsified in this manner.

14 An approach involving *substitutional quantification*, for example, would also make it hard to see how (AC) could logically bear on (MC) and the traditional theory of knowledge. This is owing to considerations broadly similar to those already discussed in the main text.

15 Proponents of this conception included Bealer (1996a; 1996b; 1998; 2000), Bengson (forthcoming), Chudnoff (2011a; 2011b), Cullison (2010), Huemer (2005), Pust (2000), Tolhurst (1998), and Tucker (2010).

16 Tolhurst's description of this alleged phenomenology is representative of what many sui generisists have in mind. He writes that intuitions and other seemings "have the feel of truth, the feel of a state whose content reveals how things really are" (1998: 288-9).

a sui generis class of mental state. Such views typically reduce intuitions in doxastic terms. Proponents of reductionism hold that intuitions are simply judgements or beliefs, inclinations to form beliefs, or perhaps some other kind of doxastic mental phenomena.[17] Sui generisism and reductionism clash over the fundamentality of intuitions. If sui generisism is true, intuitions are basic mental states distinct from all others; whereas if reductionism is true, the true theory of the mind does not admit of intuitions as distinct basic mental states. This metaphysical question has provoked a great deal of philosophical controversy. Thankfully, however, we do not need to resolve it in order to address the content problem. To be sure, we should acknowledge that there is a tight relationship between intuitions and doxastic states like beliefs; but whether the relationship is a causal one or something deeper is a matter on which it is appropriate for us to remain neutral.

### 1.3. Beyond the Gettier intuition's apparent content

So far we have found that, in order to develop an adequate model of our thought experiment, we have to move beyond the apparent content of the Gettier intuition because it is illusory. Following a recommendation of Williamson's, we may begin to do so by rethinking our earlier handling of Smith and the proposition which Smith justifiedly and truly believes yet fails to know. The suspicion is that the content problem has its roots in the perhaps naïve idea that the "objects" we imagine when we engage with hypothetical cases are genuine referents of genuine singular terms. Williamson recommends treating the apparent singular terms in hypothetical cases such as the Gettier case not as genuine singular terms but rather as "picturesque substitutes for variables" (Williamson 2007: 184). If we adopt the Williamsonian approach, we can then represent our case description using the open formula GCxp, where the variables "x" and "p" occupy the positions for, respectively, the subject and the proposition. This formula is to be understood as saying that x stands to p as described by the text of the Gettier case.

17 There are many proponents of doxastic reductionism. See among others Boghossian (2009), Dennett (1987; 1991), Erlenbaugh and Molyneux (2009), Ichikawa (MS), Ichikawa and Jarvis (2009), Lewis (1983), Lynch (2006), Ludwig (2007), Nimtz (2010a), E. Sosa (1998), van Inwagen (1997), and Williamson (2007). Van Inwagen succinctly encapsulates their general attitude toward the the metaphysical question as follows: "Our 'intuitions' are simply beliefs—or perhaps, in some cases, the tendencies that make certain beliefs attractive to us, that 'move' us in the direction of accepting certain propositions without taking us all the way to acceptance" (1997: 309).

The Gettier case itself is, of course, at the very heart of the particular thought experiment we wish to model. In carrying out the thought experiment, we bring the Gettier case before our minds and entertain it in imaginative thought. This imaginative engagement with the Gettier case would seem, furthermore, to involve an apprehension of the metaphysical possibility of its being realised by some x and some p. We take ourselves to be imagining a bona fide way the world might have been. But it is because we imaginatively engage with the Gettier case and ask ourselves about Smith's epistemic situation in it that we come to have the Gettier intuition. Indeed, the Gettier intuition is an intuition *about* the Gettier case; more specifically, it is an intuition about the distribution of the relations of knowledge and justified true belief in the Gettier case. So, in light of these considerations, we might find it tempting to think that the metaphysical possibility of some x and some p being such that GCxp should somehow be an integral component of our model—in other words, that that metaphysical possibility must at least have *something* to do with the real content of the Gettier intuition. This is one temptation which I think it is worth our while to indulge, since it holds out the prospect of a reasonable way forward. Williamson has provided us with a springboard from which we are able to launch our model-building endeavours.

## 2. The Necessity Model

A natural first stab is to interpret the Gettier intuition as really expressing a *strict conditional*, one which says, in effect, that in every metaphysically possible realisation of the text of the Gettier case someone justifiedly and truly believes, but fails to know, some proposition. Letting "T" stand for the traditional theory that the nature of knowledge is justified true belief, we would then get the following model of our thought experiment, which I shall call *the Necessity Model*:

(1) $T \rightarrow \Box\forall x\forall p\ (Kxp \leftrightarrow JTBxp)$

(2) $\Diamond\exists x\exists p\ GCxp$

($3_N$) $\Box\forall x\forall p\ (GCxp \rightarrow (JTBxp \wedge \neg Kxp))$

(4) $\neg\Box\forall x\forall p\ (Kxp \leftrightarrow JTBxp)$ From (2) and ($3_N$)

---

(5) $\neg T$ From (1) and (4)

This model shares two important properties with all the other models we are going to consider. One is that it represents our thought experiment as a valid modal argument, so the traditional theory of knowledge must be false if the premises are true. Another is that the intended real content of the Gettier intuition, which is here given by the strict conditional ($3_N$), obviously has none of the problematic existential commitments associated with the Gettier intuition's apparent content.

Although I am aware of no one who has ever actually championed the Necessity Model, it is not without its attractions.[18] An initial attraction of the model is that it would seem to do a very good job of approximating the natural progression of the mental activity we perform when we actually carry out our thought experiment. In

18 To my knowledge, the first philosopher to explicitly discuss the Necessity Model was Williamson (2005: 6-7; see also 2007: 184-5). His discussion of it is brief and rather contemptuous: he treats it as a foil for his own model and says nothing at all about any of its attractions. The Necessity Model is also discussed by Ichikawa and Jarvis (2009: 223-5) and Malmgren (2011: 273-7).

the first stage of the test procedure, the traditional theory of knowledge is taken to entail a metaphysical modal connection between the relations of knowledge and justified true belief, as represented by (1).[19] We then go on to consider the Gettier case. We apprehend in imagination the metaphysical possibility of the Gettier case, as represented by (2), and ask ourselves what we think about the epistemic status of Smith's true belief. Reflection on this question eventually gives rise to the Gettier intuition, the real content of which is given by the strict conditional ($3_N$). Thus, our imaginative engagement with the Gettier case delivers both (2) and ($3_N$), from which we can easily derive (4), the negation of the metaphysical modal connection we take to be entailed by the traditional theory of knowledge (i.e. the negation of (MC)). Finally, with (4) in hand, modus tollens gets us to (5), the conclusion that the traditional theory of knowledge is wrong. From these comments it should be clear that the Necessity Model is both elegant and sufficiently comprehensive. Each of the foregoing stages of the test procedure is modelled in a neat and clear manner; furthermore, it does not seem as though any crucial stages have been neglected.

The Necessity Model may hold considerable attraction for those sympathetic to the idea that, in some way or another, the real contents of intuitions always involve the concept of metaphysical necessity. According to Ichikawa and Jarvis (2009: 223), the idea that "intuitions such as the Gettier intuition are judgements of necessity" is one of the defining elements of the "traditional understanding of philosophical methodology." Ichikawa and Jarvis along with many other contemporary philosophers give their endorsement to this feature of traditional philosophy. George Bealer, for example, holds that when we have an intuition "it presents itself as necessary: it does not seem to us that things could be otherwise" (Bealer 1998: 207). In the same vein, Alvin Plantinga holds that having an intuition consists "in finding yourself utterly convinced that [a proposition] is not only true, but could not have been false" (Plantinga 1993: 105). Ernest Sosa, too, holds that the contents of our intuitions have a "modally strong character" (Sosa 2007: 62). And Laurence BonJour holds that in having an intuition we "'see' or apprehend that [the truth of a proposition] is an invariant feature of all possible worlds, that there is no possible world in which it is false" (Bonjour 1985: 192). Since the Necessity Model posits a strict con-

19 Sorensen gives statements such as (1) the fitting name of "modal extractors" (1998: 136).

ditional as the real content of the Gettier intuition, it constitutes one way of applying these kinds of traditional views to the real contents of intuitions generated by philosophical thought experimentation.

However, given the complete absence of the concept of necessity from the apparent content of the Gettier intuition, it might be wondered whether falling into line with tradition is really warranted. Joel Pust (2000) has pressed a worry of this sort. According to Pust, Smith's having a justified true belief without knowledge "does *not* occurrently seem necessarily true to the typical person simply presented with the Gettier case and asked whether the agent in the case does not know" (2000: 38).[20] The absence of apparent necessity in the "typical person's" Gettier intuition moves Pust to deny the involvement of the concept of necessity in its real content. As he puts it, we are able to do "a fair amount of philosophy before invoking the concept of necessary truth" (2007: 38). But Pust arrives at this conclusion only because he puts too much faith in the appearances. For the reasons already canvassed, the real content of the Gettier intuition surely does differ from its apparent content in some way or another; they cannot be identical. The appearances are in this case illusory and known to be so. In order to establish that the Necessity Model is inadequate, therefore, it is not enough to *merely* draw attention to a difference between ($3_N$) and the apparent content of the Gettier intuition. After all, we expect and indeed require there to be at least some difference between them. In addition to drawing attention to a difference, it is necessary to establish that the difference is cause for serious alarm. This may be done by bringing out its unpalatable consequences, if there are any. Pust's argument poses no problem for the proponent of the Necessity Model because it does nothing to bring out any unpalatable consequences of the involvement of the concept of necessity in ($3_N$).

It is possible, however, that Pust has a different line of argument in mind. For, in addition to thinking that the concept of necessity is *not* involved in the apparent content of the "typical person's" Gettier intuition, he also thinks that "when asked to consider whether [Smith's having a justified true belief without knowledge] seems necessarily the case, one considers whether it could be otherwise and it *then* seems necessarily the case" (2000: 38). That is, he thinks that explicit reflection on the

20 The italics in this direct quotation are Pust's.. I add no italics to any direct quotations. Noting this fact here obviates the need to clutter the text with repetitions of the phrase "italics in original".

question of whether Smith necessarily has a justified true belief without knowledge gives rise to a non-typical Gettier intuition, the apparent content of which *does* involve the concept of necessity. Let us grant these observations. Pust could then argue that the Necessity Model implies that both Gettier intuitions (i.e. the typical one and the non-typical one) have the same real content (viz. ($3_N$)), and that as a consequence the model lacks the resources to explain the difference he observes in their apparent contents. This alternative argument is certainly better than the first one; but even so, it too does not pose much of a problem for the proponent of the Necessity Model. One simple and, I think, quite plausible line of reply would be to maintain that although both Gettier intuitions do have ($3_N$) as their real content, their apparent contents differ because consideration of whether Smith necessarily has a justified true belief without knowledge makes the occurrence of the concept of necessity in ($3_N$) become manifest, presumably by drawing the mind's attention to something it is normally prone to overlook. Another possible line of reply would be to maintain that the difference between the intuitions' apparent contents is actually due to a difference in their real contents. It might be, for example, that consideration of whether Smith necessarily has a justified true belief without knowledge has the effect of bringing forth an intuition with the necessitation of ($3_N$) (i.e. the strict conditional $\Box\Box\forall x\forall p$ (GCxp $\rightarrow$ (JTBxp $\wedge$ $\neg$Kxp)) as its real content, yet when one undergoes this intuition only one of the two occurrences of the concept of necessity in its real content is detected by the mind. No doubt these approaches admit of various refinements. But whichever turns out to be the most suitable, it can hardly be said that there is a lack of explanatory resources here.[21]

If the real content of the Gettier intuition is ($3_N$), then since ($3_N$) alone obviously does not entail the negation of (MC), it is a bit misleading to speak (as I have done) of the Gettier intuition "clashing" with the traditional theory of knowledge. This is another aspect of the Necessity Model we might find problematic, at least insofar as we find it natural to speak of such a clash. As a matter of fact, however, the logical compatibility of the Gettier intuition with the traditional theory of knowledge

21 Admittedly, my two lines of response to Pust's argument are rather vague. Much more needs to be said, for example, about why occurrences of the concept of necessity in the real content of the Gettier intuition are not transparent to the mind. But I am not trying to come up with a detailed and thorough explanation of the phenomenon Pust has observed. It is enough for my purposes here if I make it plausible that there are sufficient explanatory resources for someone to do so.

is just what we ought to expect, for we find an analogous logical phenomenon in the practice of empirical experimentation. It is a familiar point from the philosophy of science that the theories subjected to testing by scientists are typically logically compatible with all manner of observations (or, more properly, observation statements). Logical incompatibilities emerge only when a theory is taken together with a description of the initial conditions of an experimental set-up. Logically, falsifications in science are always based on observation statements in combination with such descriptions, never on observation statements alone. For this reason it is always possible (though by no means always sensible) for a scientist to save his favourite theory from falsification by shifting the blame to the description of the initial conditions of the relevant experimental set-up.[22] Scientists, of course, often do speak in a natural way of observations "clashing" with theories, but this abridged parlance of theirs is plainly one of convenience. The initial conditions of experimental set-ups are often not in any doubt, which makes it superfluous for scientists to explicitly relativise "clashes" to them in many conversational contexts. To deal with the present worry, therefore, the proponent of the Necessity Model, or of any model with a similar structure, would do well to turn it to his advantage, by arguing that the roles played by intuitions and hypothetical cases in thought experimentation are plausibly analogous to the roles played in empirical experimentation by, respectively, observations and initial conditions. The metaphysical possibility of the Gettier case is not in any doubt, which is why we find it natural to speak as if there is a direct conflict between the Gettier intuition and the traditional theory of knowledge. Like scientists, philosophers avail themselves of convenient yet slightly misleading ways of reporting their experimental results.

### 2.1. Underspecification

In spite of its attractions, however, the Necessity Model must ultimately be found in-

22 This is generally known as the Duhem/Quine thesis. The person to have first raised it seems to have been Duhem (1991, first published in 1914), but Quine (1951) did much to revive interest in it. Logically, the Duhem/Quine thesis is easy to grasp. Let "T" stand for any theory and "I" for the description of the initial conditions of a relevant experimental set-up. From the conjunction T $\wedge$ I, but not from T alone, we may derive an observational prediction O. If we then observe $\neg$O, we may derive the disjunction $\neg$T $\vee \neg$ I, but we cannot derive $\neg$T unless we have I in addition to $\neg$O. In other words, we may only treat T as falsified on the basis of $\neg$O in combination with I. For further discussion of the Duhem/Quine thesis, including several good examples of how it has been applied throughout the history of science, see Chalmers (1999: 88-91).

adequate. As Williamson has pointed out, its inadequacies stem from the fact that the Gettier case is *underspecified*. He explains underspecification as follows:

> In philosophy, examples can almost never be described in complete detail. An extensive background must be taken for granted; it cannot all be explicitly stipulated. Although many of the missing details are irrelevant to whatever philosophical issues are in play, not all of them are. This applies not just to highly schematic descriptions of examples […] but even to the much richer stories Gettier and other philosophers like to tell (Williamson 2007: 185).

According to Williamson, the vignette with which we imaginatively engage when we carry out our thought experiment by no means describes a complete possible world or situation. Indeed, it is possible, while preserving the metaphysical coherence of the text of the Gettier case, to *enrich* it in an infinite number of different ways. Many of these enrichments include details which have ramifications for the distribution of the relations of knowledge and justified true belief. On the one hand, we could enrich the text in such a way as to reinforce the Gettier intuition. But we could also enrich the text in such a way as to generate an intuition about the epistemic status of Smith's true belief which would run counter to our original intuition about it. Enrichments of the latter variety may be divided into two kinds.

First, we could enrich the text of the Gettier case in such a way as to generate the intuition that Smith's true belief that there are kangaroos in the paddock is not even justified. This may be done, for example, as follows:

> … Smith looks into a nearby paddock and sees what looks exactly like a mob of kangaroos standing next to a huge rock formation. "Ah," he thinks to himself, "there are kangaroos in this paddock." **But Smith is letting his eagerness to spot wildlife get the better of his reason. He knows that the extremely powerful hallucinogens he consumed earlier in the day have been causing him to hallucinate native animals of many different kinds, including kangaroos.**

This enrichment obviously describes a metaphysical possibility. But since any metaphysically possible realisation of it would also constitute a metaphysically possible realisation of the original text, the following must hold:

(US1) $\Diamond\exists x\exists p\ (GCxp \wedge \neg JTBxp \wedge \neg Kxp)$

Second, we could enrich the text in such a way as to generate the intuition that Smith's true belief is not only justified but constitutes knowledge.[23] An example is as follows:

> … Even so, Smith's belief in the proposition that there are kangaroos in the paddock is actually true. A mob of real kangaroos is standing behind the huge rock formation, completely hidden from his view. **But Smith's belief is based not only on his perceptual experience, but also on his recollection of being told that all of the paddocks on his walk would be empty, except for one, which would contain both robotic kangaroos and real kangaroos. He had been told this information earlier by someone whom he knows to be perfectly trustworthy about such matters.**

This is another enrichment which obviously describes a metaphysical possibility. But any metaphysically possible realisation of it would also constitute a metaphysically possible realisation of the original text. So this time we get:

(US2) $\Diamond\exists x\exists p\ (GCxp \wedge JTBxp \wedge Kxp)$

We may encapsulate underspecification as follows:

(US) $\Diamond\exists x\exists p\ (GCxp \wedge ((\neg JTBxp \wedge \neg Kxp) \vee (JTBxp \wedge Kxp)))$[24]

23 In his (2007), Williamson himself only mentions enrichments of the first kind. But it is clear from the general tenor of his discussion that he would recognise enrichments of this second kind as well.

24 Of course, (US) is weaker than the conjunction of (US1) and (US2), but for our purposes here (US) will be much more convenient to work with.

The text of the Gettier case does not rule out (US). Of course, when carrying out our thought experiment, we are *not meant* to imagine the metaphysical possibility of the Gettier case in a way which establishes (US). For this reason I shall call any metaphysically possible realisation of the text of the Gettier case in which the existential statement embedded in (US) is true a *deviant world*.[25]

### 2.2. Inadequacy of the Necessity Model

We are now in a position to appreciate why the the Necessity Model is inadequate. The underspecification of the text of the Gettier case, as represented by (US), is incompatible with the strict conditional ($3_N$). For ($3_N$) requires that modal space be altogether devoid of deviant worlds, while (US) requires that there be at least one such world. They cannot both be true; one of them has to go. But since only a modicum of reflection is sufficient to impress upon our minds the truth of (US), we are rationally obliged to acknowledge the falsity of ($3_N$) and hence the unsoundness of the argument from (1), (2), and ($3_N$) to (5). This renders the Necessity Model objectionable, as Williamson (2007: 185), Ichikawa and Jarvis (2009: 224-5), and Malmgren (2011: 275-77) have all observed. One objection may be stated as follows. According to philosophical orthodoxy, our thought experiment is a paradigmatic example of a *successful* falsification of the theory it was designed to test. In particular, the vast majority of contemporary philosophers would acknowledge that the Gettier intuition—the intuition that Smith's true belief is justified yet fails to constitute knowledge—has a true content which we have justification for believing. Most would even be of the opinion that we *know* the content of the Gettier intuition to be true. Upon further investigation, I suppose, it might turn out that philosophical orthodoxy is mistaken about the truth value and epistemic status of the Gettier intuition; perhaps, contra the majority view, the Gettier intuition actually is false and our belief in its content is epistemically defective. However, even if philosophical orthodoxy is mistaken on these scores, surely pointing to the obvious truth of (US) is not enough to establish the fact; the victory of scepticism about philosophical thought experimentation cannot be *that* easy. It may be helpful to restate the point here in a different way. Epistemically, our thought experiment is about as good as it ever gets

25 More precisely, a metaphysically possible world counts as deviant if the embedded existential statement $\exists x \exists p\ (GCxp \wedge (\neg JTBxp \wedge \neg Kxp) \vee (JTBxp \wedge Kxp))$ is true at it.

in philosophy. If it does not succeed, then it is doubtful whether any of the thought experiments carried out by philosophers ever succeed. Accordingly, if the Gettier intuition really expressed ($3_N$), then since the truth of (US) is so obvious, there would be no serious debate to be had about the general epistemic potency of philosophical thought experimentation; only fools would continue the struggle against metaphilosophical scepticism. But there *is* a serious debate to be had. Indeed, many philosophers are actually having it, and it is a hard intellectual fight for everyone involved.

A second objection to the Necessity Model builds upon the first one. As we have already discussed, the truth of (US) may be established simply by enriching the text of the Gettier case in a metaphysically coherent but unintended way. Given this fact, the model now under consideration implies that underspecification must be of great relevance to the way in which philosophers use thought experimentation to investigate the nature of knowledge. In particular, it should be appropriate for a proponent of the traditional theory of knowledge to challenge our thought experiment by adducing any unintended enrichment of the text of the Gettier case he is able to think of. For, by drawing our attention to such an enrichment, he would rationally compel us to acknowledge that the Gettier intuition is false and our thought experiment unsound. More generally, deviant worlds should be taken to pose a genuine and acute threat to the success of any experiments carried out within the imagination. It should be imperative for philosophers who wish to test theories by means of thought experimentation to undertake the laborious and painstaking task of constructing texts immune from deviancy. In point of fact, however, deviancy is *irrelevant*. It is appropriate to dismiss unintended enrichments and the deviant worlds they describe as wholly beside the point. Even if some philosophers do harbour sufficient reserves of ingenuity and energy to construct texts which cannot be realised in unintended ways, they are not obligated to draw on them. The reason, I think, is that unintended enrichments are *clearly* just that: unintended. If, for example, someone were to criticise our thought experiment on the basis of an unintended enrichment of the text of the Gettier case, we would surely feel no obligation to repudiate the Gettier intuition; instead, we would explain to him that he has egregiously missed the point of what we are doing. We might say to him: “In order to properly run our thought experiment, one is *obviously* not meant to imagine a possible world of the

kind you have just described. You have failed to appreciate this fact. Hence, your criticism does nothing to undermine the negative result of our thought experiment." Contra the implication of the Necessity Model, we would be perfectly *right* to say so. The status of underspecification in the dialectics of philosophical thought experimentation is therefore exactly the opposite of what the Necessity Model makes it out to be.

## 3. The Counterfactual Model

The lesson of the previous section is that the strict conditional ($3_N$) is too strong to be the real content of the Gettier intuition. The response which immediately comes to mind is that the Gettier intuition must really express a *counterfactual conditional*. According to this line of thinking, instead of expressing the very strong claim that the Gettier case metaphysically necessitates justified true belief without knowledge, our intuition expresses the much weaker claim that if the Gettier case *were* to obtain, then there *would* be justified true belief without knowledge. The idea, as stated by Williamson (2007: 186) in the lingo of Lewis-Stalnaker semantics, is that the real content of the Gettier intuition "requires justified true belief without knowledge only in the closest realisations of the Gettier case, not in all possible realisations." Williamson (2005; 2007) himself is perhaps the most prominent proponent of such a view; others include Sören Häggqvist (1996; 2009), Christian Nimtz (2010b), and Roy Sorensen (1998).[26] Borrowing Williamson's formalisation of the relevant counterfactual, we get the following model of our thought experiment, which I shall call *the Counterfactual Model*:

(1) $T \rightarrow \Box\forall x\forall p\ (Kxp \leftrightarrow JTBxp)$

(2) $\Diamond\exists x\exists p\ GCxp$

($3_{CF}$) $\exists x\exists p\ GCxp \mathrel{\Box\!\!\rightarrow} \forall x\forall p\ (GCxp \rightarrow (JTBxp \wedge \neg Kxp))$

(4) $\neg\Box\forall x\forall p\ (Kxp \leftrightarrow JTBxp)$ From (2) and ($3_{CF}$)

---

(5) $\neg T$ From (1) and (4)

26 In this connection it is worthwhile noting that, even though his metaphilosophical work has received a lot of attention in recent times, Williamson was most certainly not the first philosopher to construe philosophical thought experimentation in terms of counterfactuals. Thus the publication dates of the books by Sorensen and Häggqvist (1992 and 1996 respectively) give the lie to Ichikawa's bizarre claim that Williamson's counterfactual approach "represents a radical departure from prior philosophical theorizing about the nature of thought experiments" (Ichikawa 2009: 436).

