REASONING ABOUT FICTION

TOM SCHOONEN & FRANZ BERTO
Institute for Logic, Language, and Computation

PREPRINT VERSION
University of Amsterdam

Abstract

In this paper we provide a suggestion on how to model reasoning about fiction. We will argue that three things happen: (1) one (sequentially) takes on board the explicit content of the fiction; (2) one imports background beliefs; this is represented by the most plausible worlds after an update triggered by the explicit content; and (3) from these most plausible worlds, an agent reasons on alternative courses of action within the fiction. We suggest to model (1) and (2) by means of a *plausibility* ordering on worlds after having updated with the explicit fictive content. Finally, we use an *inferential* ordering on worlds to model (3).

Keywords: [Truth in Fiction, Impossible Worlds, Counterfactual Reasoning, Dynamic Epistemic Logic, Plausibility Ordering]

INTRODUCTION

People often reason about fiction. For example, the truth of this sentence is hotly debated amongst *Lord of the Rings* fans:

(1) If Frodo had *flown* to Mount Doom on an eagle, rather than *walking* there in order to destroy the ring, there would have been a quick and easy victory in the war with Sauron.

On the one hand people argue that eagles are shown to be strong enough to carry men in the stories of Tolkien, and capable of beating even the Nazgul beasts in speed, so it seems like a plausible thought that this sentence is true. On the other hand, people argue that Sauron would have more easily seen them coming and that the eagles could not withstand direct confrontation with Sauron; therefore there are good reasons why this statement is, arguably, false.

This is one example of humans who reason about counterfactual circumstances *within fictional* situations. Reasoning about fiction is important because we investigate, through counterfactual speculation, how things might have gone otherwise in a fiction – e.g., when we want to expose a hole in the plot. More importantly, studying this kind of reasoning may give us clues on how mental simula-

tion and counterfactual reasoning in general work. We will argue that, in counterfactual reasoning about fiction, three things happen: (1) one (sequentially) takes on board the explicit content of the fiction; (2) one imports background beliefs, which is represented by most plausible worlds after an update triggered by the explicit content; and (3) from these most plausible worlds, an agent reasons on alternative courses of action within the fiction.

We will proceed as follows: we start by briefly exposing the formal framework of Fontaine & Rahman (2014), which we take to be our starting model. We will then extend this to deal with the question of truth in fiction. In section 2, we will add a closeness ordering in order to deal with the counterfactual reasoning within fiction. After this, we will remark on some of the formal features of the resulting model. Finally, we will reflect on the work we have done and suggest some future research in the last section.

Before we get started, some preliminary remarks. In order to capture blatantly inconsistent fictions and, in particular, for the similarity ordering, we allow for impossible worlds. Impossible worlds can be such that are inconsistent (e.g., making both φ and $\neg\varphi$ true) or incomplete (e.g., making neither φ nor $\neg\varphi$ true). For any world w, let '$|w|$' denote the truth-set of that world in a model $\mathcal{M}$, i.e., $|w| =_{df} \{\varphi \mid \mathcal{M}, w \vDash \varphi\}$. We make two assumptions: (i) worlds always have a *non-empty* truth-set and (ii) for each set of sentences of the language, there is a(n impossible) world that makes those and only those sentences true. Conversely, for any sentence φ, let '$|\varphi|$' denote the truth-set of that sentence, i.e. $|\varphi| =_{df} \{w \mid \mathcal{M}, w \vDash \varphi\}$.[1]

1 Truth in Fiction

We take as our starting point a model proposed by Fontaine & Rahman (2014) and suggest a few extensions. Fontaine and Rahman propose their model as a formalisation of an *Artifactual Theory* of fictional objects. Such a theory claims that fictional objects are existent, created abstract objects. In order to capture all the insights of Artifactual Theorists, they formalise a variety of ontological dependencies. Due to space limitations, we will ignore part of the framework that is meant to deal with these dependencies and extend the model of Fontaine & Rahman (2014) with impossible worlds and a different accessibility-relation.

Note that the set of impossible worlds, I, and the set of possible worlds, P, are

[1] This is sometimes called the proposition expressed by φ, however, we want to avoid using such a theoretically-loaded term. Also, note that it might be misleading to call situations that are incomplete 'impossible *worlds*', as the allocation 'world' seems to imply a form of maximality or completeness. However, we will follow the literature and refer to these incomplete situations also as impossible worlds (cf. Berto 2013).

jointly exhaustive of the logical space and mutually exclusive. That is, $I \cup P = W$ and $I \cap P = \varnothing$.

