Truth in Fiction via Non-Standard Belief Revision

MSc Thesis *(Afstudeerscriptie)*

written by

Christopher Badura
(born 18th April 1993 in Hamburg, Germany)

under the supervision of **Prof. Dr. Francesco Berto**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

Master of Science

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
24th June 2016	Prof. Dr. Benedikt Löwe (chair)
	Prof. Dr. Francesco Berto (supervisor)
	Dr. Maria Aloni
	Dr. Luca Incurvati

Institute for Logic, Language and Computation

Abstract

Fiction operators such as 'In the fiction f,' (In_f,) have seen applications particularly in philosophy of fiction, but more broadly in any ontological/metaphysical debate. For example there are fiction operator approaches towards modality, mathematics and morality. Giving a suitable analysis for when a sentence of the form $\ulcorner In_f, \varphi \urcorner$ is true, is hence of importance. The most famous approach has been David Lewis's analysis. However, it has certain shortcomings, especially when applied to inconsistent fictions in which not everything is true. We start by taking Lewis's (1978) Analysis 2 and give it a formal interpretation that takes into account impossible worlds and ideas from belief revision theory. Our formal framework comprises multi-agent plausibility models with a domain of possible and impossible worlds, ordered by a group plausibility ordering. This gives rise to Grove-style sphere models which are known to be models for the AGM axioms. We extend these models to an impossible world setting.

Then, a sentence of the form $\ulcorner In_f, \varphi \urcorner$ is true under our interpretation of Analysis 2 iff. for any world that is, after revising with the explicit content of the fiction, at least as plausible as any common belief world and that makes the explicit content of the fiction true, it also makes φ true.

While we read a novel, we are insane-bonkers. We believe in the existence of people who aren't there, we hear their voices...
Sanity returns (in most cases) when the book is closed.

Ursula K. Le Guin

Von den vielen Welten, die der Mensch nicht von der Natur geschenkt bekam, sondern sich aus dem eigenen Geist erschaffen hat, ist die Welt der Bücher die größte.

Hermann Hesse

Acknowledgements

First and foremost, I thank Franscesco Berto for his great supervision by replying instantly to my e-mails, always having time to meet, many helpful and motivating discussions and critical remarks. I appreciate this a lot. Moreover, thanks to Benedikt Löwe for his supportive and encouraging advice as an academic mentor - many difficult decisions became a lot easier to make. Also, I thank Tanja Kassenaar for keeping me and everyone reminded about everything important at the ILLC and always having an open door for any kind of problems.

The time at the ILLC and the MoL would not have been possible without my time at the University of Hamburg. Thus, I'd like to thank Ali Behboud and Stefania Centrone for raising and nourishing my interest in logic. Also, I thank Christian Folde for his excellent class on philosophy of fiction. I am tremendously grateful to Nathan Wildman and Benjamin Schnieder for supporting my application to the Master of Logic and their ongoing support throughout my academic life so far.

Many thanks go to all my friends–and now, thanks to the MoL, I can say from all over the world–for everything among helpful discussions, keeping in contact with me wherever I was, giving recovering distraction from all the work, the fun in- and outside of class and for just being there. In particular, I'd like to thank Julian Schlöder for proofreading this thesis.

I am especially grateful to Medea for all her acceptance, tolerance and support during all the time and particularly in the final period of the thesis. It means a lot to me.

Last but not least, I thank my parents for all their support, financially and emotionally. It's impossible for me to find words to express how much your support means to me.

Contents

Chapter 1

Introduction

1.1 The Problem of Truth in Fiction

Is the sentence 'Sherlock Holmes lives in Baker Street 221b' true (does it express a true proposition)?[1] Most people would agree, or give the proviso 'well, he does, *in the fiction*'. The question becomes even more pressing if one accepts that Sherlock Holmes does not exist, since *prima facie*, if something does not exist, then it does not have, or, in philosophy-speech, instantiate, properties. And even if it exists, some might hold that it is not the kind of thing that can have properties such as living in Baker Street.[2]

In literary studies, many claims of this form are made: Romeo loves Juliet. They both live in Verona. Frodo is a Hobbit who saved Middle Earth. And so forth. All these sentences, as for example Lewis (1978) and Künne (2007) hold, are not literally true but *true in the fiction*. And even if Holmes did not exist, *in the fiction* he does and so, *in the fiction* he can have properties such as living in Baker Street. Thus, one prefixes sentences φ like the mentioned ones by an operator 'In the fiction f', which we represent by writing 'In_f', and then evaluates whether the resulting sentence of the form $\ulcorner In_f, \varphi \urcorner$ is true.[3] We call approaches using such a strategy 'fiction-operator-strategies'. The question then is: what are the truth conditions for a sentence of the

[1] For convenience, we will often say that a sentence is true, meaning that it expresses a true proposition.

[2] A famous proponent for the existence of fictional objects who has to deal with this very issue is van Inwagen (1977). In this work we will not take a stance on the existence question, since it won't matter for our purposes: within Lewis's modal realism, fictional objects exist and are (merely) possible objects, hence Lewis would be a realist. However, his fictional-operator strategy has mainly seen applications among anti-realists, such as Künne (2007).

[3] We add a comma between the operator and the sentence for convenience.

form $\ulcorner In_f, \varphi \urcorner$?

This question is not only important in the philosophy of (literary) fiction. Such fiction-operator-strategies have recently become very popular in various areas of ontology under the label *fictionalism*.[4] To avoid ontological commitment to entities of kind K, for example Lewis's possible worlds, one employs a fiction operator strategy: in the fiction of modal realism, there are possible worlds. We can also find fictionalism with respect to (wrt) mathematical entities, see for an overview Balaguer (2015), and moral fictionalism. Hence, the sentences quantifying over entities of kind K are not literally true, but always *in the fiction* of the corresponding theory and we hence speak as if, say, there were numbers. But there are numbers only *in the fiction of mathematics*. Hence, so says the fictionalist, these theories "can be good without being true". Thus, giving truth conditions for sentences of the form $\ulcorner In_f, \varphi \urcorner$ is not only important in the philosophy of fiction because an analysis of truth in fiction can have applications in various areas of philosophy.

To find the truth conditions for sentences of the form $\ulcorner In_f, \varphi \urcorner$, it is helpful to consider what reasons we sometimes have to say that a sentence φ is true in a fiction f. Some of the sentences mentioned we take to be true in the fiction because they occur *explicitly* in the fiction and we take the narrator to be reliable. In section 1.2 we will discuss this naïve approach in more detail.

Of course, sometimes, we have reason to assume that the narrator is unreliable. This might be because the narrator contradicts himself, or because he tells us that he is a notorious liar, or because he tells us something which is then contradicted by a different (reliable) narrator or character. In this work, however, we will not model unreliable narration and focus on reliable narration.[5]

There are also sentences we deem true in a fiction, which do not explicitly occur in the fiction. For example, it is true that Holmes lives in a European city and that he has a kidney. These seem to be true because we *import* certain *background* knowledge (or beliefs) into the world of the fiction.[6] Also, some *logical consequences* of fictional truths are true in the fiction. Let '*Canon*' denote the whole of Doyle's Holmes stories:

[4] Eklund (2015) provides an overview of fictionalism, justifications and objections. Fictionalism in the philosophy of fiction, is emphasized in Sainsbury (2010), where also some versions of fictionalism in other areas are discussed.

[5] Heyd (2006) and Heyd (2011) argue that unreliable narration can be modelled pragmatically by taking into account Gricean maxims, relevance theory, and politeness.

[6] Just to anticipate, it will turn out that there is no such thing as *the* world of the fiction. It is more like a universe of many more or less plausible worlds, where the fiction is not told as fiction but as known fact.

(P1) In_{Canon}, Holmes is a detective. (explicit)

(P2) In_{Canon}, if something is a detective, it is human. (import)

(C) In_{Canon}, Holmes is human (logical consequence).[7]

We will come back to the question whether truth in fiction is, or should be, closed under logical consequence, and if so, under which logic.

Clearly, some things are false in a fiction. For instance, it is false in *The Lord of The Rings* ($LOTR$) that Sauron wins the final battle. And this is because it is true in $LOTR$ that Sauron is defeated in the final battle.

Also, for most fictions, some sentences are neither true nor false in them. That means, for some sentence φ, in the fiction, neither φ is true, nor in the fiction, the negation of the sentence, $\neg\varphi$, is true. In that case we also say that f is incomplete wrt φ. For example, it is neither true in $LOTR$ that Frodo is left-handed, nor is it false in $LOTR$ that Frodo is left-handed (because it is also not true in $LOTR$ that Frodo is right-handed). So, usually, fictions are *incomplete*, since there are sentences φ, such that neither In_f, φ, nor $In_f, \neg\varphi$.[8] In our formal representation of the operator and the logical connectives that is $\neg In_f, \varphi \wedge \neg In_f, \neg\varphi$. So for some fiction f and sentence φ, we have $\neg(In_f, \varphi \vee In_f, \neg\varphi)$.

Two remarks concerning the incompleteness of fiction are in order. First, incompleteness of a fiction wrt φ does not mean that the fiction-operator violates the law of excluded middle wrt φ and f. This would mean that for some f and φ we had $\neg(In_f, \varphi \vee \neg In_f, \varphi)$. But as we have seen, if a fiction f is incomplete wrt φ, then $\neg In_f, \varphi$ and thus, given we have the $\vee$-introduction rule, we do also have $In_f, \varphi \vee \neg In_f, \varphi$ for the particular φ and f.

Second, if f is incomplete wrt φ this should not entail that *in* f, the law of excluded middle is violated, which would mean $In_f, \neg(\varphi \vee \neg\varphi)$. Although neither In_{LOTR}, Frodo is right-handed, nor In_{LOTR}, Frodo is left-handed, it seems intuitively true that In_{LOTR}, (Frodo is left-handed $\vee$ Frodo is right-handed). However, the fact that some fictions are incomplete also does not

[7] We assume 'Holmes is human' does not explicitly occur in *Canon* and also that *modus ponens* is a valid rule of inference *in the fiction*. The issue of determining the logic in the fiction is tied to the issue of finding a deductive system/a proof theory for the operator In_f.

[8] We say 'usually' because it is an open question, whether there are fictions in which everything is true, and hence whether there are complete fictions. Folde & Wildman (under review) argue for the existence of such fiction(s). Could there be more than one such fiction? Since everything will be true in any such fiction, it seems, they would all have the same content, and thus be identical. Either one has identity criteria between fictions that go beyond content, or one needs to introduce a more fine-grained notion of content to distinguish them.

exclude the possibility of fictions violating the law of excluded middle, i.e. it might be possible that there are fiction f and sentence φ, s.t. $In_f, \neg(\varphi \vee \neg\varphi)$.

An argument against the view that there are fictions in which classical logical laws fail, or which otherwise have an impossible content, will concern us in section 2.3.1, and it is clearly related to the question whether there are sentences which are true *and* false in a fiction. If this is so, and fictional truth is closed under, say, classical logical consequence, then in a fiction in which a sentence φ and its negation are true, everything will be true in the fiction by *ex falso quodlibet*. Even worse, any inconsistent fictions f and g will have exactly the same sentences true in them, and hence, it seems, they would have the same content, which is clearly counterintuitive. Since this is due to *ex falso quodlibet*, sometimes called *Explosion Principle*, any logic that allows for this principle, e.g. intuitionistic logic, will face this problem. Consequently, closing fiction in general under any explosive logic, will yield this problem. This hints again at the question of whether logical closure is desirable and if so, how it could/should be achieved.

After this overview, the truth conditions we are after ideally account for all the aforementioned cases:

1. Explicit truths
2. Import of background knowledge/belief
3. logical consequence
4. inconsistent fictions, without trivializing

Lewis (1978) provided an influential, widely discussed proposal for truth conditions for In_f. He gives two analyses for In_f, which we will both present in chapter 2, where we also discuss objections from Proudfoot (2006) and Hanley (2004) and conclude that the main flaw of both analyses is that they cannot account for the case of inconsistent fiction without trivializing.

We discuss Lewis's (1983) proposed solution and argue that it cannot account for those inconsistent fictions, where some contradiction is essential for the plot. After an argument for preferring Analysis 2, we start extending this analysis to inconsistent fictions in a non-trivial way. The goal is to give a formal framework in which we can interpret Analysis 2 as faithful as possible to Lewis's original idea. For this, we first present the well-known AGM axioms for belief revision, which are, on most standard proposals, taken to be desirable postulates for rational belief revision.

The extension of Lewis's Analysis 2 will be given by using plausibility models and Grove sphere models but with domains of impossible worlds. We

argue how this extension can deal with the objections put forward against the original Analysis 2. Finally, in chapter 4 we discuss some philosophical issues concerning our formal framework, give an outlook on further research and conclude.

In this thesis, we will assume that there are fictions in which contradictions are true and are essential for the understanding of the fiction, for example, *Sylvan's Box* by Graham Priest. We call such fictions 'blatantly inconsistent fictions'. Moreover, we assume that every fiction has an explicit content that can be determined by a competent reader. This will be explained below in more detail. We treat propositions as sets of (im)possible worlds where they are true and sentences to express propositions.

Moreover, the approach we are giving is based on the conception of logic as a theory of reasoning and information processing. Our approach is normative in that we are giving *part of* what Priest (2012) calls *logica utens*, that is (part of) the norms of the correct practice, and that is in our case the norms of the correct practice of literary studies. The approach is normative in that it models what a good reasoner about truth in fiction *ought* to do. It just provides part of the norms of the correct practice, since we are focussing on reliable narrative literary fiction.

Methodologically, we will start out with following the common practice in semantics and the philosophy of fiction to evaluate the definitions of truth in fiction against competent speakers' intuitions about truth. As we evaluate whether our semantic analysis of, say 'and' is adequate, we look at examples and evaluate whether the truth value we get according to the analysis captures our intuitive notion of 'and'. The procedure to evaluate an analysis of truth in fiction is similar. We look at examples and test whether the resulting truth value the analysis gives for a sentence In_f, φ is the one we would intuitively assign.

In the following, whenever we speak of 'logical consequence', we mean classical logical consequence for propositional logic if not indicated otherwise. Also inconsistency amounts to classical logical inconsistency, that is φ is inconsistent iff. $\neg\varphi$ is a theorem of classical propositional logic, where φ is a sentence of classical propositional logic. We will also say that φ is a contradiction if φ is inconsistent, although the former is semantic and the latter is syntactic. We will also abuse the technical terminology and count conceptual impossibilities, such as 'x is a married bachelor', and metaphysical impossibilities, such as 'x is red and green all over', as contradictions or say that these are inconsistent. we also assume that any of those inconsistencies can somehow be made explicit by a sentence of the form $\varphi \wedge \neg\varphi$.

For better legibility, when it comes to formulas, we will be rather loose with the use/mention distinction and often omit quotation marks or Quine quotes. We use Greek lower case letters as metavariables for formulas and also as variables for sentences. We use lower case Roman letters as variables for propositionally atomic sentences.

1.2 Two Naïve Definitions of Truth in Fiction

Let us first discuss two naïve approaches to truth in fiction, which we consider to be quite natural first guesses on how to define truth in fiction. Both will turn out to be inadequate as definitions but at least one of them provides a rule of thumb/guide to truth in fiction due to its intuitive appeal and its accuracy in the case of reliable narration. These proposed definitions are:

(Explicit) A sentence of the form $\ulcorner In_f, \varphi \urcorner$ is true iff. φ occurs explicitly in f.

(Intention) A sentence of the form $\ulcorner In_f, \varphi \urcorner$ is true iff. the author of f intended φ to be true in f.

Neither of these two is necessary nor sufficient for truth in fiction. For (Explicit) we can consider the Sherlock Holmes stories and the sentence 'In the Sherlock Holmes stories, Holmes is human'. The sentence 'Holmes is human', to the best of our knowledge, does not explicitly occur in the fiction. However, it is entailed by Holmes being a detective in the fiction (compare ch. 1). Thus, there are intuitively true sentences of the form $\ulcorner Inf, \varphi \urcorner$, where φ does not occur explicitly in f. Hence (Explicit) does not provide a necessary condition.

For sufficiency, consider Nabokov's *The Eye*. In it, a sentence expressing that the protagonist Smurow is dead occurs explicitly. The narrator, Smurow, tells us this, but, according to him, as a ghost. However, as it turns out in the end (spoiler alert!), Smurow is not dead and is not a ghost. Despite having unreliable narrators, such as Smurow, we can also have ironic/sarcastic narrators. In that case, a sentence φ occurring explicitly might mean the opposite of its literal meaning and so $\neg\varphi$ would, intuitively, be true in the fiction. Moreover, independently of troublesome narrators, as Künne (2007) argues, a sentence can explicitly occur, for example, only in the antecedent of a conditional or be recited by some character. Hence, it might not be true in the fiction but occur explicitly in it. Hence, in general, (Explicit) does not

provide a sufficient condition.

For (Intention) consider, for example, the sentence 'In *LotR*, Frodo has an even (e) or an odd number (o) of hairs on his feet' (Hobbits have a lot of hair on their feet), $In_{LOTR}, (e \vee \neg e)$. This seems to be true. However, saying that Tolkien *intended* this to be true does not seem to be the case, since intention seems to be an act of consciousness and we find it at least doubtful that Tolkien performed this particular act for our example. Thus (Intention) does not give us a necessary condition for truth in fiction.

It also does not give us a sufficient condition; we just have to consider cases of really bad literature where, for example, an author intends her character to be charming, sexy and chivalric, but fails to do so and the character turns out to be an arrogant, self-centered sexual maniac.

(Intention) also faces the potential objection to be circular. For 'true in fiction' occurs on both sides of the analysis: how can the author intend something to be true in a fiction if she does not already possess the concept of truth in fiction? A short answer might be that, maybe, everyone already possesses the concept but philosophers and literary experts have so far failed to explicate the concept. However, it is worth pointing out that possessing a concept, and correctly applying it, might not require one to have a philosophical analysis of the concept at hand. Consider an analogical case: we are pretty good at determining cases of knowledge, that is we seem to possess the concept, without an agreed upon definition/analysis of the concept of knowledge. Hence, the circularity objection has to tell us more about the relation of intentionality, concept possession and defining/analysing concepts to be a convincing objection.

Although those two analyses are not suitable as definitions for truth in fiction, at least (Explicit) seems to be a reasonable rule of thumb to determine what is true in a fiction. We usually assume the narrator to be reliable and hence what the narrator is narrating ("explicitly saying") is true in the fiction. So we treat the narrator as telling us about facts (s)he has come to know about.

A potential objection against this is that sentences might not have their literal meaning within the fiction, as mentioned above: metaphors and irony are not rare in fiction. However, we assume that a reader of such a fiction is competent enough to deal with these cases, in the same way as we are able to usually deal with these phenomena in conversations: also in natural language we usually take to be true what conversational partners utter explicitly to us, unless we have reason to doubt that; for instance, because some of the Gricean maxims have been violated. Similarly, we take the narrator to tell

the truth within the fiction/about the world of the fiction, unless we have reason to doubt that, e.g. because (s)he violates some equivalent of Gricean maxims for reliable narrators.[9] A violation might be a contradiction in the narrator's story or if the narrator describes himself as a notorious liar from the beginning onwards. But the default attitude towards a narrative, fictional or non-fictional, is usually that we take everything the narrator tells us to be true.[10] Following this analogy, we also claim that it is not the purpose of a *formal semantic* analysis of truth in fiction to account for contextual violations of the Gricean maxims or their analogues, just as it is not the purpose of a formal semantics *in general* to account for Gricean pragmatic maxims and their contextual violations. Hence, it seems to us, that the cases of unreliable narration can be dealt within pragmatics. So (Explicit) is a principle we take as a guide to truth in fiction; it is often sufficient for truth in fiction, but clearly not necessary.