Williamson's ($3_{CF}$) may seem odd. The relevant English counterfactual presumably goes something as follows: "If a subject were to stand to a proposition as described by the text of the Gettier case, then he would have a justified true belief in it which does not constitute knowledge." On the face of it, though, ($3_{CF}$) says something quite different and rather less natural: "If a subject were to stand to a proposition as described by the text of the Gettier case, then anyone who stood that way to it would have justified true belief in it which does not constitute knowledge." But the relevant English counterfactual is an instance of the notorious technical problem of donkey anaphora. In general, capturing the logical form of such "donkey conditionals" is a surprisingly difficult task.[27] With regard to the particular donkey conditional of relevance to us here, it is not at all obvious how we could do better than Williamson's ($3_{CF}$) as a formalisation of it. The technical details of donkey anaphora and the question of the formal adequacy of ($3_{CF}$) need not detain our discussion, however, since none of the remarks I should like to make about the programme of modelling philosophical thought experimentation in terms of counterfactuals essentially depend on whether ($3_{CF}$) gets the formalisation exactly right. For our purposes ($3_{CF}$) is close enough.[28]

The Counterfactual Model is attractive. First of all, it is able to avoid the two problems associated with the underspecification of the text of the Gettier case. If the real content of the Gettier intuition is ($3_{CF}$), then it does not follow from the truth of (US) that our thought experiment is an epistemically defective exercise based upon a falsehood. This is because, on any plausible semantics for counterfactuals, ($3_{CF}$) does not say that its antecedent metaphysically necessitates its consequent. For all semantic theory is able to tell us, both (US) and ($3_{CF}$) could be true. Another attraction

27 Geach (1962) was the first to raise the problem of donkey anaphora. His original example was the sentence "Every farmer who owns a donkey beats it." See King (2013) for a good introduction to why this sentence poses a challenge for semanticists, as well as the various proposals which they have put forward to deal with it.

28 Williamson devotes considerable attention to the problem of donkey anaphora as it applies to the relevant English counterfactual. In addition to ($3_{CF}$) he considers the formula $\forall x \forall p$ (GCxp $\Box\!\!\rightarrow$ (JTBxp $\wedge \neg$Kxp)). To my mind this is the only other potential formalisation with any plausibility; but Williamson rejects it in favour of ($3_{CF}$), for what seem to be pretty good reasons. See his (2007: 195-9). Some may find themselves attracted to the formula $\exists x \exists p$ GCxp $\Box\!\!\rightarrow \exists x \exists p$ (GCxp $\wedge$ JTBxp $\wedge \neg$Kxp)) or something like it. But close inspection reveals this to be a highly implausible formalisation. For, as Williamson (2007: 199) explains, "it does not require the instance of the Gettier case with which we started to be an instance of justified true belief without knowledge; it is satisfied if some other instance of the Gettier case is an instance of justified true belief without knowledge. That is not enough to vindicate [the relevant English counterfactual]."

of the Counterfactual Model is that it would seem to do just as good a job as the Necessity Model of approximating the natural progression of the mental activity actually involved in carrying out our thought experiment. It furnishes an elegant and sufficiently comprehensive model of the test procedure. The steps by which we transition from our imaginative engagement with the Gettier case to the conclusion that the traditional theory of knowledge is wrong are modelled neatly and clearly, and no crucial stages of the test procedure seem to have been left out. Of course, ($3_{CF}$) will be offensive to those who endorse the traditional idea that the real contents of intuitions generated by philosophical thought experimentation must somehow involve the concept of necessity. But that idea has just been shown to be on shaky ground. And, in any event, it arguably lacks the naturalness of the counterfactual approach. This has been emphasised by Williamson, who maintains that the counterfactual approach is the "natural way to articulate what is at stake with a Gettier counterexample" (2007: 204). Its naturalness is said to be rooted in the prevalence of counterfactual thinking in ordinary life. According to Williamson, "counterfactual questions arise continually in everyday thought, whereas questions of metaphysical necessity rarely arise outside philosophy, so the burden of proof is on those who claim that our initial questions about a hypothetical case are metaphysically modal rather than simply counterfactual in nature" (2007: 204). The same suggestion is at least implicit in the work of Häggqvist and Sorensen, neither of whom even thinks it worth his while to give any consideration whatever to non-counterfactual approaches.

A further and related attraction of the Counterfactual Model, also emphasised by Williamson, is that it holds out the prospect of demystifying the power of imaginative mental activity to discover truths about the natures of philosophically interesting properties and relations. Williamson warns against postulating the existence of a special cognitive capacity reserved exclusively for philosophical investigations into the natures of things. Humankind, he points out, "evolved under no pressure to do philosophy", so we should "expect the cognitive capacities used in philosophy to be cases of general cognitive capacities used in ordinary life, perhaps trained, developed, and systematically applied in various special ways" (2007: 136). Our general capacity for counterfactual thinking would seem to fit the bill perfectly. We

deploy it extensively and frequently throughout the course of ordinary life; it has become practically indispensable for many mundane yet very important cognitive tasks. For example, counterfactual thinking is used to develop plans for the future, learn from past mistakes, and reason about causal interactions.[29] This general cognitive capacity of ours, furthermore, normally functions in a reliable manner which yields knowledge, and we understand, at least in rough outline, how it is able to do so. To evaluate a counterfactual, "one supposes the antecedent and develops the supposition, adding further judgements within the supposition by reasoning, offline predictive mechanisms, and other offline judgements" (Williamson 2007: 152-3). The counterfactual is assessed to be true if one's development of its antecedent eventually leads one to add its consequent. The imaginative mental activity which underlies this evaluative process is generally reliable because it is guided by one's overall sense of how the world works, or what Williamson (2007: 143) calls "background knowledge." As he explains, "the reliability of our cognitive faculties in their online applications across a wide range of possible circumstances induces reliability in their offline applications too" (2007: 155).

We may thus feel as though we are beginning to get a grip on how it could be possible for our thought experiment to tell us something about the nature of knowledge. Williamson's alluring idea is that "the imagination is used in verifying [($3_{CF}$)] just as it is used in verifying many everyday counterfactuals, such as 'If the bush had not been there, the rock would have landed in the lake.' There is nothing peculiarly philosophical about the way in which the counterfactual is assessed" (2007: 188). More specifically, when we imaginatively engage with the Gettier case, the imagination is led to evaluate ($3_{CF}$) as true "on the basis of of an offline application of our ability to classify people around us as knowing various truths or as ignorant of them, and as having or as lacking other epistemologically relevant properties" (2007: 188). Nothing beyond this quite ordinary capacity to engage in counterfactual thought about an obvious metaphysical possibility is needed in order to carry out our test of the traditional theory of knowledge. Williamson (2005: 15) summarises the foregoing epistemological story thus: "We have our ordinary capacities for making judgements about what we encounter, and a further capacity to evaluate counterfactuals

29 For further discussion of the various ways in which we deploy our general capacity for counterfactual thought in ordinary life, see Byrne (2005).

by running those capacities 'offline'; that is already enough to get philosophy going, without any need of a kickstart from a special faculty." Even though (as Williamson himself admits) this is little more than a rough sketch of an epistemology for philosophical thought experimentation, it does seem to steer us away from the mire of epistemological obscurities which almost always attend the postulation of an extraordinary mental faculty devoted specially to the task of philosophical cognition.

### 3.1. Contingency

Insofar as it is able to avoid the problems associated with the underspecification of the text of the Gettier case, the Counterfactual Model constitutes a significant improvement upon the Necessity Model. Even so, it too must ultimately be found inadequate. The inadequacies of the Counterfactual Model stem from the fact that ($3_{CF}$) is *contingent* (as is any other plausible formalisation of the English counterfactual "If someone stood to a proposition as described by the text of the Gettier case, then he would have a justified true belief in it which does not constitute knowledge"). It will be convenient to discuss the contingency of ($3_{CF}$) and the resultant inadequacies of the Counterfactual Model in terms of the standard Lewis-Stalnaker theory of semantics of counterfactuals. For, even though that theory has been subjected to numerous criticisms, it is both evocative and easy to grasp. And just as nothing I am going to say against the counterfactual approach essentially depends on whether ($3_{CF}$) resolves the problem of donkey anaphora, nor does any of it essentially depend on whether the Lewis-Stalnaker theory gets the overall semantics of counterfactuals exactly right. Each of the objections I shall level against the Counterfactual Model has to do with the contingency of ($3_{CF}$), and surely any decent alternative semantics for counterfactuals will make ($3_{CF}$) (as well as any other plausible formalisation of the relevant English counterfactual) come out contingent.

Very roughly, the Lewis-Stalnaker theory holds that a counterfactual $p \mathrel{\square\!\!\rightarrow} q$ is true if and only if q is true at every world *nearest* to the actual world at which p is true.[30] The structure of modal space, that is to say, the relative distances between

30 The classic discussions of this theory are to be found in Lewis (1973) and Stalnaker (1968). It should hardly need to be said that my presentation of the theory is vastly oversimplified. For one thing, the general version of the theory gives conditions under which $p \mathrel{\square\!\!\rightarrow} q$ is true at a world; whereas here I have presented a restricted version of the theory which only gives conditions under which $p \mathrel{\square\!\!\rightarrow} q$ is true, that is to say, actually true, or true at the actual world. This restriction

possible worlds, is said to be determined by comparative similarity. In general, if world w is more similar to the actual world than world v is, then w is nearer than v to the actual world in modal space. The counterfactual ($3_{CF}$) is contingent because its truth value depends on the location of the actual world in that space, in particular, on the actual world's proximity to what I have been calling deviant worlds. If, on the one hand, the worlds nearest to the actual world at which the antecedent of ($3_{CF}$) is true are deviant, then ($3_{CF}$) is false; but, on the other hand, if those worlds are non-deviant, then, on the safe assumption that knowledge without justified true belief is metaphysically impossible, ($3_{CF}$) is true. So the underspecification of the text of the Gettier case is unproblematic in and of itself. The *mere* existence of deviant worlds is not enough to impugn the truth of ($3_{CF}$). A deviant world's existence is incompatible with ($3_{CF}$)'s being true if all non-deviant worlds are further away from the actual world than it is; but no deviant world's existence is incompatible with ($3_{CF}$)'s being true if there is at least one non-deviant metaphysically possible realisation of the Gettier case nearer to the actual world than all of the deviant worlds are. Location, as they say, is everything. There is thus a marked contrast between ($3_{CF}$) and a counterfactual like "If I were taller than you, then you would be shorter than me", since the latter is presumably true no matter where the actual world happens to be located in modal space.

The reasoning underlying the contingency of ($3_{CF}$) may not be immediately apparent, so it will be helpful to run though it here in greater detail. We begin by breaking down ($3_{CF}$) into its propositional constituents, namely, its antecedent, which is an existential statement, and its consequent, which is a universally quantified conditional:

(A) $\exists x \exists p\ GCxp$

(C) $\forall x \forall p\ (GCxp \rightarrow (JTBxp \wedge \neg Kxp))$

is for the sake of convenience: it obviates the need to continually add the cumbersome qualification "at the actual world" whenever speaking of ($3_{CF}$)'s truth value. Another oversimplification is inherent in the name "the Lewis-Stalnaker theory." Lewis and Stalnaker did not in fact develop a single theory of the semantics of counterfactuals. Each developed his own theory, and they famously differ on several important points, such as the validity of conditional excluded middle. See Sider (2010: 219-21) for a good summary of the differences. None of them bear on my discussion of the Counterfactual Model.

According to the Lewis-Stalnaker theory, ($3_{CF}$) is true if and only if the nearest worlds to the actual world at which its antecedent (A) is true are also worlds at which its consequent (C) is true. For the sake of convenience let us call any world at which (A) is true an "A-world". We are going to show that the truth value of ($3_{CF}$) depends on whether or not the nearest A-worlds are deviant. We first consider the case where they are deviant, since it is the most straightforward one. If the nearest A-worlds are deviant, then the existential statement embedded in (US) is true at them:

(EM) $\exists x \exists p\ (GCxp \wedge ((\neg JTBxp \wedge \neg Kxp) \vee (JTBxp \wedge Kxp)))$

Now suppose for reductio that (C) is also true at the nearest A-worlds. Then, by elementary logic, (C) and (EM) together entail the following:

(AB1) $\exists x \exists p\ ((JTBxp \wedge \neg Kxp) \wedge ((\neg JTBxp \wedge \neg Kxp) \vee (JTBxp \wedge Kxp)))$

But since its first conjunct is inconsistent with each of the disjuncts in its second conjunct, (AB1) is an absurdity. It follows that (C) must be false at the nearest A-worlds given their deviancy. So we may conclude that if the nearest A-worlds are deviant, the antecedent of ($3_{CF}$) is true at them but the consequent of ($3_{CF}$) is false at them; which is just another way of saying that if those worlds are deviant, then ($3_{CF}$) is false.

Next, consider the more complicated case where the nearest A-worlds are non-deviant. If the nearest A-worlds are non-deviant, then the negation of (EM), which is logically equivalent to a universally quantified conditional, is true at them:

(¬EM) $\forall x \forall p\ (GCxp \rightarrow (\neg(\neg JTBxp \wedge \neg Kxp) \wedge \neg(JTBxp \wedge Kxp)))$

It is easy to see that (LT) is a logical truth, and hence true in every possible world, including the nearest A-worlds:

(LT) $\forall x \forall p ((\neg JTBxp \wedge \neg Kxp) \vee (JTBxp \wedge Kxp) \vee (JTBxp \wedge \neg Kxp) \vee (\neg JTBxp \wedge Kxp))$

This logical truth depicts every permutation of the relations of knowledge and justified true belief. But its last disjunct is generally acknowledged to be ruled out by the nature of knowledge. We may thus safely assume that, as a matter of metaphysical necessity, no one ever has knowledge without justified true belief; in other words, that in every possible world, including the nearest A-worlds, (SA) is true:

(SA) $\forall x \forall p \neg(\neg JTBxp \wedge Kxp)$

Now suppose for reductio that (C) is false at the nearest A-worlds. The negation of (C) is logically equivalent to an existential statement:

(¬C) $\exists x \exists p (GCxp \wedge \neg(JTBxp \wedge \neg Kxp))$

This together with (¬EM) and (SA) entails the following by elementary logic:

(AB2) $\exists x \exists p (\neg(\neg JTBxp \wedge \neg Kxp) \wedge \neg(JTBxp \wedge Kxp) \wedge \neg(JTBxp \wedge \neg Kxp) \wedge \neg(\neg JTBxp \wedge Kxp))$

(AB2) negates every permutation of the relations of knowledge and justified true belief for at least one subject and one proposition in each of the nearest A-worlds. As a result it is inconsistent with (LT). But anything inconsistent with (LT) is an absurdity because (LT) is a logical truth. It follows that (C) must be true at the nearest A-worlds given their non-deviancy. So we may conclude that if the nearest A-worlds are non-deviant, both the antecedent and the consequent of ($3_{CF}$) are true at them; which is just another way of saying that if those worlds are non-deviant, then ($3_{CF}$) is true. Putting both of the foregoing conclusions together, therefore, we find that the truth value of ($3_{CF}$) depends on the relative distance between the actual world and deviant worlds in modal space.

### 3.2. Inadequacy of the Counterfactual Model

The contingency of ($3_{CF}$) is the source of two objections to the Counterfactual Model. The first has been forcefully pressed by Ichikawa and Jarvis (2009: 225-6; see also Ichikawa 2009: 436-8)) as well as Malmgren (2011: 78-81), and is best introduced by considering the possibility of actual world deviancy. Let us suppose that by an amazing coincidence the actual world happens to be such that a man stands to a proposition just as described by the text of Gettier case. And let us further suppose that by another amazing coincidence the actual world happens to be just as described by one of the unintended enrichments of the Gettier case I presented earlier, so that the man's true belief is either unjustified because he knows he is under the influence of hallucinogens, or constitutes knowledge because he remembers certain pertinent information about the paddocks he is passing on his walk. Then the antecedent of ($3_{CF}$), (A), is true but ($3_{CF}$)'s consequent, (B), is false. This is sufficient to fix ($3_{CF}$)'s truth value. The actual world is always the world (or among the worlds) most similar to itself. So given any true proposition whatever, the actual world counts as the world nearest to itself at which the given proposition is true. Since (A) is true, it follows that the actual world counts as the nearest world to itself at which the antecedent of ($3_{CF}$) is true. And since (B) is false, it also follows that the consequent of ($3_{CF}$) is false at the nearest world to the actual world at which the antecedent of ($3_{CF}$) is true. But then ($3_{CF}$) is false.

However, surely the Gettier intuition would *not* be made false by actual world deviancy. For consider how we would react, and indeed appropriately react, if we were to ever find out that the actual world is deviant. The improbable nature of the coincidence would certainly astonish us, but we would feel no obligation whatever to repudiate the Gettier intuition. The reason is that we would straight away recognise the deviancy of the actual world for what it *clearly* is: deviancy. We would, in other words, straight away recognise the actual world as a deviant realisation of some unintended enrichment of the text of the Gettier case. We might say: "How extraordinary! But, as amazing as it is, it's beside the point because it's *obviously* not what we had in mind when running our thought experiment." And we would be perfectly *right* to say so. A fanatical proponent of the traditional theory of knowledge

who, upon finding out about our thought experiment, began scouring the Australian countryside for evidence of actual world deviancy would truly be on a fool's errand, and not merely because of the great improbability of his ever finding his quarry. He would be searching for something which is simply *irrelevant* to the truth value of the Gettier intuition. But if deviancy is irrelevant to the Gettier intuition's truth value even when it is exhibited by the actual world, then it is surely irrelevant to the Gettier intuition's truth value when it is exhibited by any other possible world. The proximity of the actual world to deviant worlds does not determine whether or not the Gettier intuition is true. Therefore, the real content of the Gettier intuition is not given by ($3_{CF}$) (or any other plausible formalisation of the relevant English counterfactual).

Williamson anticipates this objection to the Counterfactual Model. His reply strikes me as fantastical, but it is worth examining it in detail if only because its author's reputation may impart some credibility to it. Williamson concedes that, if it were ever discovered, actual world deviancy would indeed *seem* irrelevant to the truth value of the Gettier intuition. But he claims that our situation would be analogous to that of the person in the following example:

> Suppose that someone says "Every man in the room is wearing a tie"; I look around, see a man not wearing a tie, misidentify him as Dave (who is in fact wearing a tie), and say "Dave isn't." When it is pointed out to me that Dave is wearing a tie, I deceive myself if I insist that my original reply was correct because the man whom I had in mind was not wearing a tie; that was just not the "counterexample" I actually presented. I spoke falsely when I said "Dave isn't" (Williamson 2007: 201).

Williamson is correct to hold that in this example the person's adherence to the truth of the statement "Dave isn't" would amount to self-deception. This is because by uttering "Dave isn't" the person commits himself to the world's being a certain way, in particular, to its being such that Dave is not wearing a tie. Furthermore, the person's adhering to the truth of the statement "Dave isn't" could be straightforwardly explained as a manifestation of the desire everyone has to be right. Williamson claims

that adherence to the truth of the Gettier intuition even after finding out about actual world deviancy would likewise amount to a kind of self-deception. It would do so, he says, because acceptance of the Gettier intuition also commits us to the world being a certain way, in particular, to its being such that anyone who happens to stand to a proposition as described by the text of the Gettier case has a non-knowledge justified true belief. And here as well, according to Williamson, our adhering to the truth of the Gettier intuition could be straightforwardly explained as nothing more than a manifestation of our desire to be right or, as he (2007: 201) puts it, "the common human characteristic of reluctance to admit having been wrong." Williamson then goes on to add:

> We should not distort our account of thought experiments in order to indulge that tendency. Often purported counterexamples fail for accidental reasons and can easily be repaired. To attempt to build into the counterexample in advance all repairs which might conceivably be needed is a futile exercise. It loads the purported counterexample with complexity and in the the process weakens it in other respects. The repairs need not articulate qualifications that were in some obscure sense implicit in the thought experiment from the beginning. Rather, they genuinely modify the thought experiment, but the similarity of the new thought experiment to the old one is evidence that the old one was not far wrong. (Williamson 2007: 201).

For Williamson, therefore, we would dismiss actual world deviancy as irrelevant not because it really would be irrelevant but because we generally find it unpleasant to acknowledge our own mistakes; furthermore, adherence to the truth of the Gettier intuition in the face of actual world deviancy would be quite silly, since it may be easily dealt with by adding modifications to the original text of the Gettier case as required.

One serious defect of Williamson's reply, noted by Ichikawa (2009: 438), is that his example of misidentification is importantly disanalogous to the example of known actual world deviancy. In Williamson's example of misidentification, the person's sincere assertion "Dave isn't" *obviously* commits him to the world's being

such that Dave is not wearing a tie. It is obvious, in other words, that that assertion means that Dave is not wearing a tie; it does not mean that some guy over there is not wearing a tie. Language is a public activity governed by public rules. As it is used in the person's conversational context, "Dave" is a proper name, and according to the public rules governing the use of such linguistic items, the person cannot use it to refer to just someone or other; instead, those rules imply that in the person's conversational context "Dave" must refer to a particular object, specifically, to Dave. So, given that the particular object referred to by the proper name "Dave" is in fact wearing a tie, the person's sincere assertion "Dave isn't" is most plausibly regarded as a falsehood. In contrast, it is *not* obvious that, in the example of known actual world deviancy, the Gettier intuition commits us to the world's being such that anyone who happens to stand to a proposition as described by the text of the Gettier case has a non-knowledge justified true belief. As a matter of fact, if anything is obvious about that example, it is that we would not be committed to the actual world's being that way. The upshot of the disanalogy here is that Williamson's reply does not do much to make it plausible that actual world deviancy would turn the Gettier intuition into a falsehood.

Another serious defect of Williamson's reply is his idea that actual world deviancy would put the original text of the Gettier case in need of repair. The original text of the Gettier case may be enriched in a myriad of unintended ways and, as Williamson (2007: 185) himself admits, "[a]ny humanly compiled list of such interfering factors is likely to be incomplete." It follows that if the real content of the Gettier intuition is ($3_{CF}$), then if we were to modify the original text of the Gettier case to deal with some known actual world deviancy, the apparently repaired text could *itself* turn out to be broken because of some other actual world deviancy we do not know about. And if we were to find out about this modified text's brokenness and add further modifications to deal with it, even the resultant apparently repaired apparently repaired text could *also* turn out to be broken because of yet more actual world deviancy we do not know about. And so on. This process of textual brokenness and textual repair could continue indefinitely, since the actual world might turn out to be absolutely riddled with deviancy. If so, the process would involve running a series of thought experiments, each one of which would fail to falsify the tradition-

al theory of knowledge because it would depend upon a false intuition about a broken text. But it is surely absurd to think that a situation of that kind could ever come about, even if deviancy were to be found everywhere in the actual world. Under no circumstances would anyone need to sit around in such an absurd fashion "repairing" broken text after broken text. There would be no need to do so because actual world deviancy would not break the text of the Gettier case in the first place.[31]

The second objection to the Counterfactual Model is hinted at by Ichikawa and Jarvis (2009: 225-6) and developed more fully by Ichikawa (2009: 439-40). To set it up we must consider what it would take for us to know ($3_{CF}$). The method for evaluating counterfactuals has already been roughly delineated. In general, we evaluate a counterfactual by supposing its antecedent and then imaginatively developing the supposition in accordance with relevant background knowledge. As Ichikawa and Jarvis (2009: 226) point out, if this method is to yield knowledge of ($3_{CF}$), then since ($3_{CF}$) is contingent but obviously not contingent a priori, our imaginative development of its antecedent will have to draw on "a great deal of empirical knowledge about the actual world." The requisite a posteriori knowledge must presumably include, among other things, knowledge of certain properties typically instantiated by spotters of native wildlife and individuals and organisations which deploy sophisticated robotic machinery. We need to know, for example, whether spotters of native wildlife tend to consume powerful hallucinogens prior to going on walks in the countryside. For, if they do tend to do so, then perhaps the nearest worlds in which a man stands to a proposition as described by the text of the Gettier case are also worlds in which his true belief is unjustified because he knows he is under the influence of hallucinogens. We also need to know whether individuals and organisations which deploy sophisticated robotic machines tend to allow uninformed members of the public to wander about in their vicinity, roughly, within a medium range viewing

31 Incidentally, Williamson's reply to the first objection to the Counterfactual Model significantly detracts from that model's attractiveness relative to the Necessity Model. Williamson appeals to the obvious truth of (US), that is to say, to the obvious existence of deviant worlds, as the source of his motivation for moving from the Necessity Model to his Counterfactual Model. But if the right reaction to the discovery of actual world deviancy is to admit the falsity of the Gettier intuition and repair the text of the Gettier case, then it is surely *also* the right reaction to the discovery of the mere existence of deviant worlds, whether or not the actual world is itself deviant. It would be ad hoc for Williamson to maintain otherwise. So he must, on pain of ad hocery, allow that the essence of his reply is equally available to the proponent of the Necessity Model. Of course, this observation does not—and is not intended to—demonstrate that Williamson's reply is wrong, which is why I have relegated it to a footnote. See also Malmgren (2011: 280).

distance. For, if they do not tend to do so, then perhaps the nearest worlds in which a man stands to a proposition as described by the text of the Gettier case are also worlds in which his true belief is not only justified but constitutes knowledge because he remembers being told that the only non-empty paddock on his walk contains both real and robotic kangaroos. The knowability of ($3_{CF}$) is dependent on the knowability of these and many other empirical matters.

