Fictional Modality and the Intensionality of Fictional Contexts*

Sara L. Uckelman
Department of Philosophy
Durham University
s.l.uckelman@durham.ac.uk

June 14, 2022

Abstract

In [4], Kosterec claims to provide "model-theoretic proofs" of certain theses involving the normal modal operators $\Diamond$ and $\Box$ and the truth-in-fiction (à la Lewis) operator F which he then goes on to show have counterexamples in Kripke models. He concludes from this that the embedding of normal modal logic under the truth-in-fiction operator is unsound. We show instead that it is the "model-theoretic proofs" that are themselves unsound, involving illicit substitution, a subtle error that nevertheless allows us to draw an important conclusion about intensional contexts (such as fictional contexts) and semantic equivalences.

1

In a recent paper [4], Kosterec argued that embedding any of a variety of standard normal modal logics (NMLs) within the scope of a possible-worlds "truth-in-fiction" operator, F, results in "non-intuitive consequences" which are "shown to be contradictory" [p. 13543]. His strategy to demonstrate this is to give "model-theoretic proofs" of certain theses involving F, $\Box$, $\Diamond$, and then to provide Kripke models that falsify these theses. From this he concludes that "if the mainstream theories of truth in fiction that I will be discussing are correct, then the mainstream modal logics do not hold within them. This is a rather surprising result, and one that calls for a response" [4, p. 13544].

In the present paper, I will show how Kosterec's results fail because the "model-theoretic proofs" he gives involve an illicit move that prevents his proofs from being sound, which means that the unintuitive consequence of these theses being falsifiable on Kripke models is no longer in tension with the purported "proofs" of the theses. As a point on its own, merely showing that the proofs in one single particular paper contain an error is not especially significant, and writing something purely to point out an error borders on churlish. However, there are two important morals that can be drawn from understanding exactly

*Thanks are due to my PHIL3201 students in 2021–22, as it was for them that I first read and presented Kosterec's paper, and they were the first audience for my working through my uncertainties with his "proofs".

how Kosterec's error arises, points of more broader interest about the interaction of intensional contexts and equivalences or identities.

The plan of the paper is this: First, I summarize the problem Kosterec is trying to solve, and present his purported "proofs" and countermodels of five theses (§2). Then, I discuss some of the gaps in his approach, and show how filling them in makes clear the illicit move he makes (§3). Finally, I discuss the consequences identifying this illicit move has on our broader understanding of reasoning within an intensional or opaque context (§4).

2

The focus of Kosterec's investigation is "the overlooked area of the truth or falsity of modal sentences within fiction" [4, p. 13544]. He picks a traditional possible-worlds approach to the analysis of fiction, adopting a Lewisian-style fiction operator F. I say here "Lewisian" because Kosterec in fact considers a family of approaches in the vein of Lewis's original [5], all of which "ultimately present some sentence as true if it is true in *all possible worlds from some subset of the set of all possible worlds*" [4, p. 13548, emphasis in the original]. Remember this italicized bit: It will become important later on.

His strategy is to investigate modal fictional sentences by assuming the validity of certain normal modal logics (NMLs) within the scope of the F operator. The systems he considers are K, T (which he calls M), S4, S5, and K4B (which he calls B) [4, p. 13549]; we assume familiarity with these logics, their constitutive axioms, and basic Kripke semantics. Kosterec's investigation comprises two parts: First, he argues for the "model-theoretic validity" of an initial thesis combining the F operator and the standard modal operators:

$$Fp \to F(p \leftrightarrow \Diamond p). \tag{C}$$

Next, he argues that the following theses follow, again, model-theoretically from C in conjunction with the adoption of certain modal axioms:

Under the assumption that T **is valid**[1]

$$Fp \to F(\neg p \leftrightarrow \Box\neg p) \tag{C1}$$

$$(F\neg p \wedge F\Diamond p) \to F(\Diamond p \leftrightarrow \Diamond\Diamond p \leftrightarrow \Diamond\Box p) \tag{C2}$$

$$F(\Diamond p) \to F(\Box(\Diamond p \to p) \to (\Box\Diamond\Diamond p \to \Box p)) \tag{C3}$$

Under the assumption that S4 **is valid**

$$Fp \to F(\Box\Diamond p \to \Box\Box p) \tag{C4}$$

$$F\Diamond p \to F(\Diamond p \leftrightarrow \Diamond\Diamond p \leftrightarrow \Box\Diamond p) \tag{C5}$$

Under the assumption that S5 **is valid**

[1] Note that this assumption is necessary in order for C itself to be valid, as we see in §3.2.

$$Fp \to F(\Box(p \to \Diamond p) \to (\Box\Diamond p \to \Box p)) \tag{C6}$$

Finally, he gives model-theoretic counterexamples for each of these theses that he has purportedly shown follow from C. From these two conflicting results, he concludes that the embedding of the various NMLs considered within the scope of the F operator is unsound.

