Metafiction: the fourth wall against possible world semantics

Deniz Satık, Harvard University
deniz@g.harvard.edu

Discourse involving non-existent fictional characters, such as *Sherlock Holmes is a detective*, is prima facie a source of trouble for the idea that proper names have extensions. Lewis (1978) provides an account of such sentences in which *Holmes* refers to the individual(s) corresponding to Holmes in the possible worlds similar to the Holmes stories. This paper provides several novel arguments involving *metafictional* statements, such as *Holmes knows he does not actually exist*, or sentences which "break the fourth wall" such as *Holmes spoke to his actual audience* to argue that Lewis cannot account for discourse involving metafiction. In the spirit of Pietroski (2005), this paper argues that, given the peculiarities of discourse involving fiction which we intuitively find true, we ought to give up the idea names refer to objects in the outside world. Rather, we should assume semantic internalism, which claims that expressions do not have extensions; meanings themselves are mental representations.

Keywords: fiction, semantics, Millianism, extensions, internalism, possible worlds

1 Introduction

Discourse involving fictional characters has remained a source of confusion and difficulty for philosophers of language and even metaphysicians, and troubling for the various views that they hold. Though numerous analyses of discourse involving fiction have been proposed, philosophers have not considered, in detail, certain peculiar facts about fictional discourse which are intuitively true, and the consequences that arise as a result. In this paper, I discuss consequences that arise from sentences involving *metafictional* facts. For example, I discuss the metaphysical and semantic consequences of one particular phenomenon known as *breaking the fourth wall*.

But first, let us begin by discussing what it means for discourse to involve fiction. In fictional contexts and in ordinary conversation, people often utter sentences such as *Harry Potter is a wizard* which they take to be undoubtedly true, despite being aware of the fact that Harry Potter doesn't actually exist. If Harry Potter doesn't exist, then why aren't these sentences meaningless? Even more surprisingly, we would expect such sentences to trigger *presuppositional failure*. For example, to say that *the king of France is bald* triggers presuppositional failure, given that there is no king of France; the sentence presupposes that the king of France exists. Yet, *Harry Potter is a wizard* does not trigger presuppositional failure.

There are two ways to evaluate sentences like *Harry Potter is a wizard*. The first is what Evans (1982) dubs as the *conniving* use of such sentences, in which the sentences have merely fictional truth-conditions, where they are true merely from the standpoint of the fiction. On the conniving

use, the utterer is engaged in pretense when he or she utters that *Harry Potter survived the Avada Kedavra curse*, as this has a fictional truth-value and not an actual one.

Another use, and the use that we are primarily concerned with in this paper, is called the non-conniving use of such sentences, in which the sentences have real truth-conditions. For example, if someone asked you what Harry Potter's occupation was, according to the Harry Potter stories, to say that *Harry Potter is a wizard* has a real truth-value, and not just a fictional one. A challenge for the philosopher of language is to give us a theory of meaning in which both uses of this sentence can be true, or to give us an error theory which attempts to explain away the various intuitions that speakers have about the truth values of these kinds of sentences.

Millianism, the common sense view that the semantic value of a name is its referent, runs into an unsurprising amount of trouble when attempting to provide a meaning for sentences of this kind. We assume that a sentence is meaningful just in case every constituent is meaningful. However, the proper name *Harry Potter* seems to have no semantic value since it refers to nothing.

Yet, one very common approach, proposed by Lewis (1978), claims that when we utter such sentences, we refer to a person named Harry Potter in other possible worlds, which themselves resemble the worlds of Harry Potter. A version of this was defended at greater length by Berto and Badura (2019), by further adding the notion of *impossible worlds*, which increases the range of peculiarities–especially those involving logical impossibility in fiction–which Lewis could not originally account for. This approach has gained acceptance among formal semanticists. For example, Fintel and Heim (2002), a textbook on intensional semantics, presents Lewis's account as a semantics for discourse involving fiction.

This paper presents a commonly used literary device within fiction known as *direct address* or *breaking the fourth wall* within popular culture, which has not yet been discussed. It is a literary device in which a fictional character speaks directly to and acknowledges the reader. For example, Italo Calvino's novel, *If on a Winter's Night a Traveler*, addresses the reader directly in the second person and discusses the supremacy of the reader, because the reader is able to realize the text in her imagination.

My goal in this short paper is to raise problems for Lewis (1978)'s possible world semantics for discourse involving *meta*fiction. For example, I build a *fourth wall* sentence which intuitively seems true, but implies transworld communication, which is impossible as worlds are not spatio-temporally connected. An example is given in (1) below.