DEFINITION 1. **Reasoning about Fiction Model**
Consider the following model, $\mathcal{M}$: $\langle P, I, \mathcal{A}, \mathcal{R}_\Phi, \{\preceq_a\}_{a\in\mathcal{A}}, V\rangle$:

- A set of *possible* worlds, P, and a set of *impossible* worlds, I,
- A set of agents, $\mathcal{A}$,
- An accessibility relation, $\mathcal{R}_\Phi$ on W, s.t. $\mathcal{R}_\Phi \subseteq W \times \wp(W)$,
- A set of agent-dependent plausibility orderings, $\preceq_a$, on all $S \subseteq W$,
- A standard valuation function, V.

Note that we have an accessibility relation for any intentional verb Φ, $\mathcal{R}_\Phi$, that we use when we discuss the worlds of a fiction. As there can be impossible fictions and fictional objects, this accessibility relation ranges of *all* worlds, also the impossible ones.[2] Most importantly, we have added an agent-dependent plausibility ordering to the model: $\preceq_a$. We take this to be a well-preorder. In the next section we will explain what this is, how it works, and how we will use it in our account of truth in fiction. Finally, the valuation function, V, takes a world and a sentence and returns '1' if the world is in the truth-set of the sentence and '0' otherwise. At possible worlds the truth-values are defined through the standard recursive definitions and at impossible worlds sentences are evaluated directly, without regard for compositionality.

1.1 TRUTH IN FICTION WITH A PLAUSIBILITY ORDERING

Nichols & Stich (2000) use belief-revision in their cognitive theory of pretence; we take this as *prima facie* evidence that a realistic account of reasoning about fiction should include such a mechanism. So, our account of truth in fiction is based on Lewis' (1978) Analysis 2, extended with a belief-revision update on the beliefs of an agent.[3] The idea is something along the following lines: we take truth in fiction to be truth in the set of most plausible worlds for an agent after having

[2] This is particular to a Modal Meinongian account of fictional objects: there is a world where the fictional object exist, not as fictional object, but as the thing it is described as being by the fiction (cf. Priest 2005 and Berto 2011). Note that, if one holds that fictional objects are *necessarily* non-existent, she could opt for letting $\mathcal{R}_\Phi$ range *only* over impossible worlds. Relatedly, in order to prevent concluding that actual objects are fictional objects due to accidental similarities, we could explicitly exclude the actual world from the range of $\mathcal{R}_\Phi$ (see Priest 2005, p. 124).

[3] We sloppily use 'update' for what, e.g., **?** calls *Lexicographic upgrade*.

updated her beliefs with the sentences of the fiction (or pretended to).[4] To this end, we use an agent-dependent plausibility-order on any $S \subseteq W$: $\preceq_a$ (see Def. 1).

Here, we will briefly explain how an agents arrives at a set of worlds she finds most plausible to be the worlds of the fiction, as these worlds will provide the starting point of our analysis of reasoning about fiction.

To see how we get to the set of most plausible worlds for an agent after having read a fiction, we need to go over some notational conventions. For instance, we let '$\bar{\varphi}$' denote a sequence of sentences, $\varphi_0\varphi_1 \ldots \varphi_n$, and for any fiction, F, we let '$\mathcal{F}$' denotes the (finite) sequence of explicit sentences of that fiction. Secondly, let '$\mathcal{B}_a$' be an abbreviation of a specific instance of the accessibility relation: $\mathcal{R}^a_{\text{belief}}(@)$. That is, '$\mathcal{B}_a$' denotes the belief-set of agent a at the actual world. We will focus on the plausibility-ordering on the belief-set of an agent.

We follow Baltag & Smets (2006) in that the plausibility-order is a *well-preorder*. A *well*-preorder, is a preorder (i.e., reflexive and transitive) such that "every non-empty subset [of W] has minimal elements" (Baltag & Smets, 2006, p. 12, emphasis removed), where the minimal elements are defined as elements at least as plausible as all other elements of the set. We write '$w \preceq_a w'$' when an agent a considers world w at least as plausible as w'. Updating with a particular sentence, φ, will influence the plausibility-order. Let '$\preceq^{\varphi}_a$' denote the resulting plausibility-order after an update with φ that (i) preserves well-preorderedness; (ii) is such that $\forall w \in |\varphi|$ and $\forall w' \notin |\varphi|$, $w \preceq^{\varphi}_a w'$; and (iii) is such that the ordering within the φ-worlds and the non-φ-worlds remains the same.

