Fiction and Fictional Objects

Francesco Berto
Mark Jago
Christopher Badura

Abstract and Keywords

This chapter begins with the problem of what counts as true in a given fiction, beyond what's explicitly given in that fiction. It then considers the problem of inconsistent fictions, which are naturally handled using impossible worlds. An account of truth in fiction is presented, which develops one of Lewis's analyses into an approach which can handle inconsistent fictions with ease. The chapter then turns to the second main topic: how we should think about fictional entities. Realism and fictionalism about fictional characters are contrasted. A third option is then considered, which takes the Meinongian line that fictional characters are non-existent objects. Several versions of this idea and their various issues are discussed.

Keywords: truth in fiction, inconsistent fiction, fictional characters, realism, fictionalism, Meinongianism

11.1 Problems of Fiction

Our ability to tell fictional stories, to engage with them, to think about their characters, and to reason about the situations they find themselves in, is an important part of being human. Tales, novels, plays, and operas represent things as being such-and-so. Besides enjoying their fictional contents, we can learn a lot from them. But how can this be, given that, in general, they are no *true* representations of reality?

In this chapter, we will set out some philosophical issues of fiction and discuss to what extent impossible worlds help with them. We will tackle two problems in particular. These can be mapped to a prima facie intuitive distinction between two kinds of fiction-related discourse. Some features are taken as true of fictional characters like Heathcliff, Gandalf, and Sherlock Holmes, *within* the fictions in which they appear. It is true in the respective fictions that Sherlock Holmes is a detective and that Gandalf is a wizard. Call this *intra*-fictional discourse.

We also ascribe to those characters features they don't have in their respective fictions. We say that Heathcliff is a fictional character due to Emily Brontë and that Holmes is more famous than any real detective. These claims are not true in the stories: within the respective **(p.240)** fictions, Holmes avoids celebrity and Heathcliff is not fictional at all, but a very real, tormented hero. Call this *extra*-fictional discourse.

The distinction between intra-fictional and extra-fictional discourse may not be sharp in all cases (Pelletier 2003), but it seems robust and intuitive enough. Fictional characters enjoy a double life. In Kit Fine's words:

> On the one hand, they have certain properties within the contexts in which they appear; they love and hate, thrive and fail, and live their varied lives. On the other hand, they also relate to the real world; they are created by authors, read by readers, and compared, for better or worse, with one another and with what is real.

> (Fine 1982, 97)

We start by focusing on one side of this double life, the intra-fictional one. What does it mean that something is *true in a fiction*? Fictions are not, on the whole, true, nor do their authors generally aim for them to be. (Recall the proviso, 'any resemblance to real people or events is purely coincidental'.) But in intra-fictional discourse we can truthfully talk about what happens in a story. We can do so even when such truths are not explicitly stated in the fiction. When Heathcliff and Catherine meet for the final time, Heathcliff is dressed in the manner of an eighteenth-century country gentleman and not as a circus clown. That's an intra-fictional truth about the world of *Wuthering Heights*. Yet the text of that scene never explicitly says anything about how Heathcliff is dressed. So what is true within the fiction of *Wuthering Heights* must go beyond what's explicitly written in the text. The problem is to explain how this is so: how it is true in *Wuthering Heights* that Heathcliff was dressed in the manner of an eighteenth-century country gentleman, and not as a circus clown, at his final meeting with Catherine.

The second problem we will discuss is that of fictional entities such as Holmes, Gandalf, Catherine, and Heathcliff. As fictional characters, it seems that they exist in the fictions of *Sherlock Holmes, The Lord of the Rings*, and *Wuthering Heights*, and not in reality. If this means that there really are no such things as Holmes, Gandalf, Catherine, and Heathcliff, how can we make true claims about them **(p.241)** in extra-fictional discourse? How can we (let's assume, truthfully) say that Holmes is more famous than any real detective, that Emily Brontë created Heathcliff, and that Kate Bush's Heathcliff is the same Heathcliff as Emily Brontë's? How can it be true that Kate Bush and Emily Brontë are speaking about the same Heathcliff, if there is no such thing?

Let us start with the first issue: truth in fiction. We have a range of intuitions about what's true in a given fiction (Woodward 2011). The issue is how to systematize them.

11.2 Truth in Fiction
We commonly talk of the worlds of fiction ('in *Star Trek*'s world …', 'in the world of *The Lord of the Rings* …'). So we might understand what's true in a fiction along the lines of truth relative to a world. But in general, many worlds (including many possible worlds) will be compatible with the explicit text of *Wuthering Heights*. Which of these worlds is the world of *Wuthering Heights*?

It's not even clear that the world of the fiction must be compatible with everything stated in the text. Sophisticated 'unreliable narrators' explicitly state something that turns out later on not to be true in the fiction. The narrator may make ironic or tongue-in-cheek remarks. Further subtleties involve the author-narrator distinction and the positing of a fictional or implicit author (Currie 1990). But since our aim here isn't to attempt a full theory of truth in fiction, we're going to ignore these subtleties. We want to focus on the role impossible worlds may play in a good theory. We shall only take into account reliable and literal narration, in which a narrator's explicitly uttering that A is sufficient for A to be true in the fiction. This seems to be the default case.

(Heyd (2006) proposes to account for unreliable narration pragmatically, via Gricean maxims. Perhaps non-literal speech and genre conventions, such as characters speaking in verse (Walton 1990), or **(p.242)** mentions of '555' numbers in American television (Hanley 2004), should also be dealt with pragmatically.)

The set of sentences explicitly and literally uttered by the narrator gives the *explicit content* of a fiction. We thus endorse the following principle as a default rule:

> (EXPLICIT) If A occurs explicitly in the story f, then A is true in f (and so 'in f, A' is true).

The worlds complying with the explicit content of *Wuthering Heights* are all worlds in which Heathcliff is adopted by Mr Earnshaw, comes to own Thrushcross Grange and Wuthering Heights, marries Isabella, and so on. Although the explicit content of the novel narrows down the class of worlds to those compatible with it, we cannot get to a unique world we can call *the* world of *Wuthering Heights*. Fictions are *incomplete*. Given a fiction f, there are sentences A such that neither it nor its negation is true in f.