(1) According to the Holmes stories, Holmes spoke to his actual audience.

Furthermore, Berto and Badura (2019)'s update of Lewis's account is not able to obtain the truth of this sentence either, given that communication between the actual world and impossible worlds is still impossible.

I propose that the account that is best equipped at handling these aforementioned peculiarities is internalism on semantic meaning, which is most commonly associated with Chomsky and Smith (2000), and more recently defended by Pietroski (2018). Given that meaning, as Pietroski (2018) puts it, is just an instruction to build concepts of a special sort, an internalist approach provides a straightforward way of providing a semantics for sentences such as (1).

2 Possible world semantics

Possibilism is the view that fictional entities are entities that exist not in the actual world, but in other possible worlds. On this view, there are possible worlds in which the Sherlock Holmes stories really do happen, and Holmes really does solve cases and defeat Moriarty. The most well-known and developed of this kind is in Lewis (1978), which is successful at solving many of the problems posed to it by many philosophers, and especially by Kripke (1980).

Lewis thinks that we should not take the non-conniving use of fictional sentences such as *Holmes is a detective* at face value; rather, he believes that we should paraphrase them with an intensional operator such as *according to the Holmes stories*. The sentence with the operator is true just in case Holmes is a detective in the worlds picked out by the intensional operator. The sentence *Holmes is a detective* without the operator is either false or meaningless, depending on one's view of presuppositional failure that we have just discussed.

As Lewis points out, it is not right for the operator to just pick out all the worlds in which all of the events as described by the story are true. Let us assume that Conan Doyle wrote the stories as fiction. As Kripke points out, by pure chance, it may be that our own world could be a world in which the Holmes stories take place without Doyle knowing of it. If Sherlock Holmes happened to have existed in real life without Doyle's knowledge, this Holmes still would not have been the Holmes that Doyle wrote about. After all, it is false that in our world, the name *Sherlock Holmes* refers to someone though it is true in the stories. He does not exist.

Lewis's solution for this is as follows. We should consider the worlds in which events are qualitatively identical to the events as described in the fiction, in which the fiction is told as a story told by a particular storyteller at some time, about known facts rather than about made up stories. Such a story must be told by Conan Doyle himself. For example, if I were to tell the same story word for word, that would not make it the same story as Conan Doyle's. Different acts of storytelling lead to different stories. In these worlds, the proper name *Sherlock Holmes* behaves as a rigid designator. So, it cannot designate anyone in our world: hence, it can account for Kripke's problem that we cannot identify a fictional object with a possible object.

This analysis still has a few problems. It brings in far too many worlds in which all sorts of things are true. It is intuitively true, but never stated, that Holmes is not an alien from Mars, and that Holmes does not eat mud. But this analysis brings in all these worlds in which the opposite is the case. Lewis suggests that we analyze true fictional statements as counterfactuals, and pick out the worlds which differ the least from our own world.

We pick out the worlds which differ the least from our own world, but which world is the actual world? It is a contingent matter which possible world is the actual world. There are still many things we don't know about the actual world: it could be possible that the Holmes stories have truths which no one knows about. Instead of using the actual world as a standpoint for our fictional truths, we should use the worlds in which the overt beliefs of the community of origin are true—the community of origin being the community in which Doyle wrote the stories.

Lewis's analysis solves many well-known problems that plague possibilist accounts. For example, Kripke argues that there are too many possible worlds to pick out: there are even worlds in which Holmes is from Tatooine. As Lewis says, after restricting the possible worlds we can consider, there's no point in picking out the world in which the stories take place. There are worlds in which Holmes has the blood type A and others in which he has the blood type AB—it is absurd to suppose that the story has answers for these. The worlds of the Holmes stories are

plural, and different worlds may have different answers to these questions.

Many problems still remain, though. There are countless stories which involve logically impossible events. A good example is time travel, and impossible events involving the Grandfather paradox–what if you go back into the past and kill your grandfather when he was a little boy? Would you disappear from existence? There is also a great deal of philosophical literature which discusses this paradox, and many agree that it is logically impossible to do so.

In completely unrelated work, Lewis (1976) proposes a very interesting solution to the Grandfather paradox. He agrees that it is logically impossible, but commonplace reasons such as slipping on a banana peel might stop the logically impossible from happening. But unfortunately for Lewis, we can write such a story, where Holmes travels in time and kills his grandfather before he was born. Likely many such stories have already been written. And this seems not only intuitively true, it is impossible for Lewis's account as stated to account for, given that logically impossible events may not happen in possible worlds.