Concerning (Intention) we claim that at least in the cases relevant for us, as for example *Sylvan's Box*, it is clear that the author had certain intentions about what is true in the fiction and that not taking the intention into account *at all*s seems rather unjustified. Of course, there are cases where authors might claim too much authority about their work, as for example the big discussion about Dumbledore being gay in the *Harry Potter* fiction suggests, and there are critics about authorial intention such as Barthes (1967). Moreover, determining authorial intention is notoriously difficult, especially when the author is already dead. However, since we will mostly use (Explicit) as a guide, we could do without (Intention).

We do not claim that these two principles are the only guides to truth in fiction and also not that they are always even necessary. Nevertheless, we will assume that (Explicit), plus some considerations about pragmatics, suffice to determine the *explicit content* of the fiction. Hence, the explicit content of f is the content which is expressed explicitly, after the relevant pragmatic tweaks, by the sentences occurring in (inscriptions of) f. Note, that the pragmatic tweaks do not include logical inferences such as entailments from being a detective to being human.

[9] We cannot use the Gricean maxims alone here because narrators tend to give way more information than necessary to convey what they want to convey. They might describe a feast by giving all the details of food and clothes and so on instead of just stating that there was a feast.

[10] This clearly is particular for literary fiction which *has* a narrator. Since we are restricting this work to this particular case, this is not a problem.

Chapter 2

David Lewis on Truth in Fiction

In this chapter, we summarize Lewis's (1978) Analyses 1 and 2, discuss certain objections against them, present Lewis's attempts to solve these problems and argue why Analysis 2 is the better candidate-analysis. We first explain the rationale behind taking a possible worlds approach towards our operator In_f. It is due to the fact that it seems to be an intensional operator and possible world semantics has been shown fruitful for such operators.

2.1 How to approach the operator

A natural way to go about an analysis of truth in fiction is to treat the operator In_f as an intensional operator. We will call a sentential operator μ intensional if it is not extensional, where an operator μ is extensional if whenever sentences ψ and φ are materially equivalent, so are $\mu(\psi)$ and $\mu(\varphi)$.[1] Formally, where $\rightarrow$ is the material conditional and $\leftrightarrow$ is defined accordingly:

$$\forall\varphi\forall\psi((\varphi \leftrightarrow \psi) \leftrightarrow (\mu(\varphi) \leftrightarrow \mu(\psi))$$

For example, in classical logic, it is easy to see that negation is extensional. In modal logic, the belief operator is intensional. Although the $\rightarrow$ implication holds, the $\leftarrow$ implication fails.[2]

[1] Since 'In_f' is a one place sentential operator we characterise only one place intensional operators here.

[2] Some people think that already the $\rightarrow$ implication should not hold. It is a weaker form of the problem of logical omniscience: if ψ is a logical consequence of φ and an agent a believes φ, then the agent also believes ψ. It is weaker because the $\rightarrow$ implication tells us that only if two sentences are logically equivalent, then the agent believes the one iff. she believes the other.

An informal argument is to consider a situation of a coin flip with a fair coin. Let $\mu := B$ be a belief operator 'the agent believes that'.[3] Let $h :=$ 'The coin shows heads' and $t :=$ 'The coin shows tails'. An agent might be indifferent towards t and h and thus believe neither: $\neg B(h) \wedge \neg B(t)$. Then $B(h) \leftrightarrow B(t)$ is vacuously true. But it is not true that if the coin shows heads, then the coin shows tails (and vice versa): $\neg(h \to t)$ and $\neg(t \to h)$. Thus, $(B(t) \leftrightarrow B(h)) \not\to (t \leftrightarrow h)$.

Another common example of an intensional operator is 'It is necessary that' and there are many more. For example, many intentional operators (we saw belief already) are also intensional.[4]

Now, is In_f an intensional operator? Lewis (1978, p. 37) does believe so but does not give an argument for it. We also believe it is. Consider $p :=$ 'London is a European city' and $q :=$ 'There is a bank in 221b Baker Street'. Since both are true, $p \leftrightarrow q$ is true. But in the Doyle fictions, $\neg q$ and hence $\neg In_f, q$, unless the fiction is inconsistent, which we assume it is not. And hence, $In_f, p \leftrightarrow In_f, q$ fails.

Given the assumption that there are fictions whose logic is non-classical (that is the logic within the fiction is non classical) or that there are fictions in which everything is true, as considered by Folde (2011), it seems that In_f is even hyperintensional, that is that the principle of extensionality above fails for classically logically equivalent p and q:[5] Suppose φ and ψ are logically equivalent, for example $\neg p \vee p$ and $p \to p$. Then for a fiction f whose logic is intuitionistic, we would not, or at least should not, have $In_f, (\neg p \vee p) \leftrightarrow In_f, (p \to p)$.

We think there is enough evidence to treat In_f as an intensional operator. Moreover, it is reasonable that it is not a factive operator, i.e. In_f, φ does not, in general, entail φ, which is reasonable for fiction: just because something is true in a fiction, it is not true in the actual world. Of course for some

[3]This argument is informal since we won't specify the relevant Kripke models and hence also not the "kind" of belief involved, e.g. KD45 or some other. We just appeal to an intuitive notion here to elucidate intensionality.

[4]For necessity defined as the usual $\Box$-modality in relational semantics, it suffices to consider a Kripke model with two worlds w and v, s.t. Rwv but not Rww and let $V(p) = \{w, v\}$ and $V(q) = \{v\}$. Then $w \vDash \Box p \leftrightarrow \Box q$ because v is the only world accessible from w but $w \nvDash p \leftrightarrow q$ because $w \nvDash p \to q$ because $w \vDash p$ and $w \nvDash q$.

[5]Some might just simply deny that the logic of a certain fiction can be non-classical. One problem for this account would be a fiction that explicitly states that the logic of "its world" is, say, intuitionistic. As pointed out above, we grant that explicit occurrence of a sentence in a fiction is neither sufficient nor necessary for its truth. Nevertheless, a proponent of such a view has to tell a story why such an explicit occurrence of a statement about the fiction's logic is not sufficient for it being true in the fiction.

sentences In_f is factive, for example some parts of *Moby Dick*.

It has been fruitful to model intensional operators with possible world semantics and so it is, as said above, natural to define truth in fiction in terms of possible worlds, that is treat In_f as a (non-factive) modal operator. Thus it is reasonable to start the analysis by using possible world semantics to model truth in fiction. This is what Lewis (1978) famously did. As we pointed out above, In_f might even be hyperintensional. For hyperintensional operators, impossible world semantics has seen applications, such as by Jago (2014).[6] We will use a similar approach later.

2.2 Lewis's Analyses 1 and 2

Lewis's (1978) analysis of truth in fiction has been very influential and since it is formulated in terms of possible world semantics, it seems a good candidate for the intensional operator In_f. Other famous analyses of truth in fiction comprise, for example, Currie's (1990) narrator-based account. Walton (1990) is sceptical that a systematic account of truth in fiction can be given. However, a narrator-based account is too closely tied to literary fiction and it might be difficult to be extended to fiction without narrators.[7] We share Walton's scepticism to the degree that the systematic analysis Lewis gives is inherently *vague*, due to the notion of similarity between possible worlds, and will remain so also on our account. However, this is due to the nature of fiction, since, as we have seen, fictions usually are incomplete.

Lewis (1978) provides two analyses of which he remains indecisive which one is correct. In this section we present both analyses and discuss Lewis's own objections to them. We then go on to present and discuss objections to both analyses from Proudfoot (2006). Finally, we argue why Analysis 2 is the analysis we prefer.

2.2.1 Analysis 1

As we have seen, a possible world semantics for the fiction operator seems appropriate. But which possible worlds should we consider to determine what is true in some fiction f? Clearly, we cannot consider those, where f is true, since this would be circular. Lewis makes a subtle move to avoid circularity, which involves the different modes in which a fiction (or story)

[6] There are other approaches towards hyperintensionality, see Jespersen & Duží (2015) an the corresponding special issue *Synthese* for various positions.

[7] Although we are only focussing on literary fiction, we believe Lewis's account in our formulation can be extended to the relevant fictions.

is, or can be, told. A story can be told *as fiction* or can be told *as known fact*. Lewis (1978, p. 40) follows Searle (1975) that, in our world, the actual world, a fiction is told as *pretence*, that is, a fiction, in the actual world, is told *as* fiction:[8]

> Storytelling is pretence. The storyteller purports to be telling the truth about matters whereof he has knowledge. [...] This is most apparent when the fiction is told in the first person. Conan Doyle pretended to be a doctor named Watson, engaged in publishing truthful memoirs of events he himself had witnessed. But the case of third-person narrative is not essentially different. The author purports to be telling the truth about matters he has somehow come to know about, though how he has found out about them is left unsaid.

It is apparent from the quote, that by 'storyteller' Lewis means the author of the fiction. Moreover, storytelling is identical to, or is some form of, pretence. We thus encounter a tension in this quote which can be resolved by clearly distinguishing between author and narrator.

The tension is the following: Doyle, the author, pretended to be someone (Watson) who tells the truth about events he (Watson) has come to know about. However, storytellers pretend to be telling the truth about facts they have come to know about. So authors, the storytellers, pretend to be the ones telling the truth about something *they* have come to know *and* they pretend to be *someone*, namely the narrator (who is clearly different from the author in most cases) who is telling the truth about things *the narrator* has come to know about.

But it does not seem to be the case that Doyle pretends that *he* is telling us about things he has come to know about Holmes. Doyle pretends *to be someone*, namely Watson, the narrator, who tells us about things *he* (=Watson) has come to know about. Now, it seems *Watson*, the narrator, is the one telling us the story, whereas *Doyle* pretends to be someone who tells us a story.

Now, it is not entirely clear to us, whether Doyle actually pretended to be Watson or imagined himself as Watson or anything like that. It doesn't seem to be a necessary condition for successfully writing fiction that the author pretends to be one of her narrators or characters. An example, in which the

[8]This also avoids the problem that a story which is told as fiction comes out as accidentally true in the real world, which Lewis discusses as an objection to his previous Analysis 0, which we do not address here. That we indeed treat fiction this way, can be seen by the well known disclaimer 'All characters appearing in this work are fictitious. Any resemblance to real persons, living or dead, is purely coincidental'.

author does not pretend to be one of his characters, is *Herausgeberfiktion*, such as Umberto Eco's *The Name of the Rose*, where the author claims to have found the text and claims to publish it. Such an author does not seem to pretend that he is one of his characters.

But what seems to be the case in our previous example, and what Lewis (1978, p. 40) seems to have in mind, in distinguishing between a story being told as fiction and a story being told as known fact, is this:

> The act of storytelling occurs, just as it does here at our world; but there it is what here it falsely purports to be: truth-telling about matters whereof the teller has knowledge.

Doyle, in writing the fiction, performs an act of storytelling which is recognized by everyone as, following Walton (1990), invitation to a game of make-believe, that is pretending that someone like Watson and Holmes existed and pretend certain things to have happened. However, Watson, also performs an act of storytelling *in the fiction*, but he tells the story not as an invitation to a game of make-believe but as a story (more a report) about facts. Thus, there are two acts of storytelling happening, one which is performed in the real world with the intention to play a game of make-believe, and the other is performed in the fiction with the intention to "play a game of truth-telling". Clearly, these acts are very similar since they are expressed by literally identical expressions: what Doyle wrote down is literally identical to what Watson is saying in the fiction. However, they differ in intention. The worlds to consider are those, where we look at the act of storytelling performed by the narrator.[9]

Hence, we can figure out the possible worlds we have to consider in our analysis, as Lewis (1978, p. 40) suggests, the worlds to consider are 'the worlds where the fiction is told, but as known fact rather than fiction'.

Lewis makes an epistemological assumption here, which we will follow. Namely, that there is some way for us to determine the possible worlds where the story is told as known fact rather than told as a fiction. We assume that (Explicit) as a rule of thumb is what gives us those worlds and we will treat those worlds as the ones where the explicit content of the fiction obtains. We

[9]Another, maybe more Lewisean, way to avoid the aforementioned tension would be to consider the act A of storytelling at our world and then consider its counterpart act at those worlds where it is not the same as pretence, but told as known fact. Hence, we only consider one act A and then go to possible worlds where this act has different properties. Just as we would do it on Lewis's account with an assertion 'Hegel could have studied economics', where we would consider Hegel's counterparts in other possible worlds. But since this relies very much on Lewisean counterpart theory, we prefer the solution pointed out above.

will call those worlds 'f-worlds', where f is the relevant fiction.

Considering *only* those worlds, however, is not enough. There seem to be other reasonable restrictions on those worlds. For example, if we start out with a realistic fiction, it makes perfect sense to consider those f-worlds which are most similar to our world to account for the desideratum that beliefs and background knowledge can be imported. And even in the case of non-realistic fiction, like fantasy, we seem only to deviate as much as necessary from the real world to reach the f-worlds. For example, although $LOTR$ is a fantasy setting, Aragorn is a human and thus we accept him to be mortal. Moreover, we also accept that gravity works in that universe sufficiently similar to as it does in our world. Taking into account these ideas, Lewis presents us Analysis 1:

(Analysis 1) A sentence of the form 'In the fiction f, φ' is non-vacuously true iff. some world where f is told as known fact and φ is true intuitively differs less from our actual world, on balance, than does any world where f is told as known fact and φ is not true. It is vacuously true iff. there are no possible worlds where f is told as known fact.

This analysis has two major advantages. First, it allows import of background knowledge, since we only deviate as much as necessary from the actual world and thereby we keep fixed many of the facts about the actual world when considering the relevant f-worlds.

Second, we can deal with incompleteness of fiction. Take the case of $In_{LOTR}, (e \vee o)$. Since any $LOTR$-world where e is true differs equally from the actual world as any $LOTR$-world where $\neg e$, that is o, is true, neither In_{LOTR}, e, nor $In_{LOTR}o$ come out as true. However, it seems that any $LOTR$-world where $e \vee o$ is true differs less from the actual world than any $LOTR$-world where neither e nor o, since in such a world the law of excluded middle would fail, and, intuitively, $LOTR$ does not violate this law. Nevertheless, Analysis 1 has to face serious problems.

Problems for Analysis 1

The objection Lewis accepts as convincing against his Analysis 1 is one he attributes to Carl Gans:

> In "The Adventure of the Speckled Band" [ASB] Sherlock Holmes solves a murder mystery by showing that the victim has been killed by a Russell's viper that has climbed up a bell rope. What

> Holmes did not realize was that Russell's viper is not a constrictor. The snake is therefore incapable of concertina movement and could not have climbed the rope. Either the snake reached its victim some other way or the case remains open.

An unfortunate problem with this argument, pointed out by Folde (2011), is that the snake in *ASB* is not called 'Russell's viper' but 'swamp adder'. But it is possible to reconstruct this argument by noticing that in *ASB* it is claimed that swamp adders can hear whistling. We follow Folde (2011, p.74) in his reconstruction of the argument:

(P1) In_{ASB}, Holmes claims that swamp adders can hear whistling.

(P2) If Analysis 1 is true, then In_{ASB}, all to us [that is us in the actual world] known snakes are deaf.

(P3) In_{ASB}, the swamp adder is a to us [that is us in the actual world] known snake.

(P4) Analysis 1 is true.

(P5) $\neg In_{ASB}$, Holmes is mistaken

(C1) In_{ASB}, Holmes is mistaken

(C2) Analysis 1 is false.

(C1) materially follows from (P1)-(P4). To dismiss the argument, one could dismiss (P3). Although there is a swamp adder in the real world, in the fiction it can be claimed to have properties different from the ones it actually has, just as London has had no inhabitant called 'Sherlock Holmes'. Thus, the swamp adder referred to in the fiction is, strictly speaking, not a to us known snake because it has slightly different properties than the real one. Consequently, also the London in the fiction is, strictly speaking, not the London known to us because Holmes lives in it. However, both are sufficiently similar to the things we know to license certain imports. But since it is explicitly denied in the fiction that swamp adders cannot hear, i.e. it is asserted that they can hear, and we assume Holmes is always right, we accept this small deviation.

Another move is to deny (P5). In the end, also a Sherlock Holmes might make a mistake. Whether one choose to deny (P3) or (P5) seems to depend on which worlds one considers to differ less from the real world. Since 'swamp adder' occurs in *ASB* and also denotes a snake in the real world, one might

be inclined to accept that Holmes must have been mistaken. However, one might also argue that such a world is less similar to our world than any world where Holmes is right and swamp adders are just a little bit different. Hence, we agree with Folde (2011) that Lewis dismisses his Analysis 1 a bit too quickly based on Gans' argument.

However, Proudfoot (2006, p. 17) puts forward an objection we deem more convincing:

> Analysis 1 makes it true in the Sherlock Holmes stories, Greek comedy, and every other fictional work that Anglo-American philosophy takes a scientific turn in the late 20th century, since some f-world where Anglo-American philosophy takes a scientific turn in the late 20th century will be closer to the actual world, on balance, than any f-world, where it is false that Anglo-American philosophy takes a scientific turn in the late 20th century. This is highly counter-intuitive.

Put more explicitly, the argument goes like this:

(P1) In the actual world, Anglo-American philosophy (A-A philosophy) took a scientific turn in the late 20th century.

(P2) There are Sherlock Holmes Worlds (SH-worlds).

(P3) An SH-world in which A-A-philosophy took a scientific turn in the late 20th century is closer, on balance, than any SH-world where A-A-philosophy does not take a scientific turn in the late 20th century.

(P4) Analysis 1 is true

(C1) In_{SH}, (A-A-philosophy took a scientific turn in late 20th century)

(P5) (C1) is counter-intuitive, that is $\neg(C1)$.

(C2) Analysis 1 is false/counter-intuitive.

This generalizes and so for every fiction f, every f-world where everything which is true in the actual world is also true, should be considered as differing less from any f-world where there is a deviation from the actual world. Thus every actual truth is imported into every fiction, unless explicitly denied in the fiction.

Proudfoot (2006, fn. 16) makes a remark concerning a potential objection:

> It might be argued that the 'on balance' condition renders it unlikely that such a proposition is true in the Sherlock Holmes stories or Greek comedy. However, this points to a further difficulty for the theory, namely that of how such calculations are to be performed.

Lewis (1979) points to criteria of different priority for determining similarity between worlds (most important to less important) and hence gives some hint on 'how such calculations are to be performed':

1. Avoid big, widespread violations of laws of nature/logic.
2. Maximize spatial-temporal region between worlds throughout which perfect match of particular fact prevails.
3. Avoid small, simple violations of laws of nature/logic.
4. Secure approximate similarity of particular fact, even in matters that concern us greatly

Unfortunately for Lewis, these support (C1). The SH-worlds are going to have laws of nature very similar to ours, since we are dealing with realistic fiction.[10] But then, following all the four requirements, it seems, we need to accept that A-A philosophy takes a scientific turn in 20th century in the SH stories. If we denied it, there must have been more differences in the SH-worlds to our world than just the ones necessary to make the Holmes stories obtain as fact. This might even include changes in natural law (again given determinism) or at least 2. would be violated.

Thus, Proudfoot's claim that appeal to the 'on balance' condition wouldn't help seems convincing, at least if one accepts Lewis's criteria for the similarity relation, which Lewis most likely would.

The crucial issue in Proudfoot's argument to us is how to justify (P5). After all, the Holmes stories are more or less realistic fictions. So the question is why we should not allow for imported truths about the real future, compared to the time when the story was written or even the time the plot is supposed to be set in.

One way to justify (P5) might be appeal to (Intention). If (Intention) is necessary for truth in fiction, Doyle must have intended it to be true in the story that A-A philosophy takes a scientific turn in the late 20th century.

[10] We think Lewis would claim that laws in SH-worlds might diverge a tiny bit from the ones in our world, assuming determinism, as Lewis does, because at those worlds Holmes exists as a human being, whereas he does not in our world. Hence, there must be some difference in laws, given determinism and the fiction is realistic.

However, without going into details about intentionality, it seems safe to say that Doyle did not intend this.[11] However, as we have seen (Intention) is not, in general, necessary and it seems rather *ad hoc* as a justification for (P5).