There are two important things to note here. First of all, note that such an update makes all worlds that make φ true more plausible than all worlds that fail to make φ true; whether or not these worlds makes the negation of φ true is irrelevant. This is important as we work with impossible worlds. A stronger requirement would be that after such an update all the worlds that make φ true are more plausible than the worlds that fail to do so *and the worlds that make* $\neg\varphi$ *true*. For now, we will not impose this stronger requirement. Secondly, the ordering within the φ-worlds and the non-φ-worlds remains the same. To make things a bit more clear, consider a model with two propositional variables, p and q and a language with four sentences: $p, q, \neg p, \neg q$. Then, after an update with p, the set of most plausible worlds is as follows: $w_1 = \{p\}, w_2 = \{p, q\}, w_3 = \{p, \neg q\}, w_4 = \{p, \neg p\}, w_5 = \{p, q, \neg q\}, w_6 = \{p, q, \neg p\}, w_7 = \{p, \neg q, \neg p\}$, and $w_8 = \{p, \neg p, q, \neg q\}$.

[4] In our setting, an agent is engaging with fiction and the updates she performs are not actual, i.e., she engages in pretence. This could be incorporated formally by indexing the sentences and corresponding updates so that the agent can later 'unroll' the updates to recover her original plausibility-order. We will ignore this complication here.

We now have everything we need to define the set of most plausible worlds for an agent after having read a fiction, F. We denote this set with '$\text{BEST}_{\leq_a^{\mathcal{F}}}\mathcal{B}_a$' ('$\text{BEST}_{\leq^{\mathcal{F}}}$' for short).

$$\text{BEST}_{\leq_a^{\mathcal{F}}}\mathcal{B}_a =_{df} \{w \in W \mid \forall w' \in \mathcal{B}_a[w \leq_a^{\mathcal{F}} w']\}$$

We take it that this set of most plausible worlds captures what the agents takes to be true in a particular fiction, F. (We will not give a clause for truth-in-fiction; for a fully worked out account of this see Badura 2016.) Note that this analysis deals both with steps (1) and (2) mentioned in the introduction. That is, given this analysis we are able to deal with the explicit content of the fiction (i.e., the sequential updates) as well as the relevant imported background information on the basis of this (i.e., the plausibility-order). In the next section, when dealing with counterfactual reasoning about fiction, we will concern ourselves with the last step, (3).[5]

2 Counterfactual Reasoning about Fiction

The main novelty of our paper will be the modelling of (3): counterfactual reasoning about fiction (henceforth: reasoning about fiction). So, for example, sentences such as 'If Sherlock Holmes had access to DNA-testing, then he would have solved the case of the Baskerville Hound much faster,' or 'If Frodo would have flown on an eagle, he would have reached Middle Earth sooner'. In order to model such reasoning, we draw inspiration from simulation models of counterfactual reasoning (cf. Nichols & Stich 2000 and Williamson 2007). In particular, we aim to formalise the insight that "[f]rom the initial premise along with her own current perceptions, her background knowledge, her memory of what has already happened in the episode, [...], *the pretender is able to draw inferences about what is going on in the pretense*" (Nichols & Stich, 2000, p. 119, emphasis added).

In order to use this insight we will proceed as follows: (i) we will use $\text{BEST}_{\leq^{\mathcal{F}}}$ to select the starting point from where the agent does the reasoning about the fiction; (ii) we will use an inference-based closesness-ordering of worlds (inspired by work of Mark Jago 2009, 2014) to model the inferences the agent uses in her reasoning; (iii) and finally, we will give a semantics for a conditional, '$\boxed{\mathcal{F}}\!\!\rightarrow$ ', that is to capture this reasoning about fiction.

[5] In what follows, we assume that there is some (objective) set of most plausible worlds after reading a fiction. These worlds should make true the explicit content of the fiction and some of the relevant background knowledge. However, this leaves open the option that one has her own account of how we get to the set of such worlds. That is, the analysis below is independent of our analysis of truth in fiction; all that the former assumes is that there is a set of worlds an agents takes to be candidates for the worlds of the fiction.

So, first we need to find a suitable starting point from where the agent starts to reason about the fiction. We take it that when one evaluates a counterfactual about the fiction – which we will formulate as '$\varphi \boxed{\mathcal{F}}\!\!\rightarrow \psi$' – she first updates $\text{BEST}_{\leq^{\mathcal{F}}}$, in the usual way, with the antecedent of the counterfactual. So, the set of most plausible worlds is now based on a sequential update of all the sentences of the fiction, i.e. $\mathcal{F}$, and finally with another update of the antecedent, φ. We will denote the result of this with '$\text{BEST}_{\leq^{\mathcal{F}\varphi}}$'. Clearly, the starting point of the agents reasoning should be related to this set. What will be important for (ii) is that the agent can 'reason towards' a conclusion, which we will model with a relation from one world to another, where the latter world makes more things true than the former. This way, the move from the one world to the other represents the fact that the agent inferred new information. Therefore, what we need as the starting point for the agent is a world that makes true the explicit content and the relevant background facts (i.e., is in $\text{BEST}_{\leq^{\mathcal{F}\varphi}}$), but that makes *no other things true*. That is, the world that makes those and only those things true. (It is here where the crucial need for incomplete worlds should become clear.)