However, there would seem to be little or no realistic prospect of our ever getting all of the a posteriori knowledge required to know ($3_{CF}$). For my own part, although I think I do know that spotters of wildlife do not tend to consume hallucinogens prior to going on walks in the countryside, I confess that I have no idea whether the individuals and organisations which deploy sophisticated robotic machines tend to allow uninformed people to wander about within eyeshot of them. And there is no realistic prospect of my ever finding out. I take it that almost everyone else is in the same boat as I am in this regard. We should thus be inclined to agree with Ichikawa and Jarvis that ($3_{CF}$) is a counterfactual which, realistically speaking at least, "cannot be known at all" (2009: 226). The epistemological attractiveness of the Counterfactual Model turns out to be rather specious.[32] Note that acknowledging this fact is quite compatible with also acknowledging that we possess a general capacity for counterfactual cognition. There is no risk here of our falling into anything like universal scepticism about counterfactual knowledge. "One may," as Ichikawa explains, "admit the general capacity to evaluate counterfactuals, while remaining skeptical about one's ability to know a particular counterfactual" (2009: 440). This should come as no surprise because it is easy to see that the same thing must hold for every other general cognitive capacity of ours. To borrow an example from Ichikawa: although Williamson possesses a general perceptual capacity to identify the colours of objects, he has no way of knowing what colour shirt I am wearing at the present moment.

Since knowledge of ($3_{CF}$) is pretty much unattainable for us, it follows that if ($3_{CF}$) is the real content of the Gettier intuition, then our thought experiment is epi-

32 The difficulty of our coming to know ($3_{CF}$) has also been observed by D. Sosa (2006: 640), who writes: "It is not as if we have extensive practical experience with subjects in circumstances similar to those of the envisaged case and have found, by something like empirical investigation, that they have tended to have justified ignorance, so that we are now in a position to assert the counterfactual with confidence that rests on that experience."

stemically defective because its negative result depends on an intuition which, for all we are able to determine, may very well be false. Furthermore, we should be able make rather *major* improvements to our thought experiment by making rather *minor* modifications to the text of the Gettier case, specifically, minor modifications which compensate for our ignorance of relevant empirical facts. We could, for example, change the text to make it specify that the organisation responsible for deploying the robotic kangaroos allows uninformed members of the public to walk around near the paddock in which those sophisticated robots are enclosed; this minor textual change would completely obviate the necessity of finding out whether the individuals and organisations which deploy sophisticated robotic machines tend to allow such behaviour. We could then carry out a new thought experiment by imaginatively engaging with the modified text, which may be represented using the open formula $GC_Mxp$. If we did so, then according to Williamson we should undergo a new intuition with the following counterfactual as its real content:

$$(3_{CFM})\ \exists x \exists p\ GC_Mxp \mathrel{\Box\!\!\rightarrow} \forall x \forall p\ (GCxp \rightarrow (JTBxp \wedge \neg Kxp))$$

Knowledgeable evaluation of this counterfactual must also draw on a great deal of a posteriori background knowledge, but due to the compensation built into the text represented by $GC_Mxp$, it is going to be much easier for us to come to know $(3_{CFM})$ than $(3_{CF})$. Our new thought experiment should thus be a *much better* thought experiment than our original one. But surely it would *not* be; the aforementioned minor modification to the text of the Gettier case would simply not bring about a major improvement, or indeed any improvement at all. The fact of the matter, as Ichikawa (2009: 440) points out, is that the texts represented by $GCxp$ and $GC_Mxp$ would "function equally well [or, perhaps, equally badly] for establishing knowledge of the conclusion of the Gettier argument." We may therefore conclude that the counterfactual approach to modelling philosophical thought experimentation is liable to give rise to *distortions* of the relative merits of case descriptions. This is another reason to think that the real content of the Gettier intuition is not given by $(3_{CF})$ (or any other plausible formalisation of the relevant English counterfactual).

### 3.3. Some bad objections

The Counterfactual Model, it will be recalled, has considerable attractions, so in order to do it justice we must make sure that we condemn it on the basis of the most compelling reasons. Ichikawa and Jarvis's critique of the Counterfactual Model fails to do it justice because they rather confusingly blend some bad objections with the two good ones set forth in the previous sub-section. One of the bad objections we can extract from their work appeals to the "traditional understanding of philosophical methodology" to which they and many other contemporary philosophers give their endorsement. In addition to the idea that the real contents of intuitions like the Gettier intuition must somehow involve the concept of metaphysical necessity, another defining element of traditionalism is the idea that philosophical thought experimentation is an a priori activity. The arch-traditionalists Bealer and BonJour, for example, both uphold the latter idea. Bealer says that philosophical thought experimentation is a "procedure of a priori justification" (1996a: 4; 1996b: 122); while BonJour rather more dramatically declares that "philosophy is a priori if it has any intellectual standing at all" (1998: 106). But we have already observed that, since ($3_{CF}$) is contingent but obviously not contingent a priori, any knowledge of it has to be a posteriori. This makes ($3_{CF}$) repugnant from the very get go to the traditionalist sensibilities of Ichikawa and Jarvis. They are moved to write: "If the Gettier intuition is like this [i.e. contingent], then it is not the sort of thing that traditional philosophy takes it to be. […] [It] cannot be knowledge a priori. This is, we think, sufficient reason to look further for a treatment of thought experiment intuitions" (2009: 226). In the view of Ichikawa and Jarvis, therefore, the aposteriority of ($3_{CF}$) is *itself* enough to render the Counterfactual Model inadequate.

However, Ichikawa and Jarvis are here putting more dialectical weight on their "traditional understanding of philosophical methodology" than it is able to support —or at least more than they have *shown* it is able to support. If the aposteriority of ($3_{CF}$) is to provide us with something in the vein of a sufficient reason to reject the Counterfactual Model, then surely we require a vindication of the idea that philosophical thought experimentation is an a priori activity; that idea cannot simply be taken for granted. After all, even though there are many contemporary philosophers who endorse it, there are also many who have argued strongly against it, and they

are hardly fringe cranks. It is nowadays a matter of great controversy in mainstream philosophy whether we can have a priori knowledge of anything at all, let alone a priori knowledge of the natures of philosophically interesting properties and relations. Ichikawa and Jarvis (2009: 240-3) do contend that there is "a burden of proof argument" in their favour, but their discussion of why the burden must fall on the shoulders of non-traditionalists is sketchy at best and, in their own words, "somewhat speculative." Since they neither vindicate the apriority of philosophical thought experimentation nor tell us where such a vindication is to be found, their claim that the aposteriority of ($3_{CF}$) is sufficient reason to reject the Counterfactual Model should be taken with a grain of salt. We should also be wary about the unhelpful and unnecessary partisanship it introduces into our model-building project. Perhaps philosophical thought experimentation really is a posteriori. It may be, for example, that the real content of the Gettier intuition is a posteriori not because knowledge of it especially depends on the a posteriori knowledge required to know ($3_{CF}$), but because epistemic justification is holistic. Even if the real content of the Gettier intuition is a posteriori in this way, the Counterfactual Model is still inadequate—both for the reason that actual world deviancy would make ($3_{CF}$) but not the Gettier intuition false (our first good objection), and for the reason that that model can distort the relative merits of case descriptions (our second good objection). The perpetual war between traditionalists and radical empiricists is just a distraction.

Another bad objection which we can extract from Ichikawa and Jarvis's work attempts to build upon our first good objection. Consider a situation in which the actual world happens to be deviant and we carry out our thought experiment in ignorance of the fact. If the Counterfactual Model adequately models our thought experiment, then this situation is one in which we come to truly believe that the traditional theory of knowledge is wrong by validly deducing its negation from our true beliefs in (1) and (2) and our false belief in ($3_{CF}$). But due to the falsity of ($3_{CF}$) the thought experiment we carry out must be defective. With this in mind, Ichikawa and Jarvis (2009: 226) go on to claim: "As we all know, deducing a true belief from a false belief, even a justified one, does not confer knowledge. So Williamson's view implies that in this case […] a defective thought experiment has generated a Gettier case about the analysis of knowledge! This, we take it, is unacceptable." I shall call

the situation on which the foregoing objection is based *the ironic Gettier case*.

It is worthwhile pointing out that the ironic Gettier case threatens to infect Ichikawa and Jarvis's critique of the Counterfactual Model with incoherence. In the situation now under consideration, we end up with a true and justified belief in the negation of the traditional theory of knowledge *only if* our true beliefs in (1) and (2) and our false belief in ($3_{CF}$) are all justified. If this necessary condition is not satisfied, then our defective thought experiment does not, as Ichikawa and Jarvis claim, generate a Gettier case about the analysis of knowledge. It might be wondered whether Ichikawa and Jarvis can allow that such a condition could ever be satisfied. As we have already seen, they are of the view that ($3_{CF}$) "cannot be known at all", not even when it is true. But if ($3_{CF}$) cannot be known by us even when it is true, then we cannot have a justified belief in it either. (The reasoning behind this conditional is straightforward. In general, unless it is Gettierised, a justified belief must constitute knowledge when its content is true. And if we can have a justified belief in ($3_{CF}$), then surely we can do so without our justified belief being Gettierised. From these obviously true premises it follows that we can have a justified belief in ($3_{CF}$) only if we can know ($3_{CF}$) when it is true; but this is just the contrapositive of the relevant conditional.) So Ichikawa and Jarvis would seem to be committed to the impossibility of our having a justified belief in ($3_{CF}$), which would in turn seem to commit them to denying that the relevant necessary condition is satisfiable. They are here on the verge of incoherence. Of course, the question of whether Ichikawa and Jarvis actually do fall into incoherence depends on whether they mean to claim that it is *metaphysically* impossible for us to know ($3_{CF}$) or, as I more charitably suggested above, only that there is no *realistic* prospect of our ever coming to know it. If their claim is the former one, then the incoherence of their critique of the Counterfactual Model is blatant; if it is the latter, then they may coherently hold that the ironic Gettier case is a metaphysical possibility, albeit a very unrealistic one. They do not make it clear where they stand on this matter, however.

Most importantly, though, nothing about the ironic Gettier case would seem to be problematic for the Counterfactual Model *other than that* it describes a situation in which actual world deviancy makes ($3_{CF}$) but not the Gettier intuition false. In particular, the fact that the ironic Gettier case is a Gettier case is neither here nor there.

And this is just what we ought to expect. For we possess many cognitive capacities in addition to the imaginative ones which underlie philosophical thought experimentation (such as our various perceptual capacities, mnemonic capacities, and introspective capacities), and most if not all of them can become implicated in Gettier-style cases. It would be very surprising if it turned out that our imaginative capacities could not become implicated in Gettier cases as well. The ironic Gettier case therefore fails to build upon our first good objection; it adds nothing to that objection at all. In defence of Ichikawa and Jarvis here, it may perhaps be suggested that they intend the ironic Gettier case to merely facilitate comprehension of our first objection, rather than to add anything to it. But the ironic Gettier case is unhelpful in this regard because no help is required: our first objection is easy enough to comprehend in itself. The ironic Gettier case does nothing but generate confusion as to whether Ichikawa and Jarvis's critique of the Counterfactual Model is even coherent.

## 4. The Possibility Model

In the previous section we learned that the Counterfactual Model is also too demanding. Williamson's ($3_{CF}$) requires that the nearest A-worlds to the actual world be non-deviant, but the Gettier intuition imposes no such requirement upon modal space. One response might be to weaken ($3_{CF}$) to its mere possibility:

($3_{CFP}$) $\Diamond(\exists x \exists p\ GCxp \mathrel{\Box\!\!\rightarrow} \forall x \forall p\ (GCxp \rightarrow (JTBxp \wedge \neg Kxp)))$

This possibility claim only requires that there be some world such that the nearest A-worlds to it are non-deviant; it makes no difference whether the world is the actual world or some other one. The requirement is presumably satisfied; ($3_{CFP}$) is true. Moreover, within the modal logical system S5, ($3_{CFP}$) along with (1) and (2) provides us with a valid argument to the conclusion that the traditional theory of knowledge is wrong. But here we might complain with Williamson (2007: 202) that "it is strained to attribute the commitment to S5 to people who have never considered matter." (Neither the models we have already looked at, nor any of the ones we are yet to look at, commit us to making such a strained attribution.) And we might also complain, this time with Malmgren (2011: 281), that ($3_{CFP}$) "seems like overkill, given that there is another, simpler possibility claim in the vicinity." The possibility claim Malmgren has in mind forms the core of the following model of our thought experiment, which I shall call *the Possibility Model*:

(1) $T \rightarrow \Box \forall x \forall p\ (Kxp \leftrightarrow JTBxp)$

($2_P$) $\Diamond \exists x \exists p\ (GCxp \wedge JTBxp \wedge \neg Kxp)$

(4) $\neg \Box \forall x \forall p\ (Kxp \leftrightarrow JTBxp)$ From ($2_P$)

---

(5) $\neg T$ From (1) and (4)

On this model the intended real content of the Gettier intuition is ($2_P$), the claim that it is metaphysically possible for someone to stand to a proposition as described by the text of the Gettier case, have a justified true belief in it, but not know it.

As Malmgren has emphasised, the Possibility Model is attractive principally because it is able to avoid the problems faced by the other two models we have discussed so far. First of all, the Possibility Model avoids the problems which render the Necessity Model inadequate. It is able to do so because the underspecification of the text of the Gettier case has no impact on the truth value of ($2_P$). More specifically, the mere fact that there are deviant worlds is not enough to make ($2_P$) false; ($2_P$) is compatible with the truth of (US). The Possibility Model also avoids the problems which render the Counterfactual Model inadequate. It is able to do so because the relative distance between the actual world and deviant worlds has no impact on the truth value of ($2_P$). More specifically, if the nearest A-worlds to the actual world happen to be deviant, that fact is not enough to make ($2_P$) false; ($2_P$) is true no matter where in modal space the actual world happens to be located. All of these problems are avoided by the Possibility Model, yet it still manages to provide us with a valid argument to the conclusion that the traditional theory of knowledge is wrong. It would therefore seem as though the Possibility Model does a much better job of representing our thought experiment than the Necessity and Counterfactual Models do, at least with regard to the irrelevance of deviancy considerations for the truth value of the Gettier intuition.

In addition to these attractions, the Possibility Model may also present those sympathetic to the "traditional understanding of philosophical methodology" with an alluring compromise in light of the inadequacy of the Necessity Model. For, although ($2_P$) does not directly involve the concept of necessity, it is presumably a non-contingent truth. And there seems to be no good reason why, if we are able to know some possibility claims a priori, we could not know ($2_P$) a priori as well. Traditionalists could even modify the Possibility Model in order to bring it further into line with their desiderata. If, for example, ($2_P$) is replaced with its necessitation (i.e. $\Box\Diamond\exists x\exists p$ (GCxp $\wedge$ JTBxp $\wedge \neg$Kxp)), the result is a model on which the real content of the Gettier intuition is given by something which not only involves the concept of

necessity, but can at least arguably be known a priori. Since everything I am going to say about the Possibility Model applies equally to this traditionalist variant of it, I shall set the latter aside in what follows.

Another potential attraction of the Possibility Model is the simple structure it attributes to our thought experiment. This has been emphasised by David Sosa. According to Sosa, it is doubtful whether "there is even as much structure in the Gettier cases as Williamson supposes" (2006: 642). We should, he says, regard the Necessity and Counterfactual Models as exaggerating the structural complexity of our thought experiment, because our imaginative engagement with the Gettier case does not seem to get us to the negation of (MC) "as the result, in any way, of a *derivation* from the possibility of the case together with [a counterfactual] or [a strict conditional]" (2006: 642). Although Sosa does not elaborate on this observation, I suspect he is impressed by how natural we find it to speak of the Gettier intuition "clashing" with the traditional theory of knowledge. The naturalness of that locution certainly does make it seem as though there must be an incompatibility between the Gettier intuition and the traditional theory of knowledge. But if the real content of the Gettier intuition is ($2_P$), then there really is an incompatibility between them because ($2_P$) directly entails the negation of (MC) and hence also the negation of the traditional theory of knowledge. Neither the Necessity Model nor the Counterfactual Model is able to accommodate the naturalness of the "clash" locution in such a straightforward manner; proponents of those models are forced to treat it as misleading. So, owing to its simplicity, the Possibility Model would also seem to do a better job of representing our thought experiment in this regard. But no matter how natural it is to speak of the Gettier intuition clashing with the traditional theory of knowledge, we must take care not to overestimate the attractiveness of the Possibility Model's simplicity. For it has a significant trade-off. As I discussed earlier (in Section 2), proponents of the Necessity Model or any other model with a similar structure, such as the Counterfactual Model, are able to explain away the naturalness of the "clash" locution by appealing to the fact that, for most philosophers, the metaphysical possibility of the Gettier case is typically not in any doubt and hence not worth mentioning. This approach, furthermore, receives independent support from the practice of empirical experimentation: scientists often find it natural to speak of clashes

between observations and theories which are in all strictness compatible with one another, the explanation being that descriptions of the initial conditions of experimental set-ups are often not in any doubt and hence not worth mentioning. The Necessity and Counterfactual Models are therefore suggestive of a pleasing and plausible uniformity of structural complexity between thought experimentation and empirical experimentation. But such uniformity is uncongenial to the simplicity of the Possibility Model.

It could be tempting to think that the Possibility Model involves a redundant element which may be excised without affecting the model's adequacy. This temptation is felt by Sosa, who writes: "it is not clear that an objection to the claim that knowledge is equivalent to justified true belief needs to depend on more than the observation that, possibly, an agent could have justified true belief without knowledge" (2006: 642). In other words, Sosa entertains and may actually endorse a variant of the Possibility Model where ($2_P$) is replaced by the following:

($2_{PS}$) $\Diamond\exists x\exists p$ (JTBxp $\wedge$ $\neg$Kxp)

Sosa's variant of the Possibility Model excises ($2_P$)'s reference to the Gettier case. And, to be sure, there is no question that ($2_P$)'s reference to the Gettier case is wholly superfluous insofar as the validity of the argument from (1), ($2_P$), and (4) to (5) is concerned. However, we are trying to find the answer to the content problem, and Malmgren has shown that, given this objective of ours, the replacement ($2_P$) with ($2_{PS}$) must lead to complete disaster. She begins by pointing out "that an adequate content proposal should generalize in natural ways to intuitive judgements other than the Gettier judgement, and that it should not ride roughshod over our pre-theoretical classifications of those judgements" (2011: 283). She then continues as follows:

> But [($2_{PS}$)] does precisely that (on what looks like the only natural way to generalize the proposal). Simply put: it is hard to see how we could accept [($2_{PS}$)] —as an analysis of the Gettier judgement—without committing to giving exactly the same analysis of, for example, the intuitive judgement that might be expressed by saying 'Jill has a justified true belief but does not know that the

president has been assassinated', and of the judgement that might be expressed by saying 'Henry has a justified true belief but not know that there is a barn in front of him.' But it is absurd to suppose that the Gettier judgement is the *same judgement as*—that it has the same content as—either of those judgements (Malmgren 2011: 284).

The temptation to excise ($2_P$)'s logically superfluous reference to the Gettier case should therefore be resisted by anyone sympathetic to the Possibility Model.

### 4.1. Rational commitment

The Possibility Model has some significant advantages over the Necessity and Counterfactual Models, but even it must ultimately be found inadequate. The inadequacies of the Possibility Model stem from the fact that there are *rational commitments* associated with the practice of thought experimentation. Rationality imposes various requirements on our minds. In general, if rationality requires one to F when one is in the mental state (or collection of mental states) M, then we say that one's being in M rationally commits one to F. The requirement to avoid contradictory beliefs is a paradigmatic example: the state of believing p rationally commits one to refrain from believing p's negation. Other examples include requirements to believe what one believes to be entailed by one's beliefs, and to intend the necessary means to one's ends. A fully rational agent fulfils all of its rational commitments. For any such agent, there is no M such that the agent is in M, M rationally commits the agent to F, but the agent does not F. So, for instance, no fully rational agent both believes p and believes p's negation, because such a combination of mental states would be incoherent. But although the failure to fulfil just one rational commitment is enough to make one less than fully rational, it need not make one irrational overall. Rationality and irrationality are matters of degree. Since the phenomenon of rational commitment is a very general one, it should come as no shock to find it associated with the practice of thought experimentation.[33]

33 See Broome (1999) for the seminal discussion of rational commitment. Since Broome's paper, two interrelated problems have come to dominate the debate. One is the problem of whether the requirements of rationality are wide scope or narrow scope. The other is the problem of whether the requirements of rationality have any normative force. For a good recent overview of both problems, see Way (2010). My discussion of rational commitment in thought experimentation does not presuppose any particular answers to them, so there is no need for me to take a stand on

It is easy to appreciate how our particular thought experiment can lead to the incurrence of certain rational commitments. The first thing we do when we carry out our thought experiment is bring the Gettier case before the mind and apprehend its metaphysical possibility in imaginative thought. This imaginational state consists at least in part of a modal judgement or belief. Next, we proceed to ask ourselves about the epistemic status of Smith's true belief, which induces us to undergo the Gettier intuition, that is to say, to intuit the following:

(I1) Smith has a justified true belief that there are kangaroos in the paddock, but he does not know that there are.

The Gettier intuition, as I explained earlier, may or may not be a belief. But whatever the nature of the Gettier intuition turns out to be, when we carry out our though experiment we somehow come to form a judgement or belief which we express using (I1). Doing so amounts to *taking a stand* on the question of the epistemic status of Smith's true belief. We accept that the verdict delivered by the tribunal of our own imagination is the *correct verdict* about the distribution of the relations of knowledge and justified true belief in the Gettier case. In principle, however, we need not have undergone the Gettier intuition at all. Although the Gettier intuition is probably what most philosophers would undergo if they were to imaginatively engage with the Gettier case, other intuitions about the epistemic status of Smith's true belief are possible.[34] In particular, instead of (I1), we could have intuited either of the following:

(I2) Smith has an unjustified true belief that there are kangaroos in the paddock, and he does not know that there are.

(I3) Smith has a justified true belief that there are kangaroos in the paddock,

---

them here.

34 And it very well may turn out that there are many people who would undergo an intuition other than the Gettier intuition. After all, there is a growing body of research suggesting that non-philosophers do not react to Gettier-style cases in the way most (or many) philosophers do. See Starmans and Friedman (2012) and Weinberg, Nichols, and Stich (2001). However, there is also research suggesting the contrary; see Turri (2013).

and he knows that there are.[35]

But as it happens, we do not intuit them and so we do not come to form any judgement or belief which we would express using (I2) or (I3). Insofar as we give any consideration to these alternative distributions of the relations of knowledge and justified true belief, we reject them as *incorrect* verdicts about the epistemic status of Smith's true belief. So carrying out our thought experiment puts our minds in at least two mental states: a state of apprehending in imagination the metaphysical possibility of the Gettier case and a belief state which we use (I1) to express.