Another reason why one might consider (C1) as counter-intuitive is by pointing to the incompleteness of fiction. It is not determinate whether in the fiction, the scientific turn of A-A philosophy occurs or not. The only clues we have from the story is that past and present in the fiction are very similar to past and present of the real world. But, assuming the future in the real world is undetermined, also the future in the fiction seems undetermined due to it being a realistic fiction. Even if we allow for the future to be determined in the real world and in the fiction, since the SH-worlds are already a little different because Holmes is a human being there, this does not exclude the possibility of the future in the story to develop differently from how it actually developed from the actual 19th century on.

Moreover, it seems reasonable to claim that much of the future in the fiction is undetermined because most people in the community of origin of the Holmes stories, not only Doyle, did not hold particular *beliefs* about the development of A-A philosophy in the 19th century. It was just no matter of concern at that point and so nothing about it should become part of the plot of the Holmes stories. What is supposed to be imported is just what Doyle believed at that time and what the common beliefs at that time were, where common beliefs are what everyone believes and what everyone beliefs that everyone believes it etc. Consequently, Lewis takes overt beliefs into account in Analysis 2, which are similar to common beliefs.

2.2.2 Analysis 2

Taking into account something similar to common beliefs of the community of origin, Lewis provides an alternative to Analysis 1:

(Analysis 2) A sentence of the form 'In the fiction f, φ' is non-vacuously true iff. whenever w is one of the collective belief worlds of the community of origin of f, then some world where f is told as known fact and φ is true differs less from the world w, on balance, than does any world where f is told as known fact and φ is not true. It is vacuously true iff. there are no possible worlds where f is told as known fact.

According to Lewis (1978, p. 44) the collective belief worlds are 'those worlds where the overt beliefs of the community – those beliefs which "more

[11] If this particular example is not convincing, consider the sentence 'in 2016 people send e-mails'. Again, this comes out true in the Holmes stories under Analysis 1.

or less everyone shares ... [and] more or less everyone thinks that more or less everyone else shares" – are true'.

Proudfoot's (2006) argument does not go through for Analysis 2. Any overt belief world will be indecisive about A-A philosophy in the 20th century and hence, there does not seem to be a SH-world where A-A philosophy takes a scientific turn in the 20th century, which is closer to any overt belief world than any SH-world where this does not hold. However, this analysis has to face other objections, of which we present two.

Problems for Analysis 2

The first objection is by Proudfoot (2006, p. 22):

> On Analysis 2 it is true in the Sherlock Holmes stories that $P \rightarrow P$. Although this may be unobjectionable, the same reasoning will apply for any substitution-instance of $P \rightarrow P$. Take the proposition: if Bush was elected for a second term, then Bush was elected for a second term. Counter to intuition, this comes out true in the Sherlock Holmes stories on Analysis 2. That Analysis 2 produces extraneous truths in fiction shows only that it, like Analysis 1, is counter-intuitive. *Since these extraneous propositions are irrelevant to the plot, no logical tension is set up.* (our emphasis)

Again, Proudfoot mobilizes our intuitions to claim that the import of a sentence, this time a logical truth, is unintuitive. Although she does not relate the irrelevancy as a reason why it is counter-intuitive that $p \rightarrow p$ is true in the fiction, one could consider this a potential reason.

Let's say that if $p \rightarrow p$ is not relevant for the plot of the fiction then it should not come out as true in the fiction. Given this intuition about relevance, it seems this argument exports to the real world. Assuming that classical logic is the logic of the actual world, $p \rightarrow p$ comes out as true in any substitution instance in the actual world as well. If we grant Proudfoot (2006) her intuition based on relevance, then no instance of $p \rightarrow p$ should come out as true in the actual world, since it is usually never relevant in any conversation either. Thus the criterion of relevance is not suitable, or one denies that $p \rightarrow p$ is true in the real world, which seems absurd.

Thus, how is the intuition motivated? If one says that the Bush-instance of $p \rightarrow p$ presupposes the existence of Bush, then again, this argument exports since any instance of $p \rightarrow p$ is true in the real world, also for p with empty names, which on the presupposition assumption should not be the case.

If the Bush-instance presupposes that Bush had been elected for a first term, again, the argument exports. $p \rightarrow p$ is true in the real world, even if we consider p to be 'in 2016 Trump is elected for a second term'. $p \rightarrow p$ will be true in the real world, although the presupposition that Trump had been elected for a first term is false. Thus, the appeal to intuition in this case does not seem motivated enough.

The second objection against Analysis 2 we consider is inspired by Bonomi & Zucchi (2003). Consider Klaus Mann's *Mephisto* in which one of the characters is Dora Martin who is jewish. Let's say the commmunity of origin is Nazi Germany.[12] Consider

s: Dora Martin is of inferior race.

According to Bonomi & Zucchi (2003), '$In_{Mephisto}, s$' is true because the overt beliefs in the community of Nazi Germany were that jews were of inferior race and hence any world where s is true would differ less from such an overt belief world than any world where s is not true. However, Mann certainly did not intend s to be true in the story.

We do consider this objection convincing. However, our diagnosis is that it only works because we take into account *overt* beliefs of the community of origin. As soon as we refer to *common* beliefs of the community–what everyone believes that everyone believes that everyone believes that everyone believes etc.–we can avoid this problem.

The remaining objection to Analysis 2 we are going to address is thus the one by Bonomi & Zucchi (2003). But there is a more general objection to Analysis 2 and that is due to its 'vacuously'-clause.

2.3 Blatantly Inconsistent Fictions

Lewis's analyses both suffer from the following problem: if a fiction is (logically) inconsistent it is (logically) impossible. Hence, there is no world where it is told as known fact. But since the analysis universally quantifies over all the worlds where the fiction is told as known fact, In_f, φ comes out as vacuously true for any φ. Hence, any inconsistent fiction makes any sentence true, which is counterintuitive. Lewis (1983) proposes two ways of dealing with this issue: the method of intersection and the method of union.

[12]It could have been if Mann had not made it to exile on time and still published the novel. Historically, however, this is not accurate because the novel was published from exile and hence its community of origin is actually (a subset of) the complement of Nazi Germany. Bonomi & Zucchi (2003) use a completely fictional example but we think it is good to have an actual example and just make this assumption about the community of origin.

If f is an inconsistent fiction, one first considers all the maximally consistent fragments of f.[13] On the method of intersection, to determine whether In_f, φ, one considers whether φ is true, according to the respective analysis, in *all* maximally consistent fragments of f. On the method of union, for In_f, φ to be true, it suffices that φ is true in *at least one* of the maximally consistent fragments of f.

Thus, depending on which method one chooses, In_f, φ is true iff. for some/for all consistent fragment(s) f' of f, $In_{f'}, \varphi$, where truth in consistent fragments is defined by Analysis 1/2.

Consider, for example, the case of Watson's war wound in the Holmes stories. There is no consistent fragment f', s.t. '$In_{f'}$, (Watson has exactly one war wound and it is on his shoulder *and* his leg' is true). However there are corresponding consistent fragments for each case in particular. Hence, it is true in some consistent fragment that Watson has a war wound on his shoulder and it is true in some consistent fragment that Watson's war wound is on his leg. Hence, on the method of union, it is true in the Holmes stories that Watson's war wound is on his leg and it is true in the Holmes stories that Watson's war wound is on his shoulder. On the method neither is true and thus the fiction is incomplete wrt Watson's war wound. However, on both methods, we have it is not the case in the Holmes stories that (Watson has exactly one war wound and that is on his shoulder and his leg).

The major objection is that there are (fragments of) fictions which are inconsistent but essential for the plot and thus any world where f is told as known fact must be one, where a contradiction is told as known fact. We will call those fictions 'blatantly inconsistent fictions'. In them, a contradiction is clearly true, either because it is part of the explicit content or because two contradictory sentences are true in it and also conjunction introduction is a valid rule in this fiction. What we also claim is that in most of these fictions (at least famous ones), not everything is true and that the corresponding fragments cannot be ignored, that is they are essential for the plot and the understanding of the fiction.

At this point, we will discuss a particular argument against the possibility of such fictions, put forward by Hanley (2004).

2.3.1 Hanley's Argument against Inconsistent Fictions

Hanley (2004, p. 120) is sceptical about the possibility to write fictions in

[13] A consistent fragment f is maximal if for any consistent fragment f' such that $f \subseteq f'$, then $f = f'$.

which contradictions are true:

> [M]any think it all too easy to construct stories in which 'such-and-such' [e.g. a contradiction] is true. It is much harder than you might think.

For example, an easy proposal, very similar to (Explicit), fails, since as Hanley points out 'saying it, does not always make it fictional [i.e. true in the fiction]'. What he seems to mean here is that for φ to be true in a fiction f it does not suffice that φ occurs explicitly in f, i.e. that (Explicit) is not sufficient for truth in fiction, even if the narrator or some character in the story asserts φ. As we have seen in the beginning, (Explicit) is, in general, neither sufficient nor necessary for truth in fiction.

An alternative principle which would entail the possibility of inconsistent fictions is *The Principle of Poetic Licence*, which Hanley (2004, p. 121) believes many appeal to in claiming that blatantly inconsistent fictions are possible:

> There is a tendency in the literature on truth in fiction to subscribe to the *Principle of Poetic Licence*, 'One can always write a story in which it is true that p, for any p'. [...] Lewis must deny the principle to eliminate contradictory propositions, but I deny the principle on independent grounds.

However, given that Hanley is sceptical about the possibility of blatantly inconsistent fictions and considers the principle as one justification for the possibility of such fictions, it seems that he wants to deny the principle and thereby make a point for his scepticism. So, in fact, he denies the principle to argue against the possibility of inconsistent fictions.

The principle of poetic licence (pop) Hanley refers to and wants to argue against is thus:

(pop) One can always write a story in which it is true that p, for any p.

One remark is in order here:[14] it is not entirely clear, what p is supposed to be a variable for, either a sentence, or a proposition. The first occurrence suggests a propositional reading due to the that-clause, whereas the latter could be either.

If both occurrences stand for propositions, then it seems an unrestricted reading of (pop) immediately renders it false: there are propositions which are, for example, not expressible in first order logic. Now, one cannot write a

[14]Thanks to Francesco Berto for this remark.

story in the language of first order logic, in which p is true, simply because the proposition p is not expressible by a sentence of the language of the fiction.[15] If we restrict the principle to sentences, we need to specify the language. Thus, it seems in order to restrict the principle to the respective language and the expressive limitations of the language for either propositional or sentential or hybrid reading:

(pop$_{prop}$) For any proposition p expressible in language L, it is possible to write a fiction f in the language L, s.t. it is true in f that p.

(pop$_{sent}$) For any sentence p in language L, it is possible to write a fiction f in the language L, s.t. the sentence p expresses a proposition that is true in f.

(pop$_{hybrid}$) For any proposition p expressible in language L, it is possible to write a fiction f in the language L, s.t. the sentence expressing p expresses a proposition that is true in f.

As Hanley claims to give an argument against (pop) and we claim that his argument is actually one against (Explicit), it is worth pointing out that and how the variations of (pop) differ from (Explicit). Thus denying one does not necessarily imply denying the other.

In the discussion above about the intensionality of In_f, we mentioned the case of stories in which everything is true, call them 'universal fictions'. If such stories exist, all mentioned variants of (pop) are true: just pick for φ a story in which everything is true, hence φ is true in it. Thus, the existence of universal fiction implies the truth of all mentioned variants (pop). However, it does not imply the truth of (Explicit).

However, if the variants of (pop) and (Explicit) were equivalent, they would both be entailed by the existence of universal fiction. There are other differences independent of the existence of universal fictions:

(Explicit) is not a necessary condition for any variant of (pop). We will consider (pop$_{sent}$) as an example. So suppose (pop$_{sent}$) holds. Consider the sentence $p :=$ 'Sherlock Holmes is unmarried'. We can write a story in which p expresses a true proposition without p occurring in it. We simply write a story in which 'Sherlock Holmes is a bachelor' is true, which can be done by (pop$_{sent}$), and make sure that p does not occur explicitly in it, which is obviously possible. Given in that story being a bachelor entails being

[15] If we allow for universal fictions, then by the very definition of such a fiction, we get that p is true in f. However, since universal fictions seem to require inconsistencies in them, see Folde & Wildman (under review), and Hanley argues against inconsistent fictions, this is not a viable objection.

unmarried, which is reasonable by import of background knowledge, p comes out as true in the story without occurring explicitly in it.

However, if (Explicit) holds, then (pop_{sent}) also holds because we can just write a story in which φ occurs explicitly and since by (Explicit) it is going to be true in f, we have constructed such a fiction. As we have seen before, (Explicit) does not hold in general, which does not mean that it does not hold in some instances, as we assumed.

As mentioned, Hanley claims to show the failure of (pop) and since (Explicit) seems to entail (pop), if he shows that (pop) fails, he shows that (Explicit) fails. But since (pop) does not entail (Explicit), showing that (Explicit) fails is not sufficient for (pop) to fail. Hence, if we can show that his argument is one against (Explicit) without showing that (pop) fails, he hasn't shown that (pop) fails. We claim that his argument indeed shows that (Explicit) does not hold in general. However, it does not show that (pop) fails in all its variants. Hanley's (2004, p. 121f.) argument goes as follows:

> I assert that in contemporary American fiction, a standard mention of a '555' *never* generates the fictional truth that the number in question begins with '555'. It leaves the number in question fictionally indeterminate. The argument for this claim is a reductio. What explains the consistency in the first three digits? There seem to be two basic possibilities. First, it may be that although in the world of the fiction there is usual divergence in these digits, for some coincidental reason, every number mentioned in the fiction has the same ones. Or, second, it may be that virtually every number in the U.S.A. begins with '555', hence it is no coincidence that the ones mentioned do. But both explanations are implausible. It is no part of either story [*X-files* (XF) and *Days of Our Lives* (*Dool*)] that virtually every phone number in the U.S.A. begins with the same three digits [...] And although bizarre coincidences obtain in fictions like these two, it's no part of either story that the first three digits of all these numbers coincide– that would itself be worth an 'X-file'. The '555' number, in a standard mention, is a conventional device for communicating the (fictional) fact that someone reveals a phone number, without stipulating what that number is. [...] Standard mentions of '555' fail to generate the 'obvious' fictional truths; indeed, given the conventions governing such mentions, they *couldn't* generate such fictional truths. The author's hand are effectively tied.

Hanley means to show the failure of (pop). For that, he needs to present a sentence φ (or a proposition) for which it is impossible to write a fiction f

such that In_f, φ.

He seems to claim that sentences with standard mentions of '555' numbers are examples of such sentences and that he wants to show that for no sentence φ with a standard mention of a '555' number it is possible to write a (contemporary American) fiction f, such that In_f, φ.[16] For a reductio, he would have to assume that there is some sentence φ with a standard mention of a '555' number such that it is possible to write a fiction f, s.t. In_f, φ.

Based on the quote, what he actually assumes for his reductio seems to be something different to us. One assumption is that for (almost) every contemporary American fiction f, if In_f, there is a number revealed, then this number is presented as a number starting with '555'.[17] But this is not the assumption meant to be negated.

However, if (Explicit) holds, then it is true in those fictions that (almost) all phone numbers start with '555'.[18] But, by what it means to be a standard mention of a '555' number, it is not the case that in any contemporary American fiction In_f, φ, where f is a contemporary American realistic fiction and φ is the sentence with the standard mention of the '555' number. To get a contradiction and a valid argument, we need to assume (Explicit) and this indeed seems to be the candidate assumption to be negated by the reductio.

This argument is not sound if we use (pop) instead of (Explicit), since (pop) does not imply that explicit occurrence of a sentence with a standard mention of a '555' number is sufficient for the truth of In_f, φ. And as we argued above, (pop) does not imply (Explicit) and so denying (Explicit) is not enough for denying (pop). Thus, this particular argument, which we think is an accurate reconstruction of Hanley's argument, is not one against (pop) but against (Explicit).

If we follow Hanley and construct an argument against (pop), we need to proceed differently. Thus, let us run the argument with the reductio assumption that there is a sentence φ with a standard mention of a '555' number, such that it is possible to write a fiction f s.t. In_f, φ. The question is, do we require that φ still contains a standard mention of a '555' number with respect to this fiction f?

If we do not require this, the premise does not seem to lead to absurdity. Even for the cases Hanley mentions, we can take such a sentence from XF,

[16]The contemporary American fiction he refers to is mainly realistic films and TV shows. We will grant this assumption here.

[17]Note that the scope of In_f ends after 'revealed'.

[18]The alternative is to assume that in the USA every phone number is a '555' number and thus, since he considers mainly realistic fictions, by importing background information, in those fictions the numbers begin with '555'. However, it is a matter of fact that in the USA not all phone numbers begin with '555'.

s.t. it has a standard mention of a '555' number with respect to XF but then, in our new fiction f it just is not a standard mention anymore (because we could write a contemporary German fiction) and so we can get In_f, φ either by allowing for (Explicit) in this particular instance or have φ being entailed by other sentences occurring in f.

If we indeed require that the standard mention of the '555' number is kept in f, then it seems this premise is immediately absurd because by the meaning of 'standard mention' it is already excluded that φ expresses its literal meaning. Thus, it seems under those qualifications, the correct reductio premise and assuming the standard mention has to be fixed, we get an argument against (pop_{sent}).

However, (pop_{prop}) is not dismissed by considering sentences with standard mentions of '555' numbers. This is due to the fact that the proposition expressed by such sentences depends on the fact that we are dealing with standard mentions of the '555' number and that in turn is dependent on the sentence occurring in a fiction. The very fact that the sentence appears in a genre whose convention is to write '555' numbers whenever one wants to express that there is a phone number revealed, determines that we are dealing with a standard mention of that number. We simply cannot use the proposition expressed by the sentence with a standard mention of a '555' number as a counterexample because the proposition expressed by such a sentence is determined by the context in which this sentence occurs and it is determined in such a way that it is always going to be true. If we consider such a sentence φ, the proposition it expresses in the fiction is that a phone number is revealed. But this very proposition is automatically true in the fiction in which φ occurs.

What proposition, if any, such sentences express outside the fiction is difficult to determine. First of all, it would not make sense, outside the context of the fiction, to speak of this sentence having a standard mention of a '555' number. Moreover, these sentences might contain fictional names or terms. Depending on one's ontology those sentences might then express no proposition at all. Thus, these sentences seem not to provide the propositions we need as counterexamples to (pop_{prop}).

If (pop_{prop}) holds, which we have not shown, but at least Hanley hasn't shown it to fail with his argument, then it seems we do in fact get that there blatantly consistent fictions are possible.

There are some further remarks concerning the argument. First, although Hanley shows that (Explicit) is, in general, not sufficient for truth in fiction, this does not provide a case against the possibility of blatantly inconsistent fictions. To show that there cannot be any blatantly inconsistent fictions and

base this argument on the denial of (Explicit) Hanley needs to assume:

(Ent) The possibility of a blatantly inconsistent fiction f, where φ is the explicit contradiction, entails 'In_f, φ iff. φ occurs explicitly in f', for this particular f and φ.

Since only then one can have a modus tollens inference to 'blatantly inconsistent fictions are impossible'. However, Hanley has not provided an argument for the failure of the consequent for the particular instance. This might be due to the fact that he claims to only deny (pop) but in fact wants to make a point about inconsistent fictions.

Moreover, even if we grant that (pop_{sent}) or even (pop_{prop}) fails, to get a reductio to the possibility of blatantly inconsistent fictions, we need that the possibility of such fictions implies this principle, i.e. that the principle is the only way to generate such fictions. Hanley does not explicitly assume this. If we grant this premise, we can again, in many cases, point to the particular sufficiency of (Explicit) for truth in fiction and claim that we do not need (pop) in those cases.