In an extremely influential paper, Lewis (1978) proposes that we associate a plurality of possible worlds with the fiction and then analyse truth in the fiction as whatever is true according to all those worlds. He considers three accounts of how these worlds are selected. In the first, truth in fiction is a matter of what's true at all those worlds 'where the fiction is told, but as known fact rather than fiction' (Lewis 1978, 40):

> (FICTION$_1$) 'In fiction f, A' is true iff A is true at every world where f is told as known fact rather than fiction. (Lewis 1978, 41)

This approach generates too few truths-in-fiction. To understand *Wuthering Heights*, we need to understand something of the social customs of late eighteenth-century England. Heathcliff leaves, becomes wealthy, and returns as a gentleman. To understand the importance of those events, we need to understand something of England's attitudes to class and gender relations in the eighteenth century. Emily Brontë took for granted that her contemporary readers **(p.243)** would easily grasp these, so didn't include any explicit facts about class and gender relations in her text. But FICTION$_1$ ignores these.

There will be worlds where *Wuthering Heights* is told as known fact but eighteenth-century English social relations are turned on their heads. So, according to FICTION$_1$, it won't come out true in *Wuthering Heights* that Heathcliff exerts social power over Catherine. And there are plenty more contingent actual-world facts which should appear in the world of *Wuthering Heights*, but which are ignored by FICTION$_1$: that there's gravity, that it obeys an inverse-square law, and so on.

FICTION$_1$ generates too few truths because it includes too many worlds. Perhaps we should consider just those worlds closest to our own which are compatible with the text. This approach ignores all those which differ radically from our own, in ways not required by *Wuthering Heights*'s being told as known fact. This is Lewis's second analysis:

> (FICTION$_2$) 'In fiction f, A' is true iff some A-world where f is told as known fact differs less from the actual world than any non-A-world where f is told as known fact. (Lewis 1978, 42)

This approach seems to include too much of the actual world in the world of the fiction (Currie 1990, Proudfoot 2006). In the closest worlds to ours compatible with *Wuthering Heights* being told as known fact, Jeremy Corbyn is Labour leader in 2017. So FICTION$_2$ treats the fact of Corbyn's leadership in 2017 as a truth of *Wuthering Heights*. That's rather surprising. *Wuthering Heights* is about Catherine and Heathcliff's story, not Corbyn's. The approach also fixes too much of our physical law. Since there are in fact no ghosts, FICTION$_2$ implies that any ghost story should be understood as a story in which people merely hallucinate or imagine ghosts.

FICTION$_2$ delivers strange consequences for non-realistic fiction. What kind of world is closest to our own, whilst realizing everything explicitly described in *The Lord of the Rings*? Perhaps one in which the goings on in Middle Earth take place in a bubble universe, causally isolated from our own reality, and with its own laws of nature. We **(p.244)** might picture a world like that as a perfect duplicate of our own reality, but with a Middle Earth-universe tacked on the side. For any such world, the smaller its Middle Earth bubble is in time and space, the more similar it is to our own reality. The closest such world is probably one in which the Middle Earth bubble vanishes from existence the moment the story closes. Yet it doesn't seem right to say that it's true in the fiction that Middle Earth annihilates an instant after Sam comes home, draws a deep breath and says, 'Well, I'm back'.

A compromise involves similarity, not with actual fact, but with the 'generally prevalent beliefs … of the author and his intended audience' (Lewis 1978, 44). (Lewis speaks of 'overt beliefs': these are what almost everyone believes, almost everyone believes that almost everyone believes, and so on.) We consider a world which realizes the overt beliefs of Emily Brontë and her readers. We then move to the closest worlds at which *Wuthering Heights* is told as known fact. Whatever is the case at those worlds is what's true in *Wuthering Heights*. This is Lewis's third analysis:

> (FICTION$_3$) 'In fiction *f, A*' is true iff for each collective belief world *w* of the community of origin of *f* some *A*-world where *f* is told as known fact differs less from *w*, than does any non-*A*-world where *f* is told as known fact. (Lewis 1978, 45)

A worry here is how we balance the author's beliefs with her audience's in arriving at a 'collective belief' world (Bonomi and Zucchi 2003). Take a novel written in Nazi Germany by a progressive author, opposing the regime. Is the collective belief world one of progressive anti-Nazi feeling? Or are the author's beliefs outweighed by her audience's predominantly Nazi ideology?

We find FICTION$_3$ the most promising of Lewis's proposals. In the next section, we'll refine it using impossible worlds, together with ideas from *belief revision* theory.

(p.245) 11.3 Hyperintensional Fictions

In each of the analyses we've looked at so far, truth-in-fiction relies on possible worlds where the story is told as known fact. But an inconsistent (or otherwise impossible) story can't be known, and so can't be told as known fact. So what should we say about truth in inconsistent fictions?

There are different sources of inconsistent fiction. One is narrative oversight, in which the inconsistency is accidental to the plot. Watson's war wound is in his shoulder in some stories; in others, it's in his knee. It's also a given that he has only one wound. We might take the fictional truth to be given by whatever's true in some maximally consistent fragment of the story, or we might take it to be whatever's true in all of them (Lewis 1978, 46). On the former, Watson's wound is in his shoulder, and Watson's wound is in his knee, but it's not both. Truth-in-fiction is non-adjunctive: *A* and *B* can each be true in the fiction without *A* ∧ *B* being true in the fiction. (This is a version of the subvaluational 'fragments of belief' approach from §8.2.) On the latter approach, it's true in the fiction that Watson's war wound is either in his shoulder or his knee, but neither disjunct is true in the fiction.

Not all inconsistent fictions are due to narrative oversight. Some are intentional; some blatantly so:

> Carefully, I broke the tape and removed the lid. The sunlight streamed through the window into the box, illuminating its contents, or lack of them. For some moments I could do nothing but gaze, mouth agape. At first, I thought that it must be a trick of the light, but more careful inspection certified that it was no illusion. The box was absolutely empty, but also had something in it. Fixed to its base was a small figurine, carved of wood, Chinese influence, Southeast Asian maybe.
>
> I put the lid back on the box and sat down hard on the armchair, my mental states in some disarray. I focused on the room. It appeared normal. My senses seemed to be functioning properly. I focused on myself. I appeared normal. No signs of incipient insanity. Maybe, I thought, it was some Asian conjuring trick. **(p.246)** Gently, I reopened the box and gazed inside. … The box was really empty and occupied at the same time. The sense of touch confirmed this.
>
> (Priest 1997b, 575–6)

This is the central passage of Graham Priest's short story, *Sylvan's Box*. The narrator is Priest himself (or, a fictional version of him). Qua author, he asks us: what's true in this fiction? The most straightforward reading takes the narrator's statements at face value. It's true, in the fiction of *Sylvan's Box*, that Priest discovers a box that is both empty and not empty. The obtaining of a contradiction is essential for understanding the story. But no operation on maximally consistent fragments of the story will deliver the result that, in *Sylvan's Box*, a box is both full and empty at the same time.