3

Our conclusion is different: We will argue that Kosterec's counterexamples are in fact counterexamples, but this does not mean that there is anything problematic about embedding NMLs within the scope of the F operator. Instead, the problem lies in his purported "proofs" of the validity of various theses. Each of them relies on a crucial step which is not sound (and is *not* warranted by the inclusion of NML within the scope of F); as a result, it is unsurprising that these "proofs" purport to prove theses for which counterexamples can be found. You use bad proof methods, you get bad results.

Determining that the fault lies in the model-theoretic "proofs" is tricky because Kosterec does not provide any explicit semantics for the F operator, instead taking them to be intuitive. In order to demonstrate that his model-theoretic "proofs" are not proofs, we will make explicit a number of these things that he has left un- or underspecified.

3.1 Semantics for F

The closest that Kosterec comes to giving the truth conditions for F is this:

> the sentence p is considered true in a given fiction (here $F(p) = T$) if and only if it is true *in all possible worlds from some specific subset S of the set of possible worlds* [4, p. 13550].[2]

Let's unpack this.

The models are standard Kripke models, that is, triples $\langle W, R, V\rangle$, where W is a set of possible worlds, R a binary relation on W, and V a valuation function that says which atoms are true at which worlds. To accommodate the F operator, we must extend the models with S, the specific subset against which F is evaluated.[3] The truth condition for F (on a given model $\mathfrak{M}$) can then be specific as followed:

$$\mathfrak{M}, w \vDash Fp \quad \text{iff} \quad \text{for every } w' \text{ in } S, \mathfrak{M}, w' \vDash p$$

(Note that this makes F a *universal operator* with respect to the subset S; the accessibility relation R does not come into play).

Kosterec says that, on these models, he can prove from C "several non-intuitive consequences that arise *if we assume that*

[2] We use the notation Fp instead of Kosterec's $F(p)$ to improve readability. Note that neither $F(p) = T$ nor $Fp = T$ are well-formed logical formulas, as they conflate object- and meta-language notions. This conflation is symptomatic of the underlying confusion surrounding Kosterec's use of F.

[3] It is possible to discuss more than one fiction in a single model, in which case we would have a family of F operators each of which have their own associated S. This does not come into play in the present discussion.

1. *the investigated fiction is non-contradictory*
2. *the normal modal logic holds in the fiction*
3. *there are at least some fictional worlds*" [4, p. 13551, emphasis in the original].

That is, our models and other modal operators are the usual ones, and S is non-empty. In what follows, we adopt the same assumptions.

3.2 A proof that C *does* hold

Having specified the class of models in question more precisely, it is straightforward to show that C *is* valid on these models, because we are in fact able to show something stronger, namely that the following is a theorem of T:[4]

$$\Box p \rightarrow \Box(p \leftrightarrow \Diamond p)$$

Proof.

1	$p \rightarrow \Diamond p$	Theorem of T
2	$p \rightarrow (p \rightarrow \Diamond p)$	Propositional Logic, 1
3	$p \rightarrow (\Diamond p \rightarrow p)$	tautology
4	$p \rightarrow (p \leftrightarrow \Diamond p)$	Propositional Logic, 2, 3
5	$\Box p \rightarrow \Box(p \leftrightarrow \Diamond p)$	follows from the Derived Rule "From $\vdash \varphi \rightarrow \psi$ infer $\vdash \Box\varphi \rightarrow \Box\psi$."

□

Since this is a theorem of T, it will hold in any reflexive model, hence in any subset of any reflexive model, and hence it will remain valid when $\Box$ is replaced with F.