Hence, we will keep following our assumption about the existence of blatantly inconsistent fictions. Since we saw that Lewis's Analysis 2, even with Lewis's (1983) attempts, does not account for blatantly inconsistent fictions, we are going to extend Analysis 2 by using impossible world semantics.

Moreover, as we have seen in the discussion of Proudfoot (2006), the crucial premise (P5) relied on intuitions about what should come out as true in a fiction. We will incorporate this into our extension by appealing to plausibility orders on worlds instead of the closeness/on balance condition. We will address the issue of the plausibility order in chapter 4.1.

Chapter 3

Extending Lewis's Analysis 2

In this chapter, we extend Lewis's Analysis 2 to be able to account for blatantly inconsistent fiction. We do this by treating engagement with fiction as a process structurally analogous to belief revision, in particular, soft upgrades. We first present the AGM axioms, (named after their proponents Alchourròn, Gärdenfors & Makinson (1985)), which are supposed to be rationality postulates a revision operation should satisfy. Then we provide single agent-plausibility models and sphere models with impossible worlds and explain the revision on those models. Subsequently, we extend this to the multi-agent case and introduce common beliefs. In those sections we mainly draw on material from Baltag (2016) and Smets (2015), however these do not deal with impossible world semantics. Finally, we give the truth clause for In_f, φ on multi-agent sphere models and argue how we can account for blatantly inconsistent fictions in which not everything is true. To conclude this chapter, we discuss which of the initially presented AGM axioms might (not) fail on our account.

3.1 The AGM-Postulates for Belief Revision

Belief revision theory deals with the issue of how to incorporate new information into an already existing set of information. This is a non-trivial matter, since new information might be inconsistent with the already gathered information. In particular belief revision theory is concerned with sets of sentences an agent beliefs, so called *belief sets*. These sets are usually assumed to be consistent and logically closed. If they are not logically closed, they are called a *belief base*.

In belief revision theory, there are at least three different kinds of operations which can be performed if new information is faced: expansion (+),

contraction ($-$) and revision ($*$). Expansion amounts to adding the new information φ to the belief set and then close the resulting set under (classical) logical consequence. Contraction amounts to deleting some information ψ from the belief set and everything that entails ψ, and again close under logical consequence. Revision is an operation performed when the new information is inconsistent with the belief set. One proposal is to define revision in terms of contraction and expansion, known as the *Levi-Identity*: if B is a belief set and φ is the new information, then $B * \varphi = (B - \neg\varphi) + \varphi$.

If we have $+$ and $-$ satisfying the respective AGM axioms, then, as Gärdenfors (1988, p. 69) has shown, the Levi-Identity holds. If B is a consistent logically closed set of beliefs of an agent, φ is the new information and $*$ is the revision operator, the AGM postulates for $*$ are as follows:

1. $B * \varphi$ is consistent and logically closed.
2. $\varphi \in B * \varphi$
3. $B * \varphi \subseteq B + \varphi$
4. If $\neg\varphi \notin B$, then $B + \varphi \subseteq B * \varphi$
5. $B * \varphi$ is inconsistent only if φ is inconsistent
6. If φ is logically equivalent with ψ, then $B * \varphi = B * \psi$
7. $B * (\varphi \wedge \psi) \subseteq (B * \varphi) + \psi$
8. If $\neg\psi \notin B * \varphi$, then $(B * \varphi) + \psi \subseteq B * (\varphi \wedge \psi)$

These axioms only tell us, how a (rational) revision operator should behave but not what such an operator amounts to. In fact, there are various operators satisfying the axioms. That is, there are various *models* for the axioms, i.e. structures on which we can define an operation satisfying these axioms. One kind of models which has been fairly standard in belief revision theory are sphere models, as presented by Grove (1988). We will discuss their viability for our purposes in section3.4.

For *sphere models* Grove (1988) has shown that these models satisfy the AGM axioms. Since Lewis's Analysis 2 is inspired by his semantics for counterfactuals, which can, in turn, be expressed in terms of sphere models, and these have seen applications in belief revision theory, it is quite natural for us to consider these models as possible candidates as models we want for truth in fiction. We will use these models but with a domain comprising not only possible but also impossible worlds. However, we will start by providing

single agent plausibility models with domains containing impossible worlds. From such models we can construct sphere models. We then extend this to the multi-agent case. This gives us models on which we can give truth conditions for In_f close to Lewis's initial Analysis 2.

As mentioned, in standard belief revision theory, belief sets/bases *and their updates are consistent* and belief sets are logically closed. We are going to treat engagement with fiction as a process analogous to belief revision. Since we want to account for inconsistent fictions without trivialization, we won't follow the standard assumption that belief sets are logically closed and also allow for agents with already inconsistent beliefs. Consequently, most of the AGM axioms will fail on our models. However, given certain additional assumptions, some of them will hold. This will be discussed in 3.4. We now build up the models we are going to use to give the truth conditions for In_f, φ.

3.2 Semantics

Lewis's analysis is inspired by Lewis's (1983) analysis of counterfactuals, for which he gave a semantics in terms of a similarity order. We thus start out by considering models with an ordering. However, we will change from the notion of similarity to the notion of plausibility. We do this in order to incorporate that it is almost overall agreed upon in the literature that our intuitions are usually very good about which fictional worlds are more plausible than others. This plausibility order will thus be given externally and we assume moreover that it is influenced by the agent's (or agents') favourite theory of interpretation of fictional texts. As a result, a disagreement on the theory of interpretation might lead to a disagreement about what is true in a fiction, which seems realistic. Based on the order-models we define sphere models, which, without impossible worlds, have been shown by Grove (1988) to validate the AGM axioms. However, we will, as in the case of order-models, use impossible worlds in our domain too.

Moreover, Analysis 2 takes into account the overt beliefs of the community of origin.[1] These need to be formally represented. Additionally, we have to formally represent the worlds where f is told as known fact and to capture the similarity relation. We interpret overt beliefs as common beliefs in a

[1]Note, that the *overt* beliefs are not *just* the beliefs of the community. It is what, according to Lewis (1978, p. 44), almost everyone believes and almost everyone believes that almost everyone believes... Hence, if every agent's belief set is consistent, the overt beliefs are too. This is not true for the totality of beliefs of the community, since beliefs are fallible and thus two agents may have beliefs inconsistent with each other.

technical sense. We replace the similarity relation by a plausibility order on worlds. We then treat engagement with fiction as a soft upgrade that induces a new plausibility order on the set of worlds.

3.2.1 Single Agent Plausibility Models

First, we define a propositional language, which is our object language. We have a countably infinite set of propositional variables $Prop = \{p, q, r, ...\}$, indexed by natural numbers if needed. The set of well formed formulas $Form$ is the smallest set s.t.:

If $\varphi \in Prop$, then $\varphi \in Form$.

If $\varphi \in Form$, then $\neg\varphi, \Box\varphi, \diamond\varphi, In_f, \varphi \in Form$.

If $\varphi, \psi \in Form$, then $(\varphi \vee \psi), (\varphi \wedge \psi), (\varphi \supset \psi) \in Form$.

We will call elements of $Form$ formulas, sentences or propositions.[2] As usual we omit outer parantheses if no confusion arises.

We are going to use single agent plausibility models, similar to the ones used by Baltag & Smets (2006), but extend their domains to encompass impossible worlds and moreover assume that the plausibility ordering $\leq$ is going to be total also on the set of impossible worlds. If necessary, we refer to the agent by 'a'.

We define a single agent plausibility model $\mathfrak{M} = \langle W, \leq, R_C, V\rangle_{C\in\mathcal{C}}$, where $W = P \cup I$ is the union of a set $P \neq \emptyset$ of possible and a set $I \neq \emptyset$ of impossible worlds, s.t. $P \cap I = \emptyset$.[3]

The order $\leq$ is the agent's plausibility order on the worlds.[4] We write $u \leq v$ for $(u, v) \in \leq$. $u \leq v$ means the agent considers v to be at least as plausible as u. This order is assumed to be a conversely well-founded total preorder.[5] We write $u < v$ if $u \leq v$ and $v \not\leq u$ (not $v \leq u$). Thus, for any

[2]Strictly speaking, above we need to use Quine quotes, but for legibility reasons we omit those.

[3]We could make this a pointed model by adding the actual world but since we won't use the actual world, we just omit it.

[4]Agents can consider impossible worlds to be equally plausible to consistent worlds. Whether those agents are rational is not of concern to us here. But as Priest (2001) argues, having inconsistent beliefs does not necessarily make an agent irrational. Of course one could put restrictions on $\leq$ that disallow this. However, in our context it seems appropriate to not restrict the relation.

[5]As a preorder, it is reflexive and transitive. Converse well-foundedness means that for any subset P of W, that P has a maximum wrt the ordering $\leq$, i.e. $max_{\leq}P = \{v \in W | \forall w \in P : w \leq v\} \neq \emptyset$. This means that any subset P of W has a set of most plausible worlds wrt all the worlds in P. So worlds can tie wrt plausibility and thus, the set of maxima need not be a singleton. We do not require anti-symmetry for $\leq$.

$w, v \in W$, $w \leq v$ or $w \not\leq v$. We write $u \simeq v$ if $u \leq v$ and $v \leq u$ and read that as 'the agent considers u and v as equally plausible'. It is easy to see that $\simeq$ is an equivalence relation.

The set R_C is a set of accessibility relations between worlds, each determined by some explicit content $C \in \mathcal{C}$, where $\mathcal{C}$ is a set of explicit contents. We write $R_C wv$ instead of $(w, v) \in R_C$ and follow Berto (forthcoming) that $R_C wv$ means 'v is representationally accessible from w' via explicit content C.

$V = \langle V^+, V^- \rangle$ is a pair of valuation functions that assigns for any atomic formula $\varphi \in Prop$ the set of worlds $V^+(\varphi)$, where φ is true and the set of worlds $V^-(\varphi)$ where φ is false. We define truth/falsity at a possible world $w \in P$ in the model $\mathfrak{M}$ recursively:[6]

$w \vDash^+ \varphi$ iff. $w \in V^+(\varphi)$, for $\varphi \in Prop$
$w \vDash^- \varphi$ iff. $w \in V^-(\varphi)$, for $\varphi \in Prop$

$w \vDash^+ \neg\varphi$ iff. $w \vDash^- \varphi$
$w \vDash^- \neg\varphi$ iff. $w \vDash^+ \varphi$

$w \vDash^+ \varphi \wedge \psi$ iff. $w \vDash^+ \varphi$ and $w \vDash^+ \psi$
$w \vDash^- \varphi \wedge \psi$ iff. $w \vDash^- \varphi$ or $w \vDash^- \psi$

$w \vDash^+ \varphi \vee \psi$ iff. $w \vDash^+ \varphi$ or $w \vDash^+ \psi$
$w \vDash^- \varphi \vee \psi$ iff. $w \vDash^- \varphi$ and $w \vDash^- \psi$

$w \vDash^+ \Box\varphi$ iff. for all $v \in P$: $v \vDash^+ \varphi$
$w \vDash^- \Box\varphi$ iff. for some $v \in P$: $v \vDash^- \varphi$

$w \vDash^+ \Diamond\varphi$ iff. for some $v \in P$: $v \vDash^+ \varphi$
$w \vDash^- \Diamond\varphi$ iff. for all $v \in P$: $v \vDash^- \varphi$

$w \vDash^+ \varphi \supset \psi$ iff. $w \vDash^+ \varphi$ implies $w \vDash^+ \psi$
$w \vDash^- \varphi \supset \psi$ iff. $w \vDash^+ \varphi$ and $w \vDash^- \psi$

Note that implication is material implication.

Note that, in general, $\nvDash^+$ (not $\vDash^+$) is not the same as $\vDash^-$. For impossible worlds we allow for truth value gaps or gluts. For some $w \in I$ and some φ we have gaps: $w \nvDash^+ \varphi$ and $w \nvDash^- \varphi$ and we also have gluts for some $v \in I$ and some φ: $v \vDash^+ \varphi$ and $v \vDash^- \varphi$. Following Berto (forthcoming) we impose a classicality condition (CC) on the possible worlds, that is

[6] We usually write $\mathfrak{M}, w \vDash^{\pm} \varphi$ but if the model is clear from the context, we omit $\mathfrak{M}$.

(CC) For all $w \in P$ and all $\varphi \in Form$: either $w \vDash^+ \varphi$ or $w \vDash^- \varphi$ and not both.[7]

To account for truth at impossible worlds, we extend V^+ and V^- to any formula, i.e. $V^+(\varphi) = \{w \in W | \varphi \text{ is true at } w\}$. For $w \in P$, 'true at w' just means $\vDash^+$ as defined above. Similarly $V^-(\varphi) = \{w \in W | \varphi \text{ is false at } w\}$. For $w \in I$, we treat any formula as atomic and thus $V^{\pm}$ assigns to each formula φ directly the impossible worlds where φ is true/false. Truth at impossible worlds $w \in I$ is then defined by:

$w \vDash^+ \varphi$ iff. $w \in V^+(\varphi)$
$w \vDash^- \varphi$ iff. $w \in V^-(\varphi)$

The extension of a formula φ in a model $\mathfrak{M}$, $[\![\varphi]\!]^+_{\mathfrak{M}}$, is defined as

$$[\![\varphi]\!]^+_{\mathfrak{M}} := \{w \in W | \mathfrak{M}, w \vDash^+ \varphi\}$$

Its anti-extension $[\![\varphi]\!]^-_{\mathfrak{M}}$, is defined as

$$[\![\varphi]\!]^-_{\mathfrak{M}} := \{w \in W | \mathfrak{M}, w \vDash^- \varphi\}$$

Again we might omit the model subscript.

Logical Consequence is defined over the set of possible worlds P. Let Γ be a set of formulas and φ a formula. Then

$\Gamma \vDash \varphi$ if for any model $\langle W, \leq, R_C, V\rangle_{C \in \mathcal{C}}$ and any $w \in P$: if $w \vDash^+ \gamma$ for any $\gamma \in \Gamma$, then $w \vDash^+ \varphi$

We then say φ is a *logical consequence* of Γ. φ is a *logical truth* iff. $\vDash \varphi$ iff. $\emptyset \vDash \varphi$. If $\mathfrak{M}, w \vDash^+ \varphi$, φ is valid at a world in a model. If for all w in model $\mathfrak{M}$, $\mathfrak{M}, w \vDash^+ \varphi$ we say that φ is valid in the model. If there is a world w in model $\mathfrak{M}$, s.t $\mathfrak{M}, w \vDash^+ \varphi$, we say φ is satisfiable in the model.

We will now consider sphere models which can be constructed from plausibility models. This is done in order to allow for a more streamlined formulation in the end and it is closer to Lewis's own analysis. He defines truth in fiction similarly to truth for counterfactuals. For counterfactuals, similarity sphere models have been used. Thus, analogously, we use sphere models here as well. We started out by giving plausibility models because we think that it is a natural way to think about agents ordering (im)possible worlds and it is quite close to the actual practice in philosophy of fiction. This will be

[7] In fact, as Berto (forthcoming) correctly emphasizes, it is enough to require (CC) for propositional variables and then extend it to any formula by induction.

addressed in more detail in sec. 4.3. Before we introduce sphere models, we consider a particular set of worlds, namely that of most plausible worlds.

Since we are dealing with plausibility models, it is useful to consider the set of worlds an agent considers to be *most plausible*. For any $S \subseteq W$ the set of most plausible worlds wrt S is

$$bestS = \{w \in S | \forall v \in S : v \leq w\} = max_{\leq} S$$

where $max_{\leq} S$ is the set of maxima wrt the order $(S, \leq)$. The set of most plausible worlds in the model thus is $bestW$. The agent's set of beliefs $\mathcal{B}$ in the model is given by what is true at all the most plausible worlds in the model:

$$\mathcal{B} = \{\varphi | \forall w \in bestW : w \vDash^{+} \varphi\}$$

Note that it is possible for an agent to have inconsistent beliefs/believe contradictions without believing every sentence. Suppose an agent believes $p \wedge \neg p$, which is only the case if for any most plausible world w, $w \vDash^{+} p \wedge \neg p$, which is only possible if $bestW$ contains only impossible worlds. However, these might still be as plausible as possible compared to the real world, i.e. not everything is true in them, based on the plausibility order of the agent.[8]

Based on the plausibility order of the agent, we can define a *system of spheres*. We define a family $\mathcal{S} \subseteq \mathcal{P}(W)$ by

$$\mathcal{S} := \{w^{\leq} | w \in W\}$$

where $w^{\leq} = \{v \in W | w \leq v\}$. We call members S of $\mathcal{S}$ *spheres*. So for each world w, we get a sphere, which is a set of worlds v, which the agent considers at least as plausible w. We call $\mathcal{S}$ a *system of spheres* if it satisfies the following conditions:

(S1) The spheres are nested/totally ordered:

$$\forall S, S' \in \mathcal{S} : S \subseteq S' \text{ or } S' \subseteq S$$

(S2) For every non-empty subset $P \subseteq W$, there is a smallest sphere, $c(P)$, intersecting it. This is also called *limit assumption* and corresponds to a well founded ordering of the spheres:

$$\forall P \subseteq W (P \neq \emptyset \Rightarrow (\exists S \in \mathcal{S} (\forall S' \in \mathcal{S} (P \cap S' \neq \emptyset \Leftrightarrow S \subseteq S'))))$$

(S3) The spheres are non-empty and they cover the set of worlds:

$$\forall S \in \mathcal{S} : S \neq \emptyset \text{ and } W = \bigcup \mathcal{S}$$

[8] One could require $\mathcal{B}$ to be consistent by $\forall w \in bestP$.

It can be shown that $\mathcal{S}$ defined as above satisfies (S1)-(S3) (see appendix).

Note that given a single agent plausibility model, then $bestW \subseteq S$ for any $S \in \mathcal{S}$.[9] And hence, the spheres are centered around the agent's belief set and the beliefs contain the most plausible worlds and all other spheres contain less and less plausible worlds.

We will now introduce a notion of revision/soft upgrade on our single agent models.

3.2.2 Revision with Explicit Content of a Fiction

The worlds where F obtains, or where f is told as known fact, will now be understood in terms of our accessibility relation R. If w is a world where f is told as fiction, then the set of worlds where f is told as known fact rather than fiction/where F obtains is assumed to be the set of worlds v s.t. $R_F wv$, i.e. all worlds which are accessible via the fiction's explicit content. We will call those worlds f-worlds. We capture that f is told as fiction by saying that $\neg R_F ww$.[10]

Now, in the single agent case, if f is a fiction and F its explicit content, a revision operation $\Uparrow F$ can be performed on the model. This operation will clearly not satisfy the classical AGM axioms, introduced by Alchourròn et al. (1985) because we want to *keep* inconsistencies *without* trivializing.

When engaging with fiction, following Walton (1990), the agent make-believes or pretends, that the explicit content of the fiction obtains. In the end, we want to say that φ is true in f if after "revising" the common beliefs of the community of origin of f by the explicit content F of f, φ is true. 'Revision' has to be taken with a grain of salt here because it is, strictly speaking, pretended revision.

Hence, we will treat reading a fiction similarly to softly upgrading one's beliefs by the fictions explicit content F, also called 'lexicographic upgrades' by van Benthem (2007). By softly upgrading, no worlds are going to be deleted from the model but the worlds will be reordered, as opposed to hard updating with φ where all $\neg\varphi$-worlds are deleted from the model.[11] We will assume, however, that upgrading with the explicit content of a fiction is always successful, so there always are f-worlds.

[9] Let $w \in bestW$ and $v^{\leq} \in \mathcal{S}$ be arbitrary. Since $w \in bestW$ for any $s \in W$, $s \leq w$. Hence $v \leq w$ and thus $w \in v^{\leq}$.

[10] This condition is not sufficient for f being told as fiction at w since f can be not told at all at w.

[11] Analogously to assuming that for any proposition one can write a fiction in which it is true, we assume that for any fiction there is going to be a world where its explicit content obtains.