At the very least, one must expand the possible worlds apparatus by adding in impossible worlds–where logically impossible events happen, or logically impossible objects do exist. This is precisely what Berto and Badura (2019) proposes, to account for similar scenarios where contradictions are true. But as we will see in the next section, this is far from enough to account for all of the peculiarities we see in fiction. There are a number of metafictional truths and phenomena which, by their very virtue, are related to the actual world rather than some possible or impossible world. And it is impossible to account for these by further expanding the apparatus.

3 Metafiction

Let us now consider what the notion of *discourse involving metafiction* is. It is the central notion of this paper: very broadly, I use it to refer to discourse involving fiction in which the fictional character is aware that he or she is fictional. Furthermore, I use it to refer to discourse in which we make actually true claims about fictional characters. For example, Lewis himself admits that he cannot account for sentences such as *Holmes is a fictional character*, a problem which he leaves open.

This problem is much more serious than it appears. He does not offer solutions for further kinds of metafictional statements. For example, a metafictional sentences such as *Holmes is smarter than Gregory House*, where we compare two individuals across different possible worlds, or *Holmes appears in more than one literary work*, where we discuss a possible individual appearing in a book, do not have an account either. We cannot do this with the intensional operator. The first sentence is not true in any possible world, while the second is true in the actual world.

With this background, we can now consider more serious metafictional objections. For example, let us consider the *fourth wall* sentence such as the one in (2).

(2) Holmes spoke to his actual readers.

Let us assume for simplicity that Holmes breaks the fourth wall in the Holmes stories as Doyle wrote it, even though he does not. The reader can substitute in a story where the fourth wall is broken, if she so wishes. There are many such stories, such as *If on a Winter's Night a Traveler*. As we have discussed, following Lewis, we can read this with the intensional operator:

(3) According to the Holmes stories, Holmes spoke to his actual readers.

This is still intuitively true, but under Lewis's semantics it is false. It says that Holmes, an individual in another possible world, spoke to his actual readers, which exist in this world. Lewis himself himself points out the impossibility of this in Lewis (1986), which is unrelated work: possible worlds do not have any sort of spatiotemporal relation with each other and Lewis's counterfactual analysis of causation prevents the possibility of transworld causation.

One way out of this might be to suggest that *speaking to* is not a causal relationship between two objects. It may be a mere intentional one, as seen by sentences such as *Gary spoke to his intro to logic students, but none of them listened.* In this sentence, there is no direct causal relation between Gary and his intro to logic students but it still is intuitively true. Similarly, (2) can be true independent of any causal relation.

But suppose we had a version of (2) in which we could force the causal reading between Holmes's action in another possible world and us in the actual world:

(4) When Holmes spoke to his actual readers, he made some of them laugh.

This sentence seems intuitively true as well, given an appropriate context. What this sentence says with the intensional operator is this: according to the story, when Holmes spoke to his actual readers from the world of fiction, his action was such that it caused some of his actual readers to laugh. We are back to where we started, since this sentence implies causation across worlds.

In that case, suppose that one objected as follows: the sentence is intuitively false, because when we read the event as described in the story and laugh, it's not really Holmes that makes us laugh. It's thinking about the event as described in the story that makes us laugh, and not Holmes interacting with us across possible worlds that does. Prima facie this is correct. There seems to be a sense in which fictional characters really can make their audiences react in certain ways.

For example, it is completely felicitous to say *Holmes pissed me off in Doyle's latest book* or *Holmes made me happy when he kissed Watson on the lips*–although it is not clear what Lewis's analysis of these would be, given that it does not seem possible for the intensional operator to be present in these, let us grant them to him. Such sentences are not directly causal, so this might seem to provide a way out of the *fourth wall* objection.

But we can make a sentence which is directly causal with a slight modification: for example, *Holmes wrote a letter to Watson that pissed him off* is such a sentence, where Holmes is the cause of an event that led to a reaction in Watson. Similarly, suppose that in the canon of the Holmes stories, Conan Doyle had Holmes write a letter to Conan Doyle's second wife Jean; let us imagine that the contents of the letter made her laugh.

Now it might be objected that it is in fact Conan Doyle himself trying to communicate with his wife, rather than Holmes himself. But Conan Doyle himself could reasonably deny that he was trying to communicate with his wife; he could claim that he intended for Holmes himself to do so. Indeed, I, the author, or you, the reader of this paper, can write such a fictional story as well, and reasonably claim that it is the fictional character communicating with a certain reader. For that is the power of fiction; anything that we desire to be true in the story goes.