That is, the 'smallest' world, in terms of the number of propositions it makes true, from the set of most plausible worlds after having updated with the fictional content, $\mathcal{F}$, and the antecedent. Let us call this the *minimal-set*. Formally:[6]

$$\text{MIN}^{\varphi}_{\mathcal{F}} =_{df} \{w \in W \mid \forall w' \in \ \text{BEST}_{\leq^{\mathcal{F}\varphi}}(|w| \supset |w'|)\}$$

This minimal world nicely captures the idea of Nichols and Stich that "[s]ome [...] constraints are imposed by the details of what has gone on earlier in the pretense along with the pretender's background knowledge. [However, t]his still *leaves many options open*" (2000, p. 127, emphasis added).

Secondly, we need a way to model (ii), i.e., the step-by-step reasoning towards the consequent. We will use an inference-based closeness-ordering to model this. That is, we will order worlds based on the number of inference steps 'it takes to get there' relative to the world of evaluation. Informally, given a particular information state (i.e., world), we can count the number of inference steps it would take to get from that information state to a new information state (world). The more inference steps needed to get to a world, the further away the world is. (Here, we follow work done on such orderings by Jago 2009, 2014.)

So, we need to define the set of inferences that one is capable of making in reasoning from the minimal-world to other worlds. (Note that, as Jago (2009) already points out, having the ability to apply certain inference rules is not the same as having one's beliefs closed under those inference rules.) Let $\mathfrak{R}$ be the set

[6] This is likely to be a singleton, however, it might also be an equivalence class of worlds. For what follows this does not matter significantly and we will often talk as if it is a singleton.

of rules that one can easily make when reasoning about fiction. (For now, we will not go into the details of what rules that might be.) This set of inference rules allows us to model reasoning about fiction with the following formal tools.

First of all, given the set of rules, we can define a set of ordered-pairs of worlds such that the first world of the pairs makes true the premises of an argument and the latter world additionally makes true the conclusion that we can draw with the particular inference rule. So, let $[\![\mathfrak{R}]\!]$ be the set of ordered-pairs of worlds, then $\langle w, w' \rangle$ is in that set iff there is a rule in $\mathfrak{R}$,

$$\frac{\varphi_1, \ldots \varphi_n}{\psi} \, r \in \mathfrak{R}$$

such that $\{\varphi_1, \ldots \varphi_n\} \in |w|$, $\psi \notin |w|$, and $|w'| = \{\psi\} \cup |w|$. For what follows, we need two weak assumptions with respect to the inference rules that possibly are in $\mathfrak{R}$:

Assumption 1: For all sets of inference rules (used here), it is not the case that there is an inference rule $r \in \mathfrak{R}$ such that $\varphi \wedge \neg\varphi \vdash_r \psi$.

Assumption 2: For all sets of inference rules (used here), there is an inference rule $r \in \mathfrak{R}$ such that $\varphi \wedge \neg\varphi \vdash_r \varphi$ and $\varphi \wedge \neg\varphi \vdash_r \neg\varphi$.

We take it that these are *prima facie* reasonable assumptions. The first denies the following reasoning: if there is an object that is green all over and red all over that it follows that Sherlock Holmes never solved a crime. The second asserts that from the fact that there is something that is both round and square one can reason to the information that there is something that is round.

Given the minimal-world and the set of inference rules, we now finally have everything in place to define a function that will provide us with the closeness-order. As noted above, we assume that there is a particular antecedent of the counterfactual and this influences what is in the minimal-set, therefore the function is in a sense sentence-specific (so for every sentence of our language there will be such a function):

> Let f_φ be a partial function from worlds, W, to the natural numbers, $W \to \mathbb{N}$. Then, $f_\varphi(w) = n$ if and only if there is a sequence of worlds, $w_0 w_1 \ldots w_n$, such that (i) $w_0 = w$, (ii) $w_n \in \text{MIN}^{\varphi}_{\mathcal{F}}$, and (iii) for all $i \leqslant n, \langle w_i, w_{i-1} \rangle \in [\![\mathfrak{R}]\!]$, and there is no such sequence $w_0 w_1 \ldots w_m$ for $m < n$.

This function does exactly what we intuitively want it to do. Given a world, this function determines the smallest number of inference steps from the minimal-world to that given world.

With this closeness-order, we can now finally turn to (iii), giving a semantics for the conditional that aims to capture reasoning about fiction: $\varphi \boxed{\mathcal{F}}\!\!\rightarrow \psi$. Again, we aim to capture the insights from cognitive science and epistemology concerning counterfactual reasoning. Nichols and Stich describe the process of reasoning about pretence as follows: "[e]arly on in a typical episode of pretense, [...], one or more initial pretense premises are placed in the [...] workspace. [...] What happens next is that the cognitive system *starts to fill* the [scenario] with an *increasingly detailed* description of what the world would be like if the initiating representation were true" (2000, p. 122, emphases added). Williamson uses this description to provide an epistemology of counterfactuals and argues that "one asserts the counterfactual conditional if and only if [such] development eventually leads one to add the consequent" (2007, p. 153).