The reason we choose soft upgrades instead of hard updates is that we can still account for when the agent stops engaging with fiction and stops make-believing. Thus, she can just recover her old ordering $\leq$ on the same set of worlds, since we assumed she came equipped with it, and hence it seems reasonable she can recover it. Or, what we consider an advantage, she can reorder again based on some new insights she might have had through engaging with the fiction. This might be useful in applications in learning through fiction.[12]

We express, in the metalanguage, the upgrade by $\Uparrow \varphi$ if we upgrade with φ. If F is the explicit content of f, then by $\Uparrow F$ we mean upgrading with $\bigwedge_{\varphi \in F} \varphi$.[13] We write $w \vDash^{+} F$ if for all $\varphi \in F$, $w \vDash^{+} \varphi$. What the upgrade means is that first, after the upgrade with φ, all the worlds where φ is true become more plausible than any $\neg\varphi$-worlds. Second, among the two sets of worlds, the previous order remains. What this formally amounts to is to reorder the worlds in the model by a new plausibility order $\leq^{\Uparrow F}$.

We will stick to the first condition in the following sense: any f-world is going to be more plausible than any non-f-world, i.e.

$$\textbf{Cond 1: } \forall t \in W[t \vDash^{+} F \Rightarrow (\forall s \in W : s \nvDash^{+} F \Rightarrow s \leq^{\Uparrow F} t)]$$

However, in the case of engagement with fiction, the second condition might fail. The ordering within the two sets might change based on the agent's engagement with the fiction.

For the ordering among the f-worlds, suppose a engages with *War and Peace* (WaP) for the very first time. Before performing the upgrade, she might consider a WaP-world w where Napoleon (from WaP) is as clever as the real Napoleon more plausible than a world v where Napoleon from WaP is duller than the real one. However, after engaging with WaP, it seems reasonable for her that v is more plausible than w. Thus, after the upgrade the ordering among the f-worlds has changed. To make all the *best* f-worlds more plausible after the upgrade is no option to avoid this problem because the best worlds would be best wrt the *previous* ordering. And as just pointed out, the fiction can change this previous ordering in a way independent from what was previously considered plausible.

[12]Learning through fiction, especially wrt empathy, is of recent interests, especially when it comes to interactive fiction (video-games), as in, for example, Wildman (unpublished). Our model or an expansion of it, might be able to account for those phenomena since agents can always recover their previous plausibility order or rearrange based on the fiction they engaged with.

[13]We assume that F is finite/is expressible by a finite set of sentences. This is unproblematic for existing understandable fiction.

Also the ordering among the non-f-worlds might change, since agents might learn about the real world through fiction. For example, before engaging with *Moby Dick* one might know nothing about whale hunting in the 18th/19th century and consider worlds where it was not very cruel at least as plausible as worlds where it was. After engaging with the story, the ordering among the worlds where F does *not* obtain and which are very similar to the real world and where whale hunting in 18th/19th century was cruel will be more plausible wrt the new ordering.

Moreover, we want that any most plausible world wrt the new order $\leq^{\Uparrow F}$ is an f-world because in engagement with fiction the agent make-believes (at least) F. This yields:

$$\textbf{Cond 2: } \forall t \in W[(\forall s \in W : s \leq^{\Uparrow F} t) \Rightarrow t \vDash^{+} F]$$

Despite these two conditions, we require the order to be a total conversely well-founded preorder.[14]

Concerning the ordering among the other worlds, we take the new order among worlds to be primitive/given by the agent. Note that this might remain the same as before or might just change completely.[15]

The new order $\leq^{\Uparrow F}$ gives rise to the new model $\mathfrak{M}^{\Uparrow \mathcal{F}} = \langle W, \leq^{\Uparrow \mathcal{F}}, R_C, V\rangle_{C \in \mathcal{C}}$, where $\leq^{\Uparrow \mathcal{F}}$ is a conversely well founded total preorder on W given externally by the agent after she engaged with the fiction f. This change of ordering will be guided by the agent's knowledge about genre conventions, her competence to detect irony and interpret it correctly, and many other pragmatic criteria. Since the plausibility order changed, also the set of most plausible worlds changed, so for any $S \subseteq W$, we define $best^{\Uparrow F} S := \{w \in W | \forall v \in S : v \leq^{\Uparrow F} w\}$. It is easy to see that any $w \in best^{\Uparrow F} W$ is an f-world, by Cond 2.

We can define a new system of spheres based on our new plausibility order $\leq^{\Uparrow F}$ by

$$\mathcal{S}^{\Uparrow F} := \{w^{\leq^{\Uparrow F}} | w \in W\}$$

where $w^{\leq^{\Uparrow F}} = \{v \in W | w \leq^{\Uparrow F} v\}$.[16]

What happens here is that the original plausibility model is revised based on the explicit content of the fiction. The revision does not delete any worlds but only reorders them by a new plausibility order.

We can still define a belief set on our new model by $\mathcal{B}^{\Uparrow F} = \{\varphi | \forall v \in best^{\Uparrow F} W : v \vDash^{+} \varphi\}$. However, we will call this the 'make-belief-set' in our

[14] In the following example it is easy to see that both conditions are satisfied and that the order is a total conversely well-founded preorder: $W = P = \{s, t\}, \leq^{\Uparrow F} = \{(s,t), (s,s), (t,t)\}, F = \{p\}, V^{+}(p) = \{t\}, V^{-}(p) = \{s\}$.

[15] For concerns about this, we refer to our discussion in sec 4.3

[16] This satisfies (S1)-(S3), since $\leq^{\Uparrow F}$ still has the same properties we used for the proof before.

context because the agent does not actually change her beliefs but just make-believes the explicit content of the fiction. We could also call the revision operation 'pretended revision', as mentioned above. Although we use the technical terminology used in this area of research, we are not claiming in any way that our notions of (make-)belief or revision correspond to anything with the same name in the known frameworks. Since the process is claimed to be *similar* (and not necessarily identical) to actual belief revision, nothing hangs on this terminology.

Now Lewis's analysis makes use of the common beliefs of the community of origin. We thus have to define multi-agent plausibility models and add a common belief operation and group belief revision to finally give the semantics for the operator In_f.

3.2.3 Multi-Agent Plausibility Models

A multi-agent plausibility model is a structure $\mathfrak{M} = \langle W, \leq_a, R_C, V\rangle_{a\in\mathcal{A},C\in\mathcal{C}}$, where $\mathcal{A}$ is a finite set of agents $a_1, ..., a_n$, $\leq_a$ is the plausibility order for each agent $a \in \mathcal{A}$ as defined before and just indexed by the name a_i for the agent. We define everything, such as $bestS$ etc., as before but index it with a for each agent, i.e. for $S \subseteq W$, $best_aS = \{v \in W | \forall x \in S : x \leq_a v\}$. Truth at a world in a model is defined as before since all operators are independent of the agents.

When we have a multi-agent plausibility model $\mathfrak{M}$, we define a doxastic accessibility relation $\rightarrow_a$. Let for $w \in W$, $w(a) = \{v \in W | v \simeq_a w\}$, i.e. the set of worlds v which a considers to be equally plausible to w. Then for $s, t \in W$:

$$s \rightarrow_a t \Leftrightarrow t \in max_{\leq_a} s(a) \Leftrightarrow t \in \{v \in W | \forall x \in s(a) : x \leq_a v\}$$

That is t is doxastically accessible from s by agent a iff. t is at least as plausible as any world v considered equally plausible to s by agent a.

The overt beliefs Lewis uses in his analysis we treat as common beliefs, and hence will be a stricter concept than the one Lewis mentioned. Common belief of φ can be understood as 'everybody believes φ and everybody believes that everybody believes φ and so on' as opposed to *almost* everybody believes etc.

As Bonanno (1996) emphasizes, from a semantic perspective, common belief understood this way is unproblematic. However syntactically, it poses some problems due to being an infinite conjunction and most usually used formal languages are finitary. Since our approach is semantic, we will not address the syntactic issues here. According to Bonanno (1996), semantically,

the common belief accessibility relation $\rightarrow_{CB}$ for a group of agents $\mathcal{G} \subseteq \mathcal{A}$ is the transitive closure of the union of the individual accessibility relations $\rightarrow_a$, that is, it is the smallest transitive relation $\rightarrow$, s.t. $\bigcup_{a\in\mathcal{G}} \rightarrow_a \subseteq \rightarrow$. Thus, a formula φ is commonly believed at a world w by group $\mathcal{G}$ iff. φ is true at any world v, s.t. $w \rightarrow_{CB_\mathcal{G}} v$. We define this formally, but as an operator in our metalanguage, for $w \in P$:

$$w \vDash^+ CB_\mathcal{G}\varphi \text{ iff. for any } v \in W : w \rightarrow_{CB_\mathcal{G}} v \text{ implies } v \vDash^+ \varphi$$

$$w \vDash^- CB_\mathcal{G}\varphi \text{ iff. for some } v \in W : w \rightarrow_{CB_\mathcal{G}} v \text{ implies } v \vDash^+ \varphi$$

For $w \in I$, we again assign the truth value arbitrarily. Note that $w \vDash^+ CB_\mathcal{G}\varphi$ entails that for any $a \in \mathcal{G}$, $\varphi \in \mathcal{B}_a$.

Proof. Consider an arbitrary multi-agent plausibility model $\mathfrak{M} = \langle W, \leq_a , R_C, V\rangle_{a\in\mathcal{A},C\in\mathcal{C}}$. Suppose $w \vDash^+ Cb_\mathcal{G}\varphi$ and let $v \in best_a W$ be arbitrary for arbitrary $a \in \mathcal{G}$. Then $w \leq_a v$. Hence $v \in max_{\leq_a} w(a)$ and thus $w \rightarrow_a v$ and thus $w \rightarrow_{CB} v$. Therefore $v \vDash^+ \varphi$ and thus $\forall x \in best_a W : x \vDash^+ \varphi$. Hence $\varphi \in \mathcal{B}_a$. □

This is reasonable since everything that is commonly believed by a group should also be believed by every single member of the group.

How are we to understand a common belief *world* then? What Lewis has in mind is not all the worlds where the common beliefs at w are also commonly believed, but where the common beliefs from w are true. Hence, a common belief world is *not* just a world in $[\![CB^w_\mathcal{G}]\!]$, where $CB^w_\mathcal{G} = \{\varphi | w \vDash^+ CB_\mathcal{G}\varphi\}$. Instead, for Lewis (1978, p. 44), a common belief world is certainly a world at which all common beliefs of a certain community are true:

> Then we can assign to the community a set of possible worlds, called the *collective belief worlds* [i.e. common belief worlds] of the community, comprising exactly those worlds where the overt beliefs all come true.

In our case, this community is the community of origin of a fiction f. Usually, this community is a community which existed in the past of "our" world, or, more generally, part of the world at which we are to evaluate 'In_f, φ. We use 'our' instead of 'actual', since Lewis understands 'actual' as an indexical. So on his account our *actual* world, is not the same as the *actual* world of, say, the community of origin of *LOTR*. However, our world, without the proviso 'actual', is spatio-temporally inclusive on Lewis's (1986) account and hence comprises the community of origin of *LOTR*. Hence, their common belief worlds, will be worlds in which all the common beliefs they had in our world, are true.

If $CB^w_{\mathcal{G}} = \{\varphi | w \vDash^+ CB_{\mathcal{G}}\varphi\}$ is the set of common beliefs of group $\mathcal{G}$ at w, then we let $|CB^w_{\mathcal{G}}|$ be the set of worlds, where all those common beliefs are true. Formally, this amounts to

$$|CB^w_{\mathcal{G}}| = \{v \in W | \forall \varphi \in CB^w_{\mathcal{G}} : v \vDash^+ \varphi\} = \bigcap_{\varphi \in CB^w_{\mathcal{G}}} [\![\varphi]\!]^+$$

We now go on to discuss the revision on the multi-agent model.

3.2.4 Multi-Agent Revision

There are (at least) two different ways we could do things. First, we could introduce a group plausibility ordering and then perform the revision on it. Or we could take the single agent case as basic for each agent, perform revision for each agent individually, and then define a new plausibility ordering for the group.[17] We think, the latter way is more intuitive since any agent engages individually with the fiction, revises, and then, maybe through discussion among the group members, a plausibility ordering on all the worlds is built, based on all agents' individual orderings.

There are several ways of arriving at the group plausibility order after revision $\leq^{\Uparrow F}_{\mathcal{G}}$. These mainly depend upon whether one wants to put more weight on the orderings $\leq^{\Uparrow F}_a$ for a particular a or whether one wants to have a more democratic way of determining the ordering. A democratic way is to simply intersect all individual plausibility orderings. However, this ordering might end up not being total, that is, not every two worlds will be comparable to each other within the group.[18] Moreover, it might just be empty, if all the individuals heavily disagree.

Another way is to give preference to certain individuals' plausibility orders. So for example, depending on one's preferred theory of interpreting fiction, one could give priority to the author's plausibility ordering or that of certain experts.

We will use an approach similar to the lexicographic merge used by Andréka, Ryan & Schobbens (2002), which is following the idea that we give priority to certain agents' plausibility orders. We do this for technical reasons because the ordering is not going to be empty and still be total. Moreover, it allows, for us as modellers, to almost ignore plausibility orderings of agents which are not competent in reading fiction or should be ignored for other reasons.

[17] These two might prove to be equivalent but we will leave this to further research.

[18] The problematic case is if $w \leq_a v$ and $v \leq_b w$ for agents a, b. Are we going to say $w \leq_{\mathcal{G}} v$ or vice versa or both? If $w \leq_{\mathcal{G}} v$ (or vice versa), then we are giving preference to one agent. If we treat the worlds as equally plausible, we lose information.

The merged ordering *after* revision is defined as follows, where we assume that there is a hierarchy among the agents a, b:

$$\leq_{a/b}^{\Uparrow F} := <_a^{\Uparrow F} \cup (\simeq_a^{\Uparrow F} \cap \leq_b^{\Uparrow F})$$

That is, the group's ordering after revision is a's strict ordering, and if a considers worlds equally plausible, b's ordering is adopted.

This can be extended to the whole group of agents. Let $a_0, a_1, a_2..., a_n$ be a hierarchy of agents s.t. the left are prioritized over the right. Let $\mathcal{G}_i$ be the group of agents up to including agent a_i. Then we generalize $\leq_{\mathcal{G}}$ as follows:

$$\leq_{\mathcal{G}_0/a_1}^{\Uparrow F} = <_{a_0}^{\Uparrow F} \cup (\simeq_{a_0}^{\Uparrow F} \cap \leq_{a_1}^{\Uparrow F})$$

$$\leq_{\mathcal{G}_n/a_{n+1}}^{\Uparrow F} = <_{\mathcal{G}_n}^{\Uparrow F} \cup (\simeq_{\mathcal{G}_n}^{\Uparrow F} \cap \leq_{a_{n+1}}^{\Uparrow F})$$

If for any agent $a \in \mathcal{G}$ the individual orders are total conversely well-founded preorders that satisfy Cond 1 and Cond 2, then it can be shown that $\leq_{\mathcal{G}}^{\Uparrow F}$ satisfies those too (see appendix).

We now have everything at hand to formally capture an extension of Analysis 2 that can deal with blatantly inconsistent fictions.

3.3 Semantics for In_f

We discuss two alternative proposals for a semantics. Let $\mathfrak{M} = \langle W, \leq_a, R_C, V \rangle_{a \in \mathcal{A}, C \in \mathcal{C}}$ be a multi-agent plausibility model and let $\mathcal{G} \subseteq \mathcal{A}$ be the community of origin of a fiction f.

Following Lewis's original formulation (for all common belief worlds, there is a world ...), a truth clause for In_f on a multi-agent plausibility model $\mathfrak{M}$ could be as follows, where $w \in P$:

$\mathfrak{M}, w \vDash^+ In_f, \varphi$ iff. $[\forall v \in |CB_{\mathcal{G}}^w| \exists u \in best_{\mathcal{G}}^{\Uparrow F} W(\mathfrak{M}^{\Uparrow F}, u \vDash^+ \varphi \;\&\; v \leq_{\mathcal{G}}^{\Uparrow F} u \;\&\; (\forall s \in W : s \nvDash^+ \varphi \Rightarrow s <_{\mathcal{G}}^{\Uparrow F} u))]$ In words, for any common belief world v of the community of origin of f, there is a world u that is among the most plausible worlds after revising by F, which makes φ true and is more plausible than any f-world s that does not make φ true. Plausibility here is the group plausibility ordering obtained by merging the individual agents' plausibility orders based on the hierarchy of the agents.

However, since u is going to be among the most plausible worlds with respect to all the worlds after revision, the condition $v \leq_{\mathcal{G}}^{\Uparrow F} u$ is redundant and the quantification over common belief worlds becomes superfluous and the truth clause breaks down to:

$\mathfrak{M}, w \vDash^{+} In_f, \varphi$ iff. $\exists u \in best_{\mathcal{G}}^{\Uparrow F} W(\mathfrak{M}^{\Uparrow F}, u \vDash^{+} \varphi \;\&\; (\forall s \in W : s \nvDash^{+} \varphi \Rightarrow s <_{\mathcal{G}}^{\Uparrow F} u))$

Again this is for $w \in P$ and for $w \in I$, we treat In_f, φ as atomic. The clause for $w \vDash^{-} In_f, \varphi$ is formulated dually.

We do not consider this problematic or to deviate too much from Lewis's original analysis. We simply get his formulation as a special case. On this semantics, we get that for any $w \in P$, $w \vDash^{+} In_f, F$, which is reasonable since in every fiction its explicit content should be true.

Proof. We know $best_{\mathcal{G}}^{\Uparrow F} W$ is non-empty by converse well-foundedness. So we have some world $u \in best^{\Uparrow F} W$. Hence, for any $x \in W : x \leq_{\mathcal{G}}^{\Uparrow F} u$ and so by Cond. 2, $u \vDash^{+} F$. Suppose $s \nvDash^{+} F$. By Cond. 1, $s \leq^{\Uparrow F} u$. Suppose $u \leq^{\Uparrow F} s$ for arbitrary $s \in W$. Then, by transitivity, for any $x \in W$, $x \leq^{\Uparrow F} s$ and by Cond 1., $s \vDash^{+} F$, contradicting our assumption. Hence $u \not\leq^{\Uparrow F} s$ and thus $s <^{\Uparrow F} u$, as required. □

Thus, in particular, if a contradiction is part of the explicit content of the fiction, it is true in the fiction. The case of SB is such a case.

Now, do inconsistent fictions of this particular kind make everything true? We can phrase this as, does the entailment $In_f, (\varphi \wedge \neg\varphi) \vDash^{+} In_f, \psi$ hold for any f and ψ, where $\varphi \wedge \neg\varphi \in F$?[19] It is easy to see that it does not.

Proof. Let $I = \{u\}, P = \{w, v\}, \leq_{\mathcal{G}}^{\Uparrow F} = \{(w,w), (u,u), (w,u), \}$ and $V^{+}(p) = V^{+}(\neg p) = \{u\}$. Then $w \vDash^{+} In_f, (p \wedge \neg p)$. This is because $u \vDash^{+} p \wedge \neg p$ and for any $x \in W$, $x \nvDash^{+} (p \wedge \neg p) \Rightarrow x <_{\mathcal{G}}^{\Uparrow F} u$. The latter holds because neither $u \leq_{\mathcal{G}}^{\Uparrow F} v$ nor $u \leq_{\mathcal{G}}^{\Uparrow F} w$.

$w \nvDash^{+} q$ because there is no most plausible world where q is true, since u is the only most plausible world, and $u \nvDash^{+} q$. □

What about inconsistent fictions where the contradiction is not part of the explicit content? Suppose $w \vDash^{+} In_f, (\varphi \wedge \neg\varphi)$, where $\ulcorner \varphi \wedge \neg\varphi \urcorner \notin F$. Then there is a most plausible world u where $\varphi \wedge \neg\varphi$ is true. For $w \vDash^{+} In_f, \psi$ to hold, there needs to be another most plausible world v where ψ is true which is strictly more plausible than any world where ψ is not true. If u is the only world considered most plausible and ψ does not hold at u, then $w \nvDash^{+} In_f, \psi$. If u is not the only world, it depends on the plausibility ordering of the group whether there is such a world.