Lewis may attempt to provide another analysis of such sentences, in which they are true just in case Holmes really does speak to a future reader or an observer of his real life endeavors in the worlds of fiction. But what we care about in (2) is that Holmes spoke to his readers in the actual world, and not in any other possible world. Furthermore, the author of the story can simply make it canon that Holmes did not have any future readers or observers in the worlds in which the stories take place. For example, he can claim that Holmes travels to Mars completely alone.

There are metafictional truths apart from *Holmes is a fictional character* and *fourth wall* sentences. Indeed, Lewis's account of fictional characters runs into trouble when dealing with metafictional truths of any kind, in which fictional characters are aware of the fact that they are fictional characters, and numerous other facts that follow. Suppose the sentences below are true according to the story:

(5) Holmes knows that he doesn't exist.

(6) Holmes knows that he was created by Conan Doyle.

As long as the author writes such a story, we can attribute beliefs of this kind to any fictional character, and speakers can believe them to be true. On Lewis's view, possible worlds are just as real as the actual world, and they are no different in kind from the actual world. It would be contradictory for Holmes to think that he does not exist if this is the case. Furthermore, in the worlds where such stories are told as known fact, Holmes can at most only be aware of the fact that his story was told by Conan Doyle to other people. It would be strange for him to think that he was literally created by the actual Conan Doyle.

But there is nothing logically impossible about having such beliefs. Indeed, Lewis can even allege that Holmes may be deluded and have these beliefs. But it cannot be true for Holmes as it is for us: from the standpoint of the story, we know that Holmes is certainly not deluded (since the story says so), and he can be justified in having these beliefs as a fictional character. We cannot say that Holmes is justified in having such beliefs if we think that stories of fiction take place in other possible worlds. So possible worlds are not a good metaphysical foundation for discourse involving metafiction.

Finally, we can also consider sentences from the standpoint of the Holmes stories. We can have an utterance in the story as follows:

(7) I do not actually exist.

According to Lewis's analysis of actuality, it is an indexical whose extension is determined by the context of the speaker's utterance. In other words, it would pick out the world in which Holmes uttered this sentence. We would agree with Holmes if he were to utter this sentence, but it is plainly false under Lewis's analysis.

While Lewis's view is successful at solving many of the possibilist's problems, the fundamental problem with this approach is the existence of metafictional discourse. The author of a story can make canon whatever she chooses: she can make it so that Holmes is aware that he is a fictional character, that he does not actually exist, and that he attempts to communicate with his actual audience. But in the worlds corresponding to Holmes's story, it is absurd for Holmes to have such beliefs; he would not hold them unless he were deluded. It is hopefully the case that neither the reader nor the author of the Holmes stories believe that Holmes is deluded.

4 Semantic internalism

For reasons of space, I have not discussed several other possible solutions to discourse involving fiction. But the *fourth wall* objection raised in section 3 applies just as well to two other commonly assumed approaches for discourse involving fiction. These are Meinongianism, defended by Parsons (1980) and Zalta (1983), which posits that there are nonexistent objects with stand-in

for fictional characters, and Artifactualism, defended by Thomasson (1998) where fictional characters are abstract objects which come into being after their creator conceives of them.

The basic problem is that nonexistent and abstract objects–neither of which are present in space-time–are incapable of interacting, or communicating, with concrete objects. One account that provides a straightforward account of the *fourth wall* sentence is the antirealist approach to discourse involving fiction, under which such discourse is merely pretense of a certain kind in an authorized game, defended most effectively by Walton (1990). For example, Walton gives the example of *Tom Sawyer attended his own funeral*, which would be understood as follows:

(8) The Adventures of Tom Sawyer is such that one who engages in pretense of kind K [claiming "Tom Sawyer attended his own funeral"] in a game authorized for it makes it fictional of himself in that game that he speaks truly.

To say that *Gregor Samsa is a fictional character* would be as follows (p. 423):

(9) There may be an unofficial game in which one who says ["Gregor Samsa is a fictional character"] fictionally speaks the truth, a game in which it is fictional that there are two kinds of people: "real" people and "fictional" characters.