We follow Williamson in our semantics, namely, that the conditional is true if, when reasoning from the antecedent, one reaches the consequent before she reaches its negation. Formally:

DEFINITION 2. **Semantic-clause f-counterfactual**
Consider a fiction, F, and a world, $w \in W$, then $\mathcal{M}, w \vDash \varphi \boxed{\mathcal{F}}\!\!\rightarrow \psi$ iff $\exists w'(f_\varphi(w')$ is defined, $\mathcal{M}, w' \vDash \psi$, and $\forall w''[\mathcal{M}, w'' \vDash \neg\psi \Rightarrow f_\varphi(w') \leqslant f_\varphi(w'')])$

Let us call this conditional, '$\boxed{\mathcal{F}}\!\!\rightarrow$', the *f-counterfactual* for future reference. Note that the universal quantifier used here makes it that if there is no world that makes the negation of the consequent true, then the f-counterfactual is vacuously true. However, the existential quantifier makes it such that if there is no world where the consequent is true, the counterfactual will be false.

3 Non-Monotonicity, Reflexivity, and Modus Ponens

Before we conclude with some more philosophical reflections on our analysis, let us briefly remark three formal features of the conditional.

In general with impossible worlds semantics or logic, validity is defined as truth-preservation at all *possible* worlds. Most classical validities are remained that way. However, if we want to look at the behaviour of our conditional, non-normal worlds are crucial, as we saw. So, when we talk about what 'holds' for our conditional, we mean at all worlds, even impossible ones.

Non-Monotonicity: First of all, one of the main features of counterfactual reasoning is that it is a form of *non-monotonic* reasoning, that is, additional information in the antecedent might make false an otherwise true counterfactual. A classic

example that shows this is the following: 'if I had struck this match, it would have lit' is true, yet 'if I had struck this match and it had been wet, it would have lit' is false. It is quite straightforward to see that our conditional, $\boxed{\mathcal{F}}\!\!\rightarrow$, captures such non-monotonicity. Informally, this is explained by the fact that we use a sequential update for the set $\text{BEST}_{\leq^{\mathcal{F}}}$. This means that our ordering is sensitive to the most recent sentence with which it is updated. Thus, it is not at all necessary that the minimal-world $\text{MIN}^{\varphi}_{\mathcal{F}}$ is the same as the minimal-world $\text{MIN}^{\varphi\chi}_{\mathcal{F}}$. As these two worlds may be different, the consequences that can be drawn from these worlds need not be identical. Hence, the conditional allows for non-monotonic reasoning.

Reflexivity: Another feature that holds for our conditional is that of *reflexivity*, i.e., $\varphi \boxed{\mathcal{F}}\!\!\rightarrow \varphi$. This is also rather straightforward. Consider an arbitrary φ and we can show that $\varphi \boxed{\mathcal{F}}\!\!\rightarrow \varphi$. By definition of '$\text{MIN}^{\varphi}$' and the update procedure, we know that all worlds in MIN^{φ} are such that they make φ true and that there is at least one such world. Then, by definition of '$\boxed{\mathcal{F}}\!\!\rightarrow$ ', we know what $\varphi \boxed{\mathcal{F}}\!\!\rightarrow \varphi$ is true, for after an update with the antecedent, all worlds in MIN^{φ} already make the consequent true. Thus, the conditional is reflexive.

Modus Ponens: A feature that many theorists agree on to hold for counterfactuals is *modus ponens*, that is, the following inference should hold:

$$\varphi, \varphi \boxed{\mathcal{F}}\!\!\rightarrow \psi \models \psi$$

As we pointed out above, when talking about modus ponens, we will not talk of validity. Even more so, it is important to note that modus ponens generally does not hold at impossible worlds for *any* conditional. Consider any language and model with incomplete worlds, then we can always construct a world where the conditional and the antecedent are true, yet that is 'ignorant' on the status of the consequent. Similarly, for any language and model with inconsistent worlds, we can always construct a world where the conditional and the antecedent hold, but where the negation of the consequent is true.

So, we will here focus on a particular kind of, what we will call, 'restricted modus ponens'. That is, we will prove that there is a characteristic of this conditional that one can interpret as a version of modus ponens.[7] In order to do so, we will first prove a lemma about $\mathfrak{R}$ and then we will define a function from worlds to worlds, that given a world, delivers the closure of that world under the inference rules in $\mathfrak{R}$. First the lemma:

[7]Thanks to Mark Jago for his suggestions on modus ponens and the needed functions and thanks to Aybüke Özgün for many corrections on the formal details.