Hanley (2004) and Nolan (2007) challenge the view that the contradiction is essential to *Sylvan's Box*. Both opt for a reading under which Priest, as narrator, falsely believes that there is a simultaneously full-and-empty box. His actions are understandable, in light of his false beliefs. That's surely one possible interpretation and, as such, there's nothing wrong with it. One might read Priest's text that way if taken as an attempted statement of historical fact. But we don't see why we *must* understand the story in this way. Is an author powerless to make her characters perceive and judge veridically and her narrator speak accurately? Surely not. We think the 'Principle of Poetic Licence' (Hanley 2004, 121, attributed to Harry Deutsch) is plausible:

(PPL) For any A, one can write a story in which A is true.

Other impossible fictions suggest that the inconsistent (or otherwise impossible) reading must be available. Proudfoot (2006) discusses the comedy film *Last Action Hero*, in which the central plot device involves a teenage boy, Danny, who finds himself in the fictional world of his favourite film hero, Jack Slater. Slater's evil nemesis later escapes to the real world and Danny and Jack have to follow. Throughout the film, various human characters become fictional, and various fictional characters become human. We take it that this is metaphysically impossible: each real human is essentially human and **(p.247)** so necessarily non-fictional. But it doesn't make sense of the plot to claim that Danny is hallucinating. If the impossible reading is available here, it should be available also for *Sylvan's Box*.

Hanley (2004) challenges our assumption, EXPLICIT, that the utterances which constitute the story are all true in that story (§11.2). (We assumed this with the proviso that unreliable narration and non-literal speech should be treated pragmatically.) But whilst EXPLICIT implies that truth-in-*Sylvan's Box* is inconsistent, we don't need to rely on EXPLICIT to make that case. Even if EXPLICIT failed for each particular contradiction occurring explicitly in a story, this would not stop implicit contradictions being true in the story. There are inconsistent time travel stories in which no contradiction is presented explicitly, for example. We may need a good deal of reasoning to unpack the inconsistency. Rather, our argument above was based on intuitive judgements about what's

true in *Sylvan's Box* and in *Last Action Hero*, which are then systematized into a cohesive (but not consistent!) narrative.

Let's return to Priest's story and the morals he draws from it. It is not true in *Sylvan's Box* that Priest finds himself levitating, dressed in a tutu. Yet that would be classically entailed by a contradiction, such as the box's being empty and not empty at the same time. So what's true in this fiction isn't closed under classical entailment:

> in understanding the story one has to draw inferences – often non-monotonic ones – from what is explicitly presented, together with background information. ... Clearly, however, the deductive canons employed cannot be those of classical logic. ... The logic of the story must be paraconsistent.

> (Priest 1997b, 580)

If the logic of this fiction is paraconsistent, not supporting the inference from a contradiction to arbitrary conclusions, then the worlds of the fiction must include impossible worlds. But exactly which logic is the 'logic of *Sylvan's Box*'? It seems that *some* inferential principles must apply: *Sylvan's Box* is (explicitly) set near Canberra and so we may infer that it is set in Australia, although this is not stated explicitly. Assumed background knowledge (that Canberra is **(p.248)** in Australia) allows us to infer an implicit truth of the fiction from an explicit one. We infer this conclusion using the transitivity of 'located within': if x is located within y and y within z, then x is located in z. *Modus ponens* is required to make sense of that reasoning.

Other logical principles seem equally necessary to make sense of the story. In the story, the box is empty. It also has something in it. Those truths of the fiction are, strictly speaking, not explicit in the text. What's explicit is their conjunction (expressed with 'but'). Yet it seems clear that the conjuncts are also true in the fiction. Further exploration of the text may provide evidence that the standard introduction and elimination rules for conjunction, disjunction, and implication, plus double negation introduction and elimination, are all fine in *Sylvan's Box*. This might give reason to think that the logic of this fiction is a paraconsistent logic in which '→' is not the material conditional. One of the relevant logics of Chapter 6 may then fit the bill.

That may be the right logic for *Sylvan's Box*. But what, in general, is the right logic of the 'in fiction f' operators? It is this logic, if one exists, which deserves the label 'the logic of fiction'. We are not sure that one logic may be singled out as *the* logic of fiction. Recall the Principle of Poetic Licence (PPL) from above: for any A, we can always create a fiction in which A is true. So, for any candidate

'logic of fiction', it seems we can create a fiction which breaks the rules of that logic (Proudfoot (2018) makes a convincing case for this).

(Actually, PPL suggests, but does not imply, that conclusion. For that, we'd need a stronger principle: that for any $A_1, ..., A_n$ and C, we can create a fiction in which each A_i is true but C is not. This echoes the discussion of *Nolan's Principle* (NP), and the stronger (NP$^+$), from §8.4.)

11.4 A Formal Semantics

We'll now set out an account of truth in fiction, as developed by Badura and Berto (2018). It's based on Lewis's third analysis (FICTION$_3$ from §11.2), expanded to include impossible worlds and **(p.249)** considerations from cognitive science (Nichols and Stich 2003), which inspired our treatment of the logic of imagination in Chapter 7. The key idea is that fiction involves a kind of *belief revision*.

The semantics for doxastic and epistemic logics of belief revision involves worlds ordered (usually, totally) by some *plausibility* relation (Grove 1988, Segerberg 1995, 2001). We can think of the ordering of worlds in terms of spheres, nested around a core, as in the standard Lewis semantics for counterfactuals (Lewis 1973b). But here, the spheres do not model objective similarity. Instead, they model subjective plausibility, or degrees of belief entrenchment. The innermost sphere is the set of worlds that realize certain beliefs. The closer other worlds are to the core, the more plausible they are for the relevant agent(s). These worlds are more likely to be embraced as fallback belief-positions, after new information induces an agent to revise her beliefs.