3.3 Where the "proofs" go wrong

Given this, the problem is not with C, but with how Kosterec purports to draw consequences out of it. His method is to "depend on the axioms provided for systems of normal modal logics [and] then extend the systems with the validity entailed in the previous subsection [C]" [4, p. 13552]. The way he does this is to assume what is within the scope of F in the antecedents of each of the theses, and then purportedly show that what is within the scope of F in the consequents also holds. We give his "proof" of C2 as an illustration of the method (why I've chosen this as the example will be clear below):

$$(F\neg p \wedge F\Diamond p) \rightarrow F(\Diamond p \leftrightarrow \Diamond\Diamond p \leftrightarrow \Diamond\Box p)$$

[4]Note that this proof demonstrates that Kosterec errs in his footnote 22 when he says that "Of course, theorem C does not hold within the investigated systems of NML"; any system that includes T will be able to prove the modal analogue of C, and hence it will hold in any subset S of any model, meaning that C itself is true.

1	$\neg p$	assumption
2	$\Diamond p$	assumption
3	$p \leftrightarrow \Box p$	Double Negation and C1, 1
4	$\Diamond p \leftrightarrow \Diamond\Diamond p$	C, 2
5	$\Diamond p \leftrightarrow \Diamond p$	tautology
6	$\Diamond p \leftrightarrow \Diamond\Box p$	Substitution of Equivalents, 3 and 5
7	$\Diamond\Diamond p \leftrightarrow \Diamond\Box p$	Transitivity of Equivalence, 4 and 6
8	$\Diamond p \leftrightarrow \Diamond\Diamond p \leftrightarrow \Diamond\Box p$	Propositional Logic, 4 and 7

Note the absence of the F operator from anywhere in this proof. Kosterec omits it because he is assuming that he is "within the operator F", i.e., working solely within worlds in the set S. As it turns out, this is a problematic assumption, and it is this assumption (more precisely, it is steps in the proof that rely on this assumption) where the problem with his proofs lies.

We can see this by recasting the proof with explicit reference to F and the truth conditions defined for it in §3.1, and spelling out steps that are compressed in Kosterec's "proof". We will also build a model alongside the proof, to make explicit what is true where, and which worlds are in S and which worlds are not.

1	$F\neg p$	assumption
2	$F\Diamond p$	assumption
3	$F\neg p \rightarrow F(\neg\neg p \leftrightarrow \Box\neg\neg p)$	Uniform Substitution, C1
4	$F(\neg\neg p \leftrightarrow \Box\neg\neg p)$	modus ponens, 3 and 4

At this point, we should pause to comment on what, exactly, it is that we are assuming in lines 1 and 2. This assumption is *the assumption of truth at some world w in the set S*; if we can show for some arbitrary world $w \in S$ that if the antecedent of C2 holds, then the consequent holds, then we can conclude that C2 holds in all worlds in S. So this means we have, so far, the model in Figure 1:

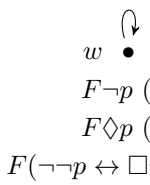

Figure 1: Steps 1–4

Because w is in S, it follows from the truth condition for F that if $w \vDash F\varphi$, then $w \vDash \varphi$, so we can strip F off all the lines:

5	$\neg\neg p \leftrightarrow \Box\neg\neg p$	def. of F, 4
6	$p \leftrightarrow \Box p$	Double Negation, 5

Continuing to build the model, this means we have the model in Figure 2:
Now we turn to our other initial assumption, and argue similarly:

w •

$F\neg p$ (1)

$F\Diamond p$ (2)

$F(\neg\neg p \leftrightarrow \Box\neg\neg p)$ (4)

$\neg\neg p \leftrightarrow \Box\neg\neg p$ (5)

$p \leftrightarrow \Box p$ (6)

Figure 2: Steps 1–6

7	$F\Diamond p \leftrightarrow F(\Diamond\Diamond p \leftrightarrow \Diamond p)$	Uniform Substitution, C
8	$F(\Diamond\Diamond p \leftrightarrow \Diamond p)$	modus ponens, 2 and 7
9	$\Diamond\Diamond p \leftrightarrow \Diamond p$	def. of F, 8
10	$\Diamond p \leftrightarrow \Diamond p$	tautology
11	$\Diamond p \leftrightarrow \Diamond\Box p$	Substitution of Equivalents, 6 and 10
12	$\Diamond\Diamond p \leftrightarrow \Diamond\Box p$	Transitivity of Equivalents, 9 and 11

Every step here is either involves a propositional tautology, is from the definition of F, or is acceptable modal reasoning (e.g., Uniform Substitution of Substitution of Equivalents). So where's the problem? Let us return to the model, now depicted in Figure 3. Because formulas prefaced by F are true *everywhere* in S, they are true at w as well. Now, the modal operator in $\Diamond p$ is an ordinary modal operator; there is nothing special about it, nor does it change how it functions because it is (here) in a fictional context. This means that for $\Diamond p$ to be true at w, there must be a world w' accessible to w where p is true. Nothing prevents us from stipulating the following two facts: (1) that w is accessible from w' (i.e., the relationship is symmetric), and (2) that w' is *not* in S.