Hence, the analysis makes explicit what the trouble for determining the correct output of Lewis's analysis is, namely our judgements which worlds are more plausible or which worlds are supported by a certain interpretation.

[19] We suppose any contradiction can be made explicit in form of some formula $\varphi \wedge \neg\varphi$.

These are choices the modeller makes, maybe based on some empirical data about agents' plausibility orderings or by discussing historical circumstances and then guessing what the orderings of the agents of the community of origin might have been. Thus, the work of the contemporary interpreter can be understood as that of the modeller. To figure out the external variables one has to input to determine what is true in the story. And thus, different theories of interpretation will not only influence the orderings of the agents of the community of origin, but also contemporary theories of interpretation will determine the external variables in different ways. And this seems to in fact model how the correct practice of literary studies should work. To have competing theories of interpretation which then tell us what is true in a story and what is not.

The semantics also still accounts for incompleteness of fiction given some assumptions about the most plausible worlds. Namely, to consider the example from above again, any most plausible $LOTR$-world where e is true will be equally plausible to any most plausible $LOTR$-world where o is true (and vice versa) and so the universal condition does not hold. Hence neither In_{LOTR}, e nor In_{LOTR}, o come out as true.

For propositional variables, we also have $\{In_f, p, In_f, q\} \vDash In_f, (p \wedge q)$.

Proof. Suppose for arbitrary $w \in P$ that $w \vDash^+ In_f, p$ and $w \vDash^+ In_f, q$. Consider the former. Then $\exists u \in best_{\mathcal{G}}^{\Uparrow^F} W$, s.t. $u \vDash^+ p$ and for any $v \in W$: $v \nvDash^+ p \Rightarrow v <_{\mathcal{G}}^{\Uparrow^F} u$. From $w \vDash^+ In_f, q$, we get $\exists s \in best_{\mathcal{G}}^{\Uparrow^F} W$, s.t. $s \vDash^+ q$ and for any $v \in W : v \nvDash^+ q \Rightarrow v <_{\mathcal{G}}^{\Uparrow^F} s$. Now, either $u \vDash^+ q$ or $s \vDash^+ p$. Suppose neither. Then $s <_{\mathcal{G}}^{\Uparrow^F} u$ and $u <_{\mathcal{G}}^{\Uparrow^F} s$. Thus $s \not\leq_{\mathcal{G}}^{\Uparrow^F} u$ and $u \not\leq_{\mathcal{G}}^{\Uparrow^F} s$, contradicting totality of $\leq_{\mathcal{G}}^{\Uparrow^F}$. □

This does not generalize, especially not for negation.

Proof. We can consider three worlds, w, v, u, where u is the only impossible world. We let $w <_{\mathcal{G}}^{\Uparrow^F} v \simeq_{\mathcal{G}}^{\Uparrow^F} u$, and so $u, v \in best_{\mathcal{G}}^{\Uparrow^F} W$. Also we let $u \notin V^+(p \wedge \neg q)$, $u, v \in V^+(\neg q)$ and $v, u \in V^+(p)$. Then $w \vDash^+ In_f, p$ because $v \vDash^+ p$ and $u \vDash^+ p$ because w is the only world where p is not true and it is strictly less plausible than both of the other worlds. For an analogous reason $w \vDash^+ In_f, \neg q$. However $w \nvDash^+ In_f, (p \wedge \neg q)$ because the only most plausible world which makes this true is v but since $v \simeq_{\mathcal{G}}^{\Uparrow^F} u$ and $u \nvDash^+ (p \wedge \neg q)$, it is not the case that $\forall x \in W : x \nvDash^+ (p \wedge \neg q) \Rightarrow x <_{\mathcal{G}}^{\Uparrow^F} v$. □

Thus, fictions which give rise to most plausible impossible worlds can invalidate even simple rules like conjunction introduction if one of the conjuncts is not an atomic formula.

Now, how can we account for the problem from Bonomi & Zucchi (2003)? Since we pay weight to authorial intention and it seems undisputable that Mann did not intend Dora Martin to be of inferior race (even more, probably he intended her not to be of inferior race), and his plausibility ordering is thus prioritized in determining $\leq_{\mathcal{G}}^{\Uparrow F}$, we know that any f-world where p is true, is not going to be among the worlds in $best_{\mathcal{G}}^{\Uparrow F}W$ and hence there is no world among the $best_{\mathcal{G}}^{\Uparrow F}W$ worlds that makes s true. Thus, $In_{Mephisto}, s$ does not come out as true, as required.

If one objects that the previous truth clause is not close enough to Lewis's analysis because it does not explicitly refer to the common beliefs, a second proposal is:

$$w \vDash^{+} In_f, \varphi \text{ iff. for all } v \in best_{\mathcal{G}}^{\Uparrow F}|CB_{\mathcal{G}}^{w}| : v \vDash^{+} \varphi$$

This takes the idea of common beliefs at face value and is not subject to the previous objection of redundant quantification. However, to get the desired result that In_f, F comes out true at any w, we need to restrict the quantification to the f-worlds among the worlds in $best_{\mathcal{G}}^{\Uparrow F}|CB_{\mathcal{G}}^{w}|$:

$$w \vDash^{+} In_f, \varphi \text{ iff. for all } v \in best_{\mathcal{G}}^{\Uparrow F}|CB_{\mathcal{G}}^{w}|(v \vDash^{+} F \Rightarrow v \vDash^{+} \varphi)$$

Again, if we have blatantly inconsistent fictions and the contradiction is in the explicit content, we get that the contradiction is true in the fiction. Also, it is easy to see that $In_f, (\varphi \wedge \neg\varphi) \nvDash^{+} In_f, \psi$. Moreover, we have $In_f, \varphi, In_f, \psi \nvDash In_f, (\varphi \wedge \psi)$:

Proof. Consider an w, u and $u \in I$, s.t. $w \leq_{\mathcal{G}}^{\Uparrow F} u$, $u \in V^{+}(F)$, $u \in V^{+}(p), u \in V^{+}(q), u \in V^{-}(p)$ and $u \in V^{-}(q)$, but $p \wedge q \notin F$. So $u, w \in best_{\mathcal{G}}^{\Uparrow F}|CB_{\mathcal{G}}^{w}|$.[20] Then $w \vDash^{+} In_f, p$ and $w \vDash^{+} In_f, q$ because $u \vDash^{+} p$ and $u \vDash^{+} q$ but since $u \notin V^{+}(p \wedge q)$, we have that there is a world, namely u, in $best_{\mathcal{G}}^{\Uparrow F}|CB_{\mathcal{G}}^{w}|$ that does not make $p \wedge q$ true. □

Note that this depends on the assumption that $p \wedge q \notin F$, which might be debatable, if for example, $p, q \in F$.

Thus, on this semantics, even very simple logical inferences fail. Given a principle of poetic licence, that does not seem surprising because a principle of poetic licence seems to allow for fictions in which very weak logics hold. And hence, the semantics should be able to account for this. However, if we have a consistent F and assume for the most plausible worlds after upgrading to

[20] Since there are no common belief worlds, also w satisfies the clause for membership.

be only possible worlds, by the classicality condition, we get that all classical logical consequences of F hold in the fiction.

Another result of the second proposal is that if we have the set $\mathcal{B}_\mathcal{G} * F = \{\varphi | \forall w \in best_\mathcal{G}^{\Uparrow F} W : w \vDash^\varphi\}$, which we call the 'make-belief set of the group', we have that $w \vDash^+ In_f, \varphi$ entails $\varphi \in \mathcal{B}_\mathcal{G} * F$

Proof. Suppose $w \vDash^+ In_f, \varphi$ and let $v \in best_\mathcal{G}^{\Uparrow F} W$ be arbitrary. By Cond 2, $v \vDash^+ F$ and since $v \in best_\mathcal{G}^{\Uparrow F} W$, also $v \in best_\mathcal{G}^{\Uparrow F} |CB_\mathcal{G}^w|$. Thus, $v \vDash^+ \varphi$ because $w \vDash^+ In_f, \varphi$. Since v was arbitrary, $\varphi \in \mathcal{B}_\mathcal{G} * F$. □

Thus, everything that is true in the fiction is make-believed by the group. However, the other direction does not necessarily hold, that is, not everything that is make-believed is true in the fiction. This makes sense because agents might make-belief φ for which f is incomplete.

Also the objection by Bonomi & Zucchi (2003) can be met by considering the set $best_\mathcal{G}^{\Uparrow F} |CB_\mathcal{G}^w|$. This set is determined by $\leq_\mathcal{G}^{\Uparrow F}$ and as we argued above, this ordering is defined taking into account a priority ordering among the agents. Thus, from a contemporary perspective, the relevant priority ordering prioritizes those agents whose plausibility ordering does not judge worlds where Dora Martin is of inferior race as more plausible than any common belief world. So the worlds in $best_\mathcal{G}^{\Uparrow F} |CB_\mathcal{G}^w|$ won't make it true that Dora Martin is of inferior race. However, from a perspective during the time when *Mephisto* was published, one might argue that the prioritization must be different and that hence Dora Martin is of inferior race in *Mephisto*. What this points to is that it depends on the world of evaluation of In_f, φ what priority ordering on the agents one imposes. Although in the model, truth in fiction depends only on the community of origin, the meta-choices we make on plausibility orders and on priority orders influence what, in the model, comes out as true in the fiction. Thus, truth in fiction is always relevant to the community of interpretation as well, since that community is the community that determines the variables external to the model, which then determine in the model what comes out as true in a fiction.

Since the second proposal allows for failure of conjunction of atomic sentences, one might consider it the better analysis if one wants to allow for fiction which can be completely anarchic with respect to logic and invalidate very simple inferences. Again, given one believes that there are fictions where such simple inferences fail, which in turn is a reasonable assumption if one accepts a principle of poetic licence. Moreover, the second proposal seems closer to Lewis's analysis by still explicitly featuring the common be-

lief worlds. We conclude the chapter by a discussion of the aforementioned AGM axioms considering the revision operation $\Uparrow F$.

3.4 AGM-Axioms Again

We consider the case in which the belief set $\mathcal{B} = \{\psi | \forall w \in bestW : w \vDash^{+} \psi\}$ of an agent is softly upgraded by a sentence φ in the explicit content F of a fiction f. The set $\mathcal{B} * \varphi$ is the set of all those sentences which are true at all the best worlds after upgrading with φ, i.e. it is the set $\{\psi | \forall w \in best^{\Uparrow \varphi} : w \vDash^{+} \psi\}$. We call this the 'make-belief set'. The set $\mathcal{B} + \varphi$ is the set of classical logical consequences of $\mathcal{B} \cup \{\varphi\}$.

1. $\mathcal{B} * \varphi$ is consistent and logically closed: $\mathcal{B} * \varphi$ need not be consistent, since we assumed there are blatantly inconsistent fictions (or $\mathcal{B}$ is already inconsistent) and hence it can be that $\varphi = \psi \wedge \neg\psi$. However, if we are given the logic of the fiction by its explicit content, we can require that $\mathcal{B} * \varphi$ is closed under *that* logic. So, for example, in *Sylvan's Box*, most likely some Priest-style paraconsistent logic L holds. Hence $\mathcal{B} * \varphi$ should be closed under L.

   However, given the logic of the fiction is not part of its explicit content, it is not clear how to have logical closure. We discuss this in section 4.2.

2. $\varphi \in \mathcal{B} * \varphi$: we have that for any $\varphi \in F$, $\varphi \in \mathcal{B} * F$. Let $v \in best^{\Uparrow F} W$. Then $v \vDash^{+} F$ and hence $v \vDash^{+} \varphi$ for any $\varphi \in F$. This is reasonable because revising with F should yield make-believing F. If we accept the following version of Cond 2:

   Cond 2$_\varphi$: $\forall \varphi \in F \forall t \in W : [(\forall s \in W : s \leq^{\Uparrow \varphi} t) \Rightarrow t \vDash^{+} \varphi)]$

   then the axiom is satisfied. This condition seems reasonable since it expresses that after upgrade with φ all most plausible worlds should be φ-worlds.

3. $\mathcal{B} * \varphi \subseteq \mathcal{B} + \varphi$: the set $\mathcal{B} * \varphi$ might not be closed under classical logical consequence and thus it is going to be a subset of $\mathcal{B} + \varphi$.[21]

[21] In general, if we do not restrict ourselves to the language we introduced, it might turn out that $\mathcal{B} * \varphi$ is closed under a logic stronger than classical logic. Then the question would be whether adding φ to classical logic and close under classical logical consequence would lead to an inconsistent set or whether it would lead to just the logic under which $\mathcal{B} * \varphi$ is closed. In those cases the axiom holds. But is it possible to have $\mathcal{B} * \varphi$ be closed under

This requirement is reasonable since we have seen that we do not want contradictions in a fiction to entail everything.

4. If $\neg\varphi \notin \mathcal{B}$, then $\mathcal{B} + \varphi \subseteq \mathcal{B} * \varphi$: since $\mathcal{B}$ can be inconsistent and not closed under classical logic, we have an easy counterexample. We have worlds and a plausibility ordering such that $\mathcal{B} = \{p, \neg p\}$ and consider q. We have $\neg q \notin \mathcal{B}$ and we can have the upgraded plausibility order result in $\mathcal{B} * q = \{p, \neg p, q\}$. Hence $\mathcal{B} + q \supset \mathcal{B} * q$.

   Since $\mathcal{B}$ is not closed under classical logic, even if it is consistent, we can have a plausibility order such that $\mathcal{B} = \{p \to q, p\}$ and then $\neg q \notin \mathcal{B}$ but $\mathcal{B} + \neg q \not\subseteq \mathcal{B} * \neg q$.

   However, as pointed out in the introduction, it would be reasonable to have, for φ consistent with the beliefs of the agent, that its logical consequences are imported into the make-belief set.[22] Whetehr this is satisfied depends on the plausibility ordering of the agent.

5. $\mathcal{B}*\varphi$ is inconsistent only if φ is inconsistent: since $\mathcal{B}$ can be inconsistent, this can be extended to $\mathcal{B} * \varphi$ is inconsistent only if φ is inconsistent or $\mathcal{B}$ is inconsistent. This is a reasonable requirement.

6. If φ is logically equivalent to ψ, then $\mathcal{B} * \varphi = \mathcal{B} * \psi$: if φ and ψ are logically equivalent according to the logic of the fiction, then this postulate holds since the plausibility ordering will be the same for both of them. In general, however, the fiction can exactly be about φ and ψ being classically logically equivalent but not logically equivalent wrt the logic in our make belief scenario. We can have an intuitionistic fiction in which we make belief $p \to p$ but not $\neg p \vee p$.

7. $\mathcal{B} * (\varphi \wedge \psi) \subseteq (\mathcal{B} * \varphi) + \psi$: if any of $\mathcal{B}, \varphi, \psi$ is inconsistent, then this postulate trivially holds because $(\mathcal{B} * \varphi) + \psi$ will be the set of all sentences.

   Suppose they are all consistent and suppose $\chi \in \mathcal{B} * (\varphi \wedge \psi)$ but $\chi \notin (\mathcal{B} * \varphi) + \psi$. Then χ is neither a classical logical consequence of $\mathcal{B} * \varphi$ nor of ψ. So there is a world $w \in P$, such that $w \vDash^{+} B * \varphi$ and $w \nvDash^{+} \chi$ and also there is a world $v \in P$, s.t. $v \vDash^{+} \psi$ and $v \nvDash^{+} \chi$. If any of those worlds is in $best^{\Uparrow(\varphi \wedge \psi)}W$, then $\chi \notin \mathcal{B} * (\varphi \wedge \psi)$, contradicting out assumption.

a logic L and $\mathcal{B} + \varphi$ is neither L nor inconsistent? Since the only reason $\mathcal{B} * \varphi$ should be closed under L is φ, it does not seem to be possible.

[22] This could be achieved by imposing more restrictions on the plausibility ordering.

If neither is in $best^{\Uparrow(\varphi\wedge\psi)}W$, we can construct a countermodel by having $u \in best^{\Uparrow(\varphi\wedge\psi)}W$, s.t. $u \vDash^{+} \chi$. However, it might be reasonable, as in axiom 4, in the case the beliefs and φ and ψ are consistent, to assume that the plausibility orderings of the agents are such that the axiom holds.

8. If $\neg\psi \notin \mathcal{B}*\varphi$, then $(\mathcal{B}*\varphi)+\psi \subseteq \mathcal{B}*(\varphi\wedge\psi)$: if $\mathcal{B}$ or $\mathcal{B}*\varphi$ is inconsistent and not logically closed, this might fail because $(\mathcal{B}*\varphi)+\varphi$ is going to be the set of all sentences and $\mathcal{B}*(\varphi\wedge\psi)$ might not be logically closed and hence might not contain all sentences, although inconsistent.

   Again, in the case we are dealing with consistent beliefs it is reasonable to have plausibility orderings that allow for validation of this axiom.

What the discussion shows is that the validity of the axioms depends on the plausibility ordering of the agents. We think this is a good result because whether one incorporates new information or, in our case, adds logical consequences of a sentence occurring in a fiction, depends on how plausible this addition is. In our case the plausibility is based on certain pragmatic criteria such as genre conventions. Moreover, it seems reasonable to us, following the idea of a cooperation principle between narrator and reader, that the reader assumes the narrator to be a classical reasoner unless there is reason to believe otherwise.

Chapter 4

Discussion, Further Research, Conclusion

4.1 Philosophical Issues

Philosophically, what could be considered troublesome about our analysis is the assumption that we can determine the explicit content of a fiction, the use of impossible worlds and the use of the plausibility ordering.

Explicit Content First, one could wonder whether there is such a thing as the explicit content of a fiction and if it exists, what it would be like. This is an ontological or metaphysical concern. Concerning the latter, we would say that it is a set of possible worlds, namely, following Lewis, those worlds where the fiction is told as known fact rather than fiction. This content can be expressed by a (finite) set of sentences. Thus, the question breaks down to whether there are possible worlds and if so, what they are. Answering this question is clearly outside the scope of this work.

Second, it might be questioned whether we can in fact determine the explicit content of a fiction. That is, can we ever know it? This is an epistemic objection. Given the explicit content of a fiction is a set of worlds, the objection again breaks down to one to possible worlds in general: how can we come to know anything of or about possible worlds? Since there is a vast debate going on about the epistemology of possible worlds, answering this objection in detail also lies outside the scope of this work.[1]

In any case, if one objects against the assumption of an explicit content of a fiction which is expressible by a set of sentences, it seems to us one has to object to possible worlds in general.

[1] See, for example, Vaidya (2015) for an overview.

Impossible worlds Following Berto (2013) impossible worlds can be understood in at least four different ways:

1. If possible worlds are ways the world could have been, then impossible worlds are ways the world could not have been.

2. Worlds where (your favourite) logic fails.

3. Worlds where classical logic fails.[2]

4. Worlds making contradictions true.

In our semantics, we have not specified which kind we use and it seems to us that for any interpretation there is, or could be, some fiction for which one of the interpretations is suited better than the others: if we are dealing with fictions in which H_2O is not water, it seems we would have to consider worlds of the first kind. Fictions whose logic is classical would, for, say, a paraconsistent logician, call for worlds of the second kind. We could also have fictions where intuitionistic logic is the right logic and thus, they would need models with impossible worlds of the third kind. Blatantly inconsistent fictions would need impossible worlds of the fourth kind.

The major argument by Berto (2013) for using impossible worlds is very similar to Lewis's (1986): they are very useful in modelling, as we can see in our approach. But also in modelling non-logically omniscient agents or just in multi-agent scenarios where we have to consider worlds which one agent considers possible but another agent already knows to be impossible, impossible worlds have played important roles.