But we do not seem to participate in pretense when we speak of the other kinds of metafictional truths that we have seen in section 3. To say that *Holmes is a fictional character* is certainly correct and actually true from our perspective, and not a matter of pretense. Further, we would even have to engage in pretense to say something as plausible as *According to the Holmes stories, Holmes is a detective*, which seems unlikely.

Only one approach seems to remain–semantic internalism–and it is the one that I favor. All of the prior approaches that we have seen are forms of Millianism, which I choose to give up given the peculiarities of discourse involving metafiction. Rather than claiming that the semantic value of a name is its referent, I follow Chomsky and Smith (2000), Pietroski (2005) and Pietroski (2018) among others, in which mind-world reference relations play no role in semantics.

According to semantic internalists, sentences may have truth-conditions, and names can have referents, but these are just potential *uses* of sentences and words rather than their *meanings*. Pietroski (2018) provides the most well-developed account of semantic internalism to date. According to this, meanings are "instructions for how to build concepts of a special sort." Concepts are mental representations: the meaning of an expression is merely an instruction to form a certain kind of mental representation. Such concepts may have extensions, but expressions themselves do not, so we must reject the very foundation of classical semantics.

Pietroski (2005) provides several arguments in favor of his approach, which are very similar to the problems I have raised in this paper. Pietroski notes that even discourse involving nonfictional entities can lead to contradictory truth-conditions:

(10) The red book is too heavy, although it was favorably reviewed, and the blue one is boring, although everyone is reading it.

The red book in (10) has a concrete property, which is to be *heavy*; and it also has an abstract property, which is to be *favorably reviewed*. It cannot be a concrete and an abstract object at the same time. Pietroski also notes the contradiction with *France* in (11), which cannot be both *hexagonal* and *a republic*:

(11) France is a hexagonal republic.

5 Conclusion

To conclude, it seems unlikely for either Lewis (1978) or Berto and Badura (2019) to account for discourse involving metafiction that I have discussed in this paper. But one obvious approach would be to deny that there is such a phenomenon such as *direct address* or *breaking the fourth wall* at all. One could think that these are merely literary devices used by the author to address the reader, and they are not really part of the story, or true in the established facts of the canon.

This is true to an extent. There are video games in which the main character tells the player which buttons to press, and the player is never mentioned again during the story. In that case, a *fourth wall* sentence would intuitively come out as false, as it is a mere tool for the author to communicate with the player. But to say that such sentences are never true in a fictional universe would be going too far. In fiction, as we have discussed, whatever the author intends to be true–or a part of the story's *canon*–is the case according to the story.

Suppose the author of the Holmes stories intended for *fourth wall* events to be true according to the story. We can imagine the Holmes stories being based on an investigation to figure out who the reader is, or on entertaining the reader. In that case, *breaking the fourth wall* is surely not a mere literary device used by the author, and it is a genuine part of the story and the established facts of the fictional universe–or what we call the canon of the fictional universe.

In explaining the truths of metafictional discourse, the way of the possibilist–and the semantic externalist in general–is difficult. We have seen that there is not currently an externalist semantics which provide an account for sentences involving metafictional truths. But if we give up the idea that expressions such as proper names do not refer to the objects which we believe are associated to them, we have a ready-made account capable of handling discourse involving metafiction.

References

Berto, F. and C. Badura (2019). Truth in fiction, impossible worlds, and belief revision. *Australasian Journal of Philosophy 97*(1), 178–193.

Chomsky, N. and N. Smith (2000, apr). *New Horizons in the Study of Language and Minds*. Cambridge University Press.

Evans, G. (1982). *The Varieties of Reference*. Oxford University Press.

Fintel, K. v. and I. Heim (2002). Notes on intensional semantics. unpublished manuscript, MIT, Cambridge.

Kripke, S. (1980). *Naming and Necessity*. Harvard University Press.

Lewis, D. (1986). *On the plurality of worlds*. Blackwell.

Lewis, D. K. (1976). The paradoxes of time travel. *American Philosophical Quarterly 13*(2), 145–152.

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

Parsons, T. (1980). *Nonexistent Objects*. Yale University Press.

Pietroski, P. (2018). *Conjoining Meanings*. Oxford University Press.

Pietroski, P. M. (2005). Meaning before truth. In G. Preyer and G. Peter (Eds.), *Contextualism in Philosophy: Knowledge, Meaning, and Truth*. Oxford University Press.

Thomasson, A. L. (1998). *Fiction and Metaphysics*. Cambridge University Press.

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

Zalta, E. N. (1983). *Abstract Objects: An Introduction to Axiomatic Metaphysics*. D. Reidel.