Now, in and of itself, there is nothing wrong with that pattern of mental states, but the mere fact that our minds exhibit it certainly has ramifications for what *else* we are able to believe about Smith's epistemic situation without falling foul of the requirements of rationality. Let us suppose that someone who carries out our thought experiment holds an (I1)-belief and then comes to hold an (I2)-belief as well. At one and the same time, this person would not only hold the beliefs that the Gettier case is metaphysically possible and that Smith's true belief is justified, but he would also believe Smith's true belief is unjustified. Or suppose instead that this person also comes to hold an (I3)-belief. Then he would not only hold the beliefs that the Gettier case is metaphysically possible and that Smith's true belief does not constitute knowledge, but he would also believe Smith's true belief does constitute knowledge. It is *obvious* that there is something intrinsically wrong with *those* two patterns of mental states. They strike us as irrational; to exhibit either pattern of mental states is to suffer from a genuine failure of reason. Indeed, if someone were to tell us that he holds one or the other of those combinations of beliefs, then although we would perhaps initially regard him as joking, or even as using words in non-standard ways, we would have to regard him as at least somewhat mentally disordered if he continued to press the point. Our belief in the Gettier case's metaphysical possibility in combination with our (I1)-belief simply cannot be made to fit coherently together with an (I2)-belief or an (I3)-belief. In virtue of holding them, we incur both a rational

35 In addition to (I2) and (I3), there is perhaps one other thing we could have intuited: (I4) Smith has an unjustified true belief that there are kangaroos in the paddock, but he knows that there are. Of course, even if it is possible to intuit (I4), the manifest incoherence of the idea of knowledge without justification makes it is highly improbable that anyone would ever do so. I shall set aside (I4) in what follows because it can only serve to add unnecessary complications to our discussion; I can make the points I want to make by dealing exclusively with (I1), (I2), and (I3).

commitment to refrain from holding an (I2)-belief and a rational commitment to refrain from holding an (I3)-belief.

### 4.2. Inadequacy of the Possibility Model

It is widely acknowledged by philosophers that rational commitments are not brute facts. They are phenomena in need of explanation. And, for at least some rational commitments, there are good explanations already available. The rational requirement to avoid holding beliefs in both p and p's negation, for example, is explicable in terms of the nature of the mental state of belief and the logico-analytical relationship between the contents p and not-p. Believing p rationally commits us to refrain from believing p's negation because, first, it is in the nature of belief to aim at truth[36] and, second, p and not-p stand in the contradictory relation to one another (i.e. one is true if and only if the other is false). Explanations of rational commitments need not be of one kind only. Some rational commitments may admit of explanations similar to the one just presented, while yet others may call for different explanatory strategies altogether. But either way there are no explanatorily brute requirements of rationality. What is most important for present purposes is that an explanation is needed for the two rational commitments we incur in virtue of holding a belief in the Gettier case's metaphysical possibility together with an (I1)-belief. No approach to modelling philosophical thought experimentation can be adequate if it fails to furnish the resources for such an explanation, or at least, if it gets in the way of the provision of such an explanation. The objection I should now like to level against Malmgren's possibility approach is that it does just that: it makes it impossible to provide a explanation of the relevant rational commitments.

It is natural to think that the most straightforward way to go about explaining those rational commitments is in terms of the nature of belief and the logico-analytical relationships between contents. An explanation of that kind immediately comes to mind because belief aims at truth, the Gettier case is a metaphysical possibility,

36 The ideas that belief has an aim and that its aim is truth are, of course, metaphorical. Although they are very often taken for granted by philosophers, in recent years a debate has erupted around the question of how to best interpret them literally. See the papers collected by Chan (2013) for some representative contributions to this debate. The explanation I am setting forth in the main text does not oblige me to attach myself to any particular literal interpretation of the metaphor; though it does depend on the assumption that there is at least one such interpretation which is both intelligible and plausible. I take this assumption to be a safe one.

and (I1), (I2), and (I3), which are claims about the distribution of the relations of knowledge and justified true belief in that metaphysical possibility, would seem to be incompatible. They may be said to stand in something like the contrary relation to one another: given any two of them, both may be false but at most one can be true. However, no explanation along such lines is available to Malmgren. Since she holds that ($2_P$) is the real content of the Gettier intuition, she is committed to holding that the real contents of (I2) and (I3) must *also* be possibility claims, in particular, possibility claims of the *same kind* as ($2_P$). For it would be exceedingly bizarre if (I2) and (I3) turned out to have real contents of a different kind to the real content of the Gettier intuition. The respective possibility claims Malmgren must give as the real contents of (I1), (I2), and (I3) are as follows:

($2_P$) $\Diamond\exists x\exists p\ (GCxp \wedge JTBxp \wedge \neg Kxp)$

(PC2) $\Diamond\exists x\exists p\ (GCxp \wedge \neg JBxp \wedge TBxp \wedge \neg Kxp)$[37]

(PC3) $\Diamond\exists x\exists p\ (GCxp \wedge JTBxp \wedge Kxp)$[38]

But due to the underspecification of the Gettier case, these possibility claims are not

37 It might be wondered why I have used $\neg JBxp \wedge TBxp$ rather than $\neg JTBxp$ in (PC2). After all, not only did I use $\neg JTBxp$ in (US1) and (US) when I was discussing the underspecification of the Gettier case, but it is a simpler formula than the conjunction $\neg JBxp \wedge TBxp$. My reason for preferring $\neg JBxp \wedge TBxp$ over $\neg JTBxp$ in the present context is that the latter formula is too coarse grained to fully capture the meaning of the English expression "unjustified true belief" which occurs in (I2). The coarseness of $\neg JTBxp$ is due to the fact that a subject's failure to have a justified true belief in a proposition does not imply that he has an unjustified true belief in it. He may have a justified false belief in the proposition, or no belief in it at all. (The converse implication, of course, does hold: a subject's having an unjustified true belief in a proposition implies that he fails to have a justified true belief in it.) The point here may be put another way. The formula JTPxp is just an abbreviation for the formula $JBxp \wedge TBxp$. So a subject-proposition pair satisfies JTBxp if and only if it satisfies $JBxp \wedge TBxp$. Suppose such a pair fails to satisfy JTBxp. Then it either fails to satisfy JBxp or it fails to satisfy TBxp. It may fail to satisfy JBxp while satisfying TBxp. But it may also satisfy JBxp while failing to satisfy TBxp, or it may fail to satisfy both JBxp and TBxp. Consequently, JTBxp is unsuitable for capturing the meaning of "unjustified true belief"; we need $JBxp \wedge TBxp$ to do the job. Note that I used $\neg JTBxp$ rather than $\neg JBxp \wedge TBxp$ in (US1) and (US) because, logically speaking, $\neg JTBxp$ is easier to work with than $\neg JBxp \wedge TBxp$, and there was no need for me to fully capture the meaning of "unjustified true belief" when I was discussing the underspecification of the Gettier case.

38 This possibility claim is the same as (US2). I have labelled it "(PC3)" here because Malmgren's association of it with (I3) is what matters most in the present context.

only *perfectly compatible*, they are also all *true*. They cannot be said to stand in anything like the contrary relation to one another: given any two of them, the truth of one does not rule out the truth of the other. Our imaginative engagement with the Gettier case, furthermore, essentially involves no mental state the content of which could be combined with ($2_P$), (PC2), and (PC3) to generate the required incompatibilities. It is at this point that Malmgren's possibility approach runs out of resources. It cannot deliver the incompatibilities which are necessary for an explanation in terms of the nature of belief and the logico-analytical relationships between contents. If the real contents of (I1), (I2), and (I3) are given by ($2_P$), (PC2), and (PC3) respectively, then an altogether different kind of explanatory strategy must be found.

Malmgren anticipates this challenge. Her response is that the relevant rational commitments tell us "something about the *grounds*, not the content, of the intuitive judgement" (2011: 291). The main idea of Malmgren's alternative explanatory strategy is that "[t]here is a certain generality to my grounds (or *reasons*) for judging that Smith has a justified true belief but does not know […] and this generality of grounds rationally constrains my options when it comes to making certain other judgements" (2011: 291-2). To introduce this idea, Malmgren asks us to consider a situation in which one of her students, student A, puts her feet up on the table and Malmgren judges that she should put them down. "Other things equal," observes Malmgren, "I can be expected to react in the same way to the next student who puts her feet up […] If I do not, then I betray some kind of inconsistency or confusion" (2011: 292). The irrationality of making a divergent judgement about whether the next student, student B, should also put her feet down is then said to be explicable as follows:

> [N]ormally, the reasons for which an (overall reasonable) person would judge that a given student should put her feet down have a certain generality—perhaps they apply to everyone in the room, or at least to everyone in the room whose feet are dirty. What my divergent judgements betray is that I did not in fact base my original judgement on reasons of the presumed generality (even though, perhaps, I *should* have done so)—*or* I did, but I failed to see that those reasons applied to the next student too. A third possibility is that my reasons

> were defeated in some non-obvious way in the latter situation. Any which way, the tension between my two judgements reflects something about the reasons on which they are based (Malmgren 2011: 292).

By way of elaboration, Malmgren adds that the grounds on which her original judgement are based include, first, "the minor premiss that [student A] put her feet on the table" and, second, "a major premiss roughly to the effect that anyone who puts their feet up in my seminar—and, perhaps, meets some further specification—should take them down" (2011: 293). The original judgement thus "results from the application of a general principle or rule to a particular instance, an instance that falls under it (and/or is taken by the subject to fall under it)" (2011: 293). A divergent judgement about student B would be irrational because, in Malmgren's words, "'if I do it once I should do it twice'—I should apply that principle or rule in any other circumstance that is *relevantly similar* to that in which I first applied it […] This point is sometimes expressed by saying that *reasons must be consistently applied*" (2011: 293).

Having in this way illustrated that the generality of grounds is a real phenomenon with the potential to do explanatory work, Malmgren continues as follows:

> [M]y suggestion is that the rational commitment that an intuitive judgement seems to 'bring on' is fundamentally the same kind of commitment that is manifest in the above examples—a commitment that can be found in any reason-based activity (in the practical as well as the epistemic domain). The nature of this commitment is by no means fully, or even particularly well, understood as yet. But it is a commitment that we are all familiar with. It is implicitly invoked at any time someone is called on to defend why she made a certain judgement (or performed a certain action) in a circumstance $C_1$—given that $C_1$ seems to match another circumstance $C_2$ in all relevant respects, and that, in $C_2$, she made a judgement (or performed an action) of a contrasting type. (Malmgren 2011: 294-5)

According to the explanatory strategy which emerges from Malmgren's discussion, when we carry out our thought experiment, we have grounds or reasons for making

a certain judgement about the epistemic status of Smith's belief, some or even all of which may come from our imaginative engagement with the Gettier case. We respond to these grounds by forming an (I1)-belief on their basis. They are grounds which have a certain generality. And it is *because of* their generality that making divergent (or "contrasting") judgements about the epistemic situation of the subjects in cases relevantly similar to the Gettier case would be irrational. For example, if we now go on to imaginatively engage with a case which is the same as the Gettier case except that it specifies that Smith has two sons, or that he dislikes cabbage, or the case features someone called "Brown" rather than "Smith", then we must refrain from making a judgement about the subject's epistemic situation which diverges from our original (I1)-belief. A divergent judgement here would be irrational because of the generality of the grounds for that original belief of ours. The rational constraint imposed by its grounds would, as it were, extend beyond our thought experiment involving the Gettier case to any thought experiment involving a relevantly similar case. But the limit of similarity is identity. At the limit, we would, if we were irrational, make a judgement about the epistemic situation of Smith in the Gettier case, and then go on to make a further and divergent judgement about the epistemic situation of the same subject (Smith) in the same case (the Gettier case). The explanation for why that pattern of mental states would be irrational—specifically, for why we are rationally committed to avoid forming either an (I2)-belief or an (I3)-belief when we already hold an (I1)-belief—is again to be given in terms of the generality of the grounds for our original (I1)-belief.

Malmgren does not say much about what she thinks the grounds or reasons for our (I1)-belief might be. She tells us that, in holding an (I1)-belief, we are "attributing justified true belief without knowledge to a subject who stands to a proposition in a certain peculiar way" and "it seems clear that we are attributing those properties to him in part *because we take him* to stand to a proposition in that peculiar way" (2011: 296). She adds that "it seems very plausible" that our imaginative engagement with the Gettier case "has more than *causal* significance—specifically, that (some or all) of the information that is explicitly stipulated in that description constitutes a *reason* for [us] to make a certain intuitive judgement" (2011: 297). This reason is then said to be the possibility claim $\Diamond\exists x\exists p$ GCxp. Malmgren acknowledges

that ◇∃x∃p GCxp by itself is an insufficient ground for believing what she takes to be the real content of (I1), in other words, for believing ($2_P$). She worries about "how this [i.e. ◇∃x∃p GCxp] could be among our *reasons* for believing [($2_P$)]—how a claim of the form 'possibly p' could be a reason for believing a claim of the form 'possibly p & q'" (2011: 296, fn.53). There must be at least one additional reason if, as Malmgren holds, the real content of (I1) is ($2_P$) and our belief in it is a justified one based on sufficient grounds. At no stage in her discussion, however, does Malmgren spell out the additional reason(s). One option is the strict conditional ($3_N$). But since ($3_N$) is false and known to be so, Malmgren would probably not wish to accept it as being among our grounds for believing ($2_P$). A second option is the counterfactual ($3_{CF}$). But since ($3_{CF}$)'s truth value depends on the location of the actual world in modal space, and its epistemic standing for us is rather doubtful, this is another conditional Malmgren probably would not wish to accept as being among our grounds for believing ($2_P$). It is by no means obvious what she would accept instead. As a consequence of this lack of detail, Malmgren's alternative explanatory strategy, which makes so much of the generality of the grounds or reasons for our belief in (I-1), is in danger of falling into obscurantism.

More important than that, however, is the fact that it is simply impossible for Malmgren to make this or any other explanation work in a coherent fashion with her other commitments. According to Malmgren's possibility approach, the three possibility claims ($2_P$), (PC2), and (PC3), each of which is compatible with the other two, are the real contents of, respectively, (I1), (I2), and (I3). In conformity with the philosophically orthodox view of our thought experiment, Malmgren acknowledges —as we do—that the metaphysical possibility of a justified true belief which is not knowledge may be demonstrated by means of imaginative engagement with the Gettier case, and hence that the first possibility claim, ($2_P$), must be true (whether or not it turns out to be the real content of the Gettier intuition). Malmgren also acknowledges—as we do—that the text of the Gettier case is obviously underspecified, and hence that both of the other possibility claims, (PC2) and (PC3), must likewise be true. So our minds as well as Malmgren's exhibit the following pattern of mental states: belief in the metaphysical possibility of the Gettier case, belief in ($2_P$), belief in (PC2), and belief in (PC3). Now, we do not find anything irrational in that pattern

of mental states; indeed, we regard ourselves as being quite reasonable in holding those beliefs. And presumably Malmgren agrees with us on this score, for she holds exactly the same combination of beliefs. But Malmgren further acknowledges—as we do—that in virtue of our belief in the Gettier case's metaphysical possibility and our (I1)-belief, we incur a rational commitment to refrain from holding an (I2)-belief and a rational commitment to refrain from holding an (I3)-belief. It is at this point that Malmgren cannot help but fall into incoherence. For, on the one hand, she herself holds beliefs in the Gettier case's metaphysical possibility and ($2_P$) while also holding beliefs in (PC2) and (PC3), and finds no irrationality in doing so; but, on the other hand, she says that holding that combination of beliefs is, in fact, irrational, and tries her best to account for this irrationality in terms of the so-called generality of grounds. It is easy enough for *us* to evade this kind of incoherence: we need only reject the possibility approach to modelling philosophical thought experimentation. Things are quite otherwise for Malmgren qua champion of the possibility approach. If Malmgren is to extricate herself from the incoherent position she has got herself into, she must reverse her opinion about the underspecification of the text of the Gettier case or else reverse her opinion about the existence of the very rational commitments she thinks it necessary to explain. Either way, the cost of extrication is ad hocery.

But even setting such ad hocery aside, neither reversal of opinion would be defensible. First, there is nothing abstruse about the underspecification of the text of the Gettier case. It is an easy fact to grasp (especially when examples of unintended enrichments are adduced to illustrate it); so easy, in fact, as to put it beyond reasonable doubt. Malmgren in any event could not deny it without undermining her objections to the Necessity and Counterfactual Models, and hence also the primary and perhaps only motivations for her alternative proposal, the Possibility Model. After all, to deny that the text of the Gettier case is underspecified is to deny that there are deviant worlds; but all of the objections which Malmgren levels against the Necessity and Counterfactual Models are dependent upon the metaphysical possibility of deviant realisations of the text of the Gettier case. (She objects to the Necessity and Counterfactual Models in broadly the same ways as I have done; see especially sections 1.4 and 1.5 of her paper (Malmgren 2011: 275-81).) Malmgren's second option

here is to deny that our belief in the Gettier case's metaphysical possibility and our (I1)-belief together rationally commit us to refrain from holding either an (I2)-belief or an (I3)-belief. If she were to go down this path, Malmgren could perhaps appeal to the fact that the alleged rational commitments seem to turn on certain logico-analytical characteristics of the apparent contents of (I1), (I2), and (I3). The idea would then be that since those apparent contents are known to be illusions, the alleged rational commitments must also be illusions brought about by putting too much faith in the appearances. But this line of argument is borne out of desperation. The real contents of the intuitions and beliefs generated by philosophical thought experimentation must differ in some ways from their apparent contents, but they cannot differ so much that it becomes difficult to make sense of our own mental lives. The firm hand of rationality really does seem to impose itself on the practice of thought experimentation, just as it does on many other activities of the mind. It is as though the beliefs we form about the distribution of properties and relations in hypothetical cases steer us away from forming divergent beliefs. Specifically, when we apprehend in imagination the metaphysical possibility of a hypothetical case and form a belief that in it a certain property or relation is distributed in a certain way, there is a compulsion to refrain from forming a further belief that in it the same property or relation is distributed in a different way. To deny the reality and rational origin of these apparent constraints is to risk losing our grip on what is going on in our own heads when we carry out philosophical thought experiments. The risk here is a big one, and it is not a risk worth running just to save Malmgren's Possibility Model.

### 4.3. Comparison with rivals

The explanatory failure of the possibility approach is compounded by the fact that each of its rivals provides sufficient resources to support an explanation of the relevant rational commitments. According to the necessity approach, our initial imaginative act of apprehending the metaphysical possibility of the Gettier case consists at least partly in a state of judging or believing the content $\Diamond\exists x\exists p\ GCxp$. We then ask ourselves about the epistemic status of Smith's true belief and undergo an intuition; we could intuit either (I1), (I2), or (I3). The proponent of this approach gives the following strict conditionals as their respective real contents:

$(3_N)$ $\Box\forall x\forall p\ (GCxp \rightarrow (JTBxp \wedge \neg Kxp))$

(SC2) $\Box\forall x\forall p\ (GCxp \rightarrow (\neg JBxp \wedge TBxp \wedge \neg Kxp))$

(SC3) $\Box\forall x\forall p\ (GCxp \rightarrow (JTBxp \wedge Kxp))$

These strict conditionals are compatible; they do not entail one another's negations. This is because they are all vacuously true if $\Diamond\exists x\exists p\ GCxp$ is false. Even so, on the safe assumption $\Diamond\exists x\exists p\ GCxp$ is true, they may be said to stand in something like the contrary relation to one another: given any two of them, both may be false but at most one can be true. So the necessity approach has the result that the combination of $\Diamond\exists x\exists p\ GCxp$ with the real content of our (I1)-belief rules out the real contents of (I2) and (I3). It is therefore congenial to an explanation of the relevant rational commitments in terms of the nature of belief and the logico-analytical relationships between contents.

The counterfactual approach also delivers the incompatibilities required for such an explanation to work. It does so in a parallel fashion. Once more, our apprehension in imagination of the metaphysical possibility of the Gettier case is said to consist at least partly in a judgement or belief with $\Diamond\exists x\exists p\ GCxp$ as its content. But the proponent of the counterfactual approach gives not strict conditionals but the following counterfactuals as the respective real contents of (I1), (I2), and (I3):

$(3_{CF})$ $\exists x\exists p\ GCxp \mathrel{\Box\!\!\rightarrow} \forall x\forall p\ (GCxp \rightarrow (JTBxp \wedge \neg Kxp))$

(CF2) $\exists x\exists p\ GCxp \mathrel{\Box\!\!\rightarrow} \forall x\forall p\ (GCxp \rightarrow (\neg JBxp \wedge TBxp \wedge \neg Kxp))$

(CF3) $\exists x\exists p\ GCxp \mathrel{\Box\!\!\rightarrow} \forall x\forall p\ (GCxp \rightarrow (JTBxp \wedge Kxp))$

Given the standard Lewis-Stalnaker semantics for counterfactuals, these counterfactuals are compatible; they do not entail one another's negations. This is because, like

the strict conditionals above, they are vacuously true if ◇∃x∃p GCxp is false. However, on the safe assumption ◇∃x∃p GCxp is true, it follows from any plausible counterfactual semantics that they too may be said to stand in something like the contrary relation to one another: given any two of them, both may be false but at most one can be true. The result is again that the combination of ◇∃x∃p GCxp with the real content of our (I1)-belief rules out the real contents of (I2) and (I3). So the counterfactual approach is also congenial to an explanation of the relevant rational commitments in terms of the nature of belief and the logico-analytical relationships between contents.

Of course, the inadequacy of both of these approaches to modelling philosophical thought experimentation has already been established. But in light of the foregoing explanatory considerations, it behoves us to take another look at the idea that the real content of the Gettier intuition is some kind of conditional which connects the Gettier case to a certain distribution of the relations of knowledge and justified true belief. What is getting in the way of making that idea work? The principal obstacles so far have had do with the metaphysical possibility of deviant realisations of the text of the Gettier case. Williamson's formula GCxp, which represents the text of the Gettier case, is complicit in this obstruction because of its metaphysical weakness. To his credit, Williamson provided us with the springboard we needed to first launch our model-building endeavours, but now we must abandon it if we are to make further progress on solving the content problem. In the next section I turn to an approach to modelling our thought experiment which does just that. The approach put forward by Ichikawa and Jarvis (2009) does away with the formula GCxp and aims to replace it with something much stronger. To anticipate: Ichikawa and Jarvis hold, first, that our imaginative engagement with the Gettier case involves apprehending a highly determinate metaphysical possibility, and second, that the real content of the Gettier intuition is a strict conditional which says, in effect, that this highly determinate metaphysical possibility is one which metaphysically necessitates justified true belief without knowledge. As we shall see, not only does this general kind of approach avoid the problems faced by the Necessity and Counterfactual Models, but there are ways of implementing it which are congenial to explaining the rational commitments associated with our thought experiment in terms of the nature of belief

and the logico-analytical relationships between contents, thereby improving upon the Possibility Model as well.

## 5. The Truth in Fiction Model

The key to Ichikawa and Jarvis's abandonment of Williamson's formula GCxp is the intuitive distinction between *texts* and *stories*. This distinction is an absolutely fundamental one in the philosophical study of the nature of fiction.[39] The set of sentences an author produces when he creates a fictional work is what constitutes a text. These sentences are sometimes called *fictive sentences*. Texts are used by the authors of fiction to tell stories. "There is", as Ichikawa and Jarvis observe, "more to a story than the literal claims of the sentences used to tell it" (2009: 227). A story is a certain enrichment or filling in of a text; it includes a very large number of sentences in addition to the ones explicitly stated in the text used to tell it.[40] These additional sentences are sometimes called *metafictive sentences*. Metafictive sentences are somehow generated from texts, that is to say, sets of fictive sentences, in accordance with the rules or principles governing fictional discourse.[41] The union of a given text with the set of metafictive sentences generated from it is what constitutes a story.[42] The elements of a story S are *fictional truths* with regard to a certain work of fiction;

39 The most influential works in this field include Currie (1990), Lewis (1978; reprinted with a postscript in his 1983), and Walton (1990). Woods (2007) is an excellent and wide-ranging introduction; see also Woodward (2011).