Roughly, we take 'in fiction *f, A*' to be true when, having revised the relevant agents' beliefs with *f*'s explicit content, all the most plausible worlds are *A*-worlds. To achieve this, we think of revisions as *soft upgrades* (Baltag and Smets 2011, Van Benthem and Liu 2007), in which the fiction's explicit content leads to a reordering of the set of worlds. Our motivation for this comes from Currie (1990), Lewis (1978), Walton (1990), and others, who claim that, in engaging with fiction, we play a game of *pretence* or *make-believe*. (Matravers (2014) challenges that this is the crucial distinction between fiction and non-fiction. However, what matters for our purposes is just that it be one feature of fiction.)

Nichols and Stich (2003) give a cognitive model of pretence that resembles ideas from belief revision theory. Our cognitive architecture comprises a 'world box', accessed when we engage in pretence via the pretence premises, and a 'belief box', from which we pick background beliefs to integrate into our pretence. We then consider those worlds where the pretence premises obtain and that are most plausible in terms of our background beliefs, adjusted to make room for the

explicit pretence premise itself. The content of the pretence is whatever holds in all such worlds.

(p.250) In the preference-based belief revision semantics, agents order the set of worlds based on how strongly they take the various worlds as candidates for actuality. Let '$v \leq_w u$' stand for 'v is at least as plausible as u with respect to w' (for a given agent). In one kind of soft upgrade, new information that A reorders the worlds so that A-worlds become the minimal elements in the new ordering, $\leq_w^A$. Then, if A is true in v but not in u, $v \leq_w^A u$.

Think of an agent reading a fiction as being faced with new information, sequentially upgrading beliefs with each sentence the agent reads. One needn't actually believe what one reads, of course. Nichols and Stich (2003) suggest a mechanism for quarantining the update, by indexing worlds by their previous position in the ordering. When the pretence ends, one will recover the initial ordering from the index.

To make sense of truth in fiction, we have to consider the beliefs of a community of agents. For Lewis's FICTION$_3$, these are the overt beliefs of the community of origin of the fiction. We depart from Lewis in two ways. First, we consider the beliefs of the community of *interpretation*, rather than the community of origin. We want to model the interpreting community's reasoning on the fiction. Second, rather than overt beliefs, we take common ones: that is, the beliefs such that everyone has them, everyone believes that everyone has them, and so on. This is a merely practical choice: it replaces the 'most' quantifier over beliefs with 'all'. When we evaluate 'in fiction f, A', we look at all worlds (possible or not) that are minimal with respect to the common belief worlds of the community of interpretation after upgrading with the explicit content of f.

Now let's present the formal semantics. (For ease of exposition, we will not take into account Nichols and Stich's (2003) considerations about quarantining.) As before, we use a propositional language $\mathscr{L}$ built in the usual way using ¬, ∧, ∨, modalities □ and ◇, the strict conditional ⥽, and, for each fiction f, an 'in the fiction' operator 'In f'.

A *multi-agent plausibility model* for a finite set of agents Ag is a tuple $\mathscr{M} = \langle\!\langle W, N, \{\,_i\!\leq_w \mid i \in Ag, w \in W\}, \rho \rangle\!\rangle$, where W is a set of worlds, $N \subseteq W$ is the set of normal worlds, each $_i\!\leq_w$ is agent **(p.251)** i's plausibility ordering on W with respect to $w \in W$, and ρ is a valuation relation (see §5.4 and §7.2), relating formulas to 1, 0, both, or neither at the various worlds. We assume each $_i\!\leq_w$ to be transitive and well-founded, so that we can always determine the most plausible worlds, $\min(_i\!\leq_w, S) = \{v \in S \mid (\forall u \in S) v_i \leq_w u\}$, for any $S \subseteq W$. Well-foundedness implies reflexivity and total ordering, so that any two worlds are comparable with respect to each $_i\!\leq_w$.

Given individual orderings, it is a non-trivial task to come up with a group ordering $_G\leq_w$ for some group of agents $G \subseteq Ag$. Such a group ordering reflects whether the agents can agree for every set of worlds on some set of most plausible (or preferred) worlds. This amounts to a voting problem among infinitely many (or at least arbitrarily finitely many) alternatives. Our approach is therefore constrained by Arrow's (1950) impossibility theorem for social choice functions (and Leitgeb and Segerberg's (2007) analogue for belief revision).

These results show that, given certain conditions on the group of agents, a group preference ordering cannot be obtained. The reasoning assumes that any preference ordering is permissible. In our case, however, an upgrade requires that the worlds where the story is told as known fact be considered at least as plausible as any world where the story is not told as known fact. So Arrow's impossibility result does not straightforwardly apply in our case. We will assume a transitive and well-founded group ordering, based on the individual orderings.

Now for some conditions on the valuation relation ρ. For every $w \in N$ and every atom p we require that $\rho_w p1$ or $\rho_w p0$ and not both. (This is the Classicality Condition from §6.1. It extends to every formula of $\mathscr{L}$ by induction and guarantees that normal worlds are maximally consistent.) The semantic clauses for the extensional connectives and the modal operators are as in the semantics of §7.2. (In particular, at non-normal worlds ρ relates formulas to truth values directly and is not subject to the Classicality Condition.)

A is commonly believed at w by a group of agents when everyone in the group believes A, believes of everyone in the group that she believes A, and so on. World w is a common belief world of a group **(p.252)** G at world w_1 if everything that's commonly believed at w_1 is true at w. Let's denote w_1's set of common belief worlds by $\mathrm{CB}\,_G^{w_1}$.

Recall that a soft upgrade with A reorders the worlds so that all the A-worlds are then considered more plausible than all the worlds where A is not true. Formally, following Van Benthem and Liu (2007), the new ordering is defined by the following conditions:

(C1) For all $u, v \in W$, if $\rho_u A1$ and it is not the case that $\rho_v A1$, then $u_i \leq_w^A v$.

(C2) Otherwise, the old ordering remains.

Given we assume a version of (NP), we have some world in the model in which A is true and some world where it's not true. So the upgraded relation will be non-empty. The upgrade also preserves transitivity and well-foundedness. In our modelling, A is going to be part of the explicit content of the fiction.