What about the ontological status of impossible worlds? Lewis (1986) is a proponent of modal realism, that is that possible worlds exist exactly like our world as concrete entities but spatio-temporally completely isolated from each other. One way, as for example Yagisawa (1988) pursues, is to add impossible worlds to this ontology. This might make everyone stare even more incredulously than they already are when faced with a proponent of modal realism. On the other hand, since modal realism is considered a rather drastic view on the ontology of modality, one might say that adding impossible worlds to it is not such a huge leap and doesn't make it more unattractive than it already is.

There are various other options to account for the ontological status of impossible worlds, surveyed by Berto (2013), of which the most promising to us seems to treat impossible worlds as linguistic ersatz constructions.

[2]These are only impossible if one accepts classical logic as the right logic and that if classical logic is the right logic, it is necessarily so.

However, this would not be acceptable for our framework because we think of propositions as sets of worlds. But on linguistic ersatzism, worlds are sets of sentences (or propositions) and thus this would be circular. We are generally fine with treating worlds, impossible or possible, as abstract objects in our ontology, which clearly would entail that there impossible objects in our ontology. Thus, one might maybe go for a Meinongian view on ontology for the philosophical interpretation of our models.

Another option is to treat talk of (im)possible worlds as pretence and be a fictionalist about modality.

A third option is to change the framework in such a way that it allows for an ersatzist interpretation. For that we had to have V be a function of worlds and sentences that assigns truth values $0, 1$ to propositional atoms, define truth at possible worlds recursively and at impossible worlds we treat any sentence as atomic. Then a world is a pair of sets of sentences, the ones which are assigned 1 and the ones which are assigned 0. We chose the present formulation because it is closer to the standard accounts which have valuation functions assigning sets of worlds and not truth values.

Plausibility Orders One of the major objections, analogously to objections about Lewis's analysis of counterfactuals, could be that the notion of similarity, or in our case, plausibility, is rather vague and also assuming that agents come equipped with such an order is a fairly strong assumption. We have three arguments why we do not think that assuming a plausibility ordering on worlds is problematic. First, it is fairly standard in contemporary epistemic logic to assume such orderings. If one objects to plausibility orderings in general, one has to object to those approaches too.

Second, and this is more of a conjecture, we seem to be able to judge whether two scenarios are equally plausible or one is more plausible than another, even if confronted with impossibilities, especially in a make-believe setting. This would be worth some empirical enquiry and is based on an observation that gives rise to our third argument.

It seems to us that in the debate about truth in fiction, everyone*is* already implicitly assuming some plausibility ordering on (im)possible worlds. Any analysis of truth in fiction is evaluated against our intuitions about what is true in a fiction. For example, as we have seen, the crucial premise in Proudfoot's (2006) argument was about intuitions concerning truth in fiction. But these intuitions and their accuracy are never justified when used to evaluate the analysis. We compare certain scenarios, where we pretend that the story obtains, and then compare whether it is intuitively true in this scenario that φ or not. But how can we say that it is intuitively not true

in the Holmes stories that A-A philosophy takes a scientific turn in the 20th century? Or that it is intuitively true in the Holmes stories that Holmes, if reasoning deductively, he is reasoning classically?[3] It is because we, as interpreters, consider one interpretation of the stories more plausible than another. Or, to stay closer to Lewis, we, as interpreters, believe that the community of origin of f considered one interpretation more plausible than another. Hence, we claim, this appeal to intuition in the debate is an appeal to plausibility among interpretations and hence an appeal to a plausibility ordering on worlds. Clearly, if one claims a good interpretation has to give us what is true in the fiction, this will become circular. But we believe that the notion of truth in fiction is inherently dependent on ones theory of interpretation. On Lewis's analysis it depends on the theory of interpretation the group of origin is using, since this determines their plausibility ordering.

So, where does the plausibility ordering come from? It comes from our intuition about truth in fiction to which almost everyone in the debate appeals for judging analyses of truth in fiction.

4.2 Further Research

The issue of truth in fiction is connected to various open branches of research. First, the question clearly is, which of the restrictions we imposed could be relaxed, that is consider non-literary fiction but also take into account unreliable narration.

Second, our approach was mainly semantic. Hence, a natural question is, if there is a reasonable axiomatization governing In_f. Especially, it is interesting to consider the distinction between the logic within the scope of the operator, as for example whether the inference from $In_f, (p \wedge q)$ to In_f, p should be valid. But also the logic governing imports into the scope, as for example whether the inference from In_f, p and In_f, q to $In_f, (p \wedge q)$ should be valid.

Third, it would be interesting to explore connections to theories of pretence in cognitive science. That is, whether our model could also be considered a formal model of certain cognitive models. We will sketch parallels between our model and the model by Nichols (2004).

4.2.1 Extending the Analysis

Non-Literary Fiction Extension to films or plays seems unproblematic in most cases since the content of films can be expressed by writing a drama

[3] In fact, Holmes often reasons inductively.

which is a form of literary fiction. But there are cases where it is nearly impossible to capture certain aspects of a film in a narrative. For example, in *Ocean's Twelve*, Julia Roberts plays Tess Ocean. Tess Ocean is, in the film, mistaken for Julia Roberts. This creates a humorous element which seems impossible to capture by a drama in a similar way, since the whole joke is based on Julia Roberts actually being shown on screen and the character played by her being mistaken for her.

What about interactive fiction, such as video games? If the game has a story line or plot, *prima facie* there is no objection to using our semantics. One potential worry might be if games have different storylines which are mutually exclusive. Are both of the storylines true in the game and thus the fiction is inconsistent? This does not seem to be the case. There are two ways out. Either we consider playing each storyline as a different fiction or we use Lewis's method of intersection and treat playing each storyline as playing one (maximally consistent) fragment of the story. If each storyline is inconsistent, the latter is again not applicable. But it seems the former is accurate. Different storylines give different stories and hence different fictions, which are comprised under the name of one game.

If we consider interactive video games which (supposedly) do not have a storyline, for example games like chess, our semantics can still be applied if we treat the explicit content as the set of rules and the moves which are made. However, one might debate whether such video games are fictions in the first place.

There are various other kinds of fictions, such as paintings, maybe pieces of music, as Walton (1990) claims, even cloud constellations can be fictions. Working out in detail whether our extension of Lewis's would work for those and if not how one can account for truth in those fictions, is left for further research.

Unreliable Narration Unreliable narration featured in some counterexamples above and we excluded it in our consideration. Characterizing unreliable narration has been of concern in literary studies and there is no clear cut definition available.[4] Thus, whether our semantics can be extended, might depend on the treatment of unreliable narration in literary studies.

Heyd (2011) gives a pragmatic modelling of unreliable narration taking into account Gricean maxims. Roughly, we would be treating narrators similarly to unreliable conversational partners and if they violate Gricean maxims without giving rise to an implicature, that is if they violate the cooperation principle, we would be facing an unreliable narrator.

[4]An overview can be found in Issue (2011).

Since narrators often give more information than needed, the maxim of quantity might be relaxed in cases of narration. We could consider the models provided by Heyd as contributing to the plausibility ordering of the agents and also in determining what is the explicit content of the fiction. Then, our analysis can be applied since the unreliable narrator has been dealt with in the pragmatic model.

4.2.2 Finding a Logic for In_f?

It would be nice to find an axiomatization for the operator In_f. As we pointed out above, semantically, certain logical entailments hold and having axioms capturing those would be desirable. Moreover, finding reduction axioms for In_f valid on the semantics would also be of interest. For inferences that do not distribute over In_f, we have that many of the usual inference rules will be valid because our logical consequence relations is defined wrt possible worlds and we imposed the classicality condition on those.

For example, if we consider modus ponens $In_f, \varphi; (In_f, \varphi \supset \psi \vdash \psi)$, where $\vdash$ is our syntactic consequence relation, we can see that it is sound wrt our logical consequence relation:

Proof. We consider an arbitrary model $\mathfrak{M} = \langle W, \leq_a, R_C, V\rangle_{a\in\mathcal{A},C\in\mathcal{C}}$ and an arbitrary world $w \in P$ and suppose $w \vDash^+ In_f, \varphi$ and $w \vDash^+ In_f, \varphi \supset \psi$. Then by the clause for implication, $w \vDash^+ In_f, \varphi$ implies $w \vDash^+ \psi$. □

Similarly, we get that conjunction elimination and conjunction introduction hold. This is due to the fact that we defined logical consequence wrt possible worlds.

For the necessitation rule for In_f, if $\vdash \varphi$, then $\vdash In_f, \varphi$, it is easy to see that the rule is not sound by considering an impossible world in $best^{\Uparrow F}|CB^w_{\mathcal{G}}|$ that does not make φ true.

However, it seems difficult to give a logic that captures inferences that distribute over In_f, as for example from $In_f, \varphi \wedge In_f$ to $In_f, (\varphi \wedge \psi)$. As we have seen, this inference might fail. Nevertheless, most fictions will usually allow for this inference step. To be able to judge whether this inference is licensed, one needs to know whether this inference is also licensed in the fiction. For that, one needs to know the logic of the fiction. This might be difficult to find if it is not part of the explicit content of the fiction what its logic is.

One way to address this issue is to have, for simplicity, classical logic as default logic for any fiction. When we engage with a blatantly inconsistent fiction we deviate as minimally as possible from classical logic to a paraconsistent logic, where "as minimally as possible" could be thought of as the

position of f's logic in the ordering of all logics. For example, if f has neither classical nor intuitionistic logic, but an intermediate logic, we can then, in principle, locate it between the two. Of course there are infinitely many but we might be able to give an interval in which we find the fiction's logic.

Another way would be to say that one allows for "miracles" in a Lewisean way. That is, instead of supposing that certain laws, in this case laws of logic, fail in general throughout the fiction, one allows for certain instances where the laws fail. So one has a miracle, since a law of logic is violated. This respects Lewis (1979) requirements on the similarity/plausibility ordering that one should avoid big widespread violations of laws and maximize the perfect match between worlds. We presume classical logic throughout the fiction but then a miraculous contradiction appears. It's true in the fiction but, for example, we only license classical inferences from each of its conjuncts.

To us it seems that finding an interesting axiomatization that captures inferences within the scope of the operator is rather difficult, if not even impossible if one takes into account that there are blatantly inconsistent fictions. Moreover, for some of these fictions, their logics might be so weak that they might not even be considered interesting logics.

4.2.3 Connections to Cognitive Science

The cognitive theory of pretence presented by Nichols & Stich (2000), Nichols & Stich (2003) and Nichols (2004) was one of our inspirations to explore the connections between belief revision, make-belief/pretence and truth in fiction.[5] We are going to present their general idea and connect it to the formal framework we used. This section is not intended to claim that their theory is a correct account of pretence, nor that our formal approach is an accurate logical modelling of their theory. It is meant to indicate that there are connections between models from cognitive science and logical models in the debate about truth in fiction and pretence and if the cognitive theory of pretence is correct that it provides a good empirical background for our logical model.

Nichols & Stich (2000, p. 121) start out by assuming, claiming this to be a rather standard assumption, that the basic cognitive architecture in normal humans is divided between beliefs and desires: 'the mind contains two quite different kinds of representational states, beliefs and desires'. Their second assumption is 'to have a belief [...] with a particular content is to have a representational token with that content stored in the functionally

[5]Thanks to Derek Matravers for pointing us to this work during a discussion.

appropriate way in the mind'.[6]

The representational tokens which correspond to the beliefs of the agent are in, what they call, 'the belief box'. On our formal account, this is the set of all the most plausible worlds.

Moreover, they assume the existence of a possible world box (PWB). This also contains representation tokens. These, so Nichols & Stich (2000, p. 122), 'represent what the world would be like given a certain set of assumptions that we may neither believe to be true nor want to be true. The PWB is a work space in which our cognitive system builds and temporarily stores representations of one or another possible world'. Although called *possible* world box, they remark:

> We are using the term 'possible world' more broadly than it is often used in philosophy (e.g. Lewis, 1986), because we want to be able to include descriptions of worlds that many would consider *impossible*. For instance, we want to allow that the Possible World Box can contain a representation with the content *There is a greatest prime number*.

The (PWB) is supposed to *contain* a representation with a particular *content*, namely the content of the proposition that there is a greatest prime number.

Let us for a moment consider the case of something possible, that of the content of the proposition that snow is white, *Snow is white*. If we say that the content of this proposition is a set of worlds, for example the set of worlds where the proposition is true (or the sentence 'snow is white' expresses a true proposition), we must have, due to the *contain* condition that the possible world box is not a set of worlds, but a set of sets of worlds, namely for each propositional content the corresponding set of possible worlds which is the representation of the content. Thus, our formal semantics with the set W of (im)possible worlds does not exactly correspond to the PWB. What corresponds to PWB is the structure $(W, V, Prop)$, or a restriction of it, since V assigns sets of worlds to propositions in $Prop$. We will give Nichols & Stich's (2000, p. 122ff.) description of how a pretence episode works taking into account the structures they posit and relate it to the structures or operations in our framework:

> Early on in a typical episode of pretense, our theory maintains, one or more initial pretense premises are placed in the PWB workspace. [...] What happens next is that the cognitive system starts to fill the PWB with an increasingly detailed description

[6] This seems very close to a Humean position.

> of what the world would be like if the initiating representation were true.[...]
>
> How does this happen? How does the pretender's cognitive system manage to fill the PWB with representations that specify what is going on in the pretense episode? One important part of the story, on our theory, is that the inference mechanism, *the very same one that is used in the formation of real beliefs*, can work on representations in the PWB in much the same way that it can work on representations in the Belief Box. In the course of a pretense episode, new representations get added to the PWB by *inferring* them from representations that are already there. But, of course, this process of inference is not going to get very far if the only thing that is in the PWB is the pretense initiating representation. [...] In order to fill out a rich and useful description of what the world would be like if the pretense-initiating representation were true, the system is going to require lots of additional information. Where is this information going to come from? The obvious answer, we think, is that the additional information is going to come from the pretender's Belief Box. So, as a first pass, let us assume that the inference mechanism elaborates a rich description of what the pretend world would be like by taking both the pretense-initiating representations *and* all the representations in the Belief Box as premises. Or, what amounts to the same thing, let us assume that in addition to the pretense initiating premise, the cognitive system puts the entire contents of the Belief Box into the Possible World Box. [Since this can lead to contradictions,] there must be a cognitive mechanism (or a cluster of them) that subserves this process. We will call this mechanism the *UpDater*. [...] The UpDater goes through the representations in the PWB eliminating or changing those that are incompatible with the pretense premises.

So, pretence on their account starts out with an initial set of pretence premisses, which we can, in our framework, capture by the explicit content of the fiction. On their account, these pretence premises determine the first worlds that will enter the PWB. On our account, this is the step of determining the f-worlds and to make them the most plausible worlds.

Nichols & Stich (2000) then also require to add worlds which make consequences which have been drawn from the pretence premisses true. Since they do allow for impossible worlds in PWB, it seems reasonable that they allow for inconsistent pretence premises. Thus, what kind of inferences are

licensed will be determined by the logic of the pretence episode. Clearly, in the inconsistent case, classical logic does not seem to be the right choice. Thus, their theory will also have to say something about how the correct logic for the pretence episode has to be determined if it is not part of the pretence premisses what the logic of the episode is.

What is also added to the PWB on Nichols & Stich's (2000) account are *all* the belief worlds. This might clearly lead to inconsistencies one does not want to have, for example that the banana which is pretended to be a telephone is a banana and a telephone at the same time, which is not an accurate representation of how pretence works in such a case.

Their proposed solution is to posit another cognitive structure, namely an UpDater. This UpDater is immediately activated when the pretence episode is initiated. They claim that this updater works on the PWB in the same way as it would on beliefs. What the UpDater ensures is that everything that is incompatible with the pretence premisses is deleted. On our account, this seems to be captured by the soft upgrade with the explicit content of the fiction which reorders the worlds wrt a new plausibility ordering and which ensures that all most plausible worlds are going to be f-worlds.

The main difference concerning the updater is that on our framework, all worlds are kept in the model, whereas on their account, the worlds are deleted from the PWB. We think that keeping the worlds and allowing the UpDater to just reorder worlds guided by plausibility considerations can help in explaining how we can quickly switch between pretence episodes and reality. If asked whether we can call someone with a banana, we will answer 'no' because we immediately update our world ordering from 'this banana is a telephone' back to reality.

Thus, we think our formal framework has an interesting correspondence to Nichols & Stich's (2000) account of pretence and that their theory provides an interesting background theory from cognitive science for our logical models.

4.3 Conclusion

In the introduction, we mentioned the following four desiderata an analysis of truth in fiction should ideally account for:

1. Explicit truths
2. Import of background knowledge/belief
3. logical consequence
4. inconsistent fictions, without trivializing

Lewis (1978) accounted for 1., 2. and 3. but had trouble with 4., especially when it came to blatantly inconsistent fiction.

Our extension of Lewis's (1978) Analysis 2 can account for 1., since we have In_f, F as a logical truth. We can account for 4. if a contradiction is part of the explicit content because we can give counterexamples to the claimed entailment $In_f, (\varphi \wedge \neg\varphi) \vDash \psi$. In case it is not part of the explicit content, the relevant plausibility order on the model for the particular fiction will determine whether the entailment holds or not.

Moreover, we account for 2. pragmatically with the plausibility ordering. We can compare $\mathcal{B}$ and $\mathcal{B}^{\Uparrow F}$. The former is the belief set, the latter the make-belief set. And depending on the plausibility ordering and its softly upgraded version, beliefs might end up also being in the make-belief set.

Also 3. can be dealt with, based on the plausibility ordering. However, for this, we need to be able to determine the logic of the fiction, which might be difficult if it is not part of the explicit content which logic that is.

The main differences between our extension and Lewis's original proposal are that we a) use impossible worlds, b) put way more emphasis on the ordering of the worlds and changed the interpretation to plausibility rather similarity orderings, c) take into account common beliefs rather than overt beliefs and thus have a stricter notion of truth in fiction.

Moreover, since our analysis incorporates plausibility orderings of agents, our interpretation of Lewis's analysis is more pragmatic and makes truth in fiction rather group-dependent. Lewis seems to believe that there is a fact of the matter wrt truth in fiction. Although this might be true for the explicit content, in general it seems a rather strong assumption, since we have seen that fiction is usually incomplete. We have also remarked that we capture the different judgements about truth in fiction resulting from different theories of interpretation because the modeller's theory of interpretation will influence, what she thinks the plausibility orderings of the agents in the community of origin might be (might have been).

The semantic analysis we provided was concerned with a narrow class of fictions: reliable narrative fiction. Extending it to cases of unreliable fiction and also to different kinds of fiction, seems possible, however needs to be investigated further. Since we only provided only a semantic analysis, it is also of concern what inference rules are governing In_f. We argued that the logic of f is going to be a good candidate, however it remains problematic to determine this logic if it is not part of the explicit content of f what this logic is.

Appendices

Appendix A

Proofs

Claim: $\mathcal{S} = \{w^{\leq} | w \in W\}$ satisfies (S1) - (S3):

Proof.

(S1) Let $w^{\leq}, v^{\leq} \in \mathcal{S}$ be arbitrary. Suppose for reductio $w^{\leq} \not\subseteq v^{\leq}$ and $v^{\leq} \not\subseteq w^{\leq}$. Then there is $s \in w^{\leq}$, s.t $s \notin v^{\leq}$ and so $w \leq s$ and $v > s$. Also there is $t \in v^{\leq}$, s.t $t \notin w^{\leq}$ and so $v \leq t$ and $w > t$. Since $\leq$ is total, $s \leq t$ or $t \leq s$. If $s \leq t$, then $w \leq s < v \leq t$. Hence, by transitivity, $w \leq v$. But then for any $x \in v^{\leq} : w \leq x$, by transitivity. And thus, for any $x \in v^{\leq} : x \in w^{\leq}$, i.e. $v^{\leq} \subseteq w^{\leq}$, contrary to our assumption. If $t \leq s$, by a similar argument, $w^{\leq} \subseteq v^{\leq}$, contradicting our assumption. Hence (S1) holds.