40 Stories are also known as "maximal accounts". See Parsons (1980).

41 The terms "fictive" and "metafictive" are due to Currie (1990). He also introduced the term "transfictive". This applies to sentences which relate the characters and events of a fictional work to things existing outside of that work. Note that, although the set of a story's fictive sentences and the set of its metafictive sentences are always disjoint, neither of those sets need be disjoint with the set of transfictive sentences about the story's fictional characters and events.

42 This, perhaps, is a bit simplistic. One complication arises around the literary device of unreliable narration. If an author makes use of that device, then at least some of the sentences in the text he uses to tell his story will not count as elements of the story. For a recent discussion of unreliable narration, see Nünning (2005). Another potential complication arises around inconsistent texts, that is to say, texts which contain contradictory sentences. Some philosophers, such as Lewis (1983: 277-8), hold that no story is internally inconsistent. If they are right, then no story corresponding to an inconsistent text can contain every sentence of that text. But although these two complications merit further attention, doing justice to them here would require a rather lengthy and distracting digression from the main topic of discussion. Furthermore, unreliable narration and contradictory texts are pretty atypical anyway. This is especially so with regard to the case descriptions used in philosophical thought experimentation. In light of such considerations, I shall set the foregoing complications aside in what follows. My simplistic characterisation of stories will be enough for our purposes.

they are said to be *true in S*. The concept of truth in fiction is of course not the same as the concept of truth simpliciter. A sentence's being true does not imply that it is true in any particular story; conversely, a sentence's being true in some story does not imply that it is true. But that is not to say that fictional truths always make claims which fail to be true of the world. Although stories generally do contain many fictional truths which are not true, they also generally contain a great number of fictional truths which are true simpliciter as well.

We are more or less competent with fictional discourse. This competency of ours consists mainly in a tacit grasp of the rules or principles governing the enrichment of texts. Although those principles have proved very difficult to formulate explicitly, we nonetheless have a general capacity to apply them to a given text and thereby come to recognise whether something is true in the story, false in the story, or indeterminate in the story.[43] Let us consider, for example, Thackeray's novel *Vanity Fair* (2001), which portrays the vices and follies of middle and upper class British society during the Napoleonic era and its aftermath. The text of *Vanity Fair* comprises many thousands of sentences, among which are the following (see Thackeray (2001: 16 and 374-5)):

- 'Revenge may be wicked, but it's natural,' answered Miss Rebecca.

- Towards evening, the attack of the French, repeated and resisted so bravely, slackened in its fury.

Since these are fictive sentences of *Vanity Fair*, they count as fictional truths with regard to that work; they are, as we say, true in the *Vanity Fair* story.[44] But there is much more to the story of *Vanity Fair* than the mere text Thackeray used to tell it. We are able to identify many of its metafictive sentences. They presumably include:

43 See Woodward (2011) for a recent discussion of the difficulty of giving explicit formulations of the principles of fictional discourse. Woodward is doubtful whether we will ever be able to formulate those principles explicitly. Similar doubts are expressed by Currie (1990) and Walton (1990).

44 As for truth simpliciter, the two fictive sentences arguably diverge. The first of them may be held to be not true on the grounds that Becky Sharp never really existed and hence nothing was ever said by her about the naturalness of revenge. It may very well be, however, that the second sentence truthfully describes how things stood at the Battle of Waterloo in the tense moments leading up to the famous attack of Napoleon's Imperial Guard.

- Becky Sharp had a central nervous system.

- The French were defeated at the Battle of Waterloo by Britain and its allies.

Despite the fact that these sentences are nowhere explicitly stated in the text of *Vanity Fair*, they too count as fictional truths with regard to that work. The principles of fictional discourse rule them in; they are just as much elements of the *Vanity Fair* story as the above fictive sentences are.[45] In addition to this story's metafictive sentences, we are also able to identify many sentences which fall into neither the fictive nor the metafictive category. They presumably include:

- Several of Becky Sharp's ancestors were wizards from another galaxy.

- Becky Sharp had a mole on her lower back.

Although each of these sentences is consistent with the text of *Vanity Fair*, neither counts as a fictional truth with regard to that work. The first of them is plainly a ridiculous way to enrich or fill in the text of *Vanity Fair*. The principles of fictional discourse rule it out; it is false in the *Vanity Fair* story. The second sentence, in contrast, is not ridiculous, but seems to be neither ruled in nor ruled out by the principles of fictional discourse. Its fictional truth value with regard to *Vanity Fair* is indeterminate.[46]

Ichikawa and Jarvis assimilate the case descriptions deployed in philosophical thought experimentation to standard works of fiction. Accordingly, just as we distinguished between the text of *Vanity Fair* and the *Vanity Fair* story, Ichikawa and Jarvis distinguish between the text of the Gettier case and what they call *the Gettier*

45 And as for truth simpliciter, these two metafictive sentences are also like the above fictive sentences in that they at least arguably diverge. The first of them may be held to be not true on the grounds that Becky Sharp never really existed and hence no central nervous system has ever belonged to her. The second sentence, however, is a truth of European military history.

46 This time it is at least arguable that, as for truth simpliciter, the status of the two sentences is the same. They may both be held to be not true. After all, Becky Sharp never really existed. So there has never been anything biologically related to her and no mole has ever been located on her lower back.

*story*.[47] There is much more to the Gettier story than the mere text used to tell it. This is so even though the Gettier story's metafictive sentences are generated from a very small set of fictive sentences. As with other stories, our tacit grasp of the principles which govern the enrichment or filling in of texts puts us in a position to recognise many of its fictional truths, fictional falsehoods, and fictional indeterminacies. First of all, we are able to identify many of the Gettier story's metafictive sentences. They presumably include:

- Smith is conscious.

- There is life on Earth.

These sentences are nowhere explicitly stated in the text of the Gettier case, but they are ruled in by the principles of fictional discourse. They count as true in the Gettier story just like its fictive sentences do. We also have the ability to identify many non-fictive and non-metafictive sentences. They presumably include:

- Smith is one nanometre tall.

- Smith has brown hair.

The first of these sentences may or may not be consistent with the text of the Gettier case, but even if it is consistent, it seems to be ruled out by the principles of fictional discourse. It counts as false in the Gettier story. The other sentence, in contrast, is plainly consistent with the text of Gettier case, but it seems to be neither ruled in nor ruled out by the principles of fictional discourse. Its fictional truth value with regard to the Gettier story is indeterminate.

According to Ichikawa and Jarvis (2009: 227), the distinction between texts and stories is "just what we need" to deal with the obstacles presented by the metaphysical possibility of deviant realisations of the text of the Gettier case. The general idea

47 Lewis is surely the inspiration for Ichikawa and Jarvis's introduction of the distinction between texts and stories into our model-building project. In the postscript to his influential paper on truth in fiction, he observes that "[f]iction might serve as a means of discovery of modal truth. […] note that the philosophical example is just a concise bit of fiction" (Lewis 1983: 278).

is that when we carry out our thought experiment, our competency with fictional discourse somehow guides the imaginative activity of the mind. We imaginatively engage not only with the text of the Gettier case but with the much richer Gettier story. An integral component of this imaginative mental activity is the mind's recognition that the principles of fictional discourse *rule out* enrichments of the text of the Gettier case which are conducive to deviancy. More specifically, the mind recognises that such enrichments are unintended; even if they are consistent with the text of the Gettier case, they are *not* how the Gettier case is meant to be filled in. Note that this approach has a certain naturalness about it. The text of the Gettier case does tell a story about a man's encounter with robotic kangaroos in the Australian countryside. And, in general, our tacit grasp of the principles of fictional discourse does provide guidance to the mind when we imaginatively engage with fictional works. So we may indeed have found just what we need.

Ichikawa and Jarvis, however, do not simply equate the practice of thought experimentation with the practice of reading standard fictional works such as novels, narrative poems, and the like. They regard the proposal that the Gettier intuition is just another judgement about what is true in a fiction as rather unhelpful.[48] And, it seems to me, they are right to do so. First of all, there is no standard analysis of the concept of truth in fiction; so it is not even clear what the real content of the Gettier intuition would be if the Gettier intuition were nothing but an ordinary fictional judgement.[49] Second, and relatedly, it is questionable whether the contents of such judgements have anything much to do with the metaphysical modalities; so it could very well turn out that they have no logical bearing at all on the the truth or falsity of philosophical theories such as the traditional theory of knowledge.[50] In short, the

48 As confirmed by the fact that they do not discuss the proposal, not even just to dismiss it. The two reasons I go on to mention for why Ichikawa and Jarvis seem to regard it as unhelpful are reasons I have gleaned from the general tenor of their discussion.

49 Woods (2007) discusses most of the competing analyses of truth in fiction. If any of them have any claim to being the standard one, Walton's (1990) pretence analysis does and perhaps Lewis's counterfactual analyses (in particular, his Analysis 1 and Analysis 2) do as well. But these analyses have also come in for severe criticism. Woods (2007) delivers a powerful broadside against Walton's pretence analysis. As for Lewis's counterfactual analyses, Lewis himself raises several objections against them (Lewis 1978: 42-6). See also Currie (1990), Byrne (1993), Phillips (1999), Le Poidevin (1995), Proudfoot (2006), Walton (1990), Wolterstorff (1980), Woods (2007), and Woodward (2011). For a vigorous defence of Lewis's analyses, see Hanley (2004).

50 There are several analyses of truth in fiction on which this would be so. Walton's (1990) influential pretence analysis, for example, analyses fictional truth in terms of what is authorised by the games of make-believe we (allegedly) engage in when reading fictional works. If something like Walton's pretence analysis is right, then it is hard to see how the contents of ordinary fictional

proposal that the Gettier intuition is simply an ordinary fictional judgement is quite uninformative and may in fact be fundamentally misguided. Ichikawa and Jarvis therefore explore several alternative ways of implementing the truth in fiction approach to modelling philosophical thought experimentation. We shall now consider each of them in turn.

### 5.1. Mark 1

Taking the logical structure of the Necessity Model as their point of departure, Ichikawa and Jarvis start out by developing a model which directly invokes the concept of truth in fiction without making the Gettier intuition into just another fictional judgement. Williamson's formula GCxp is replaced with the formula $GC_{TF}xp$, which is to be understood as saying that x stands to p in the relation in which it is true in the fiction that Smith stands to the proposition that there are kangaroos in the paddock. This replacement yields the following model of our thought experiment, which I shall call *the Truth in Fiction Model Mark 1*:

(1) $T \rightarrow \Box\forall x\forall p\ (Kxp \leftrightarrow JTBxp)$

($2_{TF\text{-}M1}$) $\Diamond\exists x\exists p\ GC_{TF}xp$

($3_{TF\text{-}M1}$) $\Box\forall x\forall p\ (GC_{TF}xp \rightarrow (JTBxp \wedge \neg Kxp))$

(4) $\neg\Box\forall x\forall p\ (Kxp \leftrightarrow JTBxp)$ From (2) and ($3_{TF\text{-}M1}$)

---

(5) $\neg T$ From (1) and (4)

On this model, truth in fiction is invoked at the level of content. Our imaginative engagement with the Gettier case involves apprehending the truth of ($2_{TF\text{-}M1}$), the claim that it is metaphysically possible for a subject to be related to a proposition in the same way that, in the fiction, Smith is related to the proposition that there are kangaroos in the paddock. The real content of the Gettier intuition is given by ($3_{TF\text{-}M1}$), the claim that, as a matter of metaphysical necessity, anyone so related to a proposition has a justified true belief in it which fails to constitute knowledge.

judgements could bear on the truth or falsity of philosophical theories. The same goes for Woods's (2007) analysis of fictional truth in terms of labels.

The Truth in Fiction Model Mark 1 would seem to be a big step in the right direction. First, we now begin to see how the obstacles presented by deviant worlds might be overcome. The principles of fictional discourse are said to generate an enrichment of the text of the Gettier case which precludes Smith's true belief that there are kangaroos in the paddock from being either unjustified or knowledge. Since the formula $GC_{TF}xp$ pertains to that enrichment, that is to say, to the Gettier story, it should carry with it the clout to metaphysically necessitate justified true belief without knowledge. So if the Gettier intuition has $(3_{TF\text{-}M1})$ as its real content, considerations of deviancy should have nothing to do with its truth or falsity, just as we have been insisting all along. Another attraction of this model is that there are enough resources in the vicinity to support an explanation of the rational commitments associated with our thought experiment. The explanation parallels the ones discussed in Section 4.3 above. On the present way of implementing the truth in fiction approach, the respective real contents of (I1), (I2), and (I3) are given by the following strict conditionals:

$(3_{TF\text{-}M1})$ $\Box\forall x\forall p\ (GC_{TF}xp \rightarrow (JTBxp \wedge \neg Kxp))$

(TFM1-2) $\Box\forall x\forall p\ (GC_{TF}xp \rightarrow (\neg JBxp \wedge TBxp \wedge \neg Kxp))$

(TFM1-3) $\Box\forall x\forall p\ (GC_{TF}xp \rightarrow (JTBxp \wedge Kxp))$

These strict conditionals are compatible; they do not entail one another's negations. This is because they are all vacuously true if $\Diamond\exists x\exists p\ GC_{TF}xp$ is false. But if we assume the truth of $\Diamond\exists x\exists p\ GC_{TF}xp$, then they may be said to stand in something like the contrary relation to one another: given any two of them, both may be false but at most one can be true. As a result, the combination of $\Diamond\exists x\exists p\ GC_{TF}xp$ with the real content of our (I1)-belief rules out the real contents of (I2) and (I3). So the relevant rational commitments are straightforwardly explainable in terms of the nature of belief and the logico-analytical relationships between contents.

Despite its attractions, Ichikawa and Jarvis find the Truth in Fiction Model Mark 1 inadequate. They level three objections against it. First, they doubt whether

($3_{TF\text{-}M1}$) is knowable a priori given that it directly invokes truth in fiction. The worry, as Ichikawa and Jarvis put it, is that "everything would depend on the correct theory of truth in fiction; the Gettier intuition would seem to be a priori only if we can access fictional truths, given texts, a priori; and this is far from clear" (2009: 227). To illustrate, they adduce one of David Lewis's (1978) proposed analyses of the concept of truth in fiction, on which something is true in a story S if and only if it would be true if S were told as known fact. This analysis of Lewis's has it that truth in fiction claims are really counterfactuals, many of which are contingent and knowable only by a posteriori means, if they are knowable at all.[51] According to Ichikawa and Jarvis, the Lewis view would turn the present model into something "similar to Williamson's, with respect to the epistemic status of the Gettier intuition" (2009: 228). They go on to add that although "there are good reasons to think that the Lewis view is not right […] the correct theory of truth in fiction will very likely share this feature of Lewis's: it will have it that fictional truths cannot be known a priori" (2009: 228). If this aposteriority contaminates ($3_{TF\text{-}M1}$), then ($3_{TF\text{-}M1}$) is the real content of the Gettier intuition only if one of the main tenets of the "traditional understanding of philosophical methodology" is mistaken. Ichikawa and Jarvis uphold the tenets of traditionalism, and since they anticipate empirical contamination of ($3_{TF\text{-}M1}$), they do not think it can be what we are searching for.

However, this first line of objection is defective. One problem is that Ichikawa and Jarvis are strangely confused about the potential for truth in fiction to lead to empirical contamination. Let us grant that fictional truths are only knowable a posteriori. Then it is an empirical question as to how, in the fiction, Smith is related to the proposition that there are kangaroos in the paddock. But whatever the answer turns out to be, it plainly does not follow that it is also an empirical question as to whether a subject's being so related to a proposition metaphysically necessitates his having a non-knowledge justified true belief. The point here holds generally. For example, it is an empirical question as to how I am related to my mug; but whatever

51 The analysis is only one of several proposed by Lewis in his (1978). The statement I have given of it here is a simplification of Lewis's original. Note that Lewis himself does not endorse the analysis; instead, he seems inclined toward one of his others, the one on which something is true in a story S if and only if, in the worlds nearest to the collective belief worlds of the community of origin of S, it is true and S is told as known fact. Ichikawa and Jarvis do not discuss this analysis, but it would certainly lead them to worry about ($3_{TF\text{-}M1}$) just as much as the analysis which they do discuss leads them to worry about it, for the question of which worlds constitute a given community's collective belief worlds is presumably an empirical one.

the answer turns out to be, it plainly does not follow that it is also an empirical question as to whether a subject's being so related to a mug metaphysically necessitates the existence of at least one drinking utensil. Indeed, if anything is knowable a priori, then it is surely knowable a priori that, as a matter of metaphysical necessity, if a subject is related to a mug in the way I am related to my mug, then at least one drinking utensil exists. So Ichikawa and Jarvis need to clarify why it is that they are doubtful about the a priori epistemic status of ($3_{\text{TF-M1}}$); as it stands, their doubt seems wholly devoid of motivation. Another problem is that, even if ($3_{\text{TF-M1}}$) is empirically contaminated, only those who harbour traditionalist sympathies need be unhappy about the fact. Traditionalism is not mandatory. Furthermore, I have already pointed out (in Section 3.3), first, that Ichikawa and Jarvis say nothing to vindicate traditionalism, and second, that their devotion to it introduces an unhelpful and unnecessary element of partisanship into our model-building project.

Ichikawa and Jarvis's second objection to the Truth in Fiction Model Mark 1 is that it fails to establish enough distance between the Gettier intuition and ordinary fictional judgements. This alleged failure has to do with the model's invocation of truth in fiction at the level of content. Ichikawa and Jarvis acknowledge that since many philosophers are of the view that "our ordinary sentences about fictional characters include an unnoticed elliptical 'it is true in the fiction' operator […] it is perhaps not terribly worrying that our Gettier intuitions don't *feel* like judgements about fictions" (2009: 228). What is said to be terribly worrying is that Gettier-style cases known or believed to be actual (i.e. non-fictional) "work just as well as fictional ones to set up the Gettier conclusion" (2009: 228). More specifically, Ichikawa and Jarvis tell us that if our particular Gettier case had been presented to us as fact rather than fiction, then we would "have gone through the Gettier reasoning in just the same way" (2009: 228). Insofar as I understand them, Ichikawa and Jarvis are here claiming that, in terms of real content, intuitions about Gettier-style cases taken to be fictional are very much akin to intuitions about Gettier-style cases taken to be actual. Since the latter intuitions pertain to what we take to be actual, they are obviously not fictional judgements of any kind. Truth in fiction is not an essential conceptual constituent of such intuitions; we are able to reason from their contents to the negation of the traditional theory of knowledge without invoking truth in fiction. For Ichi-

kawa and Jarvis, therefore, "it seems wrong to suppose that invocation of fiction is an essential part of Gettier reasoning […] it does not seem as though the concept of truth in fiction can play a role at this level" (2009: 228).

This is another defective line of objection. It is hard to know what to make of Ichikawa and Jarvis's assertion that we would "have gone through the Gettier reasoning in just the same way" if the Gettier case had been presented to us as fact rather than fiction. No elaboration of the import of this assertion is provided. On the one hand, Ichikawa and Jarvis may intend for us to take it at face value, as I do above. Then part of its import would seem to be that the real content of our intuition about the epistemic status of Smith's true belief is the same whether the Gettier case is taken to be fact or taken to be fiction. But this interpretation makes Ichikawa and Jarvis's assertion highly problematic. If the Gettier case were a factual description and we had taken it to be so when we originally asked ourselves about the epistemic status of Smith's true belief, then the content problem would not even have arisen. This is because our intuition would have been just another intuition or judgement about the actual epistemic situation of an actual individual, and we would not have baulked at the existential commitments of its apparent content. Instead, we would have found it perfectly natural to identify our intuition's real content with its apparent content. The upshot is that it actually matters a great deal whether the Gettier case is taken to be fact or taken to be fiction. On the other hand, Ichikawa and Jarvis may intend for us to interpret their assertion in some other way. (After all, it is perhaps uncharitable to take it at face value.) But if Ichikawa and Jarvis do have something else in mind, then the onus is on them to spell it out, because it is not obvious what it could be.[52]

Ichikawa and Jarvis's third objection to the Truth in Fiction Model Mark 1 is somewhat more pressing than the preceding ones. The strict conditional which the model gives as the real content of Gettier intuition, ($3_{TF\text{-}M1}$), makes use of the formula $GC_{TF}xp$. We understand the formula $GC_{TF}xp$ to say that x stands to p in the relation in which it is true in fiction that Smith stands to the proposition that there are kangaroos in the paddock. Ichikawa and Jarvis (2009: 228) ask: "Which relation, however, is *the* relation?" This simple question has the effect of casting doubt on our

52 Malmgren also expresses puzzlement at Ichikawa and Jarvis's second objection; she calls it "confused" (2011: 304, fn.69).

very understanding of the formula $GC_{TF}xp$ and hence also on the potential of that formula to do serious theoretical work. For, as Ichikawa and Jarvis go on to point out, Smith "stands in many relations to the proposition in question. […] Specifying one is difficult, but without specifying one we have not fully offered a formulation [of the real content of the Gettier intuition]" (2009: 228). To put the objection another way, the formula $GC_{TF}xp$ is shot through with ambiguity. Unless this ambiguity is resolved, it cannot be said that ($3_{TF\text{-}M1}$) is the solution to the content problem. We must disambiguate the formula $GC_{TF}xp$ before making any such pronouncement. This will necessarily involve specifying which relation is *the* relation and then somehow restating the formula $GC_{TF}xp$ in terms of it. Ichikawa and Jarvis rightly draw attention to the fact that the task here is a difficult one, but they do not say much about what the difficulty consists in. Presumably, what they have in mind is that any satisfactory disambiguation of the formula $GC_{TF}xp$ should guarantee the truth of ($3_{TF\text{-}M1}$) without trivialising it. This requirement is imposed by the philosophically orthodox view of our thought experiment, on which the Gettier intuition is both non-trivial and true. There is no straightforward way to meet the requirement. Since, therefore, they find the prospects of coming up with a satisfactory disambiguation of the formula $GC_{TF}xp$ to be rather dim, Ichikawa and Jarvis set the present model aside as inadequate and explore other ways of implementing the truth in fiction approach.

### 5.2. Marks 2 and 3

Ichikawa and Jarvis "suggest that a better application of truth in fiction in a Gettier formulation will make use of the notion less directly" (2009: 229). They put forward the proposal that our tacit grasp of the principles governing the enrichment of texts is "useful for *picking out* and *thinking about* propositions that are key to the Gettier argument" (2009: 229). The propositions Ichikawa and Jarvis have in mind here are propositions corresponding to the Gettier story's fictive and metafictive sentences.[53]

53 Ichikawa and Jarvis's use of the terms "story", "sentence", and "proposition" is undisciplined. In the philosophical study of the nature of fiction, stories are normally held to be sets of sentences, and fictional truth is normally discussed with regard to sentences. But sometimes Ichikawa and Jarvis talk of the Gettier story as though it is a set of sentences, and at other times as though it is a set of propositions. And they slide freely between talk of sentences being true in the Gettier story and propositions being true in the Gettier story. This makes it difficult to elucidate the ways in which they have implemented the truth in fiction approach. The two models I present in this sub-section seem to me to do the best job of clarifying what Ichikawa and Jarvis are trying to get at. See the next footnote for further clarificatory remarks.