We can add additional constraints on the upgraded relation: for instance, that (non-normal) worlds which make A true *and* false are less plausible than those making A only true. Or, if the agent follows some pragmatic rules of interpretation, certain worlds obeying those rules might be considered more plausible (for such a pragmatic-based approach, see Bonomi and Zucchi 2003). But for simplicity, let's just stick with the above conditions.

Next we define, for every $S \subseteq W$, the set of most plausible worlds for agent i after the upgrade:

$$\min({}_i \leq_w^A, S) = \{v \in S \mid \forall u \in S : v_i \leq_w^A u\}$$

For the corresponding definition for a group of agents G, we simply substitute 'G' for 'i'.

Finally, here's our definition of truth in fiction. Let F be the explicit content of fiction f. For $w \in N$:

(S$_1$ In f) ρ_w(In f A)1 iff, for every world $w_1 \in \min({}_G \leq_w^F, \text{CB}\,{}_G^w)$ and every $B \in F$: $\rho_{w_1} B1$ only if $\rho_{w_1} A1$; and

(S$_2$ In f) ρ_w(In f A)0 otherwise.

(p.253) Thus, A is true in fiction f iff every world considered by group G most plausible with respect to their common beliefs, revised in line with the explicit content F of the fiction, makes A true.

This view has some advantages with respect to Lewis's possible-worlds-only approach, for the most plausible worlds can be impossible. Engaging with *Sylvan's Box* might not require us to revise our ordering much, up to the point where the empty-and-not-empty box appears. Nolan and Hanley say that the most plausible worlds are those where Priest, as the narrator, has false beliefs (relative to the fiction). We take the most plausible worlds as those that comply with the obvious intended reading of *Sylvan's Box*. Priest has true beliefs, relative to the fiction, in which there genuinely is an empty-and-not-empty box. This does not trivialize the theory: the inference from In $f(A \wedge \neg A)$ to In f B does not go through.

So we can deal with blatant contradictions that are essential to the plot. What about the case of accidental contradictions? Our approach ensures the truth-in-the-fiction of each sentence of the explicit content, even if one contradicts another. But their conjunction need not be true-in-the-fiction. Since we have impossible worlds around, Conjunction Introduction might fail for this particular case. But it need not. It might be that the initial order prefers paraconsistent worlds over all other inconsistent words, in such a way that contradictory sentences generate their inconsistent conjunction, but without explosion. So

what a model says about accidental contradictions depends very much on its initial ordering.

This approach faces a couple of worries. The notion of a plausibility ordering is rather vague (just as similarity between worlds is). Our assumption that agents each come equipped with a plausibility ordering has to be viewed as something of an abstraction. Badura (2016) and Badura and Berto (2018) provide responses to this worry. Here, we merely note that the assumption is standard in contemporary epistemic-doxastic logics.

Another objection is that, on our analysis, truth-in-fiction depends too heavily on the community of interpreters. We allow what's true in a given fiction to change over time, as the communal beliefs of **(p.254)** its readers change. Lewis, by contrast, insists that 'what was true in a fiction when it was first told is true in it forevermore' (1978, 44). One might even object that our notion isn't a notion of *truth*, since it fails the governing norm that truth is stable over time (Wright 1992).

On our story, the explicit content of a fiction is treated in an objective, fact-of-the-matter way. One can always check what the author wrote. The worry concerns a fiction's implicit content. Do our interpretations of the fiction aim to discover objective facts about its implicit content, or do they somehow influence that content, as on the model we've presented?

If we go to the literary studies department, we'll always find differing interpretations, with differing claims about what's true in a given fiction. Plausibly, those theorists are each taking something from the fiction's explicit content and evaluating it from the standpoint of their own (shared) knowledge, background, and context. Theirs is a partly creative enterprise, constrained by communal beliefs and norms, as well as by the fiction's explicit content. Differing interpretations of a fiction need not be in strict competition, when given in different historical and social contexts.

Ours is a *contextualist* notion of truth in fiction. An analogy with contextualism about knowledge (DeRose 1992, 2002, Lewis 1996) is helpful. You're in a long queue to post a letter. 'They're open until 6', the person in front of you says, 'and the queue is always shorter then'. What matters to you is whether your informant really knows that the post office is open until 6. For contextualists, that depends on how high the stakes are *for you*, the ascriber of knowledge. If it's vital that you get the letter in the post today then, in your context, there are more possibilities of error floating around and so it's correspondingly harder for you truthfully to ascribe knowledge. It's even harder for the students in the epistemology class to ascribe knowledge to the person in the queue, for they might (for all they know) be brains in vats. In these ways, the truth of 'she knows' may vary with the context of the ascriber.

Our view of 'true in the fiction' is analogous. The truth of '… is true in the fiction' may vary with the context of the interpreter. Lewis (1978, 44) **(p.255)** objects that 'that would mean that what is true in a fiction is constantly changing'. But not so. What changes is the property picked out by 'true in the fiction'. It's not that what the person in the queue knows changes with the importance of your letter (or with how much epistemology you've been reading). What changes is what property 'knows' picks out. Similarly, different communities of interpretation pick out a different property with 'true in the fiction'.

We think the analogy between truth in fiction ascriptions and knowledge ascriptions is more than superficial. 'Knows that *A*' means something like '*A* is true in all contextually unelimated scenarios' (Lewis 1996, 551). Similarly, '*A* is true in the fiction' roughly means '*A* is true in all the contextually most plausible scenarios'. In both cases, the relevant context is the context of the ascriber (or interpreter). Had the Lewis of 'Elusive Knowledge' (1996) revisited 'Truth In Fiction' (1978), perhaps he would have arrived at a theory similar in spirit to ours?

11.5 Realism and Fictionalism about Fictional Entities

Let's move on to the ontology of fictional objects. (We draw on Berto and Plebani 2015, chapter 13.) Heathcliff is a fictional character created by Emily Brontë. Sherlock Holmes is more famous than any real detective. This is *extra-fictional discourse*: seriously asserted, believed, and seemingly true, in which we seem to refer to fictional characters. We may also quantify over fictional characters, in ways which are difficult to paraphrase away (Van Inwagen 1977):

> (11.1) There are characters in some nineteenth-century novels who are presented with a greater wealth of physical detail than is any character in any eighteenth-century novel.