LEMMA 1. If an ordered pair of worlds, $\langle w, w' \rangle$, is in the set $[\![\mathfrak{R}]\!]$, then all propositions made true by the former world will also be made true by the latter, i.e., $\forall \varphi \in |w| [\varphi \in |w'|]$.

Proof. Let w and w' be such that $\langle w, w' \rangle \in [\![\mathfrak{R}]\!]$. It follows by definition of $\mathfrak{R}$ that the truth-set of w' is the same as the truth-set of w, with the addition of one sentence, ψ, that is not in w. So, $|w| \subset |w'|$. Hence, all truths from w are preserved in w'. □

This lemma will be used later on when we prove our form of modus ponens, but first, we need to define a function on worlds, that gives us the closed version of worlds under the rules of $\mathfrak{R}$:

> Let $f^{\mathfrak{R}}$ be a function from worlds, W, to worlds, $W \to W$. Then, $f^{\mathfrak{R}}(w) = w'$ if and only if there is a sequence, $w_0 w_1 \dots w_n$, such that (i) $w_0 = w$; (ii) for all $i < n$ it is the case that $\langle w_i, w_{i+1} \rangle \in [\![\mathfrak{R}]\!]$; (iii) w_n is such that $\forall \varphi (|w_n| \vdash_{r \in \mathfrak{R}} \varphi \Rightarrow \varphi \in |w_n|)$; and (iv) $w' = w_n$.

This function takes a world and gives the 'closure-world' of thta world under the rules of inference in $\mathfrak{R}$. Note that $f^{\mathfrak{R}}(\cdot)$ need not give us consistent worlds as the rules of inference in $\mathfrak{R}$ are non-monotonic, nor need they be complete, as some facts might be left out from the start and/or be non-inferable.

Before we can start our proof, note that in the definition of f_φ, the ordering is 'from the world with the larger truth-set *to the world with the smaller* truth-set'. On the other hand, in the definition of $f^{\mathfrak{R}}$, the ordering is the other way around, that is, 'from the world with the smaller truth-set *to the world with the larger* truth-set'. Luckily, it is easy to see that, *given* a sequence, we can re-label that sequence: instead of naming the starting world w_0 and the last world w_n, we call the starting world w_n and the last world w_0. That is, we 'flip' the ordering, but only of the labels we use for the worlds. The 'objective' order of the worlds is unaffected by this. As the labels are mere names for the worlds, this is fine as long as we then reverse the definition of which worlds follow from which worlds. So:

LEMMA 2. If there is a sequence of worlds, $w_0 w_1 \dots w_n$ such that (i) $w_0 = w$, (ii) $w_n \in \text{MIN}^{\varphi}_{\mathcal{F}}$, and (iii) for all $i \leqslant n, \langle w_i, w_{i-1} \rangle \in [\![\mathfrak{R}]\!]$, then there is a sequence of worlds, $w_0 w_1 \dots w_n$ such that (i) $w_n = w$, (ii) $w_0 \in \text{MIN}^{\varphi}_{\mathcal{F}}$, and (iii) for all $i \leqslant n, \langle w_i, w_{i+1} \rangle \in [\![\mathfrak{R}]\!]$.

Proof. Follows immediately from re-labelling. □

We will give a small toy-example to make this intuitive. Consider worlds u, v, w such that $[\![\mathfrak{R}]\!] = \{\langle u, v \rangle, \langle v, w \rangle\}$. If we now reason from u to w, there is a sequence,

$w_0w_1w_2$ such that $w_0 = u$, $w_2 = w$, and for each i in the sequence $\langle w_i, w_{i+1}\rangle \in [\![\mathfrak{R}]\!]$. However, we can reason similarly from w to u. In that case, there is a sequence, $w_0w_1w_2$ such that $w_0 = w$, $w_2 = u$, and for each i in the sequence $\langle w_i, w_{i-1}\rangle \in [\![\mathfrak{R}]\!]$.

With all this in place, it is quite straightforward to show what kind of 'restricted modus ponens' holds for our conditional.

THEOREM 1. For all models, $\mathcal{M}$, all fictions, $\mathtt{F}$, and all worlds, w, if $w \in \text{BEST}_{\leq^{\mathcal{F}}}$, then, if $w \vDash \varphi, \varphi \boxed{\mathcal{F}}\!\!\rightarrow \psi$, $f^{\mathfrak{R}}(w) \vDash \psi$.