In carrying out our thought experiment, we entertain the set of those propositions, "and in particular, the proposition that every member of that set is true, and then subsequently reason with that proposition to the conclusion of the thought experiment" (2009: 229). The cardinality of the set is very large, probably infinitely so. Due to normal cognitive limitations, we "cannot, of course, entertain each proposition of this infinite set individually, but [we] can think about the set containing all of them, and refer to it" (2009: 229). Ichikawa and Jarvis go on to postulate that the mind baptises the set for referential convenience. The mind gives it a name, say, "STORY", and then entertains the proposition g, that every element of STORY is true. Putting all of this together yields the following model of our thought experiment, which I shall call *the Truth in Fiction Model Mark 2*:

(1) $T \rightarrow \Box\forall x\forall p\ (Kxp \leftrightarrow JTBxp)$

($2_{TF\text{-}M2}$) $\Diamond g$

($3_{TF\text{-}M2}$) $\Box(g \rightarrow \exists x\exists p\ (JTBxp \wedge \neg Kxp))$

(4) $\neg\Box\forall x\forall p\ (Kxp \leftrightarrow JTBxp)$ From (2) and ($3_{TF\text{-}M2}$)

---

(5) $\neg T$ From (1) and (4)

On this model, there is no invocation of truth in fiction at the level of content. Our imaginative engagement with the Gettier case involves apprehending the truth of ($2_{TF\text{-}M2}$), the claim that it is metaphysically possible for all of the elements of STORY to be true.[54] The real content of the Gettier intuition is given by ($3_{TF\text{-}M2}$), the claim

54 The elements of STORY, as I have said, are propositions "corresponding" to the Gettier story's fictive and metafictive sentences. But there is a question as to what this means with regard to the sentences in the Gettier story which involve either a fictional proper name—e.g. the sentence "One day Smith is walking through the Australian countryside"—or a fictional pronoun—e.g. the sentence "He is an avid and experienced spotter of native wildlife who has the ability to identify animals such as kangaroos, wallabies, wombats, koalas, and so on, with a very high degree of reliability." Either such sentences express propositions or they do not. Let us first assume they do not (maybe they do not express propositions because, as some philosophers say, fictional proper names and fictional anaphoric pronouns are empty). Then Ichikawa and Jarvis will have to introduce certain descriptive propositions to "correspond" to them. This is straightforward enough. But now let us assume that the sentences in the Gettier story which involve either a fictional proper name or a fictional pronoun do express propositions. Should Ichikawa and Jarvis then identify STORY with the set of propositions expressed by the Gettier story's fictive and metafictive sentences? Kripke (1980: 157-8) has argued that fictional characters like Smith are essentially fictional. As Ichikawa and Jarvis (2009: 229, fn.13) point out, it would seem that, if Kripke's view is right, then given our assumption, all sentences in the Gettier story which involve either a

that the elements of STORY metaphysically necessitate justified true belief which fails to constitute knowledge.

The Truth in Fiction Model Mark 2 points to a very specific role for truth in fiction in our test of the traditional theory of knowledge. Truth in fiction is deployed by the mind only as a kind of medium for imaginatively engaging with the Gettier case. We draw on our tacit grasp of the principles of truth in fiction in order to pick out a certain set of propositions in imaginative thought. In doing so, we put ourselves in a position to entertain a certain proposition about that set, viz. the proposition that all of the set's elements are true. No further use of truth in fiction is made by the mind. Ichikawa and Jarvis, who endorse this broad characterisation of truth in fiction's role in the performance of our thought experiment, elaborate on it as follows:

> Here, our invocation of fictional truth explains how we come to entertain the proposition [g], but the concept [of truth in fiction] does not enter into the Gettier reasoning itself. Competence with truth in fiction is an important step in engaging with the thought experiment, but its role is exhausted before the actual invocation in reasoning of [($3_{TF\text{-}M2}$)], the Gettier intuition. Put another way, one's ability to grasp a story told through a text serves only as a *means to grasp* a certain proposition, which will figure into the intuition; the content of the intuition itself makes no use of the notion of truth in fiction (Ichikawa and Jarvis 2009: 230).

Like its predecessor, this indirect way of implementing the truth in fiction approach would seem to have what it takes to overcome the obstacles presented by deviant worlds. According to it, the set STORY corresponds to an enrichment of the text of the Gettier case which precludes Smith's true belief that there are kangaroos in the paddock from being either unjustified or knowledge. The Gettier intuition is a manifestation of the mind's recognition of this preclusion. The mind comes to recognise

---

fictional proper name or a fictional pronoun will express metaphysically impossible propositions. Consequently, simply identifying STORY with the set of propositions expressed by the Gettier story's fictive and metafictive sentences could make ($2_{TF\text{-}M2}$) come out false. This is potentially problematic because our thought experiment surely involves the apprehension in imagination of a genuine metaphysical possibility. To deal with it without giving offence to Kripke, Ichikawa and Jarvis will once again have to introduce certain descriptive propositions to "correspond" to the sentences in the Gettier story which involve either a fictional proper name or a fictional pronoun.

the preclusion when, guided by our tacit grasp of the principles of fictional discourse, it picks out the set STORY and apprehends the metaphysical possibility of all of STORY's elements being true. Considerations of deviancy have nothing to do with the Gettier intuition's truth or falsity because this metaphysical possibility is such a highly determinate one.

The Truth in Fiction Model Mark 2 constitutes an advance on its predecessor insofar as the set STORY is free from much of the obscurity which mars the formula $GC_{TF}xp$. As Ichikawa and Jarvis point out, we have no clear and distinct conception of what the formula $GC_{TF}xp$ is supposed to amount to. In the absence of any satisfactory disambiguation of it, we cannot appeal to that formula for illumination of the real content of the Gettier intuition. From the standpoint of model-building, it would seem that the formula $GC_{TF}xp$ is not much more than an amorphous placeholder. In contrast, stories and the sets of propositions corresponding to them are familiar to most of us from ordinary life. We are accustomed from childhood to picking out such sets and thinking about them. We are able to bring them before the mind, deploy them in our reasoning, and thereby come to adopt various intellectual and emotional attitudes toward fictional works, for example, attitudes of admiration, hatred, satisfaction, disappointment, contempt, puzzlement, condemnation, and so on. This is something we have the ability to do despite the fact that, for almost any given set of propositions corresponding to a story, we lack the cognitive power to identify all of the set's elements and entertain each of them individually. More generally, sets of very large cardinality are quite within reach of the mind. For example, in addition to the set STORY, we are able to pick out and think about the set of stars, the set of sub-atomic particles, and the set of numbers. This is so despite the fact that it is impossible for creatures like us to identify all of the elements of those sets and entertain each of the elements individually. The familiarity and generality of such mental activity greatly enhances the theoretical utility of the set STORY relative to the formula $GC_{TF}xp$. The set STORY is not an amorphous placeholder; it has the potential to do serious theoretical work.

The very large cardinality of the set STORY may, however, occasion the worry that the Truth in Fiction Model Mark 2 threatens to epistemically compromise our test procedure, in particular, the stage of our test procedure at which we commence

our imaginative engagement with the Gettier case. Although we are able to pick out and think about the set STORY, we lack the cognitive power to bring each of its elements before the mind and combine them into a representation in imaginative thought. The construction of such an immensely complex representation is beyond us. Only a creature of much greater cognitive power than ourselves could ever hope to determine by that means whether the elements of the set STORY are metaphysically compossible. But it would be hasty to conclude from this that ($2_{TF\text{-}M2}$) is simply unknowable for us. Many genres of fiction, including the one to which the text of the Gettier case belongs, are governed by a principle to the effect that metaphysical coherence is to be preserved whenever it can be. According to this principle, if there are no metaphysical incoherencies in a text, then the story it is used to tell should not contain any either. And even if there are some metaphysical incoherencies in a text, the story it is used to tell should contain no additional ones which are of no relevance to them. We tacitly grasp this principle. Given that we do, it would seem that if we have good reason for thinking that a certain text contains no metaphysical incoherencies (say, because a careful and thorough examination of it has failed to bring any to our attention), then, plausibly, we also have good reason for thinking that the propositions corresponding to the story it is used to tell must be metaphysical compossible. This makes room for optimism about the knowability of ($2_{TF\text{-}M2}$). In carrying out our thought experiment, the mind runs through the text of the Gettier case, searching for metaphysical incoherencies in it. Of course, none are found. We thus have good reason for thinking that the text of the Gettier case contains no metaphysical incoherencies. That fact, together with our tacit grasp of the foregoing principle of fictional discourse, plausibly provides us with good reason for thinking that the elements of the set STORY are metaphysically compossible.

Ichikawa and Jarvis hold that the Truth in Fiction Model Mark 2 is in keeping with the traditionalist tenet that the intuitions generated by philosophical thought experimentation are a priori. Although they admit that "the ability to grasp stories through texts involves (perhaps tacit) a posteriori knowledge", they also suggest that this a posteriori knowledge "does not prevent the thought-experiment reasoning from being a priori" (2009: 230). Its role is said to be analogous to the role of a posteriori knowledge of the English language in reasoning. To understand a piece of

reasoning in English (without the aid of a translator or translating device), an agent needs to have a posteriori knowledge of the relevant English words and the relevant fragments of English grammar. The fact that such a posteriori knowledge is required need not make the reasoning a posteriori. Ichikawa and Jarvis give the following sequence of sentences as an illustration:

- If Julius Caesar is the successor of the number one, then Julius Caesar is identical to the number two.

- Two is a prime number.

- If Julius Caesar is the successor of the number one, then Julius Caesar is a prime number.

We are able to understand reasoning which proceeds from the first two of these sentences to the third one. This involves drawing on whatever a posteriori knowledge is required to interpret the sentences. Our a posteriori knowledge of the English language *enables* us to bring the propositions expressed by those sentences before the mind; it is deployed "as a means to get to that content" (Ichikawa and Jarvis 2009: 230). But it plays no other role in addition to this merely enabling one. Furthermore, the reasoning is plausibly immune from empirical contamination from elsewhere. We run through it in an a priori manner; "[t]he reasoning is a priori because of the a priori status of [the premise] and the a priori entitlement to move from [the premise] to the conclusion" (Ichikawa and Jarvis 2009: 230). Even if, as radical empiricists would claim, the reasoning is in fact a posteriori, it is doubtful whether our a posteriori knowledge of the English language has much to do with it.

Similarly, say Ichikawa and Jarvis, the a posteriori knowledge involved in our grasp of the principles of fictional discourse need not empirically contaminate ($3_{TF\text{-}M2}$). They point out that there would be empirical contamination "[o]nly if the a posteriori knowledge were deployed as *warranting*" (2009: 231). But they insist that it is not so deployed, adding: "It is merely deployed *as a means to come to grasp the propositions involved in reasoning involving the thought experiment intuition*. The a

posteriori knowledge serves as a […] *casual enabler* for the reasoning process; it does not play a warranting role within the reasoning itself" (2009: 231). Up to this point, Ichikawa and Jarvis may very well be in the right, but they straight away fall back into their strange confusion about the potential for truth in fiction to lead to empirical contamination. They point out that although we may find it tempting to think of g descriptively as the proposition that all of the fictional truths of the Gettier story are true, g "does not have that descriptive (Fregean) sense; it is merely true in all the same possible worlds" (2009: 231). Allegedly, this is "crucial to note" because if things were otherwise the concept of truth in fiction "would be a constituent in the thought-experiment intuition, and one's a posteriori knowledge of what is true in the Gettier fiction would stand in the warrant relation to the conclusion that knowledge is not necessarily justified true belief" (2009: 231). But Ichikawa and Jarvis fail to tell us why invocation of truth in fiction at the level of content would have such an epistemic consequence. Furthermore, it is not obvious what they could have in mind. Let p be any empirical proposition whatever, even one which has truth in fiction as a constituent. It plainly does not follow that, if the strict conditional $\Box(p \rightarrow q)$ is knowable at all, then it is only knowable a posteriori. There is an abundance of plausible counterexamples.[55] In short, Ichikawa and Jarvis should give up their fixation on the a posteriority of truth in fiction. It is unmotivated and it contributes nothing to the furtherance of their traditionalist agenda.

Although Ichikawa and Jarvis regard the Truth in Fiction Model Mark 2 as basically correct, they worry that some may find its appeal to baptismal reference objectionable on phenomenological grounds (2009: 231). For, in carrying out our thought experiment, it does not obviously seem as though we baptise the set of propositions corresponding to the Gettier story with a proper name. To deal with this worry, Ichikawa and Jarvis advert to demonstrative reference, suggesting that "in many cases the Gettier intuition may come to many people with a demonstrative in

55 And they are very easy to think up. Thus: It is an empirical question as to whether I am taller than you, but it is plausibly knowable a priori that, necessarily, if I am taller than you then you are shorter than me. Here is a counterexample involving truth in fiction: It is an empirical question as to whether it is true in the *Vanity Fair* story that Becky Sharp is a sociopath, but it is plausibly knowable a priori that, necessarily, if it is true in the *Vanity Fair* story that Becky Sharp is a sociopath then it is true in the *Vanity Fair* story that she has a personality disorder. Of course, there are also many trivial counterexamples. Thus: It is an empirical question as to whether the Earth is round, but it is plausibly knowable a priori that, necessarily, if the Earth is round then the Earth is round.

the content" rather than a proper name (2009: 231-2). They put forward the proposal that when we imaginatively engage with the Gettier case, we pick out the set of propositions corresponding to the Gettier story and proceed to think and reason about it not via the proposition g, but via another proposition d, that things are like *that*. In the proposition d "that" functions as a demonstrative which directly refers to how things are according to the Gettier story. Our competence with fictional discourse, Ichikawa and Jarvis explain, "goes toward fixing and apprehending the reference of the demonstrative" (2009: 232). Replacing g with d, we get the following model of our thought experiment, which I shall call *the Truth in Fiction Model Mark 3*:

(1) $T \rightarrow \Box\forall x\forall p\ (Kxp \leftrightarrow JTBxp)$

($2_{TF\text{-}M3}$) $\Diamond d$

($3_{TF\text{-}M3}$) $\Box(d \rightarrow \exists x\exists p\ (JTBxp \wedge \neg Kxp))$

(4) $\neg\Box\forall x\forall p\ (Kxp \leftrightarrow JTBxp)$ From (2) and ($3_{TF\text{-}M3}$)

---

(5) $\neg T$ From (1) and (4)

On this model, there is once again no invocation of truth in fiction at the level of content. Our imaginative engagement with the Gettier case involves apprehending the truth of ($2_{TF\text{-}M3}$), the claim that it is metaphysically possible for things to be like *that*. The real content of the Gettier intuition is given by ($3_{TF\text{-}M3}$), the claim that things' being like *that* metaphysically necessitates justified true belief which fails to constitute knowledge.

It might be wondered whether the Truth in Fiction Model Mark 3 constitutes much of an advance on its predecessor. For one thing, the failure of baptismal reference to accord with the phenomenology of our thought experiment is not especially worrisome. We know that at least some of what actually happens during the performance of our thought experiment cannot be as it seems. In particular, we know that when we undergo the Gettier intuition we enter into a mental state the real content of which must differ from its apparent content. Given this fact, it should come as no surprise if our model-building endeavours lead us to postulate something phenomenologically incongruous. Of course, that is not to say that considerations of

phenomenology carry no theoretical weight in model-building. If we postulate something which makes it difficult to make sense of our own mental lives, then we should probably get rid of it. But the postulate that the mind deploys baptismal reference to pick out and think about the set of propositions corresponding to the Gettier story can hardly be said to make us lose our grip on what is going on in our own heads. Furthermore, even if baptismal reference does have that consequence, it cannot simply be taken for granted that demonstrative reference accords any better with the phenomenology of our thought experiment. After all, what Ichikawa and Jarvis say about baptismal reference may equally be said about demonstrative reference as well: in carrying out our thought experiment, it does not obviously seem as though we deploy "that" to demonstratively refer to the set of propositions corresponding to the Gettier story. If Ichikawa and Jarvis wish to persuade us that, of the two kinds of reference, the demonstrative one does the best phenomenological job, then they need to back it up, for example, by providing a suitable description of what it is like to carry out our thought experiment. But they provide no such thing. For all they say about the phenomenology of our thought experiment, there is little to choose between baptismal reference and demonstrative reference.

In any event, the Truth in Fiction Model Marks 2 and 3 both turn out to be inadequate in a way that the Truth in Fiction Model Mark 1 is not. They are undermined by the same objection which undermines Malmgren's Possibility Model: there are insufficient resources in their vicinity to support an explanation of why our (I1)-belief rationally commits us to both refrain from holding an (I2)-belief and refrain from holding an (I3)-belief. Ichikawa and Jarvis do not have the option of explaining these rational commitments in terms of the nature of belief and the logico-analytical relationships between contents. The respective real contents of (I1), (I2), and (I3) are to be given by strict conditionals with existential statements as their consequents:

$(3_{\text{TF-M2/M3}})$ $\Box(g/d \rightarrow \exists x \exists p\ (JTBxp \wedge \neg Kxp))$

(TFM2/M3-2) $\Box(g/d \rightarrow \exists x \exists p\ (\neg JBxp \wedge TBxp \wedge \neg Kxp))$

(TFM2/M3-3) $\Box(g/d \rightarrow \exists x \exists p\ (JTBxp \wedge Kxp))$

These strict conditionals are compatible; they do not entail one another's negations. This is because they are all vacuously true if ◇g/d is false. But unlike the other sets of conditionals (strict and counterfactual) which we have looked at so far, no incompatibilities emerge between the conditionals in this set even under the assumption of the metaphysical possibility of their antecedents. Assuming the truth of ◇g/d, they cannot be said to stand in anything like the contrary relation to one another: given any two of them, the truth of one does not rule out the truth of the other. This is because the consequents of the strict conditionals are jointly consistent existential statements.[56] Our imaginative engagement with the Gettier case, furthermore, essentially involves no mental state the content of which could be combined with ◇g/d and the strict conditionals to generate the required incompatibilities. If, therefore, the Truth in Fiction Model Marks 2 and 3 are to be saved, Ichikawa and Jarvis need to find an alternative explanatory strategy. But as with Malmgren's Possibility Model, there do not seem to be any good ones available.

### 5.3. A dilemma for Marks 1, 2, and 3

Ichikawa and Jarvis do not explore any other ways of implementing the truth in fiction approach to modelling philosophical thought experimentation. Even so, our model-building endeavours have made a great deal of headway. With help from Ichikawa and Jarvis, we have finally begun to get a handle on how to overcome the obstacles presented by the metaphysical possibility of deviant realisations of the text of the Gettier case. The general idea of their approach, as I have explicated it, is that our competency with the principles governing the enrichment of texts somehow guides the imaginative activity of the mind, leading it to recognise that the text of the Gettier case is not to be enriched in a manner conducive to deviancy. This approach is very promising, but the discussion of the previous two sub-sections sug-

56 Indeed, not only are those existential statements consistent with one another, it may well be that they are all true in the actual world. Regarding ∃x∃p (¬JBxp ∧ TBxp ∧ ¬Kxp), there are surely actual subjects who hold some unjustified true beliefs. Regarding ∃x∃p (JTBxp ∧ Kxp)), only sceptics would venture to deny it; everyone else will surely admit that there are actual subjects who hold some justified true beliefs which constitute knowledge. As for ∃x∃p (JTBxp ∧ ¬Kxp), its actual truth also strikes me as quite probable. After all, there are many hundreds of millions of subjects in the actual world, and Gettier-style cases need not be recherché or outlandish. Gettier-style cases involving stopped watches or other faulty timepieces are just one kind of example. As Williamson (2007: 192) observes, "sometimes a stopped watch really does show the right time."

gests that Ichikawa and Jarvis have struggled to come up with an adequate implementation of it. How, then, should our model-building endeavours proceed? The way forward emerges from another line of objection, one which threatens to undermine the Truth in Fiction Model Marks 1, 2, and 3 equally. Those models are caught in a dilemma. As we shall see presently, the dilemma's horns are dangerous, but they do not show that Ichikawa and Jarvis have thrown our model-building endeavours wildly and hopelessly off-course. Instead, the dilemma leads us naturally toward the development of a new and better implementation of the truth in fiction approach.

The first horn of the dilemma has to do with a certain *subjective* aspect of our thought experiment. According to the Truth in Fiction Model Marks 1, 2, and 3, when we imaginatively engage with the text of the Gettier case, our tacit grasp of the principles of fictional discourse guides the mind to form a representation of a metaphysical possibility which accords with what is true in the Gettier story. The content of this imaginational state is said to be given by ($2_{\text{TF-M1}}$), ($2_{\text{TF-M2}}$), or ($2_{\text{TF-M3}}$). Surely, however, our tacit grasp of the principles of fictional discourse need not prevent the imaginative activity of our minds from straying beyond what is true in the Gettier story. It is possible, in other words, for us to add *extra details* to the Gettier story by filling in at least some of its many fictional indeterminacies. We might imagine, for example, that Smith has brown hair, that he is wearing blue jeans and a grey t-shirt, that there are no clouds in the sky, that it is two o'clock in the afternoon, that the grass in the paddock is lush and green, that the paddock contains ten robotic kangaroos and ten real kangaroos, that the rock obscuring the real kangaroos from Smith's view is made of granite, that there is a barbedwire fence around the paddock, and that Smith, in addition to being an avid and experienced spotter of native wildlife, is a scholar of Latin. Or we might imagine the opposite, or perhaps not imagine anything at all with regard to these particular fictional indeterminacies. The latitude of the mind's imaginative activities is very wide even within the constraints imposed by our tacit grasp of the principles of fictional discourse. Of course, since we have this ability to fill in the fictional indeterminacies of the Gettier story in imaginative thought, it is perfectly natural to think that any extra details which we *do* happen to add to the Gettier story must be somehow captured by the content of our imaginational state. This is so, despite the fact that such details are strictly unneces-

sary for falsifying the traditional theory of knowledge. But Ichikawa and Jarvis's models cannot do justice to this natural thought. For nothing is represented in ($2_{TF\text{-}M1}$), ($2_{TF\text{-}M2}$), or ($2_{TF\text{-}M3}$) which goes beyond what is true in the Gettier story.

The foregoing considerations put pressure on Ichikawa and Jarvis to subjectivise their models, in particular, to reconstrue the formula $GC_{TF}xp$ and the propositions g and d so that ($2_{TF\text{-}M1}$), ($2_{TF\text{-}M2}$), and ($2_{TF\text{-}M3}$) accord not merely with what is true in the Gettier story, but with one's privately filled in version of it. In the case of the Truth in Fiction Model Mark 1, this requires understanding the formula $GC_{TF}xp$ to say that x stands to p in the relation that, in one's privately filled in version of the Gettier story, Smith stands to the proposition that there are kangaroos in the paddock. In the case of the Truth in Fiction Model Marks 2 and 3, it requires understanding the proper name "STORY" and the demonstrative "*that*" to pick out one's privately filled in version of the Gettier story. One corollary of Ichikawa and Jarvis's models would then be that different people can enter into imaginational states with different contents. Obviously, however, any such subjectivisation must flow straight through to the strict conditionals ($3_{TF\text{-}M1}$), ($3_{TF\text{-}M2}$), and ($3_{TF\text{-}M3}$), so another corollary would be that the Gettier intuition itself can have different real contents for different people. This brings us to the second horn of the dilemma, which has to do with a certain *objective* (or inter-subjective) aspect of our thought experiment. It is perfectly natural to think that our intuitional state is *shared* with people who, like us, report their intuition about the epistemic status of Smith's true belief by uttering "Smith has a justified true belief, but does not know, that there are kangaroos in the paddock." If someone carries out our thought experiment and then utters that sentence, we do not wonder whether he means the same by it as we do—whether, that is, he has intuited what we have intuited. It would be silly and bizarre for us to ask him to clarify his meaning (unless we already have reason to regard him as using words non-standardly). But Ichikawa and Jarvis's subjectivised models cannot do justice to this second natural thought. For since different people not only can, but in all probability do, fill in the Gettier story's fictional indeterminacies in different ways, those models will pretty much guarantee that our intuitional state is not shared with anyone else.

To sum up: the Truth in Fiction Model Marks 1, 2, and 3 ride roughshod over

either a subjective aspect of our thought experiment or an objective one. Although it is open to Ichikawa and Jarvis to bite one or the other of these bullets, we should look on that course of action as a last resort. After all, the bullets are rather unpalatable. And even if Ichikawa and Jarvis are happy to go through with it, their models still face the problems discussed in the previous two sub-sections. The best course of action for us is to seek a new and better way of implementing the truth in fiction approach. We should focus our efforts first and foremost on extricating ourselves from the horns of the foregoing dilemma, since it poses the most general threat to the truth in fiction approach. Why, it is worth asking, are the Truth in Fiction Model Marks 1, 2, and 3 equally vulnerable to it? My diagnosis is that their common vulnerability has its source in the assumption that the Gettier intuition is an intuition about the distribution of the relations of knowledge and justified true belief in a specific scenario. This assumption is understandable and widespread, but I counsel its abandonment for the sake of making further progress on solving the content problem. In the next section, I develop and defend an implementation of the truth in fiction approach on which undergoing the Gettier intuition involves entering into an intuitional state about the distribution of the relations of knowledge and justified true belief in all scenarios of a certain general kind.