> (11.2) Some characters in novels are closely modelled on actual people, while others are wholly products of the literary imagination, **(p.256)** and it is usually impossible to tell which characters fall into which of these categories by textual analysis alone.

If these are literally true, there must be something in the world that makes them true. *Realist abstractionists* about fictional objects, including Peter Van Inwagen, Saul Kripke, Tatjana von Solodkoff, Amie Thomasson, and Nicholas Wolterstorff, take extra-fictional discourse at face value. According to them, fictional objects are real, existent, abstract entities. That they're abstract explains why we couldn't stumble upon Holmes or Heathcliff.

Some versions of realist abstractionism have it that fictional objects are abstract *productions* (Fontaine and Rahman 2014, Salmon 1998, Thomasson 1999, Voltolini 2006). They are artefacts, brought into reality by the creative activity of

authors as they write. Unlike other abstract objects like functions or sets, Heathcliff was created by Brontë and depends on her ontologically: had she not written *Wuthering Heights*, there would be no Heathcliff. Other versions of realist abstractionism differ on this. According to Wolterstorff (1961), Heathcliff exists before Brontë thinks about him: she selects, rather than producing, her characters. Van Inwagen (1977) is non-committal between the creationist and the non-creationist variant.

The creationist version faces issues (Yagisawa 2001). We have some idea of what creation involves for concrete things: what it means for a craftsman to create a chair, or for a mother to create her baby. But it is mysterious what happens when an abstract object is created. Creation seems to involve causation. How can this be, if created abstract objects are devoid of causal features?

One may worry that the creative process looks too fuzzy. Exactly when did Holmes begin to exist? When Conan Doyle set out to write a novel involving a detective? When he wrote down the first sentence of *A Study in Scarlet*? Or perhaps when he wrote the whole first novel, or the first two? Or does Holmes's reality involve collective intentionality? If so, how many readers are needed? Ten or ten thousand? One way to respond to these worries is to point out that creation of physical artefacts is also fuzzy: just when did the table **(p.257)** or the sculpture begin to exist? If there is similar vagueness in both cases, then it can't be a problem specifically for the creationist view of fictional objects.

A more serious worry is that creationists seem unable to get the modal profile of fictional entities right (Jago 2014a, §5.5). If Holmes is a detective, he is inessentially so: he could have been a chemist, or a violinist. He might have done all the things Watson did and vice versa. But if a fictional entity is created by an author's description of them, then it seems that that description will be essential to the character. In writing about a detective living at 221b Baker Street, Conan Doyle creates an entity whose identity is built from those properties. But then, Holmes will be essentially a detective living at 221b Baker Street. That's the wrong result. Of course, Conan Doyle could have written about a detective living elsewhere, or about a violinist. But that character would have been created from a different piece of writing, and so wouldn't have been Holmes.

The non-creationist version of the view faces issues too. Sainsbury (2010) raises the *selection problem*: how could Brontë select one specific abstract entity, rather than another, to be Heathcliffe? We'll discuss the issue in §11.6, where we explore Meinongian theories of fictional characters, which face the exact same issue.

Realist abstractionists of all kinds face a worry with negative existentials: sentences in which we deny the existence of something. Sensible adults deny that Holmes, Heathcliff, or the Big Bad Wolf exist. A father who reassures his

son that the Big Bad Wolf doesn't exist aims to speak the strict and literal truth. Negative existentials about fictional objects typically count as extra-fictional discourse, not as intra-fictional ascriptions: in *Wuthering Heights*, Heathcliff very much exists. That's why 'the Big Bad Wolf doesn't exist' needs a treatment along the same lines as 'Heathcliff is one of Brontë's fictional characters'. But realist abstractionists must break this intuitive uniformity: they say, seriously and literally, that Holmes, Heathcliff, and the Big Bad Wolf exist.

Van Inwagen (1977, 308) admits that what to do with 'Mr. Pickwick does not exist' is 'a very complicated question'. Perhaps **(p.258)** these non-existence claims should be read as implicitly restricted quantifications. We are not really claiming that there is no such thing as Heathcliff, on this view, but rather that Heathcliff is not to be found in the realm of *concreta* (Goodman 2004, Thomasson 1999).

We don't think that's plausible. Uses of 'there is' are often restricted, as in, 'there's no wine left'. But restriction is much harder with 'does not exist', which naturally takes the unrestricted reading (Walton 2003). Restricted readings still make sense when made explicit: 'there's no wine left in the house'. But 'Donald Trump does not exist in California' seems nonsense. So we should take 'Heathcliff does not exist' as an unrestricted claim, which realist abstractionists will have to deny.

In the light of all these issues, antirealism about fictional characters becomes attractive. Fictionalism about a discourse (§2.8) takes claims in the area as valuable and useful but not generally true, for the things seemingly talked about are really not there. Talk of fictional objects seems to be a paradigmatic case in point. According to Walton (1990), we play collective games of make-believe when engaging with fiction, involving prescriptions that such-and-such must be pretended.

But as we've already seen, we can't analyse all talk of fictional characters using the fictionalist's 'in the fiction' operator (Van Inwagen 2003). We can't say:

(11.3) In *Wuthering Heights*, Heathcliff is a fictional character due to Brontë.

(11.4) In *Wuthering Heights*, Heathcliff is played by Laurence Olivier in the classic 1939 movie.

The embedded sentences are external to Brontë's fiction, yet still seemingly about Heathcliff. The problem extends to quantification over characters, as in,

(11.5) If no character appears in every novel, then some character is modelled on another character (Van Inwagen 2003, 137).

 (p.259) How can we understand this literal truth, if there literally are no fictional characters?

One option for fictionalists about fiction goes as follows (Jago 2014a, §5.5). Names like 'Heathcliff' surely have a meaning (else we couldn't truthfully say, 'in the fiction, Catherine marries Heathcliff'). So they have a semantic value, to which we can refer and over which we can quantify. But that semantic value isn't a fictional person. 'Heathcliff' doesn't refer to its semantic value. It refers to nothing, since Heathcliff doesn't exist. Brontë's text associates various properties with that semantic value, which can then be accessed by the 'in *Wuthering Heights*' operator.