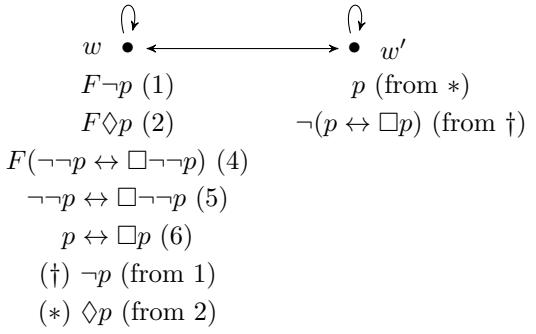

Figure 3: Expanded model

But now we run into trouble with (11) $\Diamond p \leftrightarrow \Diamond\Box p$; for $\Diamond p$ is true at w $(*)$, but at no world in the model is $\Diamond\Box p$ true, because every world can see a world where p is false, so the right-hand side of the "equivalence" fails. This means something has to have gone wrong in the "proof" of (11), and precisely what

is wrong with it is made explicit by the above model: the "equivalence" at line (6) is not a genuine equivalence; it is true only at w and not at w'. While the equivalence will be true *everywhere in* S, this is not sufficient for us to be able to apply the Substitution of Equivalents rule, because the equivalence holds only in the fictional context, and not generally.

If Kosterec had taken his model-theoretic counterexamples more seriously, he would have gone back to his model-theoretic "proofs" to examine them more closely. The fact that the equivalence between p and $\Box p$ holds only within S was masked by his dropping any reference to F in his model-theoretic "proofs", which meant that it was easy to "forget" about the fact that there may be worlds outside of the fictional context that are nevertheless relevant for evaluating formulas within the fictional context.

Similar explanations can be given of the other purported proofs; each involves an illicit substitution of things which are not in fact equivalent across the model as whole, but only within the fictional context.

4

Kosterec comes *extremely* close to recognizing the problem in his proof, but falls just short. He admits that there is something fishy about his move (6) in the "proof" of C2, saying:

> The substitution in step 6 needs some commentary. If we assume $\neg p$ and $\Diamond p$ is *true in the fiction*, then by theorem C we can prove that the equivalence $\Diamond p \leftrightarrow \Diamond\Diamond p$ is *also true in the fiction*; i.e., it holds in every world that is relevant to establishing truth in the fiction. Therefore, *as far as the fiction is concerned*, the equivalency is not just material but logical—there is no world that is *relevant to establishing the truth in the fiction* that invalidates the implication [4, p. 13553, emphasis in the original].

He is correct about what the fictional worlds think about the equivalence: Within the fictional context, p will be "logically" equivalent to its possibility. The problem is that the fictional worlds may *think* this equivalence is a logical truth, but they are mistaken.

There's two interesting morals to be drawn from this, about intensional (or "referentially opaque" to use Quine's term [6, p. 142]) contexts. It's a well-known feature of intensional contexts that equivalences or identities from outside the contexts cannot be brought inside them; in fact, we can take "intensional context" to simply mean any sentential context which is not *extensional*, that is, where coreferential or coextensive terms or semantically equivalent sentences cannot be substituted *salve veritate*. Traditional examples of intensional contexts are things like belief and knowledge; and literature on the problems that intensional contexts give rise to go back at least to Frege's classic Hesperus/Phosphorus example [1] (translated into English in [2]).

What the foregoing illustrates is the reverse phenomenon, which has received far less attention: equivalences within an intensional context cannot be taken out of that context. That is our first moral, then: Just as we cannot take things which are identical or equivalent *outside* of an intensional contexts and, bringing them into the intensional context, assume that they still hold, we also cannot

take things which are identical or equivalent *inside* of an intensional context, and bringing them outside the intensional context, assume that they still hold. For what may be equivalent if one restricts one's attention to only a subset of the model may not be equivalent when the model as a whole is considered. This point is one that, to my knowledge, has not been made in the literature to date.