Proof. Take an arbitrary world, w, and assume that $w \vDash \varphi$ and that $w \vDash \varphi \boxed{\mathcal{F}}\!\!\rightarrow \psi$. By definition of the plausibility-ordering, the fact that w makes true φ, and the $w \in \text{BEST}_{\leq^{\mathcal{F}}}$ it follows that $w \in \text{BEST}_{\leq^{\mathcal{F}\varphi}}$. From this, and the definition of $\text{MIN}^{\varphi}_{\mathcal{F}}$, it follows that the truth-set of all worlds, w', that are in $\text{MIN}^{\varphi}_{\mathcal{F}}$, are a subset of the truth-set of w. That is, $\forall w' \in \text{MIN}^{\varphi}_{\mathcal{F}}(|w'| \subseteq |w|)$.

Now, from the fact that $w \vDash \varphi \boxed{\mathcal{F}}\!\!\rightarrow \psi$ and the semantics of '$\boxed{\mathcal{F}}\!\!\rightarrow$', it follows that there is a world, let's name it u, such that $u \vDash \psi$ and such that there is a sequence, $w_0w_1 \ldots w_n$ such that (i) $w_0 = u$, (ii) $w_n \in \text{MIN}^{\varphi}_{\mathcal{F}}$, and (iii) for all $i \leqslant n, \langle w_i, w_{i-1}\rangle \in [\![\mathfrak{R}]\!]$. By Lemma 2, we can 'flip' this ordering for ease of exposition such that there is a sequence of worlds $w_0w_1 \ldots w_n$ such that (i) $w_n = u$, (ii) $w_0 \in \text{MIN}^{\varphi}_{\mathcal{F}}$, and (iii) for all $i \leqslant n, \langle w_i, w_{i+1}\rangle \in [\![\mathfrak{R}]\!]$. What this shows is that u is 'reachable' from $\text{MIN}^{\varphi}_{\mathcal{F}}$ with the rules of $\mathfrak{R}$ and, as we saw, $u \vDash \psi$.

By the definition of $f^{\mathfrak{R}}$, $f^{\mathfrak{R}}(w)$ gives us a sequence of worlds, $w_0w_1 \ldots w_n$, such that (i) $w_0 = w$, (ii) for all $i \leqslant n$ it is the case that $\langle w_i, w_{i+1}\rangle \in [\![\mathfrak{R}]\!]$; (iii) w_n is such that $\forall\varphi(|w_n| \vdash_{r\in\mathfrak{R}} \varphi \Rightarrow \varphi \in |w_n|)$; and (iv) $f^{\mathfrak{R}}(w) = w_n$.

Now, remember that the truth-set of all the worlds in $\text{MIN}^{\varphi}_{\mathcal{F}}$ are a subset of the truth-set of w. So, if we can $\mathfrak{R}$-derive u from $\text{MIN}^{\varphi}_{\mathcal{F}}$, it follows that we can also $\mathfrak{R}$-derive u from w. And, given the 'closedness' of $f^{\mathfrak{R}}(w)$ under the rules in $\mathfrak{R}$, we know that there is a $i \leqslant n$, such that in the relevant sequence, $w_i = u$.

Given Lemma 1, we know that $|u| \subseteq |f^{\mathfrak{R}}(w)|$. This, combined with the fact that $u \vDash \psi$, gives us that $f^{\mathfrak{R}}(w) \vDash \psi$. □

Remember that this is *not* the classical modus ponens as the consequent is not made true by the same world as the antecedent and the conditional. However, if the world in question is a *possible* world, this would be a version of classical modus ponens; because possible worlds are deductively closed, so for any possible world, $w = f^{\mathfrak{R}}(w)$.[8]

[8]However, note that this requires getting into the debate concerning the possibility of possible worlds being the most plausible worlds as the 'fictional world'. We will not engage with this here,

We do believe that this version of modus ponens is intuitive enough though. For consider what one does when she reasons about fiction, she does indeed start to reason from the world she took to be the most plausible 'fiction world'. However, given the new input of the antecedent, this world might no longer be the most plausible one (also given the interaction of the antecedent with the relevant background beliefs).

Also, briefly note why we need the assumptions mentioned above. As we already mentioned, it might very well be that $f^{\mathfrak{R}}(w)$ results in an inconsistent world such that $f^{\mathfrak{R}}(w) \models \psi$ and $f^{\mathfrak{R}}(w) \models \neg\psi$. However, given **Assumption 2**, it still follows that $f^{\mathfrak{R}}(w) \models \psi$. Conversely, if there is an inconsistent world in the sequence, we do not want that from that moment onwards anything can be inferred. **Assumption 1** secures this.

Conclusion: Counterfactual Reasoning and $\mathfrak{R}$

The above model for counterfactual reasoning about fiction is far from full-fledged qua technical details. However, we believe that it provides a valuable proof of concept. That is, we think that this type of analysis of counterfactuals matches an intuitive account of the epistemology and cognitive processes involved in counterfactual reasoning (cf. Nichols & Stich 2000 and Williamson 2007, Ch. 5). We hope to explore such accounts of reasoning more fully in the future, especially in relation to counterfactual reasoning in general, but also other forms of conditional reasoning (see for example, Solaki & Berto 2017, who apply a similar analysis to dual-processing theories of reasoning).

We will conclude this paper by making some remarks on the set of inference rules, $\mathfrak{R}$, which does much of the heavy lifting on our account.

An important question that remains unanswered is what (kinds of) rules of inferences are in $\mathfrak{R}$. This, we take it, is a very difficult question, one that we cannot hope to settle here. We want to briefly remark two things with regards to it – one methodological note and one suggesting an avenue of relevant research – hoping to relieve some of the pressure of this question.

First of all, note that the lack of specific content of $\mathfrak{R}$ does not present a knock-down argument against the account presented here. Surely, more needs to be said about the specifics of $\mathfrak{R}$ and the details of the current proposal probably need to be adapted in accordance. However, one might wonder whether this is an issue for logicians to address. In order to have a cognitively realistic model, determining what inferences people make, especially in the context of fictive counterfactual reasoning, should be left to cognitive scientists. The logician should have a

we only want to note this technicality.

framework ready in which she can incorporate the parameters provided by the cognitive scientists (see also Jago 2014 on this).

These remarks only relieve the pressure of the question a bit and we can in fact make some suggestions for future research. Let us mention two. First of all, it seems clear that people, in general, do not reason by the standards of classical logic, let alone when reasoning about fictions. For example, Priest (2005) notes that it is likely that paraconsistent logic "is the default logic for reasoning about a fictional situation" (p. 122). Similarly, relevant logics, non-monotonic logics, conditional logics, etc., might all prove to involve certain inferences that agents make in the context of fiction. Secondly, it seems clear that agents do not make inferences on the basis of their information *unrestrictedly* (cf. Jago 2014; Solaki 2017; Solaki & Berto 2017). That is, there are *cognitive costs* to making such inferences – if only the fact that human agents are finite beings. Solaki (2017) provides a wide range of models (most based on dynamic epistemic logics with impossible worlds) that keep track of such cognitive costs (see also Solaki & Berto 2017). Interestingly, the models she uses also involve rational agents making step-by-step inferences from their information. We take it that research into (i) non-classical logics and (ii) the cognitive costs of such inference steps will provide valuable insights that are needed to further work out the details concerning $\mathfrak{R}$.

We take the above not to be vices of the analysis we provide here, but as exciting new venues of research that will hopefully lead to a new accounts of counterfactuals, counterpossibles, and conditional reasoning in general that closely matches findings from areas of cognitive theories and the epistemology of such reasoning.

ACKNOWLEDGEMENTS

The authors would especially like to thank Aybüke Özgün, who has been an incredible help with the details of this paper. Thanks are also due to the audience at Logica 2017, Ilaria Canavotto, Thom van Gessel, and Anthia Solaki.

REFERENCES

Badura, C. (2016). *Truth in Fiction via Non-Standard Belief Revision*. Master's thesis, University of Amsterdam, Amsterdam, The Netherlands.

Baltag, A., & Smets, S. (2006). Dynamic Belief Revision over Multi-Agent Plausibility Models. In G. Bonanno, W. van der Hoek, & M. Wooldridge (Eds.) *Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision (LOFT2006)*, (pp. 11–24). University of Liverpool.

Berto, F. (2011). Modal Meinongianism and fiction: the best of three worlds. *Philosophical Studies*, *152*(3), 313–334.

——— (2013). Impossible Worlds. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: CSLI Publications.

Fontaine, M., & Rahman, S. (2014). Towards a semantics for the artifactual theory of fiction and beyond. *Synthese*, *191*(3), 499–516.

Jago, M. (2009). Logical information and epistemic space. *Synthese*, *167*, 327–341.

——— (2014). *The Impossible*. Oxford: Oxford University Press.

Lewis, D. K. (1978). Truth in fiction. *American Philosophical Quarterly*, *15*(1), 37–46.

Nichols, S., & Stich, S. (2000). A cognitive theory of pretense. *Cognition*, *74*, 115–147.

Priest, G. (2005). *Towards Non-Being; the logic and metaphysics of intentionality*. Oxford: Oxford University Press.

Solaki, A. (2017). *Steps out of Logical Omniscience*. Master's thesis, Institute for Logic, Language, and Computation, Amsterdam, The Netherlands.

Solaki, A., & Berto, F. (2017). The Logic of Fast and Slow Reasoning.

Williamson, T. (2007). *The Philosophy of Philosophy*. Oxford: Blackwell Publishing.