Extra-fictional discourse is then understood by paraphrase, in terms of the semantic values of fictional names. So, (11.3) says something like: Brontë created the meaning associated with 'Heathcliff'; and (11.4) says something like: Olivier played a role based on the meaning Brontë associated with 'Heathcliff'. Van Inwagen's (11.5) says something like: if no semantic value of a name-in-fiction is shared by all novels, then some semantic value of a name-in-fiction is modelled on some other. Quite a mouthful. No wonder we talk, loosely, in terms of characters! But don't let this loose talk fool you, says the fictionalist. There are no fictional characters.

A third position on the ontology of fictional objects, differing from both realist abstractionism and fictionalism, is the Meinongian option. Let's take a look.

11.6 Non-existent Objects and Impossible Worlds
According to Meinongians (§2.3), fictional characters are non-existent objects. Meinongianism is a form of realism about fictional objects, in that such things are taken as being parts of reality, available for reference and quantification. But, for Meinongians, this is not the same as existing. Meinongians agree with the folk that 'Heathcliff **(p.260)** does not exist' is literally true. No convoluted story about negative existentials is needed.

Non-existent fictional entities straightforwardly accommodate extra-fictional ascriptions, such as 'Heathcliff is a purely fictional character'. But things get more complicated with intra-fictional discourse. To appreciate the issue, we need to say something about the *Comprehension Principle* problem. Meinongians need some 'principle of comprehension' for their objects, telling us which objects are admitted in the theory and what properties they can bear. A naïve version of Meinongianism subscribes to what Parsons (1980) calls the 'Unrestricted Comprehension Principle' for objects:

(UCP) For any condition $A[x]$ with x free, some object satisfies $A[x]$.

(It's unclear whether even Meinong ever endorsed something like this.) Here, '*A*[*x*]' is a condition such as '*x* is a detective, *x* lives at 221b Baker Street, *x* is a cocaine addict, *x* is Moriarty's arch-enemy, …'. (UCP) guarantees that some object is characterized by the condition: call it 'Sherlock Holmes'. Then *A*[Sherlock Holmes] is literally true. But this can't work, for one can prove anything whatsoever from (UCP) (Priest 2005, xix). Let *A*[*x*] be *x* = *x* ∧ *B*, with *B* an arbitrary formula. By the (UCP), something, *b*, is such that *b* = *b* ∧ *B*, from which *B* follows by Conjunction Elimination.

Nuclear Meinongians like Jacquette (1996), Parsons (1980), and Routley (1980) limit the principle to a restricted vocabulary. They make a distinction between two kinds of predicates, *nuclear* and *extranuclear* (with corresponding properties). Nuclear predicates deal with ordinary features of objects: 'is blue', 'is tall', 'kicked Socrates', 'was kicked by Socrates', 'kicked somebody', 'is golden', and 'is a mountain' (Parsons 1980, 22–3). Extranuclear predicates express logical, ontological, or intentional notions like 'is fictional', 'is possible', 'is thought about by Meinong', 'is consistent' (Parsons 1980, 22–3). Only nuclear predicates are allowed to deliver objects. It is essential that existence be extranuclear, so that one cannot stipulate things into existence by including '*x* exists' in a condition.

(p.261) Both naïve and nuclear Meinongianism entail *literalism* about intra-fictional discourse (Fine 1982). Fictional non-existents like Holmes literally have the properties they are characterized as having. Holmes literally is a detective and literally lives at 221b Baker Street. But this is problematic. Until 2002, 221 Baker Street (there's really no 221b) hosted the Abbey Road Building Society. It has never been the home of any detective. It is literally *false* that that 221 (or 221b) is (or was) Holmes's residence. In one of the stories, Holmes has tea with William Gladstone. But Gladstone never had tea with Holmes, or with any other non-existent entity.

Did Holmes live in a non-existent counterpart of 221(b) Baker Street? Did he have tea with a non-existent counterpart of Gladstone? That seems inconsistent with the data. We want to say that some things in fiction (like Napoleon in *War and Peace*) exist, just as much as we want to say that Holmes doesn't exist. There's no evidence that 'Napoleon' is ambiguous between an existent person and a non-existent fictional character.

Besides, literalism severs intuitive connections between properties. If Holmes really is a detective because he is so characterized (and the feature of being a detective is nuclear), why can't we shake hands with him? We fail to accommodate the insight that if something really is a detective, then it must be a concrete object, with a spatiotemporal address and causal powers. It should, in particular, *exist*. This is, however, denied to Holmes by Meinongians of all kinds.

Dual copula theorists like Zalta (1983, 1988) fare better than nuclear Meinongians. Recall from §2.3 that they propose a distinction between two ways in which things can be ascribed properties. There's ordinary predication, expressing property-instantiation or exemplification, and there's encoding. Encoding a property does not in general entail exemplifying it. Non-existent objects can encode features of any kind (except properties like *encoding F*: see Rapaport 1978). On this view, Holmes encodes, but does not exemplify, *being a detective*. So this view is free from literalism.

Priest (2005) offers a third Meinongian option, with a Qualified Comprehension Principle (see Berto 2012):

> **(p.262)** (QCP) For any condition $A[x]$, with x free, some object satisfies $A[x]$ *at some world*.

When object o is characterized as $A[x]$, $A[o]$ need not hold at the actual world (though it may). It holds at some world or other. Berto (2008) calls this view *modal Meinongianism*. There is no restriction at all on $A[x]$ in (QCP). And since 'satisfying' expresses ordinary property-instantiation, there is no need for Zalta's dual copula.

The world which realizes the characterized condition may be an impossible world. The modal Meinongian theory needs impossible worlds, just as accounts of imagination (Chapter 7), epistemic and doxastic contents (Chapter 10), and truth in fiction (§11.3) do. With impossible worlds in play, logically, metaphysically, and mathematically impossible contents can all be employed in the characterization of objects, via (QCP).

Modal Meinongianism puts a limit on what Meinong called the Principle of Independence of *Sosein* (the having of properties by objects) from *Sein* (their existential status). For modal Meinongians, some but not all properties are independent from existence. Those which involve the having of causal features, or spatiotemporal location all entail existence (at the actual world). Non-existents like Holmes may bear only those properties that fail to entail existence, such as *being self-identical*, or *being a fictional character*, or *being thought about by Mark on a Monday morning*. But they cannot bear existence-entailing properties, such as *being a detective*, or *being a cocaine addict*.