(S2) Let $^{P} = w^{\leq} \in \mathcal{S}$ be an arbitrary non-empty subset of W. We let $S = max_{\leq} w^{\leq}$. This exists and is non-empty because $\leq$ is conversely well-founded. Let $S' = v^{\leq} \in \mathcal{S}$ be arbitrary.

($\Rightarrow$): Suppose $w^{\leq} \cap v^{\leq} \neq \emptyset$. Then there is s, s.t. $w \leq s$ and $v \leq s$. Let $t \in max_{\leq} w^{\leq}$ be arbitrary. Hence for any $x \in w^{\leq}$, we have $x \leq t$. Thus, in particular $s \leq t$. But since $v \leq s$ by transitivity of $\leq$, $v \leq t$ and hence $t \in v^{\leq}$.

($\Leftarrow$): Suppose $max_{\leq} w^{\leq} \subseteq v^{\leq}$. Since $max_{\leq} w^{\leq} \neq \emptyset$, by converse well-foundedness, we consider some arbitrary $t \in max_{\leq} w^{\leq}$, which is also in $v^{\leq}$. Note that $\leq$ is reflexive and thus $w \in w^{\leq}$. Thus, since for any $x \in w^{\leq} : x \leq t$, we particularly have $w \leq t$ and hence $t \in w^{\leq}$ (and hence $max_{\leq} w^{\leq} \subseteq w^{\leq}$) and thus $w^{\leq} \cap v^{\leq} \neq \emptyset$.

(S3) As noted in proof of (S2), we have $max_{\leq} w^{\leq} \subseteq w^{\leq}$, for arbitrary $w^{\leq} \in \mathcal{S}$. And since $max_{\leq} w^{\leq} \neq \emptyset$, also $w^{\leq} \neq \emptyset$. So all spheres are non-empty.

$\bigcup \mathcal{S} = \bigcup \{w^{\leq} | w \in W\} = \bigcup_{w \in W} \{w^{\leq}\} = W$ by reflexivity of $\leq$. $\square$

Claim: The relation $\leq_{\mathcal{G}}^{\Uparrow F}$ satisfies totality, converse well-foundedness (cfw), reflexivity and transitivity, Cond 1 and Cond 2, given all individual relations for any $a \in \mathcal{G}$ satisfy these conditions. We omit $\Uparrow F$ for better legibility:

Proof. By induction on the size of $\mathcal{G}$.

Base case: $\mathcal{G} = \{a_0\}$. This is trivial since it is just $\leq_{a_0}$ again.

Induction hypothesis (IH): for groups up to size n, the claim holds.

Induction step: Let $\mathcal{G}$ be a group of agents of size n and $\mathcal{G}_1$ be $\mathcal{G} \cup \{a_{n+1}\}$ and assume that $\leq_{a_{n+1}}$ satisfies the respective conditions too, as we required for any of our models. We check that for $\leq_{\mathcal{G}_1}$ the claim holds. We write $\leq_{\mathcal{G}_1}$ as $\leq_{\mathcal{G}/a_{n+1}}$.

Reflexivity: $w \simeq_{\mathcal{G}} w$ and $w \leq_{a_{n+1}} w$ by reflexivity of $\simeq_{\mathcal{G}}$ (IH) and $\leq_{a_{n+1}}$.

Transitivity: Suppose $w \leq_{\mathcal{G}/a_{n+1}} v$ and $v \leq_{\mathcal{G}/a_{n+1}} s$. We need to show that $w \leq_{\mathcal{G}/a_{n+1}} s$, i.e. $w <_{\mathcal{G}} s$ or ($w \simeq_{\mathcal{G}} s$ and $w \leq_{a_{n+1}} s$). There are four cases:

1. $w <_{\mathcal{G}} v$ and $v <_{\mathcal{G}} s$. By transitivity of $<_{\mathcal{G}}$, we have $w <_{\mathcal{G}} s$.
2. $w <_{\mathcal{G}} v$ and $v \simeq_{\mathcal{G}} s$ and $v \leq_{a_{n+1}} s$. So $w \leq_{\mathcal{G}} s$ by transitivity of $<_{\mathcal{G}}$ and $\leq_{\mathcal{G}}$.
3. $w \simeq_{\mathcal{G}} v$ and $w \leq_{a_{n+1}} v$ and $v <_{\mathcal{G}} s$. Then by transitivity of $\leq_{\mathcal{G}}$ and $<_{\mathcal{G}}$, $w <_{\mathcal{G}} s$.
4. $w \simeq_{\mathcal{G}} v$ and $w \leq_{a_{n+1}} v$ and $v \simeq_{\mathcal{G}} s$ and $v \leq_{a_{n+1}} s$. Then by transitivity of $\leq_{\mathcal{G}}$, we have $w \simeq_{\mathcal{G}} s$. And by transitivity of $\leq_{a_{n+1}}$ we have $w \leq_{a_{n+1}} s$.

Totality: Suppose for contradiction that $s \not\leq_{\mathcal{G}/a_{n+1}} t$ and $t \not\leq_{\mathcal{G}/a_{n+1}} s$ for some $s, t \in W$. Again we have four cases:

1. $s \not<_{\mathcal{G}} t$ and $t \not<_{\mathcal{G}} s$ and $s \not\simeq_{\mathcal{G}} t$. Since $s \not\simeq_{\mathcal{G}} t$, either $s \not\leq_{\mathcal{G}} t$ or $t \not\leq_{\mathcal{G}} s$. If the former, since $\leq_{\mathcal{G}}$ is total, $t \leq_{\mathcal{G}} s$. But since $t \not<_{\mathcal{G}} s$, either $t \not\leq_{\mathcal{G}} s$ or $s \leq_{\mathcal{G}} t$. If the former, we have a contradiction. If the latter, we have $s \leq_{\mathcal{G}} t$ and $t \leq_{\mathcal{G}} s$, hence $s \simeq_{\mathcal{G}} t$, contradicting our assumption. Analogously for $t \not\leq_{\mathcal{G}} s$.
2. $s \not<_{\mathcal{G}} t$ and $t \not<_{\mathcal{G}} s$ and $s \not\simeq_{\mathcal{G}} t$ and $t \not\leq_{a_{n+1}} s$. The same argument as in 1.
3. $s \not<_{\mathcal{G}} t$ and $t \not<_{\mathcal{G}} s$ and $s \not\simeq_{\mathcal{G}} t$ and $s \not\leq_{a_{n+1}} t$. The same argument as in 1.

4. $s \not\leq_{\mathcal{G}} t$ and $t \not\leq_{\mathcal{G}} s$ and $s \not\leq_{a_{n+1}} t$ and $t \not\leq_{a_{n+1}} s$. Since $\leq_{a_{n+1}}$ is total, we have a contradiction.

Cfw: We have to show for any subset P of W, the set $max_{\leq_{\mathcal{G}/a_{n+1}}} P = \{v \in W | \forall x \in P : x \leq_{\mathcal{G}/a_{n+1}} v\}$ is non-empty. By IH we have that $max_{\leq_{\mathcal{G}}} P \neq \emptyset$ and by our assumption about the individual agents, we know that $max_{\leq_{a_{n+1}}} P \neq \emptyset$. Hence we have s, s.t. for all $x \in P$: $x \leq_{\mathcal{G}} s$ and we have t s.t. for all $x \in P$: $x \leq_{a_{n+1}} t$. Now consider any $r \in P$.[1] Then $r \leq_{\mathcal{G}} s$ and $r \leq_{a_{n+1}} t$.

We claim that $r \leq_{\mathcal{G}/a_{n+1}} s$ or $r \leq_{\mathcal{G}/a_{n+1}} t$, We know, by totality and IH, that either $s \leq_{\mathcal{G}} t$ or $t \leq_{\mathcal{G}} s$. Also either $s \leq_{a_{n+1}} t$ or $t \leq_{a_{n+1}} s$. Again several cases are to check:

- $s \leq_{\mathcal{G}} t$ and $s \leq_{a_{n+1}} t$. Then by transitivity, $r \leq_{\mathcal{G}} t$. Now, either $t \leq_{\mathcal{G}} r$ or $t \not\leq_{\mathcal{G}} r$.

  $t \leq_{\mathcal{G}} r$: Then $r \simeq_{\mathcal{G}} t$. But then since $r \leq_{a_{n+1}} t$, we have that $(r, t) \in \simeq_{\mathcal{G}} \cap \leq_{a_{n+1}}$ and so $r \leq_{\mathcal{G}/a_{n+1}} t$.

  $t \not\leq_{\mathcal{G}} r$: Then, $r <_{\mathcal{G}} t$ and hence $r \leq_{\mathcal{G}/a_{n+1}} t$.
- $s \leq_{\mathcal{G}} t$ and $t \leq_{a_{n+1}} s$. So $r \leq_{\mathcal{G}} t$. Either $t \leq_{\mathcal{G}} r$ or $t \not\leq_{\mathcal{G}} r$. If the latter, $r <_{\mathcal{G}} t$ and hence $r \leq_{\mathcal{G}/a_{n+1}} t$. If the former, then $r \simeq_{\mathcal{G}} t$ and since $r \leq_{a_{n+1}} t$, we have $r \leq_{\mathcal{G}/a_{n+1}} t$
- $t \leq_{\mathcal{G}} s$ and $s \leq_{a_{n+1}} t$. Either $s \leq_{\mathcal{G}} r$ or $s \not\leq_{\mathcal{G}} r$. If the latter $r <_{\mathcal{G}} s$ and hence $r \leq_{\mathcal{G}/a_{n+1}} s$. If the former, then $r \simeq_{\mathcal{G}} s$. Either $t \leq_{a_{n+1}} s$ or $t \not\leq_{a_{n+1}} s$. If the latter, then $t <_{a_{n+1}} s$ and so $r \leq_{a_{n+1}} s$ and hence $r \leq_{\mathcal{G}/a_{n+1}} s$. If the former then $t \simeq_{a_{n+1}} s$ and since $r \leq_{a_{n+1}} t$, by transitivity, also $r \leq_{a_{n+1}} s$ and thus $r \leq_{\mathcal{G}/a_{n+1}} s$.
- $t \leq_{\mathcal{G}} s$ and $t \leq_{a_{n+1}} s$. Either $r \leq_{\mathcal{G}} t$ or $t \leq_{\mathcal{G}} r$ by IH and totality.
  - $r \leq_{\mathcal{G}} t$: Then $r \leq_{\mathcal{G}} t \leq_{\mathcal{G}} s$. Either $t \leq_{\mathcal{G}} r$ or $t \not\leq_{\mathcal{G}} r$. If the latter $r <_{\mathcal{G}} t$ and hence $r \leq_{\mathcal{G}/a_{n+1}} t$. If the former, then $r \simeq_{\mathcal{G}} t$ and since $r \leq_{a_{n+1}} t$, then $r \leq_{\mathcal{G}/a_{n+1}} t$.
  - $t \leq_{\mathcal{G}} r$: Then $t \leq_{\mathcal{G}} r \leq_{\mathcal{G}} s$. Either $s \leq_{\mathcal{G}} r$ or $s \not\leq_{\mathcal{G}} r$. If the latter, $r <_{\mathcal{G}} s$ and hence $r \leq_{\mathcal{G}/a_{n+1}} s$. If the former, then $s \simeq_{\mathcal{G}} r$ and since $r \leq_{a_{n+1}} t$ and $t \leq_{a_{n+1}} s$, we have $r \leq_{a_{n+1}} s$. Thus $r \leq_{\mathcal{G}/a_{n+1}} s$.

Thus, $\forall x \in P(x \leq_{\mathcal{G}/a_{n+1}} s \vee x \leq_{\mathcal{G}/a_{n+1}} t)$. Moreover, it follows that $\forall x \in P(x \leq_{\mathcal{G}/a_{n+1}} s) \vee \forall x \in P(x \leq_{\mathcal{G}/a_{n+1}} t)$: let $r \in P$ be arbitrary. We

[1] If $P = \emptyset$, the set $max_{\leq_{\mathcal{G}/a_{n+1}}} P$ is trivially non-empty.

know that $r \leq_{\mathcal{G}/a_{n+1}} s \vee r \leq_{\mathcal{G}/a_{n+1}} t$. In either case, by totality, either $s \leq_{\mathcal{G}/a_{n+1}} t$ or $t \leq_{\mathcal{G}/a_{n+1}} s$ and thus either t is maximal or s is maximal. Hence, the set $max_{\leq_{\mathcal{G}/a_{n+1}}} P = \{v \in W | \forall x \in P : x \leq_{\mathcal{G}/a_{n+1}} v\}$ is non-empty.

Cond 1: To Show: $t \models^+ F \Rightarrow (\forall s \in W : s \not\models^+ F \Rightarrow s \leq_{\mathcal{G}/a_{n+1}} t)$. Suppose $t \models^+ F$ and for arbitrary $s \in W$ that $s \not\models^+ F$. Then by IH $s \leq_{\mathcal{G}} t$ and $s \leq_{a_{n+1}} t$. Either $t \leq_{\mathcal{G}} s$ or $t \not\leq_{\mathcal{G}} s$. If the latter, $s <_{\mathcal{G}} t$ and hence $s \leq_{\mathcal{G}/a_{n+1}} t$. If the former, then $s \simeq_{\mathcal{G}} t$ and since $s \leq_{a_{n+1}} t$, also $s \leq_{\mathcal{G}/a_{n+1}} t$.

Cond 2: To show: $\forall t \in W[(\forall s \in W : s \leq_{\mathcal{G}/a_{n+1}} t) \Rightarrow t \models^+ F]$. Suppose for arbitrary t that $\forall s \in W : s \leq_{\mathcal{G}/a_{n+1}} t$. Then either for all s, $s \leq_{\mathcal{G}} t$ or ($s \simeq_{\mathcal{G}} t$ and $s \leq_{a_{n+1}} t$). In either case $t \models^+ F$ because both relations $\leq_{\mathcal{G}}$ and $\leq_{a_{n+1}}$ satisfy Cond 2.

□

Bibliography

Alchourròn, C., Gärdenfors, P. & Makinson, D. (1985). On the Logic of Theory Change: Partial Meet Contraction and Revision Functions, *Journal of Symbolic Logic* **50**: 510—530.

Andréka, H., Ryan, M. & Schobbens, P.-Y. (2002). Operators and Laws for Combining Preference Relations, *Journal of Logic and Computation* **12**: 13–53.

Balaguer, M. (2015). Fictionalism in the Philosophy of Mathematics, *in* E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, summer 2015 edn, The Metaphysics Research Lab: Stanford. `http://plato.stanford.edu/archives/sum2015/entries/fictionalism-mathematics/`.

Baltag, A. (2016). Dynamic Epistemic logic. unpublished lecture slides for the class *Dynamic Epistemic Logic*, held in 2016 at the *Institute for Logic, Language, and Computation*.

Baltag, A. & Smets, S. (2006). Dynamic Belief Revision over Multi-Agent Plausibility Models, *in* G. Bonanno, van der Hoek Wiebe & M. Wooldridge (eds), *Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision (LOFT 2006)*, University of Liverpool, pp. 11–24.

Barthes, R. (1967). The Death of the Author, *Aspen* **5–6**.

Berto, F. (2013). Impossible Worlds, *in* E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, winter 2013 edn, The Metaphysics Research Lab: Stanford. `http://plato.stanford.edu/archives/win2013/entries/impossible-worlds/`.

Berto, F. (forthcoming). Impossible Worlds and the Logic of Imagination, *Erkenntnis* .

Bonanno, G. (1996). On the Logic of Common Belief, *Mathematical Logic Quarterly* pp. 305–311.

Bonomi, A. & Zucchi, S. (2003). A Pragmatic Framework for Truth in Fiction, *dialectica* **57**: 103–120.

Currie, G. (1990). *The Nature of Fiction*, CUP.

Eklund, M. (2015). Fictionalism, *in* E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, winter 2015 edn, The Metaphysics Research Lab: Stanford. `http://plato.stanford.edu/archives/win2015/entries/fictionalism/`.

Folde, C. (2011). *Was ist Wahrheit in Fiktion? Eine Auseinandersetzung mit David Lewis*, Magisterarbeit, Universität Hamburg.

Folde, C. & Wildman, N. (under review). Fiction Unlimited. Under review in *Philosophical Quarterly*, accessed on 27th February 2016, available under `https://nwwildman.files.wordpress.com/2015/08/24-08-15-fiction-unlimited.pdf`.

Gärdenfors, P. (1988). *Knowledge in Flux: Modelling the Dynamics of Epistemic States*, MIT Press.

Grove, A. (1988). Two Modellings for Theory Change, *Journal of Philosophical Logic* **17**: 157–170.

Hanley, R. (2004). As Good As It Gets: Lewis on Truth in Fiction, *Australasian Journal of Philosophy* **82**: 112–128.

Heyd, T. (2006). Understanding and Handling Unreliable Narratives: A Pragmatic Model and Method, *Semiotica* **162**: 217–243.

Heyd, T. (2011). Unreliability. The Pragmatic Perspective Revisited, *Journal of Literary Theory* **5**: 3–17.

Issue, S. (2011). Unreliable Narration.

Jago, M. (2014). *The Impossible*, OUP.

Jespersen, B. & Duží (2015). Introduction, *Synthese* **192**: 525–534.

Künne, W. (2007). *Abstrakte Gegenstände. Semantik und Ontologie*, 2nd edn, Klostermann.

Lewis, D. (1978). Truth in Fiction, *American Philosophical Quarterly* **15**: 37–46.

Lewis, D. (1979). Counterfactual Dependence and Time's Arrow, *Noûs* **13**: 455–476.

Lewis, D. (1983). Postscripts to Truth in Fiction, *Philosophical Papers*, OUP, pp. 276–280.

Lewis, D. (1986). *On the Plurality of Worlds*, Blackwell.

Nichols, S. (2004). Imagining and Believing: The Promise of a Single Code, *Journal of Aesthetics and Art Criticism* **62**: 129–139.

Nichols, S. & Stich, S. (2000). A Cognitive Theory of Pretense, *Cognition* pp. 115–147.

Nichols, S. & Stich, S. P. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*, OUP.

Priest, G. (2001). Paraconsistent Belief Revision, *Theoria* **67**: 214–228.

Priest, G. (2012). Revising logic, MCMP lecture.

Proudfoot, D. (2006). Possible World Semantics and Fiction, *Journal of Philosophical Logic* **35**: 9–40.

Sainsbury, R. M. (2010). *Fiction and Fictionalism*, Routledge.

Searle, J. (1975). The Logical Status of Fictional Discourse, *New Literary History* **6**: 319—-332.

Smets, S. (2015). Logic, Knowledge and science. unpublished lecture slides for the class *Logic, Knowledge and Science*, held in 2015 at the *Institute for Logic, Language, and Computation*.

Vaidya, A. (2015). The epistemology of modality, *in* E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, summer 2015 edn, The Metaphysics Research Lab: Stanford. `http://plato.stanford.edu/archives/sum2015/entries/modality-epistemology/`.

van Benthem, J. (2007). Dynamic Logic for Belief Revision, *Journal of Applied Non-Classical Logics* **17**: 129–155.

van Inwagen, P. (1977). Creatures of Fiction, *American Philosophical Quarterly* **14**: 299–308.

Walton, K. (1990). *Mimesis as Make-Believe: on the Foundations of the Representational Arts*, Harvard University Press.

Wildman, N. (unpublished). Interactive Fiction, Participation, and Knowledge through Fiction. accessed 13th May, 2016 under `https://nwwildman.files.wordpress.com/2015/11/15-11-15-interactive-fiction-participation-and-learning-from-fiction.pdf`.

Yagisawa, T. (1988). Beyond Possible Worlds, *Philosophical Studies* **53**: 175—-204.