The second moral is that fictional contexts are a type of intensional or opaque context. This is a much less novel claim—it can be found in the original introduction by Lewis of the F operator [5], and it has been made by others before, e.g., by Jacquette [3][5]—but what we have here is further evidence for it, and an important reminder of the problems that arise when we forget or ignore this moral. Part of what makes fiction such an interesting phenomenon to study is that it *doesn't work like reality*; if fictional contexts did not provide us with something like opacity, it would be difficult to see how they could give us a relevant analysis *of fiction* as opposed to, say, counterfactual reality.

Keeping these two morals in mind, let us talk a broader look at what Kosterec is doing. He divides possible worlds models into fictional and non-fictional contexts, and uses fictional contexts to make sense of statements of fictional modality, e.g., "Possibly, Sherlock Holmes is a dentist" [4, p. 13544]. Since this is a statement within the Sherlock Holmes fiction, there are three options:

1. The world that is the witness to this possibility is in the fictional context in which Sherlock is a detective.

2. The witness world is not in the Sherlock-fiction, but is within another fiction.

3. The witness world is not in the Sherlock-fiction, and is not in any fiction at all.

Each of these raises its own particular issues.

If we take being a dentist to be incompatible with being a detective (as Kosterec appears to), then (1) is incompatible with the "truth-in-fiction" operator that Kosterec is considering. For if it is true in the Sherlock-fiction that Sherlock is a detective, then he is a detective in every world in the Sherlock-fiction, and hence there is no world in the Sherlock-fiction where he is a dentist. (If we take being a dentist and being a detective as being compatible, then there is no issue; but a similar issue will arise with other claims that are incompatible.)

Suppose (3) that the witness world is not in any fiction at all, that is, it is either the actual world or some other "real" world. First, it would be surprising indeed to find Sherlock Holmes in that world, or in any other non-fictional world, given that he is a fictional character. Second, even if we could find him (a possibility we might not tenet for Sherlock, but we might for characters in historical fiction books), its seems strange that real possibilities should have any bearing on fictional possibilities (beyond the barest way, by indicating that the possibilities are genuine possibilities). So either he's non-existent, in the real world, or he's irrelevant.

This leaves option (2) as the most palatable, and probably acceptable to many people. One complaint about this option is that the shift of context itself

[5]Though others, like Voltolini, explicitly contrast fictional contexts with intentional [sic] ones, though his method for distinguishing them (fictional contexts are those which have "at least one fictional parameter", i.e., something that involves pretense [7, p. 27]) leave open the possibility that fictional contexts are a *subset* of intentional contexts.

feels somewhat illicit; if you say "Possibly Sherlock Holmes is a dentist (and not a detective)," then you've changed fictions; you're not longer talking about the original Sherlock-fiction, and why should we care—if we are genuinely interested in possibilities of the original Sherlock-fiction—what happens in other fictions? These are, to some extent, questions that I imagine a reader may be asking themself; they are not necessarily questions *I* am asking. The issue of "how much can you change a fiction and have the resulting new fiction be relevant to the understanding of the old fiction" is at the heart of understanding how fanfiction functions, and so no quick answer should be given.

Of the three options, it is (3) that Kosterec opts for

> Straightforwardly, [*possibly p*] is true in the fiction if it holds in every world from the subset (call it S). But this can be so even if p Does not hold in any of these worlds. It is enough for p to hold in some possible world outside S, to which all the worlds from within S are connected via the accessibility relation [4, p. 13560].

And therein lies Kosterec's contradiction: If he accepts that worlds outside of the fiction are relevant to what is possible within the fiction, then he cannot substitute formulas that are equivalent only within the fiction and not outside it *salve veritate*. He can either stick within the fiction, and maintain the equivalences; or he can allow genuine possibilities that reach outside the fiction, but then he must give up the "equivalences" as being no longer equivalent.

References

[1] Gottlob Frege. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100, 1892.

[2] Gottlob Frege. Sense and reference. *Philosophical Review*, 57(3):209–230, 1948.

[3] Dale Jacquette. Intentional semantics and the logic of fiction. *British Journal of Aesthetics*, 29(2):168–176, 1989.

[4] Miloš Kosterec. On modality in fiction. *Synthese*, 199:13543–13567, 2021.

[5] David Lewis. Truth in fiction. *American Philosophical Quarterly*, 15(1):37–46, 1978.

[6] Willard van Orman Quine. *From a Logical Point of View*. Harvard University Press, 2nd, revised edition, 1980.

[7] Alberto Voltolini. Fiction as a base of interpretation contexts. *Synthese*, 153(1):23–47, 2006.