Crane (2013) proposes a similar framework. He also distinguishes between existence-entailing features, which he connects with Lewisian-Armstrongian natural properties, and non-existence-entailing, 'pleonastic' features. Non-existents like Holmes or Pegasus may instantiate features only of the latter kind and may be represented by intentional agents as instantiating features only of the former kind. Holmes is not a real detective. But he really is a fictional

detective and is so insofar as he is represented as a detective by Conan Doyle and his readers.

(p.263) Modal Meinongianism with (QCP) faces the following issue: what if we characterize something as being *actually* thus-and-so? If we characterize a character as being an *actual detective*, does (QCP) deliver an entity that's a detective at the actual world? Hopefully not, for we would soon fall into absurdity. We could characterize an actual detective just like Holmes, which conflicts with the facts about our world. Or worse still, we could characterize an actual round square, which cannot be part of any possible world.

How can (QCP) avoid this worrying implication? It seems difficult, given that 'actual', 'actually', and 'at the actual world' function by latching on to the actual world, @. Statements embedded within 'actually' always refer back to @ for their evaluation. As a consequence, all truths are necessarily actually true. Obama is necessarily actually the first president of colour (even though he isn't necessarily the first president of colour). 'Actually' is a *rigidifier* for descriptions. So, if 'x is actually F' is to be true at some world, then something must be F in the actual world. Now let our characterizing condition $A[x]$ be of the form: 'actually, $B[x]$'. Then (QCP) entails that something satisfies 'actually, $B[x]$' at some world and hence that something satisfies '$B[x]$' in the actual world. Since this works for any B, we soon run into absurdity.

Impossible worlds to the rescue! The argument above relies on the semantic clause for the 'actually' operator:

(SACT) 'Actually, A' is true at world w iff A is true at @.

Like all semantic clauses we've considered, this should hold of necessity: it applies to all possible worlds w. But it should not apply to impossible worlds. (SACT) tells us that it is impossible for 'actually, A' to be true and for 'A' to fail to be actually true. So there should be an impossible world where 'actually, A' is true, for any A which isn't actually true. So we can infer from 'actually, A' being true at w to A being actually true only when w is a possible world. But the world invoked by (QCP) need not be a possible world. The argument breaks down and the problematic conclusion is blocked (Berto 2012, Priest 2005).

(p.264) Many realist theories of fictional objects, including non-creationist realist abstractionism and all the Meinongian views we've discussed, face the *selection problem* (Sainsbury 2010), mentioned in §11.5. On all these views, fictional characters exist as abstract or nonexistent entities, independently of the author's intentions and actions. In writing her fiction, she *selects* some of these abstract or nonexistent entities to be her characters. She pins a name on them, but does not endow them with new properties: Holmes was a fictional

detective (even though he hadn't appeared in any fiction) before Conan Doyle came along.

How is this achieved? Not through some causal interaction, since we have no causal connection with abstract or non-existent entities. Can an author single out a specific fictional object using a definite description? The problem with this suggestion is that, if a Comprehension Principle delivers Heathcliff, it also delivers other entities extremely similar to Heathcliff. There'll be a whole host of abstract or non-existent entities, differing in (fictional) height by a few millimetres, or in their time of birth by a few seconds, but otherwise alike and as described by Brontë. Which of these is the real Heathcliff? How could Brontë pick just one of them out? Which one is the Heathcliff she refers to?

Perhaps Brontë's Heathcliff is an incomplete object, having no exact height (to the millimetre) and no precise time of birth (to the second)? Perhaps Brontë's Heathcliff has the (fictional) properties she explicitly attributes to him in *Wuthering Heights*, and no more? Heathcliff is (fictionally) 'tall, athletic, well-formed', but is he neither (fictionally) 6ft, nor (fictionally) taller, nor (fictionally) shorter? We do not want to say that it is true in the fiction that Heathcliff is a vague, fuzzy entity with no particular height, weight, time of birth, and so on. In the fiction, he's human, and it's a feature of all humans that they have a precise height at each time.

One bold reply to the selection problem, endorsed by Priest (2016b, chapter 11), has it that we select non-existents via acts of mental pointing. Brontë focuses on Heathcliff through her 'primitive intentionality', singling the character out via pure thought. But it's **(p.265)** hard to see how this can help with the problem of distinguishing between all the candidate Heathcliffs, each 'tall, athletic, well-formed' but differing slightly in their height. Presumably Brontë didn't specify, in her mind, Heathcliff's height to the millimetre. So it isn't clear how she could then intend one rather than another of the candidate Heathcliffs. It could be that 'mental pointing' selects one of the candidates by pointing at random. Then Heathcliff will have (fictional) properties, such as his precise height, which go beyond what we can discover.

An alternative Meinongian response to the selection problem, sketched by Fine (1982) and developed by Berto (2012, chapter 9) and Priest (2016b, chapter 14), drops the realist assumption. On this view, some non-existents are dependent on the mental activities of authors. Heathcliff is available for reference and predication thanks to Brontë's creative skills. Brontë didn't have to select Heathcliff from a pre-established domain. Rather, she somehow produced him and named her product 'Heathcliff'. We then use 'Heathcliff' with the intention of sticking to that original reference.

The view is similar to creationist abstractionism (§11.5). It faces the same problems and seems to add some new ones for good measure. If an author creates something, doesn't it follow that what she created exists? The Oxford English Dictionary has it that 'to create' means 'to bring into existence'. This is fine for realist abstractionists. For Meinongians, though, Heathcliff does not exist. Claiming that one can create non-existents seems to challenge the ordinary meaning of words.

Chapter Summary
We began with the problem of what counts as true in a given fiction, which goes beyond what's explicitly given in that fiction (§11.2). We then considered the problem of inconsistent fictions, which are naturally handled using impossible worlds (§11.3). We presented an account of truth in fiction, which develops one of Lewis's analyses **(p.266)** into an approach which can handle inconsistent fictions with ease (§11.4). We then turned to our second main topic: how we should think about fictional entities. We contrasted realism and fictionalism about fictional characters (§11.5). We found the problems for realism to be serious. We then considered a third option, which takes the Meinongian line that fictional characters are non-existent objects (§11.6). We considered several versions of this idea and their various issues.

Access brought to you by: