Extracting fictional truth from unreliable sources*

Emar Maier
University of Groningen

Merel Semeijn
University of Groningen

Abstract A fictional text is commonly viewed as constituting an invitation to play a certain game of make-believe, with the individual sentences written by the author providing the propositions we are to imagine and/or accept as true within the fiction. However, we can't always take the text at face value. What narratologists call 'unreliable narrators' may present a confused or misleading picture of the fictional world. Meanwhile there has been a debate in philosophy about so-called 'imaginative resistance' in which we are inclined to resist imagining (or even accepting as true in the fiction) what's explicitly stated in the text. But if we can't take the text's word for it, how do we determine what's true in a fiction? We propose an account of fiction interpretation in a dynamic setting (a version of DRT with a mechanism for opening, updating, and closing temporary 'workspaces') and combine this framework with belief revision logic. With these tools in hand we turn to modelling imaginative resistance and unreliable narrators.

Keywords: fiction; dynamic semantics; discourse; DRT; belief revision; unreliable narration; imaginative resistance

1 Introduction

It is a fiction author's prerogative to decide what's true in the fictional worlds she creates. After all, it's her words that create this world, by saying what it's like in there. When Tolkien wrote that "Frodo had a very trying time that afternoon" it automatically became true in the *Lord of the Rings* saga that Frodo had a very trying time on a particular afternoon. This line of thinking can be summed up in the principle of Authorial Authority:

(1) Authorial Authority: If s is part of text T (and not a quotation), then the proposition expressed by s is true in the world of T.

This principle seems to hold for all fiction, and only for fiction. If a historian or journalist writes that Napoleon was 1.47m tall, this does not thereby become 'true in

* This research is supported by NWO Vidi grant 276-80-004 (Maier). Many thanks to ...

the world of the historical text'.[1] What the historian writes is true or false depending on what the actual world is like. At first sight then, Authorial Authority promises to help pin down what fiction is, and how it differs from non-fiction.

Unfortunately, truth in fiction is not always so straightforward. First, as Lewis and many subsequent authors in philosophy and narratology have observed, there are many propositions that are true in a given fiction beyond the ones that make up the text. In the Harry Potter books, milk comes from cows, water is H_2O, and people are annoyed if you cut in line. More interestingly, the opposite is also true. There are cases where a text says that p, but that fails to be true in the world of the fiction. These cases of Authorial Authority breakdowns are likewise much discussed in both literary studies and philosophy, but in rather different terms. The first group talks about unreliable narrators, i.e. narrators that misinform or misjudge because they are trying to deceive, are prejudiced, naïve, or confused. For instance, in *The Adventures of Huckleberry Finn* the narrator, Huck, gives the following report on a dinner with the widow Douglas:

(2) The widow rung a bell for supper, and you had to come to time. When you got to the table you couldn't go right to eating, but you had to wait for the widow to tuck down her head and grumble a little over the victuals, though there warn't really anything the matter with them

Huck reports that Douglas grumbles over her food before eating as if she were unhappy with it but the reader realizes that Huck fails to understand that actually she was praying. Hence even though the text states that Douglas grumbled over her food, this is not true in *The Adventures of Huckleberry Finn*. Rather, it is true in the fiction that the widow prayed before taking her meal and that Huck mistook this for dissatisfied grumbling.

Somewhat independently from literary scholars' debates about unreliable narration, there is a now long standing debate in philosophy about so-called 'imaginative resistance', a phenomenon whereby readers of a fictional text resist imagining and/or accepting a part of a story. Consider the story *Fish Tank*:

(3) Sara never liked animals. One day, her father caught her kicking the neighbor's dog. He got really angry and she was grounded for a week. To get back at her father she poured bleach in the big fish tank, killing all the beautiful fish that he loved so much. Good thing that she did, because he was really annoying.

1 Though see for instance Zucchi (2020) for an opposing view on which any discourse or text T makes it true that 'in/according to T, φ' (for any φ in T). We suggest that there's a crucial difference between 'in' and 'according to' here, but leave a semantic analysis of these operators for a future occasion.

Readers can go along imagining a sadistic protagonist kicking her neighbor's dog and killing her father's fish, but when they arrive at the evaluative statement "Good thing that she did" they resist. Even though the text explicitly states that it was good that she did this, readers report that they can't or won't imagine that this is so, nor do they accept that it is true in the story.

In sum, cases of imaginative resistance and unreliable narration alike constitute clear prima facie counterexamples to the intuitive Authorial Authority principle for fiction.

But if we can't trust the author's words to give us the fictional truths, how *do* we know what's true in the story? How do readers of stories like the above figure out what the fictional world is like? In this paper we provide a dynamic semantic account of fiction interpretation that takes into account the role of the (unreliable) narrator and the phenomenon of imaginative resistance. We first introduce a basic framework for interpreting fiction and non-fiction, extending Discourse Representation Theory with insights from Matravers' (2014) philosophical account of the nature of fiction (Section 2), and from belief revision logic (Section 3). We then apply these tools to the two concrete examples of unreliable narration and imaginative resistance above (Section 4).

2 A framework for interpreting fiction and non-fiction

2.1 Discourse Representation Theory

Discourse Representation Theory (DRT) is a dynamic approach to meaning developed by Kamp (1981) that models how a (multi-sentence) discourse updates a context, as represented by a so-called Discourse Representation Structure (DRS). Historically, there has been some disagreement between those that interpret DRS's and related notions of dynamic context as representations of a Stalnakerian common ground (e.g. Heim 1982; Groenendijk & Stokhof 1991; van der Sandt 1992) and those that interpret DRS's as representations of an agent's individual mental state (e.g. Geurts 1999; Kamp 2015, but see Hamm et al. 2006 for a conciliatory view). We're assuming a mentalistic interpretation of DRT in which DRS's represent part of the mental state of the interpreter of the discourse, viz. the interpreter's beliefs about what is common ground between herself and the speaker (i.e., what Stalnaker would call the hearer's presuppositions).

One of the reasons we forego the simple common ground picture in favor of a more mentalistic interpretation is that eventually we are interested here in not only cooperative exchanges where abstract common grounds are updated monotonically with every utterance, but also in 'defective contexts' where speaker and hearer's conceptions of the common ground essentially diverge. Consider for instance what

happens when an addressee believes it to be common ground that only birds fly and then a trusted source says "You do know that bats are not birds, right?". In that case the addressee's presuppositions do not coincide with the actual common ground – the speaker may have never believed or even accepted that only birds fly – and what we're interested in capturing is how the hearer's presuppositions rather than the common ground are revised in light of the new information. We return to revision and unreliable conversation partners in section 3.

We start with a demonstration of the basic DRT framework modeling a straightforward cooperative, face-to-face exchange. Suppose a speaker says (4a). As a result, the addressee considers it common ground that Pedro owns a donkey. In DRT we represent this as follows:

(4) a. Pedro owns a donkey.

b.

x y
pedro(x) donkey(y) owns(x,y)

The top part of the DRS in (4b) introduces two discourse referents. We can think of this part as a form of existential quantification: there are two individuals, x and y. The bottom part contains DRS conditions, specifying properties of and relations between these discourse referents: x is named Pedro and y is a donkey, and x owns y.

Now for the dynamics. Suppose the speaker continues the discourse as in (5a). The DRS (representing the hearer's beliefs about what is common ground) is updated with the information compositionally encoded in that sentence:

(5) a. He beats it.

b.

x y
pedro(x) donkey(y) owns(x,y) beats(?,?)

In this DRS, the contributions of the pronouns 'he' and 'it' are not yet resolved, as indicated by the question marks. We have to resolve these anaphoric links, binding the question mark arguments to suitable discourse referents previously established. In this case, the subject pronoun 'he' binds to x, while 'it' binds to y, so we replace the question marks accordingly and arrive at the following final representation of the semantic contribution of the entire two-sentence discourse:

(6)

x y
pedro(x) donkey(y)
owns(x,y)
beats(x,y)

In general, anaphora resolution, and presupposition resolution more generally, involves finding appropriate and accessible discourse referents in a way that leads to a maximally coherent final output DRS.[2]

The interpretation of DRS's as representing what the hearer considers common ground between herself and the speaker works well for face-to-face communication, where speaker and hearer are both present and arguably both continually update their beliefs about what they both believe etc. A first qualification involves the reliance on the attitude of belief. As Stalnaker (1984) already points out, the common ground is best thought of not as what is commonly known or believed, but what is commonly accepted, i.e. commonly known to be taken to be true (in a given context) (Stokke 2013). A lesser known problem with the above characterization of the common ground becomes apparent with we consider forms of communication that are more one-sided than the face-to-face back-and-forth inquiry. When I read Tolkien, for instance, I have an idea of what he'd accept as true, but he surely had no idea about what I'd accept, as that would amount to him having a de re attitude about me. That is clearly wrong: Tolkien at best has some purely descriptive, *de dicto* attitudes about 'whoever reads this text'. One way out would thus be to define the relevant common ground in terms of universal quantification over engagers, where engaging includes uttering, hearing, writing, signing, reading, interpreting, translating, commenting on the text or discourse in question:

(7) φ is common ground among a community of engagement iff whoever engages with the discourse accepts φ; whoever engages with the discourse knows that whoever engages with the discourse accepts that φ; whoever engages with the discourse knows that whoever engages with the discourse knows that whoever engages with the discourse accepts φ etc.

Henceforth we'll take our DRS boxes to be representations of what the reader considers common ground among the community of engagement in this sense.[3]

2 For a more detailed introduction to basic DRT syntax and semantics, presupposition resolution, and anaphora, see e.g. Kamp & Reyle 2011; Geurts et al. 2016.

3 Cf. Badura & Berto (2018); Zucchi (2020) for related common ground adaptations. We thank Chris Badura, Sandro Zucchi and Bart Geurts for discussion on this point.

2.2 Quarantining fiction

If we model typical indicative statements in a work of fiction straightforwardly as assertions (i.e., as updates of the context DRS, representing the interpreter's conception of the common ground), we quickly run into difficulties. Suppose a reader picks up Heinlein's classic sci-fi novel *Stranger in a Strange Land*. At this point she believes (among other things) that it is common ground between her and Heinlein and any other engagers that someone named Heinlein wrote a book called *Stranger in a Strange Land*.

(8)

x y
heinlein(x) book(y)
wrote(x,y) stranger.in.strange.land(y) ...

Now she reads the first sentence. Analyzing this as a simple update of the DRS in (8) would give:

(9) a. Once upon a time when the world was young there was a Martian named Smith.

b.

x y z
heinlein(x) book(y)
wrote(x,y) stranger.in.strange.land(y) ...
martian(z) smith(z)

In other words, the reader now considers it common ground that in addition to an author named Heinlein there exists also a Martian named Smith (here and throughout we mostly ignore the intricacies of time/tense in our analyses of concrete examples).

This is clearly the wrong result: the reader probably doesn't take it to be common ground between Heinlein and herself that there ever were Martians. We don't mix fact and fiction like this. The content of a fictional narrative is to be somehow kept separate from the 'official' common ground (Stokke 2013; Bonomi & Zucchi 2003). In the following we formalize this idea by introducing the notion of a temporary workspace to DRT (Semeijn 2017).

2.3 The workspace account

Inspired by the aforementioned 'unofficial common ground' proposals and Matravers's (2014) critique of the 'consensus' analysis of fiction in terms of 'invitations

to imagine' (Walton 1990; Currie 1990),[4] Semeijn's (2017) workspace account refines the Stalnakerian common ground update view by introducing a temporary common ground, a 'workspace', alongside the permanent official common ground. Interpretation is modeled as a three-step procedure: (i) the interpreter opens a new workspace; (ii) she updates this workspace, more or less as in standard DRT, building up a representation (what Matravers calls a 'mental model') of what the incoming discourse or story is about; (iii) when she puts down the book or ends the conversation she performs a *closure operation* on the updated workspace, bringing the quarantined information back to the official common ground.

Let's flesh out these three steps in a bit more detail and then discuss two concrete examples: one of non-fiction and one of fiction interpretation.

Step 1: Opening a workspace

We assume that a new unupdated workspace is a copy of the current official common ground. This enables us to resolve anaphoric links in the workspace, so when a novel or newspaper report mentions terrorist attacks in Paris, we already have a discourse referent for that city, and all kinds of background information predicated thereof.[5] For non-fiction this just embodies the central tenets of Stalnakerian context dependence and presupposition satisfaction. For fiction it amounts to the idea that we're never interpreting a text in a vacuum, but understand it against a background or importation of factual information about the actual world, as in Lewis's (1978) counterfactual analyses of truth in fiction, Ryan's (1981) Principle of Minimal Departure, Walton's (1990) Reality and Mutual Belief Principles, and especially Friend's (2017) Reality Assumption. In all these theories, fictional worlds are assumed to be as much as possible like (or in Lewisian terms: As close as possible to) the real world as the story permits. For instance, because I know that actually water is H_2O and boils around 100^{circ}C at sea level, I assume that the same is true in the fictional worlds of *The Hobbit* or *Harry Potter* even though this is never explicitly stated.[6]

As we'll explain below, a workspace remains open for the duration of a specific

4 See also Maier 2017 and Kamp 2020 for modern semantic implementations of the consensus view, in a rather different, purely mentalistic version of DRT.

5 Strictly speaking we end up with a copy of our actual Paris representation in our workspace, so the resulting workspace is arguably not really *de re* about Paris, but about something with a lot of the same properties in a fictional world. This is in line with Wieland's (2020) proposal, but against the dominant view championed by Friend (2011) and many others (including Maier 2017) who take *A Tale of Two Cities* to be literally about the actual city of Paris in our world. Eventually, perhaps, we might want to bring the mental compartmentalization and anchoring of Maier's (2017) DRT account over to the current workspace model, but for now we want to set this issue aside, as it is orthogonal to the main topics of this paper.

6 See Franzén (2020) for an in depth defense and discussion of the reality principle.

conversation or book reading session, and will then be closed, i.e. transferred back into a new, updated official common ground. In the case of fiction, when we return to our book we don't start from scratch with a new copy of the common ground. We'll propose an operation of 'fictive opening', retrieving from the official common ground the final state of the relevant earlier fiction workspace as our current active workspace. We discuss fictive opening in 2.6.

Step 2: Updating a workspace

Once we have our initial workspace set up we start updating with the incoming information. Following Matravers (2014) the idea is that this update process is uniform across fiction and non-fiction (we'll revisit this point later when we discuss belief revision and entrenchment). Fiction however does pose some extra constraints on our implementation of this process, having to do with different types of context revision in light of inconsistencies (belief revision, accommodating unreliable narrators, cautious updates). Capturing nonmonotonic workspace updates for fiction and non-fiction uniformly is the main aim of this paper. We return to this step below and in the following sections.

Step 3: Closing a workspace

Now in the final step of the algorithm the difference between fiction and non-fiction will play a role. At the end of interpreting the possibly multi-sentence discourse the updated workspace of entertained propositions must be transferred somehow back to the official common ground. We propose two distinct closure operations. In the case of non-fiction, we perform 'assertive closure', i.e. we simply replace the official common ground with the current workspace. Note that in this case the net result is exactly the same as for standard DRT updating without opening and closing workspaces. In the case of fiction, however, we perform 'fictive closure', i.e. the workspace enters the common ground embedded under a relevant Lewisian modal fiction operator $\Box_x$, defined as follows:

(10) For any DRS K and discourse referent x (referring to a text or other contentful medium), $\Box_x K$ is a well-formed DRS condition and $[\![\Box_x K]\!]^{f,w} = 1$ iff in all possible worlds w' compatible with $f(x)$: $[\![K]\!]^{f,w'}$.

We'll leave it open what exactly it means for a possible world to be compatible with a story. Our goal instead is to describe what kind of information ends up in the embedded DRS K, and how, thus capturing the dynamics of the interpretation process.

2.4 Example: non-fiction

Suppose I pick up an article in *The New York Times* about Trump and further suppose that I consider it to be common ground that (among other things) Trump is the president of the U.S.. First, I open up a new workspace containing all information in the current common ground. As for notation, we officially represent interpretation contexts now as a pair consisting of an official common ground and an active workspace. Informally, we'll consistently use only primed discourse referents (x', y',...) in workspaces, and in some examples just display the current workspace, relying on context and primed discourse referents to indicate this.

(11) $\left\langle \begin{array}{|c|}\hline x \\ \hline \text{trump(x)} \\ \text{president(x)} \\ \ldots \\ \hline \end{array}, \begin{array}{|c|}\hline x' \\ \hline \text{trump(x')} \\ \text{president(x')} \\ \ldots \\ \hline \end{array} \right\rangle$

I then update the workspace as I build a mental model based on the propositions expressed by the article:

(12) a. Trump cheats at golf

b. $\left\langle \begin{array}{|c|}\hline x \\ \hline \text{trump(x)} \\ \text{president(x)} \\ \ldots \\ \hline \end{array}, \begin{array}{|c|}\hline x' \\ \hline \text{trump(x')} \\ \text{president(x')} \\ \ldots \\ \text{cheats.at.golf(x')} \\ \hline \end{array} \right\rangle$

As soon as the discourse (i.e., my reading of the article, considered to be non-fiction) ends, we perform assertive closure, whereby the content of the official common ground is replaced by the content of the workspace, leaving us with a new official common ground:

(13) $\left\langle \begin{array}{|c|}\hline x' \\ \hline \text{trump(x')} \\ \text{president(x')} \\ \ldots \\ \text{cheats.at.golf(x')} \\ \hline \end{array}, \emptyset \right\rangle$

In other words, after reading the article I believe it is now common ground between me and the author that Trump cheats at golf. Note that the addition of the workspace to DRT didn't really do anything. The payoff lies in the way it allows us to model

the essential similarities and dissimilarities between non-fiction and fiction.

2.5 Example: fiction

Our starting assumption is that, apart from the different closure operations, reading a fictional narrative involves the same interpretative processes as reading non-fiction. Let's revisit the Heinlein example.

I believe it to be common ground (amongst other things) that Heinlein wrote a sci-fi novel *Stranger in a Strange Land* of which I'm holding a copy. When I pick up the book I open a copy of this common ground as the new workspace and update that workspace with the information conveyed by the first line (14a), resulting in (14b):

(14) a. Once upon a time when the world was young there was a Martian named Smith.

b. $\langle$ | *x* *y* | , | *x'* *y'* *z'* |
| --- | --- |
| heinlein(x) book(y)
wrote(x,y)
stranger.in.strange.land(y) … | heinlein(x') book(y')
wrote(x',y')
stranger.in.strange.land(y') …
martian(z') smith(z') |

$\rangle$

As discussed above, starting our interpretation of the novel with a copy of the official common ground captures the general assumption that when we interpret a fiction, everything that I consider common ground between me and the author holds in the fictional world described by the novel as well, unless and until it's overridden by the story (or genre conventions, see 2.6). In other words, this copying is our implementation of the aforementioned Reality Assumption.

As soon as I put down the book, I perform fictive closure, i.e. the content of the workspace is added to the common ground under a suitable Lewisian modal fiction operator:

(15) $\langle$

x *y*
heinlein(x) book(y) wrote(x,y) stranger.in.strange.land(y) …
$\Box_y$

x' *y'* *z'*
heinlein(x') book(y') wrote(x',y') stranger.in.strange.land(y') … martian(z') smith(z')

, $\emptyset$ $\rangle$

Compare this output to the problematic result of standard common ground updating

in section 2.2 above. We no longer predict that the existence of Martians is common ground between me and Heinlein; what is common ground is just that *in the story* there's a Martian. Opening a workspace and performing fictive closure has effectively quarantined the content of the fictional narrative from the official common ground.

However, the inconsistent common ground problem now rears its head in a different place: Since the new workspace started off as a complete copy of the entire common ground, it will definitely include some uncontroversial information about the history of space travel, hidden somewhere in the '...', that conflicts with the existence of Martians named Smith. This means that we actually have an inconsistent workspace while reading the novel (and afterwards, under the modal operator in the new official common ground). This is unsatisfactory. We need a story about how to easily give up background assumptions about actual space travel and the non-existence of Martians from the workspace background, when confronted with a fictional text that asserts or entails otherwise. In the next section we explore how belief revision logic can help us achieve this.

But before we turn to revision, there is another striking feature of our output DRS that requires comment. According to (15) it is common ground that in the fictional world of the book there is a writer named Heinlein who wrote a book named *Stranger in a Strange Land*. In other words, somewhere in the fictional universe there exists not only a Martian named Smith but also some guy named Heinlein writing about Martians. Revision might help us avoid this counterintuitive consequence in cases where the content of the story conflicts with the existence of a human fiction author writing a fictional book. However, in many cases there may be nothing in the story to contradict the existence of an author with a certain name, somewhere in the background in a remote (in space and time) corner of the universe, away from the main events of the story. In the case at hand, some official common ground assumptions about Heinlein will perhaps have to be given up to maintain consistency.[7] For instance, it is highly unlikely that someone published a famous science fiction story about a Martian named Valentine Michael Smith, which much later, years in the future (after World War III) turned out to happen exactly as described. On the other hand, the mere fact that there was a sci-fi author (in the distant past) named Heinlein seems less controversial and may well survive revision, yielding something like the output in (15), with a discourse referent for Heinlein in the representation of the fiction. Note that this fictional Heinlein counterpart in (15) is not to be equated with the narrator of the story. For one, the narrator is temporally located at some unspecified time after the events, while this fictional Heinlein lives in the 1960's. Moreover, the fictional narrator by definition tells the story 'as known

7 Consistency need not be understood as mere logical consistency. It's best thought of in terms of a gradable, context-dependent notion of possibility, coherence and/or plausibility. We won't formalize this notion here.

fact' (Lewis 1978), while the fictional Heinlein, like the real one, wrote the story as pure fiction. We return to the status of the narrator in section 4.

Summing up, on our account some common ground facts surrounding the real-world author and book may be imported into the representation of the fictional world. Starting from the common ground among author and arbitrary readers already removes the most counterintuitive importation predictions of Lewis's Reality Principle (like the prediction that it should be true in the Sherlock Holmes story that it is raining in Groningen on June 19, 2019, which is true but not common knowledge among everybody who engages with the fiction). Revision, especially if based on plausibility, may remove further unwanted imports (like the fact that the content of Heinlein's 1961 fiction happens to match post WW3 reality). For the remaining imports, like those depicted in (15), we'll follow Walton's (1990) lead: without any textual or other evidence to the contrary, we assume that there was a 20th century author named Heinlein in the world of *Stranger in a Strange Land*. But as this information is completely irrelevant to the events in the story, none of the fictional characters nor the narrator will ever refer to the Heinlein discourse referent, so it will quickly fade into the background, as will presumably be captured by a more realistic processing model of regular common ground updating that tracks the salience of discourse referents and/or associated conditions.

2.6 The role of genre

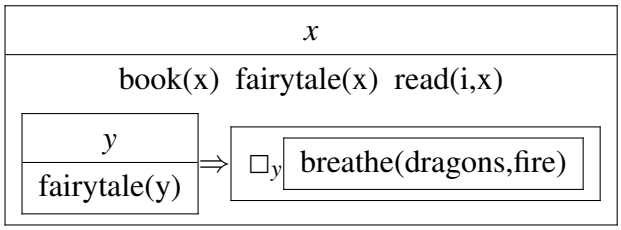
Up until this point we have presented a *tabula rasa* interpretation of fiction where the reader has no prior beliefs about what is true in the fiction before starting to engage with it (except that the fiction – any fiction – conforms to our common ground based version of the Reality Assumption). In reality, there will often be what Lewis (1978) calls 'inter-fictional carry-over' of fictional truth, i.e. additional fictional truths may derive from prior knowledge about what is true in other fictional stories. This can for instance happen because the story is part of a larger canon (the Harry Potter series, the Star Wars Expanded Universe), but more general genre conventions can add fictional truths as well;[8] In a typical fairy tale about a knight going on a quest to slay a dragon we may anticipate that the dragon breathes fire, even if that has not (yet) been stated explicitly. The question is how the information that dragons breathe fire enters the workspace since it is neither stated explicitly in the text, nor part of the official common ground (assuming that it is common ground that there are no dragons).

We propose that genre conventions are imported in a way similar to fictional truths derived from a previous book reading session, i.e. in a process we call fictive

8 Genre conventions may also influence whether fictional statements are judged reliable and how they update the workspace (See 4.2).

opening. As mentioned above, when I continue reading a fictional narrative after a break that triggered fictive closure, I obviously don't start from scratch, as that would render all previous discourse referents inaccessible. Instead I re-create the last known state of the workspace, K, from a parafictional condition of the form $\Box_x K$, generated in the official common ground after closing a previous reading event. Genre expectations may be stored in the official common ground in terms of parafictional conditions as well.[9] Consider the unknown fairy tale from before. I pick up the book, and on the basis of the cover picture and first few lines ("Once upon a time in a faraway land there lived a knight…") I decide that I'm dealing with a fairy tale. At the same time, it's common ground that, say, 'in fairytales, dragons breathe fire'.

(16)



The workspace that we open to represent the story at hand should be a copy of the common ground, as usual, but also contain the information that holds in stories of this type, as stored in quantified parafictional statements like in (16). We can define the fictive opening mechanism to take care of continued reading and genre assumptions uniformly: when starting to interpret a text x, make a copy of the official common ground and merge that with all K such that $\Box_x K$ is part of (or can be inferred on the basis of, as in (16)) the official common ground. This merge will likely cause significant contradictions (the information that dragons breathe fire will probably contradict some basic common knowledge of physics and biology), so it will have to involve revision rather than classic monotonic DRS update. As we've hinted at the need for nonmonotonic updates a few times already, we now turn to that.

3 Fiction updates and belief revision

3.1 Introducing belief revision

In the 1980's, around the same time as linguists started developing dynamic semantics, researchers in computer science, AI, and philosophy of science started developing logical tools to describe how a system of beliefs reacts to an influx of

9 Here we only consider conventions related to different genres of fiction. Non-fiction genre conventions (e.g. In a news report a family taking tea at their dining-room table means that the family is 'normal') are analysed as unprefixed stereotypic knowledge in the common ground. See Matravers (2014) and Zucchi (2020) for a uniform treatment of fiction and non-fiction genre conventions.

new, possibly conflicting information. Unlike regular dynamic semantics, belief revision describes also nonmonotonic updates, i.e. removing previously established information from the context or belief state because the new information conflicts with those previously held beliefs.

As in dynamic semantics, there are 'representational' (or 'syntactic') versions of the theory, where a belief is a set of sentences in some logical language, and more semantic versions, where a belief is modeled as a set of possible worlds (Grove 1988). Since we're already using the representational framework of DRT, we'll adopt a version of the former, classic belief revision theory, known as the AGM model (Alchourron et al. 1985). More specifically, we'll adopt a version with beliefs modeled as belief bases, which contain only the agent's core beliefs, rather than the logically closed belief sets that contain everything that the agent is arguably committed to on the basis of their belief base and general principles of rationality (Hansson 1994, 1998; Nebel 1998). For instance, if I believe that it's raining, I'm committed to believing that it's raining or sunny, but that particular disjunction is not usually part of my core belief base.

The basic operation in AGM is contraction of a belief base K with a statement φ, that is, reducing the set K in such a way that it no longer entails φ. AGM spells out a number of postulates to axiomatize well-behaved contraction operations. One way of constructing such a well-behaved contraction operation is on the basis of a given 'epistemic entrenchment' order: $\varphi < \psi$ iff φ is less entrenched than ψ, i.e. ψ has more epistemic worth (e.g. because it derives directly from a trusted knowledge source) and therefore is less easily given up than φ. Natural and moral laws for instance may be considered to be more entrenched than concrete contingent facts, especially if based on hearsay rather than direct perception. An agent's belief base is fully characterized by a set of statements K and an entrenchment ordering $<$ (again, satisfying certain axioms of rationality, like transitivity and the fact that logical consequences of φ are at least as entrenched as φ itself, Gärdenfors 1988) on the set of well-formed formulas of the language. Contracting K with (a non-tautology) φ (notation: $K \div \varphi$) now means that we chose a $K' \subseteq K$ such that K' does not entail φ. Epistemic entrenchment helps us single out an optimal such K', for instance with the following definition of entrenchment-based contraction:

$$(17) \qquad K \div \varphi = \left\{ \psi \in K \,\middle|\, \varphi < (\varphi \vee \psi), \text{ or } \varphi \text{ is a tautology} \right\}$$

Gärdenfors & Makinson (1988) show that (17) generates a well-behaved contraction operation, obeying all their rationality postulates. However, when we consider only finite belief bases, rather than logically closed belief sets, assuming a full entrenchment order on the entire language seems like overkill. Hence, Williams (1994), for instance, introduces the notion of an 'ensconcement', which is essentially

a finite entrenchment on the formulas in the base. We refer to Nebel (1998) for in-depth study of entrenchment and related notions (e.g., 'prioritized base revision') applied to belief bases rather than belief sets. Below we'll continue to use the familiar term 'epistemic entrenchment', requiring only an intuitive understanding of how an entrenchment relation on a finite set of statements K can guide the process of contracting K with φ, by letting it eliminate from K as few as possible of the least entrenched conditions as needed to avoid entailing φ. Concretely, we just start from the lowest rank and then move up to the next if that doesn't help us get rid of φ.

Once we have contraction, AGM defines belief revision with p as the process of first contracting with $\neg p$ and then adding ('expanding with') p. But since we're dealing with belief bases, which, unlike belief sets, need not be consistent, we can also do it the other way around: first expand with p and then contract with $\neg p$. Either way, the resulting belief base always entails p, i.e. new incoming information trumps all previous beliefs. In so-called semi-revision we level the playing field and treat old and new information on a par, with only epistemic entrenchment as the guiding factor. Formally, this amounts to adding p and then removing the contradiction:

(18) semi-revision: $(K \cup \{p\}) \div \bot$

Below we implement this kind of revision in our DRT update mechanism so we can deal with the pervasive nonmonotonic updating required to incorporate fictional statements that contradict the initial common ground copy. See Badura & Berto (2018) for a similar application of belief revision to fiction, but in a more semantic possible (and impossible) worlds approach.

3.2 Belief revision in the DRT workspace

First, consider a mini-discourse that leads to an inconsistent workspace in the domain of non-fictional conversation. Consider a conversation between a speaker and a hearer who believes it to be common ground that there is a person called Maxine who is vegan and who loves animals. We open a new workspace with an exact copy of this information. The speaker now says (19a), resulting in an updated workspace (19b).

(19) a. Maxine owns a donkey. She beats it

b. ⟨	x	,	x' y'
	maxine(x) vegan(x)		maxine(x') vegan(x')
	loves.animals(x) …		loves.animals(x') …
			donkey(y') own(x',y')
			beat(x',y')
⟩			

Given certain background assumptions about the relation between loving and beating,

and donkeys being animals, conveniently hidden in the '…' in (19), the conjunction of loving animals and beating donkeys may well entail a contradiction.[10] Depending on for instance how much the hearer trusts her own background information about Maxine (and about donkey keeping) and how reliable she takes the speaker to be, she will then want to revise the workspace, giving up some piece of information in order to restore consistency. We can implement the central insights from AGM belief revision introduced above to model this.

Note first that instead of belief bases we now have DRS's (pairs of sets of discourse referents and DRS conditions). To incorporate the entrenchment order we number the DRS conditions and add a third DRS compartment specifying a partial order on the conditions via these number labels. This models the epistemic entrenchment of the various bits of information that make up the DRS and thereby guide the process of resolving inconsistencies.[11] Concretely, the official common ground representation at the start of the vegan discourse may now look like this, modeling a situation where the hearer quite strongly supposes it to be common ground between himself and the speaker that Maxine indeed bears that name, and is less invested in it being common ground that she's vegan and loves animals:

(20)

x
1:maxine(x) 2:vegan(x) 3:loves.animals(x) …
1>{2,3}

Second, we assume that instead of a classic dynamic DRS update of the workspace with incoming utterance information we perform semi-revision; New information is added to the workspace, presuppositions are resolved, and the new information is assigned a position in the epistemic entrenchment ordering. We then contract with Falsum to remove any contradictions entailed by the updated DRS K' (as per (18)). Intuitively, we do this by eliminating as few as possible of the least entrenched conditions until the DRS is consistent again.

To continue our example, let's assume that it's common ground that the speaker is a close friend of Maxine, and appears to have no reason to deceive the hearer. We

10 We might eventually want to incorporate plausibility and/or coherence metrics into our model and replace 'contradiction' with 'low plausibility/coherence', i.e., a score below a certain contextually determined plausibility threshold.

11 For convenience, we number only the conditions, not the discourse referents. This is just a technical hack: if a DRS is inconsistent we can always restore consistency by merely removing conditions, because in the extreme case a discourse referent that does not occur in any conditions doesn't contribute any real information (except that the domain is non-empty). The set of DRS conditions thus plays the role of the belief base.

can capture this by placing the new information relatively high, say just below the information that the person under discussion bears the name Maxine, but above the animal-loving-veganism. (From here on we'll display only the current workspace, leaving out the official common ground DRS.)

(21) a. Maxine owns a donkey

b.

x' y'
1:maxine(x') 2:vegan(x') 3:loves.animals(x') ... 4:donkey(y') 5:own(x',y')
1>{4,5}>{2,3}

When we encounter the next sentence we again start by updating the DRS, resolving anaphora, and extending the epistemic entrenchment ordering. Let's say the current speaker's contributions about Maxine are all assumed to be of equal epistemic value:

(22) a. She beats it

b.

x' y'
1:maxine(x') 2:vegan(x') 3:loves.animals(x') ... 4:donkey(y') 5:own(x',y') 6:beat(x',y')
1>{4,5,6}>{2,3}

Given some very general, uncontroversial background assumptions (hidden in the '...' or kept in a separate encyclopedic knowledge compartment of a full representation of context (Kamp 2016)) about the relationships between animal loving and donkey beating, this DRS is arguably inconsistent. And even if not logically inconsistent it's questionable as a representation of the common ground, as it's unlikely that Maxine is both an animal lover and a donkey beater. The least entrenched conditions are 2 and 3. Elimination of condition 3 ('loves.animals(x')') is already sufficient to make the DRS consistent again:

(23)

x' y'
1:maxine(x) 2:vegan(x') ~~3:loves.animals(x')~~ 4:donkey(y') 5:own(x',y') 6:beat(x',y')
1>{4,5,6}>{2,~~3~~}

After processing this mini-discourse, we perform assertive closure, which turns the workspace in (23) into the new, updated common ground.

3.3 Cautious update

In the above instance of semi-revision the addressee considered the speaker to be very reliable, i.e. the incoming information was assigned a place high in the epistemic entrenchment ordering. We have thus essentially modeled a non-monotonic generalization of what Eckardt (2014) calls a Trust Update, i.e. we add the information of the speaker's utterance directly to the common ground. However, as Eckardt also notes, we do not always trust the speaker. Suppose that the hearer actually knows Maxine really well and assumes it is definitely common ground that Maxine is vegan and also definitely loves animals. However, she also knows that the speaker may not be very reliable when it comes to Maxine, so what she says may be based on shaky assumptions, or lies, and hence shouldn't automatically become established common ground. In terms of entrenchment, the incoming information about Maxine's donkey, conditions 4-6 in the pre-contraction DRS (22b), now instead dangle at the bottom of the entrenchment hierarchy: {2,3}>{4,5,6}. Since this DRS again represents an inconsistent common ground, we need to remove a low ranked condition to restore consistency. On the current ranking, revision will simply cancel the latest update, 6:

(24)

x y
1:maxine(x') 2:vegan(x') 3:loves.animals(x') … 4:donkey(y') 5:own(x',y') ~~6:beat(x',y')~~
1>{2,3}>{4,5,~~6~~}

In other words, the speaker's last utterance ('she beats it') is inconsistent with previous, more entrenched information and is therefore essentially ignored by the hearer.

This not quite right. Especially when we'll be trying to extract meaning from unreliable narrators in fiction, we can't completely ignore speakers just because we don't trust them. Even if we do not trust a speaker's assertion that *p*, we can still extract valuable information from the utterance, viz. the information that the speaker themselves believed that *p*, or at the very least, in case they are lying, that they asserted that *p* and are thereby committed to *p*. Although the distinction between these two kinds of unreliability is important, not least in making sense of literary unreliable narrators, we'll lump them together here and use the uniform weak Stalnakerian attitude verb of acceptance (≈ treating a proposition as true, see section

2.1 above) to describe the hedged information we can extract from an unreliable speaker.

We suggest incorporating this Cautious Update (Eckardt 2014) into the non-monotonic workspace update mechanism: whenever semi-revision leads us to cancel part of the semantic contribution of the current speech act, we instead replace the offending condition φ with a suitably hedged version under a modal operator: $ACCEPT_x\varphi$, depending on whether the hearer considers the speaker to be misinformed or deceptive, with x a discourse referent picking out the current speaker. In this case we'll assume there's been a discourse referent s' and a condition 0 representing the speaker in the workspace (and official common ground) all along. Note also that the hedged condition, here 7, will be assigned a new, typically higher, place in the entrenchment ranking.

(25)

x' y' s'
0:speaker(s') 1:maxine(x') 2:vegan(x') 3:loves.animals(x') ... 4:donkey(y') 5:own(x',y') ~~6:beat(x',y')~~ 7:ACCEPT$_{s'}$ [beat(x',y')]
{0,1}>{2,3}>7>{4,5,~~6~~}

Interestingly, because the DRS represents the addressee's beliefs about what is common ground between her and the speaker, once the DRS is updated with the hedged version (i.e. that the speaker accepts that Maxine beats her donkey) we can't really keep the information that Maxine loves animals. Although it is not strictly contradictory or even prima facie implausible that Maxine loves animals while the speaker accepts that she beats her donkey, it cannot be common ground between speaker and hearer that this is so. To see this, note that φ being common ground entails that it is commonly known that everyone (so, in particular, the speaker) accepts φ. Thus, the assumption that (25) is common ground will entail that it is commonly known that the speaker accepts (25). Since (25) is essentially a conjunction of the various conditions therein and common ground and acceptance operators distribute over conjunction, it follows that (i) it is common ground that condition 7 holds (i.e. it is common ground that speaker accepts that Maxine beats her donkey), and (ii) it is common ground that speaker accepts condition 3 (i.e. it is common ground that speaker accepts Maxine loves animals). This would mean that it is now common ground that the speaker is inconsistent, which is clearly not the case here (regardless of whether she's lying or confused about the facts, she's not logically insane). Instead, we have to give up condition 3, the assumption that

Maxine loves animals. Note that the hearer herself probably really believed condition 3 to be true, and took it to be common ground. In fact her personal belief in condition 3 will likely remain unaffected, but after the speaker's assertion it can no longer be considered part of the common ground, for it has become clear that they don't share a commitment to this information. The end result thus will be:

(26)

x' y' s'
0:speaker(s') 1:maxine(x') 2:vegan(x') ~~3:loves.animals(x')~~ ... 4:donkey(y') 5:own(x',y') ~~6:beat(x',y')~~ 7:ACCEPT$_{s'}$ beat(x',y')
{0,1}>{2,~~3~~}>7>{4,5,~~6~~}

As we will see below, this reasoning plays out somewhat differently for fiction.

We can incorporate all the above reasoning into a definitive nonmonotonic, cautious workspace update algorithm along the following lines:

(27) Update a workspace K with a preliminary DRS representation φ (of incoming utterance), notation: $K + \varphi$

a. expansion: $K \uplus \varphi$ = merge K with φ, resolve all anaphora and presuppositions, and extend the epistemic entrenchment ranking to the new conditions.

b. contraction: $(K \uplus \varphi) \div \bot$ = remove as many low ranked conditions from $K \uplus \varphi$ as needed to ensure that $\mathrm{CG}_E(K \uplus \varphi)$ is consistent (where E denotes in every world the set of people engaged in the discourse in that world, en $\mathrm{CG}_E K$ entails $\forall x \in E(ACCEPT_x \varphi)$, $\forall x \in E(ACCEPT_x(\forall x \in E(ACCEPT_x K)))$, ...)

c. caution: for any condition ψ added in the expansion phase but subsequently removed in the contraction phase, update with the corresponding hedged condition $\mathrm{ACCEPT}_{s?}\psi$ (where s? is an anaphor that needs to be bound to the current speaker), i.e. $((K \uplus \varphi) \div \bot) + \boxed{\mathrm{ACCEPT}_{s?}\psi}$

Continuing with our example. After assertive closure, (26) will become the official common ground, i.e. Maxine the vegan animal-lover owns a donkey and the speaker accepts (perhaps even believes, but in any case commits herself to it by asserting it) that Maxine beats her donkey. Note again how this correctly captures the hearer's conception of the common ground, but not the speaker's, because the speaker might actually consider it to be common ground that Maxine beats her donkey, nor does

it accurately capture either the speaker's or hearer's private beliefs about Maxine. To align all these mental states and conceptions of the common ground, the hearer would have to manifest her distrust and renegotiate an aligned common ground with the speaker.

4 Interpreting fiction

Now that we have introduced our basic framework we will apply it to the interpretation of fiction. We show how various interpretation strategies emerge from our workspace account, allowing us to model the various ways of constructing imaginative story worlds from fictions featuring impersonal and personal, reliable and unreliable narrators.

We start with a simple face value interpretation of a fictional text with a reliable, impersonal narrator. Then we turn to cases involving Cautious Update triggered by unreliable narrators as in *The Adventures of Huckleberry Finn*. Lastly we apply our framework to a typical case of imaginary resistance.

4.1 Authorial Authority revisited: face value interpretation by shielding

The algorithm we sketched in (27) for updating a workspace works the same with fiction as with non-fiction. In section 1 we identified one prima facie difference between fiction and non-fiction: the principle of Authorial Authority (i.e. whatever the text asserts, is true in the fiction). We can reformulate this principle now in terms of epistemic entrenchment: conditions derived from interpreting a fiction outrank pre-existing conditions in the workspace (derived from copying the official common ground and genre conventions).

For instance, reconsider the workspace we get by expanding an input context with the Heinlein opening passage, i.e. (14), but now with an epistemic entrenchment ordering on its conditions. For expository purposes we've also included a condition 5 to abbreviate the previously hidden cluster of commonly known scientific facts about the evolution of life in our solar system that would clash with the existence of a martian named Smith.

(28)

x' y' z'
1:heinlein(x') 2:book(y') 3:wrote(x',y') 4:stranger.in.strange.land(y') 5:mars.is.uninhabited ... 6:martian(z') 7:smith(z')
{6,7}>5>{1,2,3,4}

In the context of a factual inquiry, condition 5 would outrank almost anything you can tell me. If you tell me, factually, you saw a Martian the other day, I'd sooner assume you're joking,[12] speaking metaphorically, or lying, than remove some of the basic scientific assumptions underlying 5 from the workspace.[13] When it's understood as fiction, information deriving from the text may well outrank basic science, as illustrated in (28). Those fictional statements are effectively 'shielded' from contraction, i.e. they will never be given up, even if they are inconsistent with some other seemingly uncontroversial statement that is part of our general background knowledge.

In (28), eliminating one or several of the least entrenched conditions (1-4) will not make the workspace consistent. Next up in the epistemic entrenchment ordering is condition 5, whose elimination makes the workspace consistent. We end up with a workspace where some facts about human space travel and life in our solar system are no longer valid.[14] Unlike the destructive copy operation of Assertive Closure, Fictive Closure however doesn't remove these retracted assumptions from the official workspace. If after reading a few more pages we close this workspace, the result is as follows:

(29)

x *y*
1:heinlein(x) 2:book(y) 3:wrote(x,y) 4:stranger.in.strange.land(y) 5:mars.is.uninhabited …
6:$\Box_y$
5>{1,2,3,4,6}

x' *y'* *z'*
1:heinlein(x') 2:book(y') … ~~5:mars.is.uninhabited~~ … 6:martian(z') 7:smith(z') …
{6,7}>~~5~~>{1,2,3,4}

We call the interpretation strategy of assigning the highest possible epistemic rank to information deriving from a text, a face value interpretation of a fictional text (Matravers 2014; Altshuler & Maier 2018; Badura & Berto 2018), i.e., an

12 Perhaps joking is a form of narrative fiction, in which case we'd no longer be engaging in factual inquiry but fiction.

13 Talking to a young child, crazy person, or time traveller may make me remove 5 from the workspace, if it becomes clear to me that these basic facts are really not common ground between us. We've discussed the reasoning behind such revisions triggered by Cautious Update in the vegan example in section 3.3.

14 Recall, Cautious Update is not triggered because we're retracting only old information (see (27)).

interpretation line with the principle of Authorial Authority.

Face value interpretations are appropriate in many cases (e.g. it gives us the desired result that it is not true in *Stranger in a Strange Land* that Martians don't exist). However, as pointed out in section 1, in some cases even in fiction we cannot blindly trust the speaker. In the remainder of this paper we will discuss the interpretive processes at work in making sense of such narratives, starting with unreliable first person narration.

4.2 Unreliable narrators

Consider again our central example of unreliable narration from *The Adventures of Huckleberry Finn*, abbreviated from (2):

(30) The widow rung a bell for supper . . . When you got to the table you couldn't go right to eating but you had to wait for the widow to tuck down her head and grumble a little over the victuals . . .

Let's assume that before engaging with the novel the reader takes it to be common ground between her and Twain that the latter produced a novel called *The Adventures of Huckleberry Finn* and a lot of other background information, including some information that entails that when people in 19th century Missouri bow over their food and mumble a bit before eating they are saying a prayer. As before we represent this rather trivial cultural background assumption as as a deeply entrenched condition in the official common ground:

(31)

x' y'
1:twain(x') 2:author(x') 3:adventures.of.huckfinn(y') 4:wrote(x',y') 5:mumble.before.dinner.is.prayer . . .
5>{1,2,3,4}

When we open the book and start reading, we open a copy of (31) as our workspace. Unlike the Heinlein story, the story is written in the first person, featuring Huck Finn as the narrator. This means the reader quickly accommodates a discourse referent for the first person pronouns, representing a speaker named Huck Finn who is (fictionally) asserting the sentences that constitute the story, and who is thereby committed to their truth. By contrast, note that in the Heinlein story there was no 'I', no sign of any personal character telling the story, and hence no need to accommodate a discourse referent for a speaker.

Since Huck is evidently a naive young boy, we don't always trust his assertions, just as we don't always trust our face-to-face interlocutors. When a conflict arises

between the reader's background knowledge, as imported from the common ground, and the text, we might therefore want to revise the contribution of the text rather than the background. In other words, since the text is considered to be the assertions of a child, the semantic contributions of the text should not generally end up at the top of the epistemic entrenchment ranking. More generally, for first person narratives, i.e. narratives where we accommodate a discourse referent for a first person speaker, we relax the principle of Authorial Authority, by giving up the requirement that new information is shielded from revision by automatically ranking it at the top.

When the reader arrives at the mumbling passage, (30), she first updates the workspace with the unproblematic fictional statements (e.g. that Douglas rung the bell and that Huck had to wait for dinner etc) that do not conflict with any background information. When we get to the statement that Huck had to wait for Douglas to grumble over her food, a conflict arises. The general knowledge that when people bow over their food and speak before dinner they are praying implies that Douglas was praying rather than grumbling over the food.

(32)

x' y' u' s'
1:twain(x') 2:author(x') 3:adventures.of.huck.finn(y') 4:wrote(x',y') 5:mumble.before.dinner.is.prayer … 6:huck(s') 7:narrator(s') … 8:douglas(u') 9:rang.bell(u') 10:wait.for.dinner(s') 11:grumbling.over.food(u') …
5>{6,7,8,9,10,11}>{1,2,3,4}

Eliminating one or several of the least entrenched conditions (1,2,3 or 4) will not make the workspace consistent. Hence we move up in the epistemic entrenchment ordering. Eliminating only condition 11 (Douglas grumbled over the food) will make the workspace consistent. But note that this is one of the new contributions, so we have to be cautious and update with the hedged variant, i.e. that the speaker asserted, and therefore accepted as true, that Douglas grumbled over the food.

(33)

x' y' u' s'
1:twain(x') 2:author(x') … 5:mumble.before.dinner.is.prayer … 6:huck(s') 7:speaker(s') … 8:douglas(u') 9:rang.bell(u') 10:wait.for.dinner(s') ~~11:grumbling.over.food(u')~~ 12:ACCEPT$_{s'}$ grumbling.over.food(u')
5>{6,7,8,9,10,~~11~~,12}>{1,2,3,4}

Interestingly, unlike with the cautious update in the vegan case from section 3.3, the workspace can retain the generic background information that when people bow over their food and softly speak before dinner they are praying, even after the cautious update with the hedged information that the speaker, Huck Finn, takes Douglas to be grumbling. This is because the workspace derives from the (reader's conception of the) official common ground between Twain and his readers, not that between Huck Finn and his (fictional) narratee. Hence, the workspace that we copied off this common ground still represents – if anything – what Twain and his readers commonly accept, if only temporarily while entertaining the content of the fiction at hand. This explains how in the case of fiction, a reader and author can have what Booth (1961) calls a communion behind the narrator's back; It is as if the reader and Twain are listening to the narrator together and it is common ground between them that the narrator believes something false.

If we apply fictive closure to (33) we update the official common ground with this information embedded under the relevant fiction operator. Hence after reading this passage the reader takes it to be common ground between her and Twain and other engagers that in *The Adventures of Huckleberry Finn* the widow Douglas rung a bell for supper and muttered something before dinner, and that Huck (mistakenly) took her to be grumbling unhappily over her food.

Generalizing beyond this particular example it is worth stressing that cultural or other background information, like our condition 5 about prayer and dinner customs in (33), need not outrank incoming fictional statements, even in a first person narrative (for instance if the narrator were judged more mature and reliable, or if we're dealing with a fantasy story about an alien civilization without religion or prayer). How highly entrenched certain background information is relative to incoming textual information heavily depends on genre, i.e. with respect to what clusters of facts the fiction is expected to be realistic (Ryan 1991). For instance, if we are aware of the genre of *A Christmas Carol* as a 19th century gothic horror story, we may expect it to be realistic with respect to geographical facts (i.e. where countries and cities are located) but not necessarily with respect to all taxonomic facts (i.e. what species exist and how they are individuated). Therefore a statement such as "He rode into London, the capital of France" would trigger an unreliable narrator interpretation (we are reluctant to cancel our geographical background information that London is the capital of England). But a statement such as "[H]e looked the phantom through and through, and saw it standing before him" will not trigger an unreliable narrator interpretation in the context of *A Christmas Carol*. On the other hand, knowledge of for instance crime novel genre conventions may lead us to expect *A Study in Scarlet* to be realistic with respect to both geographical and taxonomic

facts. Hence the same two fictional statements as part of *A Study in Scarlet*, would both trigger an unreliable narrator interpretation where Watson is hallucinating or otherwise mistaken about the location of London and the existence of ghosts.

4.3 Imaginative resistance

In section 1 we discussed a seemingly related case of Authorial Authority breaking down, viz. in the story called *Fish Tank* which is a typical example of what philosophers call Imaginative Resistance:

(34) Sara never liked animals. One day, her father caught her kicking the neighbor's dog. He got really angry and she was grounded for a week. To get back at her father she poured bleach in the big fish tank, killing all the beautiful fish that he loved so much. Good thing that she did, because he was really annoying.

4.3.1 Face value interpretation

Let's explore what a face value interpretation of this story would look like. Suppose that the reader of the story believes it is common ground between her and the author that killing animals (for no good reason) is wrong. Moreover, she takes this moral law to be quite deeply entrenched, i.e. she'll be quite reluctant to give up the assumption that it is part of the established common ground between her and the author on the basis of new information and experiences. At the start of the discourse the workspace is a copy of this common ground:

(35)

x' y'
1:author(x') 2:fish.tank(y') 3: wrote(x',y') 4:killing.animals.is.wrong ...
4>{1,2,3}

The workspace is updated with the statements that Sara never liked animals, kicked the dog, got grounded, and poured bleach in the fish tank. Since there doesn't seem to be a personal, first person narrator the reader might assume an impersonal, omniscient narrator and, per Authorial Authority, assign these statements the highest possible ranking in the epistemic entrenchment ordering. Now we expand the workspace with the statement that Sara did a good thing, and still treat that as equally ranked with the rest of the text. We get a conflict between the moral law about killing animals and the fact that it's a good thing she killed her father's fish, which will be resolved by eliminating the moral law.

(36)

x' y' u' v'
1:author(x') 2:fish.tank(y') 3:wrote(x',y') ~~4:killing.animals.is.wrong~~ . . . 5:sara(u') 6:father(v',u') 7:¬ like.animals(u') 8:pour.bleach(u') 9:did.good(u')
{5,6,7,8,9}>~~4~~>{1,2,3}

Fictive closure leads to an output where the reader takes it to be common ground in the community of engagement that killing animals is wrong, and there's a story called *Fish Tank* in which a girl called Sara poured bleach in a fish tank because she's annoyed and this was a good thing, since apparently in this fictional world moral laws are such that killing animals for trivial reasons is okay.

4.3.2 Unreliable narrator interpretation

Intuitively the face value interpretation is unsatisfactory for *Fish Tank*; Many readers – even non-philosophers – feel that even though it is explicitly stated that Sara did a good thing, she actually did *not* do a good thing in the fictional world she inhabits. Empirical studies like Kim et al. (2018) support this intuition, suggesting that fictional statements, at least for some people, are not always shielded from contraction and that moral truths are really quite hard to give up.[15] The flexibility of our model allows us to model this alternative interpretation by simply adopting a different entrenchment ranking strategy on the incoming textual information.

Concretely, for the non-face-value reader, the initial updates are the same: incoming information gets a high rank by default, as we're dealing with fiction and there is no reason to distrust the fictional speaker, in fact no reason to assume the presence of a narrating source at all. When we get to the final statement the reader may reconsider this assumption, because it will lead to the unwanted face value interpretation. So instead let's rank the final sentence contribution below the obviously deeply entrenched moral law. Our update algorithm, as spelled out in (27), then leads to the elimination of the final contribution followed by a hedged update:

15 For whatever reason, see e.g. Gendler (2000), Yablo (2002), or Weatherson (2004) for philosophical investigations of what makes moral truths especially hard to give up, and Andow (2019) for an empirical investigation.

(37)

x' y' u' v'
1:author(x') 2:fish.tank(y') 3: wrote(x',y') 4:killing.animals.is.wrong ... 5:sara(u') 6:father(v',u') 7:¬ like.animals(u') 8:pour.bleach(u') ~~9:did.good(u')~~ 10:ACCEPT$_{s?}$ did.good(u')
{5,6,7,8,10}>4>~~,9~~}>{1,2,3}

One difference with our previous examples of cautious updating (Maxine the vegan and Huckleberry Finn) is that the indexical/anaphoric element in the hedging operator '*the current speaker* accepts that φ' has no obvious antecedent; there is no discourse referent for a salient current speaker in the workspace universe, as there has been no sign of a first person narrator.[16] And it is not at all clear who this speaker should be, i.e., who is it that asserts and thereby commits themselves to Sara doing a good thing?

A first option is to take the actual author of the story to be the speaker:

(38)

x' y' u' v'
1:author(x') 2:fish.tank(y') 3: wrote(x',y') 4:killing.animals.is.wrong ... 5:sara(u') 6:father(v',u') ... 10:ACCEPT$_{x'}$ did.good(u')
{5...10}>4>{1,2,3}

The resulting interpretation in (38) corresponds to what Gendler (2000) calls a 'pop-out' interpretation. The value judgement in the closing statement is not really a fictional statement but rather an 'intrusion' from the author, breaking the fourth wall to comments on the story he is telling.[17] A reader may resist a pop-out interpretation

16 We might take the evaluative construction *good thing* itself as lexically presupposing a first person judging agent, i.e. good = good according to me. We refrain from going down this path to stay neutral with respect to the semantics and pragmatics of evaluative terms. On our account, following Altshuler & Maier (2018), it is the cautious update itself that pragmatically triggers the accommodation of a fictional first person committed to this moral judgment.

17 Recall from our discussion in section 2.1 that, strictly speaking, x' in this workspace is a copy of the discourse referent x that is representing the author in the official common ground. After fictive closure, the semantic values of the copy and the original come apart, though they'll likely still share many common ground properties and may be considered counterparts. We've also noted there that it may be possible to write a story that conflicts with the existence of the author, in which case this

based on overriding background assumptions about the author (e.g. she knows with what purpose the author wrote *Fish Tank* and that he shares her moral values).

Alternatively, the anaphoric subject of the hedge ACCEPT$_{s'}$ may bind to one of the fictional characters. However, in this particular text there are no textual clues that either of the salient available fictional characters (Sara or her father) is to be understood as uttering these evaluative words (out loud or silently in thought). We see none of the (sometimes subtle and ambiguous) textual and contextual clues that would license a free indirect discourse or protagonist projection interpretation here (Eckardt 2014; Hinterwimmer 2017; Altshuler & Maier 2018; Stokke 2020; Abrusán 2020).

What we're left with is the option of accommodating a new fictional character, s', who is presumed to be offering this evaluation in a speech act: a fictional speaker/narrator responsible for telling the story, or at least this final part of it:[18]

(39)

x' y' u' v' s'
1:author(x') 2:fish.tank(y') 3: wrote(x',y') 4:killing.animals.is.wrong … 5:sara(u') 6:father(v',u') … 10:ACCEPT$_{s'}$ [did.good(u')] 11:narrator(s')
{5…11}>4>{1,2,3}

After fictive closure on this workspace the reader considers it common ground between her and the author that there's a story called *Fish Tank* in which there is a girl named Sara who kills her father's fish. Moreover, in this story killing animals is morally wrong, as in the real world, and finally, this story is partly told from the perspective of a fictional narrator who claims that Sara did a good thing killing the fish. In other words, we started out interpreting the text on a par with the Heinlein story, i.e. as a third person omniscient or rather impersonal narration, every statement to be taken at face value without the mediation of a personal narrator, and then switched to an interpretation along the lines of our Huckleberry Finn interpretation, i.e. as a first person narration, all statements weighed against the available contextual and textual evidence and potentially treated as representing merely the point of view of the fictional character narrating the story.

interpretation route may be blocked.

18 Altshuler & Maier (2018) coin the term 'narrator accommodation' and argue that this is what causes the disruptive experience that is inherent in the phenomenon of imaginative resistance.

5 Conclusion

We can't always take a text at face value – so-called unreliable narrators may present a confused or misleading picture of the fictional world, and in cases of imaginative resistance readers refuse to accept parts of a text as true in the corresponding fictional world. We have proposed a way to model the interpretive processes that allow readers to extract fictional truths from such fictional narratives. Our starting point has been that these processes should apply uniformly across fiction and non-fiction.

We started with basic DRT and added a temporary workspace to ensure a proper separation but close connection between fiction interpretation and the official common ground. The process of interpretation is indeed analyzed uniformly across fiction and non-fiction, except for the final step, where we close the workspace and translate the information gained by the discourse interpretation process back into an official common ground update. We then incorporated insights from belief revision theory to deal with unreliable information sources in both fiction and non-fiction. Combining these two theoretical additions to the established DRT framework allowed us to describe precisely the various interpretation strategies readers can choose when interpreting different types of narratives. On our analysis, the 'epistemic entrenchment' of certain background assumptions relative to incoming information from the discourse or text, as well as the presence or absence of a personified speaker/narrator are the key factors in determining the kinds of readings available.

References

Abrusán, Márta. 2020. Signals of perspective shift in narrative. In Emar Maier & Andreas Stokke (eds.), *The Language of Fiction*, this volume. Oxford: OUP.

Alchourron, Carlos E., Peter Gardenfors & David Makinson. 1985. On the Logic of Theory Change: Partial Meet Contraction and Revision Functions. *Journal of Symbolic Logic* 50(2). 510–530. https://projecteuclid.org/euclid.jsl/1183741857.

Altshuler, Daniel & Emar Maier. 2018. Death on the Freeway: Imaginative resistance as narrator accommodation. In Ilaria Frana, Paula Menéndez-Benito & Rajesh Bhatt (eds.), *Making Worlds Accessible: Festschrift for Angelika Kratzer*, Amherst.

Andow, James. 2019. Why don't we trust moral testimony? *Mind & Language* http://dx.doi.org/10.1111/mila.12255. https://onlinelibrary.wiley.com/doi/abs/10.1111/mila.12255.

Badura, Christopher & Francesco Berto. 2018. Truth in Fiction, Impossible Worlds, and Belief Revision. *Australasian Journal of Philosophy* 0(0). 1–16. http://dx.doi.org/10.1080/00048402.2018.1435698. https://doi.org/10.1080/00048402.2018.1435698.

Bonomi, Andrea & Sandro Zucchi. 2003. A pragmatic framework for truth in fiction. *Dialectica* 57(2). 103–120. http://dx.doi.org/10.1111/j.1746-8361.2003.tb00259.x.

Booth, Wayne. 1961. *The Rhetoric of Fiction*. Chicago: UCP.

Currie, Gregory. 1990. *The Nature of Fiction*. Cambridge: CUP.

Eckardt, Regine. 2014. *The Semantics of Free Indirect Speech. How Texts Let You Read Minds and Eavesdrop*. Leiden: Brill.

Franzén, Nils. 2020. Truth in fiction: In defense of the Reality Principle. In Emar Maier & Andreas Stokke (eds.), *The Language of Fiction*, this volume. Oxford: OUP.

Friend, Stacie. 2011. The great beetle debate: a study in imagining with names. *Philosophical Studies* 153(2). 183–211. http://dx.doi.org/10.1007/s11098-009-9485-4. http://link.springer.com/10.1007/s11098-009-9485-4.

Friend, Stacie. 2017. The Real Foundation of Fictional Worlds. *Australasian Journal of Philosophy* 95(1). 29–42. http://dx.doi.org/10.1080/00048402.2016.1149736. https://doi.org/10.1080/00048402.2016.1149736.

Gendler, Tamar Szabó. 2000. The Puzzle of Imaginative Resistance. *Journal of Philosophy* 97(2). 55–81.

Geurts, Bart. 1999. *Presuppositions and Pronouns*. Amsterdam: Elsevier.

Geurts, Bart, David I. Beaver & Emar Maier. 2016. Discourse Representation Theory. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University spring 2016 edn. https://plato.stanford.edu/archives/spr2016/entries/discourse-representation-theory/.

Groenendijk, Jeroen & Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and Philosophy* 14(1). 39–100. http://dx.doi.org/10.1007/BF00628304.

Grove, Adam. 1988. Two modellings for theory change. *Journal of Philosophical Logic* 17(2). 157–170. http://dx.doi.org/10.1007/BF00247909. https://doi.org/10.1007/BF00247909.

Gärdenfors, Peter. 1988. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. Cambridge: MIT Press.

Gärdenfors, Peter & David Makinson. 1988. Revisions of Knowledge Systems Using Epistemic Entrenchment. In *Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning About Knowledge* TARK '88, 83–95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=1029718.1029726. Event-place: Pacific Grove, California.

Hamm, Fritz, Hans Kamp & Michiel Van Lambalgen. 2006. There is no opposition between Formal and Cognitive Semantics. *Theoretical Linguistics* 32(1). 1–40. http://dx.doi.org/10.1515/TL.2006.001. http://www.reference-global.com/doi/abs/10.1515/TL.2006.001.

Hansson, Sven Ove. 1994. Taking Belief Bases Seriously. In Dag Prawitz & Dag

Westerståhl (eds.), *Logic and Philosophy of Science in Uppsala: Papers from the 9th International Congress of Logic, Methodology and Philosophy of Science* Synthese Library, 13–28. Dordrecht: Springer Netherlands. http://dx.doi.org/10.1007/978-94-015-8311-4_2. https://doi.org/10.1007/978-94-015-8311-4_2.

Hansson, Sven Ove. 1998. Revision of Belief Sets and Belief Bases. In Didier Dubois & Henri Prade (eds.), *Handbook of Defeasible Reasoning and Uncertainty Management Systems Volume 3: Belief Change*, Dordrecht: Springer. https://doi.org/10.1007/978-94-011-5054-5.

Heim, Irene. 1982. *On the Semantics of Definite and Indefinite Noun Phrases*. Amherst: UMass PhD Thesis. http://semanticsarchive.net/Archive/Tk0ZmYyY.

Hinterwimmer, Stefan. 2017. Prominent protagonists. *Journal of Pragmatics* http://dx.doi.org/10.1016/j.pragma.2017.12.003. http://www.sciencedirect.com/science/article/pii/S0378216617308342.

Kamp, Hans. 1981. A theory of truth and semantic representation. In Jeroen Groenendijk, Theo Janssen & Martin Stokhof (eds.), *Formal Methods in the Study of Language*, 277–322. Amsterdam: Mathematical Centre Tracts.

Kamp, Hans. 2015. Using Proper Names as Intermediaries Between Labelled Entity Representations. *Erkenntnis* 80(2). 263–312. http://dx.doi.org/10.1007/s10670-014-9701-2. http://link.springer.com.proxy-ub.rug.nl/article/10.1007/s10670-014-9701-2.

Kamp, Hans. 2016. Articulated Contexts. Unpublished Ms. Stuttgart/Austin.

Kamp, Hans. 2020. Sharing real and fictional reference. In Emar Maier & Andreas Stokke (eds.), *The Language of Fiction*, this volume. Oxford: OUP.

Kamp, Hans & Uwe Reyle. 2011. Discourse Representation Theory. In *Semantics*, vol. 1, 872–923. Berlin: De Gruyter. https://dx.doi.org/10.1515/9783110226614.872.

Kim, Hanna, Markus Kneer & Michael T. Stuart. 2018. The Content-Dependence of Imaginative Resistance. In Florian Cova & Sébastien Réhault (eds.), *Advances in Experimental Philosophy of Aesthetics*, 194–224. Bloomsbury.

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

Maier, Emar. 2017. Fictional names in psychologistic semantics. *Theoretical Linguistics* 43(1-2). 1–42. http://dx.doi.org/10.1515/tl-2017-0001.

Matravers, Derek. 2014. *Fiction and Narrative*. Oxford University Press. http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199647019.001.0001/acprof-9780199647019.

Nebel, Bernhard. 1998. How Hard is it to Revise a Belief Base? In Didier Dubois & Henri Prade (eds.), *Belief Change* Handbook of Defeasible Reasoning and Uncertainty Management Systems, 77–145. Dordrecht: Springer Netherlands. http://dx.doi.org/10.1007/978-94-011-5054-5_3. https://doi.org/

10.1007/978-94-011-5054-5_3.
Ryan, Marie-Laure. 1981. The pragmatics of personal and impersonal fiction. *Poetics* 10(6). 517–539. http://dx.doi.org/10.1016/0304-422X(81)90002-4.
Ryan, Marie-Laure. 1991. *Possible Worlds, Artificial Intelligence, and Narrative Theory*. Indiana University Press. http://dl.acm.org/citation.cfm?id=531694.
van der Sandt, Rob. 1992. Presupposition projection as anaphora resolution. *Journal of Semantics* 9(4). 333–377. http://dx.doi.org/10.1093/jos/9.4.333.
Semeijn, Merel. 2017. A Stalnakerian Analysis of Metafictive Statements. In *Proceedings of the 21st Amsterdam Colloquium*, ILLC. http://events.illc.uva.nl/AC/AC2017/Proceedings/.
Stalnaker, Robert. 1984. *Inquiry*. Cambridge: MIT Press.
Stokke, Andreas. 2013. Lying and Asserting. *Journal of Philosophy* 110(1). 33–60.
Stokke, Andreas. 2020. Protagonist projection, character focus, and mixed quotation. In Emar Maier & Andreas Stokke (eds.), *The Language of Fiction*, this volume. Oxford: OUP.
Walton, Kendall L. 1990. *Mimesis as Make-Believe: On the Foundations of the Representational Arts*. Cambridge: Harvard University Press.
Weatherson, Brian. 2004. Morality, Fiction, and Possibility. *Philosopher's Imprint* 4(3). http://hdl.handle.net/2027/spo.3521354.0004.003.
Wieland, Nellie. 2020. Metalinguistic acts in fiction. In Emar Maier & Andreas Stokke (eds.), *The Language of Fiction*, this volume. Oxford: OUP.
Williams, Mary-Anne. 1994. On the logic of theory base change. In Craig MacNish, David Pearce & Luís Moniz Pereira (eds.), *Logics in Artificial Intelligence* Lecture Notes in Computer Science, 86–105. Springer Berlin Heidelberg.
Yablo, Stephen. 2002. Coulda, Woulda, Shoulda. In Tamar S. Gendler & John Hawthorne (eds.), *Conceivability and Possibility*, 441–492. Oxford University Press.
Zucchi, Alessandro. 2020. On the generation of content. In Emar Maier & Andreas Stokke (eds.), *The Language of Fiction*, this volume. Oxford: OUP.