# 6. The New Truth in Fiction Model

The presentation of my own model requires the introduction of some further apparatus for theorising about fiction. An enrichment or filling in of a given text is constituted by that text's fictive sentences together with any non-fictive sentences whatever. The text of *Vanity Fair* together with, for example, the sentence "Becky Sharp had a central nervous system" constitutes an enrichment of that text. The text of *Vanity Fair* together with the sentences "Becky Sharp had a central nervous system" and "Becky Sharp had a mole on her lower back" constitutes another one. And there are also much more detailed enrichments of the text of *Vanity Fair*, like the *Vanity Fair* story. In general, texts can be enriched in an infinite number of different ways. We may divide the enrichments of a given text into two disjoint and exhaustive categories: the category of *fictionally permissible* and the category of *fictionally impermissible*. An enrichment of a text is fictionally permissible if no conjunction of two or more of its elements is ruled out by the principles of fictional discourse; otherwise, the enrichment is fictionally impermissible. Enriching a text with one fictionally false sentence is always enough to violate fictional permissibility. For example, adding the sentence "Several of Becky Sharp's ancestors were wizards from another galaxy" to the text of *Vanity Fair* results in a fictionally impermissible enrichment. Since that sentence is ruled out by the principles of fictional discourse, its conjunction with any other sentence must be ruled out by them as well. But even enriching a text with fictionally indeterminate sentences can lead to a violation of fictional permissibility. The sentences "Becky Sharp had a mole on her lower back" and "Becky Sharp had no moles on her lower back" are fictionally indeterminate, but if both are added to the text of *Vanity Fair* the result is a fictionally impermissible enrichment, since their conjunction is ruled out by the principles of fictional discourse.

We may also divide the enrichments of a given text into two other disjoint and exhaustive categories: the category of *fictionally complete* and the category of *fic-*

*tionally incomplete*. An enrichment of a text is fictionally complete if every sentence ruled in by the principles of fictional discourse is a member of it, that is to say, if the story told by the text is one of the enrichment's subsets; otherwise, the enrichment is fictionally incomplete. Stories are always fictionally complete enrichments of the texts used to tell them. This, of course, is because every story is trivially a subset of itself. But stories are, as it were, limiting cases of fictional completeness. The union of a story with any set of sentences whatever counts as a fictionally complete enrichment of the text used to tell the story. For example, there is a fictionally complete enrichment of the text of *Vanity Fair* comprising the sentences of the *Vanity Fair* story together with the fictionally indeterminate sentence "Becky Sharp had a mole on her lower back". There is another such enrichment comprising the sentences of the Gettier story together with the fictionally false sentence "Several of Becky Sharp's ancestors were wizards from another galaxy". There is even a fictionally complete enrichment of the text of *Vanity Fair* comprising the sentences of the *Vanity Fair* story together with all of that story's fictional indeterminacies and fictional falsehoods. The categories of fictional completeness and incompleteness obviously overlap with the categories of fictional permissibility and impermissibility. Each enrichment of a given text must therefore fall into one and only one the following four groups: fictionally permissible and complete, fictionally permissible and incomplete, fictionally impermissible and complete, or fictionally impermissible and incomplete. The *Vanity Fair* story is one example of a fictionally permissible and complete enrichment of the text of *Vanity Fair*, but there are many others.

The final piece of apparatus I need to introduce is that of *fictional roles*. Most texts and their enrichments purport to attribute properties and relations to particular entities or things, such as people, animals, plants, machines, buildings, furniture, and so on. The text of *Vanity Fair*, for example, purports to attribute thousands of properties and relations to a particular person called "Rebecca Sharp". These include a purported attribution of the property of uttering "Revenge may be wicked, but it's natural". Although Thackeray's novel is a relatively long one, the large number of purported attributions found in the text of *Vanity Fair* is dwarfed by the even larger numbers of purported attributions found in many of the text's enrichments, such as the *Vanity Fair* story. Becky Sharp is depicted in much greater detail by the *Vanity*

*Fair* story than by the text used to tell it. Presumably, this depiction includes a purported attribution of the property of having a central nervous system, as well as purported attributions of the properties of having arteries and veins, muscles and bones, a heart and lungs, a liver, kidneys, intestines, fingers and toes, a belly button, and an alimentary canal. There are, of course, other enrichments of the text of *Vanity Fair* which depict Becky Sharp in even greater detail than the *Vanity Fair* story does. Fictional roles are specifications of the properties and relations purportedly attributed to particular entities or things by texts and enrichments. They may be characterised roughly as follows. We start by taking a text or an enrichment and conjoining its elements. Then we existentially quantify over each of the particular entities or things to which properties and relations are purportedly attributed. Finally, we remove one of the existential quantifiers. The resultant open formula is, or expresses, a certain a fictional role, and anything which satisfies the open formula may be said to play that role. Thus, the text of *Vanity Fair* has a Becky-role associated with it, as does the *Vanity Fair* story and every other enrichment of the text of *Vanity Fair*. As a matter of fact, there are many different Becky-roles. None of them are played by things in the actual world, but at least some of them are probably played by things in non-actual metaphysically possible worlds.[57]

57 The (admittedly rough) way in which I have characterised fictional roles is intended to guarantee that no fictional proper names or fictional pronouns are ever used in them. All uses of such names and pronouns are to be replaced with variables. Consider, for example, the sentence "Becky Sharp attended the same academy for young ladies as Amelia Sedley." This is a metafictive sentence of the *Vanity Fair* story. (For evidence, see the first chapter of Thackeray (2001).) The fictional proper names "Becky Sharp" and "Amelia Sedley" are used in it. To form the Becky-role associated with the *Vanity Fair* story, we must conjoin that metafictive sentence with the rest of the *Vanity Fair* story's elements (i.e. with the *Vanity Fair* story's fictive sentences and its other metafictive sentences). We must then replace "Becky Sharp" (as well as "Becky", "Rebecca", "Rebecca Sharp", "Miss Sharp", etc.) and any relevant fictional pronouns with an existentially quantified variable throughout, say, the variable x. We must also replace "Amelia Sedley" (as well as "Amelia", "Miss Sedley", etc.) and any relevant fictional pronouns with an existentially quantified variable throughout, say, the variable y. And, of course, we must do likewise for every other use of a fictional proper name and fictional pronoun. The final step in forming the Becky-role associated with the *Vanity Fair* story is to remove the existential quantifier quantifying over the variable x. The resultant fictional role will include among its many conjuncts the formula "x attended the same academy for young ladies as y", where x is unbound and y is bound. (Of course, if we were forming the Amelia-role associated with the *Vanity Fair* story, we would instead remove the existential quantifier quantifying over the variable y, leaving y unbound and x bound.) The elimination of all uses of fictional proper names and fictional pronouns in the formation of fictional roles allows for at least some fictional roles to be played by things in metaphysically possible worlds. This is so even if, as Kripke (1980; 157-8) has argued, fictional characters like Becky Sharp and Amelia Sedley are essentially fictional. Kripke's doctrine of essential fictionality is problematic here only if we countenance fictional roles in which fictional names or fictional pronouns are used. Suppose that, in the foregoing example, we fail to eliminate "Amelia Sedley", resulting in a Becky-role with the formula "x attended the same academy for young

All of this apparatus straightforwardly applies to the Gettier case and the other fictions deployed in philosophical thought experimentation. An enrichment of the text of the Gettier case is constituted by the text's fictive sentences together with any non-fictive sentences whatever. There are enrichments of the text of the Gettier case which add only a little bit of detail to it, such as the enrichment consisting of the text together with the sentence "Smith is conscious". There are other enrichments, like the Gettier story, which add much more detail to it. We may exhaustively divide the enrichments of the text of the Gettier case into the disjoint categories of fictionally permissible and fictionally impermissible. The text of the Gettier case together with the sentence "Smith is conscious" is a permissible enrichment, as is the Gettier story, and the Gettier story together with the sentence "Smith has brown hair". But the text of the Gettier case together with the sentence "Smith is one nanometre tall" is presumably a violation of fictional permissibility. We may also exhaustively divide the enrichments of the text of the Gettier case into the disjoint categories of fictionally complete and fictionally incomplete. The Gettier story is fictionally complete, as is the the union of the Gettier story with any set of its fictional indeterminacies or fictional falsehoods. But if an enrichment lacks one or more of the Gettier story's fictive or metafictive sentences, then it must be fictionally incomplete, no matter how much detail it contains. Since the categories of fictional permissibility and impermissibility can overlap with the categories of fictional completeness and incompleteness, each enrichment of the text of the Gettier case must fall into one and only one of four groups. For the purposes of modelling our thought experiment, the most important of them is the group of fictionally permissible and complete enrichments. There are many enrichments in this group. The Gettier story is one of them. The other enrichments in the group all have the Gettier story as a subset because they are fictionally complete. Since they are also fictionally permissible, it follows that they cannot contain any fictional falsehoods and must therefore amount to fillings-in of some or all of the Gettier story's fictional indeterminacies.

The text of the Gettier case and its enrichments all have fictional roles associated with them. This is because they purport to attribute properties and relations to

ladies as Amelia Sedley" as one of its conjuncts. Assuming Kripke is right, it would then follow that this Becky-role is not played in any metaphysically possible world, for it would be metaphysically impossible for anything to attend the same academy as Amelia Sedley. But I do not countenance any such fictional roles.

particular entities or things. One of those things is a particular person called "Smith". The text of the Gettier case, for example, purports to attribute several properties and relations to Smith. These include a purported attribution of the property of walking through the Australian countryside and a purported attribution of the property of being an expert spotter of native wildlife. But there are enrichments of the text of the Gettier case which purport to attribute many more properties and relations to Smith. The Gettier story, for example, depicts Smith in much greater detail than the text used to tell it. Presumably, this depiction of Smith includes a purported attribution of the property of being conscious, as well as purported attributions of the properties of being alive, being in space and time, and being capable of locomotion. Another entity or thing to which the text of the Gettier case and its enrichments purport to attribute properties and relations is the proposition that there are kangaroos in the paddock. For example, in the text of the Gettier case, this proposition is purportedly related to Smith, in particular, it is a proposition to which Smith purportedly stands in the relation of belief. Some enrichments of the text of the Gettier case depict the proposition in greater detail. For example, in the Gettier story, Smith purportedly stands to the proposition in both the relation of justified true belief and the relation of ignorance (i.e. the complement of the relation of knowledge). Thus, the text of the Gettier case has a Smith-role and a kangaroo-proposition-role associated with it, as does the Gettier story and every other enrichment of the text of the Gettier case. There are many different such roles. For the purposes of modelling our thought experiment, the most important are those associated with fictionally permissible and complete enrichments. They include not only the Smith-role and the kangaroo-proposition-role associated with the Gettier story, but also the ones associated with enrichments that, in addition to containing all of the Gettier story's fictional truths, fill in some or all of its fictional indeterminacies without entailing any of its fictionally false conjunctions.

My implementation of the truth in fiction approach is based on two key ideas about the mind's deployment of the conceptual apparatus of fictional permissibility, fictional completeness, and fictional role in our test of the traditional theory of knowledge. The first idea is that, guided by our tacit grasp of the principles of fictional discourse, we pick out one of the many fictionally permissible and complete

enrichments of the text of the Gettier case, and enter into an imaginational state representing a scenario in which there are entities or things which play the Smith-role and the kangaroo-proposition-role associated with that particular enrichment. The second idea is that, after we have brought this scenario before the mind in imaginative thought, we enter into an intuitional state about the distribution of the relations of knowledge and justified true belief in any scenario of the same general kind. According to this way of implementing the truth in fiction approach, our imaginational state pertains to a *highly specific* scenario, viz. a scenario which accords with what is true in the Gettier story and our own private filling in of the Gettier story's fictional indeterminacies; whereas our intuitional state pertains to something *much more generic*, viz. the kind of scenario in which there are entities or things which play the Smith-role and the kangaroo-proposition-role associated with some fictionally permissible and complete enrichment of the text of the Gettier case. To state these two ideas another way, the traditional theory of knowledge is tested against a scenario of considerable specificity, but the tribunal of the imagination delivers a verdict—the Gettier intuition—which applies to a whole family of kindred scenarios.

We can represent something's playing the Smith-role associated with a fictionally permissible and complete enrichment using the open formula $R_S xe$, where the variable "x" occupies the position for the thing playing the role and the variable "e" occupies the position for the enrichment. This formula is to be understood as saying that x plays the Smith-role associated with e. Similarly, we can represent something's playing the kangaroo-proposition-role associated with a fictionally permissible and complete enrichment using the open formula $R_K pe$, where the variable "p" occupies the position for the thing playing the role and the variable "e" once again occupies the position for the enrichment. This second formula is to be understood as saying that p plays the kangaroo-proposition-role associated with e. We can use "i" as a singular term for the particular enrichment we happen to pick out when we carry out our thought experiment. Of course, additional singular terms (i′, i′′, etc.) will be required to stand for different fictionally permissible and complete enrichments picked out by other people. With this symbolism in hand, I am now in a position to put forward my own model of our thought experiment, which I shall call *the New Truth in Fiction Model*:

(1) $T \rightarrow \Box\forall x\forall p\ (Kxp \leftrightarrow JTBxp)$

($2_{NTF}$) $\Diamond\exists x\exists p\ (R_S xi \wedge R_K pi)$

($3_{NTF}$) $\Box\forall x\forall p\forall e\ ((R_S xe \wedge R_K pe) \rightarrow (JTBxp \wedge \neg Kxp))$

(4) $\neg\Box\forall x\forall p\ (Kxp \leftrightarrow JTBxp)$ From ($2_{NTF}$) and ($3_{NTF}$)

---

(5) $\neg T$ From (1) and (4)

On this model, our imaginative engagement with the Gettier case involves apprehending the truth of ($2_{NTF}$), which affirms the metaphysical possibility of a scenario in which entities or things play the Smith-role and the kangaroo-proposition-role associated with the fictionally permissible and complete enrichment i. The real content of the Gettier intuition is given by ($3_{NTF}$), which affirms the metaphysical necessitation of non-knowledge justified true belief by every scenario of that general kind.

At first glance, the New Truth in Fiction Model may seem overly elaborate and contrived, especially in comparison with some of the other models we have looked at. More specifically, when we imaginatively engage with the Gettier case and undergo the Gettier intuition, it does not obviously seem as though the mind deploys anything as baroque as the conceptual apparatus of fictional permissibility, fictional completeness, and fictional roles. The contrast between the baroqueness of the strict conditional ($3_{NTF}$) and the simplicity of the apparent content of the Gettier intuition is particularly marked. Before proceeding any further, therefore, it is important for me to say something to assuage the worry that I am taking our model-building endeavours down the wrong path. There are three points I should like to emphasise in this connection. The first point is one that I have already made several times over the course of the discussion, but it bears re-emphasising here. The apparent content of the Gettier intuition is an illusion; it must differ from the Gettier intuition's real content in some way or another. Since this fact obliges us to considerably temper our faith in the appearances, the postulation of something phenomenologically incongruous (such as the mind's deployment of the foregoing apparatus) need not be objectionable in and of itself. To be sure, as we go about pursuing our model-building endeavours, we should avoid postulating things which make it difficult for us to

make sense of our own mental lives. But even if the apparatus I have introduced in the present sub-section is rather baroque, it can hardly be said to make us lose our grip on what is going on in our own heads when we carry out our thought experiment.

The second point I should like to emphasise is that the concepts of fictional permissibility, fictional completeness, and fictional role are simply refined statements of certain notions already familiar to us from ordinary life, in particular, notions familiar to us from the practice of reading novels and other standard kinds of fictional works. The concept of fictional permissibility is a refinement of the ordinary notion that some ways of filling in a given text are wrong while other ways are not wrong. The concept of fictional completeness is a refinement of the ordinary notion that some ways of filling in a given text fall short of the whole story while other ways do not fall short and even go beyond it. And the concept of a fictional role is a refinement of the ordinary notion that story-telling involves descriptions of characters and other particular entities or things. Since these concepts are in these ways rooted in the practice of reading standard fictional works, it is by no means implausible that our tacit grasp of the principles of fictional discourse brings with it a tacit grasp of them as well. The third, and perhaps most important, point for me to emphasise here is that none of the existing models of philosophical thought experimentation have turned out to be adequate. We must look elsewhere if we are to make further progress on solving the content problem. This, furthermore, may require setting aside the sorts of misgivings which might otherwise induce us to dismiss a novel approach like mine without giving it much consideration. The upshot is that worrying about the apparatus I have introduced in this sub-section is not very helpful at this late stage of our discussion. We should be open to the possibility of it performing serious theoretical work.

### 6.1. Adequacy of the New Truth in Fiction Model

Having put forward my new way of implementing the truth in fiction approach to modelling philosophical thought experimentation, I now turn my efforts to the task of supporting its adequacy. The New Truth in Fiction Model avoids all of the problems which render the other models we have looked at inadequate. To begin with,

unlike the Necessity and Counterfactual Models, the New Truth in Fiction Model overcomes the obstacles presented by the metaphysical possibility of deviant realisations of the text of the Gettier case. On my model, when we imaginatively engage with the Gettier case, we pick out a fictionally permissible and complete enrichment i. This particular enrichment represents a highly specific scenario in which there are entities or things which, together with the particular enrichment i, satisfy the open formulae $R_S$xe and $R_K$pe. Since $R_S$xe and $R_K$pe respectively stand for the Smith-roles and the kangaroo-proposition-roles associated with fictionally permissible and complete enrichments, any world at which they are satisfied is a world at which each of the Gettier story's fictional falsehoods is false and each of its fictional truths is true. The Gettier story precludes Smith from justifiedly and truly believing, but failing to know, that there are kangaroos in the paddock. In other words, at any world at which $R_S$xe and $R_K$pe are satisfied, the entity or thing which plays the Smith-role is precluded from standing in either the relation of unjustified true belief or the relation of knowledge to the entity or thing which plays the kangaroo-proposition-role. The mind is led to recognise the preclusion by first apprehending the metaphysical possibility of there being entities or things which, together with the particular enrichment i, satisfy $R_S$xe and $R_K$pe, and then reflecting on the general kind of scenario to which this highly specific scenario belongs. The mind's recognition of the preclusion is manifested by the Gettier intuition. Considerations of deviancy have nothing to do with the truth or falsity of the Gettier intuition because no world at which $R_S$xe and $R_K$pe are satisfied is a deviant world.

Next, unlike the Possibility Model and the Truth in Fiction Model Marks 2 and 3, there are enough resources in the vicinity of my model to support an explanation of why the (I1)-belief we come to form when we carry out our thought experiment rationally commits us to both refrain from holding an (I2)-belief and refrain from holding an (I3)-belief. The explanation of these rational commitments parallels the explanations given in connection with the Necessity and Counterfactual Models and the Truth in Fiction Model Mark 1. On my implementation of the truth in fiction approach, the respective real contents of (I1), (I2), and (I3) are given by the following strict conditionals:

$(3_{NTF})$ $\Box\forall x\forall p\forall e\,((R_Sxe \wedge R_Kpe) \rightarrow (JTBxp \wedge \neg Kxp))$

(NTF-2) $\Box\forall x\forall p\forall e\,((R_Sxe \wedge R_Kpe) \rightarrow (\neg JBxp \wedge TBxp \wedge \neg Kxp))$

(NTF-3) $\Box\forall x\forall p\forall e\,((R_Sxe \wedge R_Kpe) \rightarrow (JTBxp \wedge Kxp))$

There is no incompatibility between these strict conditionals; they do not entail one another's negations. This is because they are all vacuously true if there is no metaphysically possible world at which some entities or things together with some enrichment satisfy $R_Sxe$ and $R_Kpe$. But under the safe assumption of the truth of $\Diamond\exists x\exists p\,(R_Sxi \wedge R_Kpi)$, the strict conditionals may be said to stand in something like the contrary relation to one another: given any two of them, both may be false but at most one can be true. The result is that the combination of $\Diamond\exists x\exists p\,(R_Sxi \wedge R_Kpi)$ with the real content of our (I1)-belief rules out the real contents of (I2) and (I3). So my implementation of the truth in fiction approach supports a straightforward explanation of the rational commitments associated with our thought experiment in terms of the nature of belief and the logico-analytical relationships between contents.

Finally, the New Truth in Fiction Model constitutes an advance on the Truth in Fiction Model Marks 1, 2, and 3 because, unlike its predecessors, it can simultaneously do justice to the natural thoughts underlying the dilemma I levelled against them in Section 5.3. We find it natural to think, on the one hand, that any extra details we add to the Gettier story by filling in its fictional indeterminacies should somehow be captured by the content of our imaginational state. But on the other hand we also find it natural to think that we share our intuitional state with people who, like us, report their intuition about the epistemic status of Smith's true belief by uttering "Smith has a justified true belief, but does not know, that there are kangaroos in the paddock." The Truth in Fiction Model Marks 1, 2, and 3, as I have already shown, are all caught in this dilemma; they ride roughshod over either a subjective aspect of our thought experiment or an objective (or inter-subjective) one. The New Truth in Fiction Model, however, has been expressly designed to avoid both horns. It avoids the first horn of the dilemma because it subjectivises the con-

tent of our imaginational state. The content of that state, on my model, is fully determined by whatever fictionally permissible and complete enrichment we happen to pick out in imaginative thought. If our privately filled in version of the Gettier story is different from someone else's, then the content of his imaginational state must diverge from the content of our own. This is so even if the difference between the two privately filled in versions of the Gettier story is very small. It could, for example, amount to nothing more than that we imagine Smith with brown hair while the other person imagines Smith with black hair. A divergence in content is still guaranteed, since we will have picked out one particular enrichment in imaginative thought (viz. the enrichment i) and he will have picked out another (e.g. i′). As for the second horn of the dilemma, the New Truth in Fiction Model avoids it as well. This is because the strict conditional ($3_{NTF}$) is given as the real content of everyone's Gettier intuition, regardless of the divergent contents of different people's imaginational states.

In addition to avoiding the problems which render its rivals inadequate, the New Truth in Fiction Model avoids offending against the "traditional understanding of philosophical methodology." By giving the strict conditional ($3_{NTF}$) as the real content of the Gettier intuition, the New Truth in Fiction Model automatically falls in line with the traditionalist tenet that the intuitions generated by philosophical thought experimentation always involve the concept of metaphysical necessity. And nothing prevents traditionalists like Ichikawa and Jarvis from combining the New Truth in Fiction Model with the other traditionalist tenet that philosophical thought experimentation is an a priori activity. Setting aside radical empiricism, there is no *special* reason to worry that ($3_{NTF}$) is empirically contaminated. (In particular, from the fact that it is an empirical question as to whether the open formulae $R_S$xe and $R_K$pe are satisfied, it does not follow that it is also an empirical question as to whether their being satisfied metaphysically necessitates justified true belief without knowledge.) Ichikawa and Jarvis should therefore be willing to defend the a priori knowability of ($3_{NTF}$), given that they are willing to defend the a priori knowability of the strict conditionals ($3_{TF\text{-}M2}$) and ($3_{TF\text{-}M3}$). Of course, for anyone uncommitted or unsympathetic to the traditionalist tenet of the apriority of philosophical thought experimentation, its compatibility with the New Truth in Fiction Model will have little

or no bearing on that model's adequacy. But as there are many contemporary philosophers who do give their endorsement to traditionalism, it is worthwhile noting that commitment to this particular traditionalist tenet is no impediment to acknowledging that, on the whole, the New Truth in Fiction Model does a better job of modelling our Gettier-style thought experiment than any of the existing models.

### 6.2. In defence of the truth in fiction approach

We are almost in a position to bring our model-building endeavours to a conclusion. So far we have found that the New Truth in Fiction Model outdoes all of its rivals. Malmgren (2011), however, has subjected the truth in fiction approach to a series of objections the intention of which is to undermine every possible way of implementing it, including my own. The overall thrust of her attack is that the assimilation of the practice of thought experimentation to the practice of reading standard fictional works is specious. There are *significant disanalogies* between testing theories in imaginative thought and reading novels, narrative poems, and the like. The identification of these disanalogies is something which, in Malmgren's view, the truth in fiction approach simply cannot survive. If, therefore, the adequacy of my (or any other) implementation of the truth in fiction approach is to be upheld, it is incumbent on me to defend it from Malmgren's series of objections. In the present context, undertaking such a defence will also serve the useful purpose of further elucidating both the truth in fiction approach and my own way of implementing it.

The first of Malmgren's objections appeals to Kripke's (1980) doctrine of the essentially fictional nature of fictional characters. According to that doctrine, there is no metaphysically possible world at which some woman is identical to Thackeray's Becky Sharp. For Thackeray wrote the novel *Vanity Fair* as a work of fiction. The discovery that, by an amazing coincidence, there was once an actual woman who completely satisfied Thackeray's description of Becky Sharp would not show that he was writing about *that* woman. And if this actual woman could not be said to be identical to Becky Sharp, then surely no non-actual yet metaphysically possible woman satisfying Thackeray's description of Becky Sharp could be said to be identical to Becky Sharp either. Becky Sharp is fictional and she is *essentially* so. As Malmgren suggests, one consequence of this doctrine of Kripke's is that an ordinary

fictional judgement about Becky Sharp expressed by uttering, say, "Becky Sharp has a mole on her lower back" would not be verified (or falsified) by the existence of an actual woman who both satisfied Thackeray's description and did (or did not) have a mole on her lower back. "In contrast," Malmgren (2011: 298) goes on to argue, "an intuitive judgement—for example, the Gettier judgement—*does* seem to be verifiable/falsifiable by actual realisations of the relevant problem case [e.g. the text of the Gettier case]. We might put the point by saying that fictional characters are essentially fictional, whereas characters in philosophical problem cases are not." A character like Becky Sharp, in other words, is significantly disanalogous to a character like Smith. Becky Sharp is nowhere to be found within the sphere of the metaphysically possible, but there are many metaphysically possible worlds at which some man is identical to Smith—indeed, by an amazing coincidence, it could even turn out that the actual world is among them.

This initial line of objection, however, is defective. The most the alleged disanalogy can show is that the Gettier intuition is not just another judgement about what is true in a fiction. But as I have already explained, the truth in fiction approach is not about equating the Gettier intuition with an ordinary fictional judgement. Ichikawa and Jarvis's models as well as my own model are all implementations of the truth in fiction approach, yet none of them pretend to equate the Gettier intuition with an ordinary fictional judgement. They are members of a family of models exemplifying the general idea that the imaginative mental activity which constitutes our thought experiment is guided by our tacit grasp of the principles or rules governing fictional discourse. It is this general idea which underlies the truth in fiction approach; but the disanalogy Malmgren claims to have identified poses no threat to it. This is because, even if it is granted that the characters in standard fictional works are essentially fictional while those in case descriptions are not, it obviously does not follow that our tacit grasp of the principles of fictional discourse (or perhaps our tacit grasp of principles akin to them) cannot guide the mind during the performance of our thought experiment. Malmgren, at any rate, does not tell us why such guidance of the mind would be impossible under such conditions. The truth in fiction approach must therefore remain unscathed. Note, parenthetically, that in attempting to establish the disanalogy between characters like Becky Sharp and characters like

Smith, Malmgren is led into incoherence. The culprit is her assertion that the Gettier intuition contrasts with ordinary fictional judgements by being "verifiable/falsifiable" by actual realisations of the text of the Gettier case. Malmgren cannot coherently maintain this assertion because, as I mentioned in Section 3.2, she joins Ichikawa and Jarvis and myself in rejecting Williamson's Counterfactual Model on the grounds that the Gettier intuition would not be falsified by the existence of an actual man who realised the text of the Gettier case while failing to have a justified true belief without knowledge.

Malmgren's second objection against the truth in fiction approach stems from her observation that "what counts as a permissible interpretation of a case description seems to depend, at least in part, on the specific *use* to which the case is put; more precisely, the tacit constraints [or principles] that govern our interpretation of a case description are sensitive to the *target* of the given thought experiment—to what theory is being tested" (2011: 299). To illustrate what she has in mind here, Malmgren compares the Gettier case with Foot's (1967) and Thomson's (1976) famous trolley cases. Normally, the Gettier case is used to test the traditional theory of knowledge, whereas trolley cases are used to test utilitarianism (and the doctrine of double effect). Malmgren points out that in a test of the former kind "we may not suppose that the subject in the Gettier case has more than one route to knowledge available, but we may suppose that killing him would cause a riot in which fifty other people die" (2011: 299). The opposite goes for the potential victim in a trolley case used to test utilitarianism. In the latter kind of test, says Malmgren, "we may not suppose that killing him would cause a riot in which fifty people die, but we may indeed suppose that he has more than one route to knowledge available (here: knowledge of any proposition of our choice)" (2011: 299). Malmgren then asks us to consider the case description that results from combining the Gettier case with a trolley case. The resultant *Gettier-Trolley case* will not be permissibly enrichable in the same ways as its constituent case descriptions are permissibly enrichable (when they are put to their normal uses). Instead, "[w]hat we may and may not import into the Gettier-Trolley case once again depends on *what we are using it for*: whether we are using the case to test the JTB theory, utilitarianism, both—or something else together" (2011: 300). But the principles governing the enrichment of the texts of standard

fictional works exhibit no analogous sensitivity, according to Malmgren, "unless, of course, the text is being hijacked for the purpose of a thought experiment" (2011: 300).

This second line of objection is also defective. Malmgren has at most shown that there are different kinds of fictional discourse governed by different sets of principles or rules. But this fact about fictional discourse should come as no surprise to anyone, for we are all familiar with what are commonly referred to as *genres*. Thackeray's novel *Vanity Fair*, for example, belongs to the genre of broadly realistic fictional discourse, in that the way things are described to be in *Vanity Fair* does not constitute a great departure from the way things are (or at least once were) in reality. The characters and events in *Vanity Fair* are realistic characters and events; indeed, some of the characters, like Napoleon, actually existed and many of the events, like the Battle of Waterloo, actually happened. Of course, there are other genres in addition to realistic fictional discourse, such as science fiction, surrealist fiction, fantasy fiction, and so on. Importantly, we find that the set of principles governing the enrichment of texts in one genre often differ quite a lot from the set of principles governing the enrichment of texts in another genre. Thus, realistic fictional discourse is presumably governed by principles which blanketly prohibit the attribution of telepathic and telekinetic powers to humans, whereas the set of principles governing the discourse of science fiction presumably contains no blanket prohibition against telepathy and telekinesis. In light of this important observation, it is perfectly natural to regard philosophical thought experimentation as also constituting its own kind of fictional discourse governed by its own set of principles. Although it is an interesting question how the set of principles governing philosophical thought experimentation might differ from the sets governing other kinds of fictional discourse, it is not necessary for me to address it here in order to defend the truth in fiction approach. For Malmgren's alleged disanalogy does nothing more than draw attention to one of the potential differences, viz. the sensitivity of that set of principles to whatever theory a given case description is being used to test.

The third objection Malmgren levels against the truth in fiction approach is that the actual world (or perhaps the way we take the actual world to be) is more involved in fixing what is true in realistic stories like the *Vanity Fair* story than it is in

fixing what is true in the Gettier story. As Malmgren puts it, the principles governing realistic fictional discourse "impose a greater degree of overall similarity to the actual world (or better: the actual world as the author and her immediate audience *take it to be*) on the fiction, than the corresponding constraints impose a problem case" (2011: 300). For example, it is presumably false in the *Vanity Fair* story that Becky Sharp has the telekinetic power to move cups across tables, but it would seem to be indeterminate in the Gettier story whether Smith has such a power. We would thus rightly complain against any enrichment of the text of *Vanity Fair* in which Becky Sharp is able to move cups across tables using only her mind. In contrast, if someone carrying out our thought experiment imagined Smith with this telekinetic power, then, says Malmgren, "we might well complain that she has filled out the given description in *idiosyncratic* or *irrelevant* ways […] But I do not think we would complain that her embellishments are *illegitimate*, or that the envisaged realization of the case is *deviant*—as indeed we would if she had filled it out in such a way that, say, Smith's justification is defeated" (2011: 300, fn.64). Having thus claimed to have identified yet another significant disanalogy, Malmgren goes on to admit that "fantasy, science fiction, and surreal fiction are associated with constraints that are more permissive" than those governing realistic fictional discourse (2011: 300); but she straight away dismisses this observation as immaterial to the matter at hand on the alleged grounds that "it would be very implausible to assimilate problem cases to, say, outré science fiction" (2011: 300, fn.63). In Malmgren's view, consideration should only be given to the genre of realistic fiction.

This third line of objection fails as well. There is no warrant for Malmgren's insistence that the truth in fiction approach can survive only if the principles governing philosophical thought experimentation give the same (or nearly the same) weight to the actual world as the principles governing realistic fictional discourse do. As I pointed out in my reply to Malmgren's second objection, there are several different kinds of fictional discourse, and the sets of principles governing them can differ from one another quite a lot. Furthermore, in the previous paragraph, I quoted Malmgren herself as admitting that one way in which those sets can differ has to do with the degree of overall similarity to the actual world which they impose on the enrichment of texts. An example is that the principles governing the enrichment of

texts belonging to genres such as fantasy, science fiction, and surreal fiction are, in Malmgren's words, "more permissive" than the principles governing the enrichment of texts belonging to the genre of realistic fiction. Since, therefore, philosophical thought experimentation is naturally regarded as constituting its own kind of fictional discourse governed by its own set of principles, Malmgren has at most shown that the principles governing the enrichment of case descriptions are likewise more permissive than the principles governing the enrichment of texts belonging to the genre of realistic fiction. But even though the fact that they exhibit this greater permissiveness is an interesting one, it obviously does nothing to undermine the truth in fiction approach or the adequacy of my particular implementation of it. To insist otherwise (as Malmgren does) is to arbitrarily privilege the genre of realistic fiction in our model-building endeavours.

# 7. Conclusion

The project of building a model of philosophical thought experimentation is propaedeutical. It is necessary to undertake it, as I said in my introductory remarks, in order to prepare the ground for serious discussion of the epistemological challenge which threatens the legitimacy of contemporary philosophical methodology. The present monograph advances the propaedeutical work already done by Williamson, Malmgren, and Ichikawa and Jarvis. Following the lead of those philosophers, I have focused on a single Gettier-style thought experiment and the problem of identifying the real content of the Gettier intuition. My contribution to the project of model-building is twofold. First, I have established the inadequacy of all of the existing models. The Necessity Model is inadequate because the strict conditional ($3_N$) is made false by the existence of deviant worlds while the Gettier intuition is not. Williamson's Counterfactual Model is inadequate because the truth value of the counterfactual ($3_{CF}$), but not that of the Gettier intuition, depends on the relative distance between the actual world and deviant worlds in modal space. Malmgren's Possibility Model is inadequate because there are rational commitments associated with our Gettier-style thought experiment which the possibility claim ($2_P$) renders inexplicable. And Ichikawa and Jarvis's Truth in Fiction Model Marks 1, 2, and 3 are inadequate because they ride roughshod over either a subjective aspect of our Gettier-style thought experiment or an objective one.

Ichikawa and Jarvis, however, do not exhaust every way of implementing the general idea behind the truth in fiction approach. The second of my contributions to the project of model-building springs from my appreciation of this fact. I have taken the general idea behind the truth in fiction approach and used it to develop a new model of my own. The New Truth in Fiction Model is very attractive. I have shown, first of all, that it does a better job of modelling our Gettier-style thought experiment than any of the existing models. I have also defended it from Malmgren's attempt to undermine every possible implementation of the truth in fiction approach. Further-

more, I have been unable to find any major defects peculiar to it. (The worry that the New Truth in Fiction Model is overly elaborate and contrived, which I mentioned above, is rather minor in and of itself; not only that, it rapidly fades away upon consideration of the New Truth in Fiction Model's attractions.) We should therefore feel confident that the New Truth in Fiction Model is an adequate representation of our Gettier-style thought experiment, and more particularly, that the solution to the content problem is given by the strict conditional ($3_{\text{NTF}}$) or at least something very much like it. We should also feel confident that the New Truth in Fiction Model is generalisable, for our Gettier-style thought experiment is paradigmatic of the phenomenon to be modelled. Philosophical thought experimentation is guided by the principles of truth in fiction.

# References


Alexander, J. 2012. *Experimental Philosophy: An Introduction*, Polity Press, Cambridge.

Bealer, G. 1996a. "A priori knowledge and the scope of philosophy", *Philosophical Studies*, 81(2-3): 121-42.

Bealer, G. 1996b. "On the possibility of philosophical knowledge", *Noûs*, 30, Supplement: Philosophical Perspectives, 10, Metaphysics: 1-34.

Bealer, G. 1998. "Intuition and the autonomy of philosophy", in M. DePaul & W. Ramsey, eds., *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*, Rowman & Littlefield, Lanham.

Bealer, G. 2000. "A theory of the a priori", *Pacific Philosophical Quarterly*, 81(1): 1-30.

Bencivenga, E. 1986. "Free logics", in D. Gabbay & F. Guenthner, eds., *Handbook of Philosophical Logic*, Reidel, Dordrecht.

Bengson, J. forthcoming. "The intellectual given", *Mind*.

Boghossian, P. 2009. "Virtuous intuitions: Comments on Lecture 3 of Ernest Sosa's *A Virtue Epistemology*", *Philosophical Studies*, 144(1): 111-9.

Bird, A. 2007. *Nature's Metaphysics: Laws and Properties*, Clarendon Press, Oxford.

Bishop, M.A. & Trout, J.D. 2005. *Epistemology and the Psychology of Human Judgement*, Oxford University Press, Oxford.

Bonjour, L. 1985. *The Structure of Empirical Knowledge*, Harvard University Press, Cambridge, MA.

Bonjour, L. 1998. *In Defense of Pure Reason: A Rationalist Account of A Priori Justification*, Cambridge University Press, Cambridge.

Bonjour, L. 2002. *Epistemology: Classic Problems and Contemporary Responses*, Rowman & Littlefield, Lanham.

Broome, J. 1999. "Normative requirements", *Ratio*, 12(4): 398-419.

Brown, J. & Fehige, Y. 2011. “Thought experiments”, in E.N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*, Fall 2011 Edition, pdf version of the entry, http://plato.stanford.edu/archives/fall2011/entries/thought-experiment/

Byrne, R. 2005. *The Rational Imagination: How People Create Alternatives to Reality*, MIT Press, Cambridge, MA.

Byrne, A. 1993. “Truth in fiction: the story continued”, *Australasian Journal of Philosophy*, 71(1): 24-35.

Chalmers, A. 1999. *What is this Thing Called Science?*, University of Queensland Press, St Lucia.

Chan, T. ed., 2013. *The Aim of Belief*, Oxford University Press, Oxford.

Chisholm, R. 1989. *Theory of Knowledge*, Prentice Hall, Englewood Cliffs.

Chudnoff, E. 2011a. “The nature of intuitive justification”, *Philosophical Studies*, 153(2): 313-33.

Chudnoff, E. 2011b. “What intuitions are like”, *Philosophy and Phenomenological Research*, 82(3): 625-54.

Cullison, A. 2010. “What are seemings?”, *Ratio*, 23(3): 260-74.

Currie, G. 1990. *The Nature of Fiction*, Cambridge University Press, Cambridge.

Dennett, D. 1987. *The Intentional Stance*, MIT Press, Cambridge, MA.

Dennett, D. 1991. *Consciousness Explained*, Little, Brown and Company, Boston.

Duhem, P. 1991. *The Aim and Structure of Physical Theory*, Princeton University Press, Princeton.

Erlenbaugh, J. & Molyneux, B. 2009. “Intuitions are inclinations to believe”, *Philosophical Studies*, 145(1): 89-109.

Fales, E. 1990. *Causation and Universals*, Routledge, London.

Foot, P. 1967. “The problem of abortion and the doctrine of double effect”, *Oxford Review*, 5: 5-15.

Geach, P. 1962. *Reference and Generality: An Examination of Some Medieval and Modern Theories*, Cornell University Press, Ithaca.

Gettier, E. 1963. “Is justified true belief knowledge?”, *Analysis*, 23(6): 121-3.

Häggqvist, S. 1996. *Thought Experiments in Philosophy*, Almqvist & Wiksell International, Stockholm.

Häggqvist, S. 2009. "A model for thought experiments", *Canadian Journal of Philosophy*, 39(1): 55-76.

Hanley, R. 2004. "As good as it gets: Lewis on truth in fiction", in F. Jackson & G. Priest, eds., *Lewisean Themes: The Philosophy of David K. Lewis*, Oxford University Press, New York.

Hetherington, S. 2001. *Good Knowledge, Bad Knowledge: On Two Dogmas of Epistemology*, Oxford University Press, Oxford.

Hetherington, S. 2011. *How to Know: A Practicalist Conception of Knowledge*, Wiley-Blackwell, Malden.

Huemer, M. 2005. *Ethical Intuitionism*, Palgrave Macmillan, New York.

Ichikawa, J. 2009, "Knowing the intuition and knowing the counterfactual", *Philosophical Studies*, 145(3): 435-43.

Ichikawa, J. MS. "Intuition and Begging the Question".

Ichikawa J. & Jarvis, B. 2009. "Thought experiment intuitions and truth in fiction", *Philosophical Studies*, 142(2): 221-46.

Jackson, F. 1982. "Epiphenomenal qualia", *The Philosophical Quarterly*, 32(127): 127-36.

King, J. 2005. "Anaphora", in E.N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*, Summer 2013 Edition, pdf version of the entry, http://plato.stanford.edu/archives/sum2013/entries/anaphora/

Knobe, J. & Nichols, S. 2008. *Experimental Philosophy*, Oxford University Press, Oxford.

Kornblith, H. 1994. *Naturalizing Epistemology*, MIT Press, Cambridge MA.

Kornblith, H. 1999. "In defense of a naturalized epistemology", in J. Greco & E. Sosa, eds., *The Blackwell Guide to Epistemology*, Blackwell, Malden.

Kornblith, H. 2002. *Knowledge and Its Place in Nature*, Oxford University Press, Oxford.

Kornblith, H. 2006. "Appeals to intuition and the ambitions of epistemology", in S. Hetherington, ed., *Epistemology Futures*, Oxford University Press, Oxford.

Kripke, S. 1980. *Naming and Necessity*, Harvard University Press, Cambridge, MA.

Lambert, K. 2001. “Free logics”, in L. Goble, ed., *The Blackwell Guide to Philosophical Logic*, Blackwell, Malden.

Le Poidevin, R. 1995. “Worlds within worlds? The paradoxes of embedded fiction”, *British Journal of Aesthetics*, 35(3): 227-38.

Lewis, D. 1973. *Counterfactuals*, Basil Blackwell, Oxford.

Lewis, D. 1978. “Truth in fiction”, *American Philosophical Quarterly*, 15(1): 37-46.

Lewis, D. 1983. *Philosophical Papers: Volume 1*, Oxford University Press, Oxford.

Lewis, D. 1990. “Noneism or allism?”, *Mind*, 99(393): 24-31.

Ludwig, K. 2007. “The epistemology of thought experiments: First person versus third person approaches”, *Midwest Studies in Philosophy*, 31(1): 128-59.

Lynch. M.P. 2006. “Trusting intuitions”, in P. Greenough & M. P. Lynch, eds., *Truth and Realism*, Oxford University Press, Oxford.

Malmgren, A. 2011. “Rationalism and the content of intuitive judgements”, *Mind*, 120(478): 263-327.

Nimtz, C. 2010a. “Saving the doxastic account of intuitions”, *Philosophical Psychology*, 23(3): 357-75.

Nimtz, C. 2010b. “Philosophical thought experiments as exercises in conceptual analysis”, *Grazer Philosophische Studien*, 81: 189-214.

Nünning, A.F. 2005. “Reconceptualizing Unreliable Narration: Synthesizing Cognitive and Rhetorical Approaches”, in J. Pheelan & P. J. Rabinowitz, eds., *A Companion to Narrative Theory*, Blackwell, Malden.

Parsons, T. 1980. *Nonexistent Objects*, Yale University Press, New Haven.

Phillips, J. F. 1999. “Truth and inference in fiction”, *Philosophical Studies*, 94(3): 273-93.

Plantinga, A. 1993. *Warrant and Proper Function*, Oxford University Press, New York.

Proudfoot, 2006, “Possible worlds semantics and fiction”, *Journal of Philosophical Logic*, 35(1): 9-40.

Pust, J. 2000. *Intuitions as Evidence*, Garland Publishing, New York.

Putnam, H. 1975. “The meaning of ‘meaning’”, *Minnesota Studies in the Philosophy of Science*, 7:131-93.

Quine, W. V. O. 1951. “Two dogmas of empiricism”, *Philosophical Review*, 60(1): 20-43.

Quine, W. V. O. 1969. *Ontological Relativity and Other Essays*, Columbia University Press, New York.

Routley, R. 1980. *Exploring Meinong’s Jungle and Beyond*, Australian National University, Canberra.

Searle, J. 1980. “Minds, brains and programs”, *Behavioral and Brain Sciences*, 3(3): 417-424.

Shoemaker, S. 1980. “Causality and properties”, in P. van Inwagen, ed., *Time and Cause*, Reidel, Dordrecht.

Shoemaker, S. 1998. “Causal and metaphysical necessity”, *Pacific Philosophical Quarterly*, 79(1): 59-77.

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

Sidelle, A. 2002. “On the metaphysical contingency of the laws of nature”, in T. S. Gendler & J. Hawthorne, eds., *Conceivability and Possibility*, Clarendon Press, Oxford.

Sider, T. 2010. *Logic for Philosophy*, Oxford University Press, New York.

Sorensen, R. A. 1998. *Thought Experiments*, Oxford University Press, New York.

Sosa, D. 2006. “Scepticism about intuition”, *Philosophy*, 81(4): 633-48.

Sosa, E. 1998. “Minimal intuition”, in M. DePaul & W. Ramsey, eds., *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*, Rowman & Littlefield, Lanham.

Sosa, E. 2007. “Intuitions: Their nature and epistemic efficacy”, *Grazer Philosophische Studien*, 74: 51-67.

Stalnaker, R. 1968. “A theory of conditionals”, in *American Philosophical Quarterly Monograph Series*, 2: 98-112.

Starmans, C. & Friedman, O. 2012. “The folk conception of knowledge”, *Cognition*, 124(3): 272-83.

Swoyer, C. 1982. “The nature of natural laws”, *Australasian Journal of Philosophy*, 60(3): 203-223.

Thackeray, W. M. 2001, *Vanity Fair,* Penguin, London.

Thomson, J.J. 1973. “Killing, letting die and the trolley problem”, *The Monist*, 59(2): 204-17.

Tolhurst, W. 1998. “Seemings”, *American Philosophical Quarterly*, 35(3): 293-302.

Tucker, C. 2010. “Why open-minded people should endorse dogmatism”, *Philosophical Perspectives*, 24(1): 529-45.

Turri, J. 2013. “A conspicuous art: Putting Gettier to the test”, *Philosopher’s Imprint*, 13(10): 1-16.

van Inwagen, P. 1997. “Materialism and the psychological-continuity account of personal identity”, *Noûs*, 31, Supplement: Philosophical Perspectives, 11, Mind Causation, and World: 305-19.

Walton, K. 1990. *Mimesis as Make-Believe*, Harvard University Press, Cambridge, MA.

Way, J. 2010. “The normativity of rationality”, *Philosophy Compass*, 5(12): 1057-68.

Weatherson, B. 2003. “What good are counterexamples?”, *Philosophical Studies*, 115(1): 1-31.

Weinberg, J., Nichols, S. & Stich, S. 2001. “Normativity and epistemic intuitions”, *Philosophical Topics*, 29(1-2): 429-60.

Williamson, T. 2005. “Armchair philosophy, metaphysical modality and counterfactual thinking”, *Proceedings of the Aristotelian Society*, 105(1): 1-23.

Williamson, T. 2007. *The Philosophy of Philosophy*, Blackwell, Oxford.

Wolterstorff, N. 1980. *Worlds and Works of Art*, Clarendon Press, Oxford.

Woods, J. 2007. “Fictions and their logic”, in D. Jacquette, ed., *Philosophy of Logic*, Elsevier, Amsterdam.

Woodward, R. 2011. “Truth in Fiction”, *Philosophy Compass*, 6(3): 158-67.