Breaking the fourth wall and (meta)fictional reference *

Merel Semeijn[1]

Institut Jean Nicod (École normale supérieure), Paris, France
semeijn.merel@gmail.com

Abstract

I investigate statements in fiction that 'break the fourth wall' (i.e., statements through which a fictional character somehow acknowledges the fictionality of their world) and suggest that they are a mirror image of 'parafictional statements', i.e., statements about what is true in some fiction. I explore two possible analyses, according to which statements that break the fourth wall are either a type of fictional statement or are a type of metafictional statement, and propose a synthesis of these two analyses.

We talk about fictions and fictional entities in a variety of meaningful ways. Following Recanati's [11] terminology, we can distinguish between at least three different kinds of discourse that involve reference to fictional entities: 'fictional', 'metafictional' and 'parafictional'. Consider the following statements:

(1) Frodo had a very trying time that afternoon

(2) Frodo was invented by Tolkien

(3) Frodo is a fictional character

(4) In *The Lord of the Rings*, Frodo lives in the Shire

(1) is a direct quote from Tolkien's *The Lord of the Rings* and is a fictional statement, i.e., a statement that is part of a fiction. Fictional statements are often analysed as pretend-assertions that are merely pretend-true. Fictional names (i.e., names of fictional entities) in fictional statements are likewise analysed as pretend-referring to actual objects. For instance, the name 'Frodo' in (1) involves pretend-reference to a flesh and blood hobbit called Frodo and (1) is only pretend-true (rather than really true). (2) and (3) are metafictional statements. Metafictional statements are statements about fictional entities *as fictional entities*. Metafictional statements are standardly analysed as regular assertions that are really true or false. Fictional names in metafictional statements are often analysed as referring to actually existing abstract objects. For instance, (2) is a true assertion because it is really true (not merely pretend-true) of the fictional character (or abstract object) Frodo that it was invented by Tolkien.

(4) is a parafictional statement, i.e., a report about the content of some fiction, i.e., a report on what is true in or according to some fiction.[1] As recently described by Recanati [11], the semantic analysis of parafictional statements raises some interesting questions. One option is to analyse them as a species of fictional statements (e.g. Evans [4]): 'Frodo' in (4) pretend-refers to a flesh and blood individual and (4) is a continuation of the pretence initiated by Tolkien and hence merely pretend-true. However, this is not entirely intuitive. It seems that parafictional statements are *really* true or false, depending on what the world is like. Given how Tolkien wrote his books, (4) is *actually* true (not just pretend-true) and its negation is *actually* false. Alternatively, we may analyse parafictional statements as a species of metafictional statements

*Many thanks to three anonymous Amsterdam Colloquium reviewers and to the audiences at the Milan MELT seminar, PLM6 and the *Fiction & Narrative across Media* workshop.

[1]This is how theorists usually characterize parafictional discourse. I have argued elsewhere [10] that we need to distinguish between truth *in* and truth *according to* some fiction.

(e.g. Zalta [15] or Inwagen [14]): Parafictional statements such as (4) are a type of assertion and hence can be really true or false. However, this kind of analysis seems to lead to the so-called 'problem of the wrong kind of object' (see Klauk [7]): On this analysis, 'Frodo' in (4) refers to an abstract object, so doesn't it then imply that in *The Lord of the Rings*, some abstract object lives in the Shire? Surely, that cannot be right either. An abstract object is not the right kind of object to live in some region.[2]

The following table from Rouille (p.c.) may shed light on why the analysis of parafictional statements is challenging:

	Genuine assertion	*Pretend assertion*
Reference to a flesh and blood individual	Parafictional	Fictional
Reference to an abstract object	Metafictional	?

Fictional statements (e.g., (1)) are classified as being made from a perspective internal to the fiction (they are pretend assertions and only pretend-true) and fictional names in them standardly involve internal pretend-reference to flesh and blood individuals. Metafictional statements (e.g, (2)) are made from a perspective external to the fiction (they are genuine assertions and really true or false) and fictional names in them involve real reference to abstract objects. Parafictional statements (e.g., (4)) at least *prima facie* seem to mix perspectives that are internal and external to the fiction: They are made from a perspective external to the fiction (they are genuine assertions that can be really true or false) but fictional names in them standardly involve internal pretend-reference to flesh and blood individuals, rather than to abstract objects.

This paper investigates what type of statement could be placed in the blank cell that mirrors the parafictional cell, i.e., statements that are (at least *prima facie*) internal to the fiction but involve metafictional external reference. I suggest that statements in fiction that 'break the fourth wall' (BtFW-statements), i.e., statements through which a fictional character somehow acknowledges their own fictionality, are such a mirror case. After a brief introduction of the phenomenon of breaking the fourth wall (section 1), I will explore two possible analyses, according to which BtFW-statements are either a type of fictional statement or are a type of metafictional statement, and propose a definition of BtFW-statements which is a synthesis of these two analyses (section 2).

1 Breaking the Fourth Wall

The concept of a 'fourth wall' originates in Diderot's [1] writings on theatre. On a traditional stage, if characters are in a room, then it is true in the fiction that they are surrounded by four walls. These walls are solid for the fictional characters on stage. However, one of these walls – the fourth wall – is transparent for the audience. Traditional theatre convention dictates that the theatre audience can 'peer into the fictional world' through this fourth wall whereas the fictional characters are oblivious to this audience. Generalising this concept to other media, the term 'fourth wall' denotes the ontological barrier or separation between the fictional and the actual world. In film, the fourth wall will often be a point of view that can move around in the fictional world and coincides with the location of the cameras that were used during filming. The question of where the fourth wall is to be located in other kinds of media (e.g. novels, pretend play/LARPing or videogames[3]) can become complicated. However, although it may be

[2] See Semeijn and Zalta [12] for a recent defence of the metafictional analysis against this objection.

[3] Some game scholars have argued that the inherent interactivity of videogames entails that there is no fourth wall in the case of video games. See Mosselaer [13] for an overview and the opposite view.

difficult to physically locate it, I assume that for any fiction (across different media) there is in fact an ontological divide (a 'fourth wall') between the fictional world and actual world.

This ontological barrier can seemingly be 'broken'. The common pretheoretical and intuitive understanding of the phenomenon of breaking the fourth wall is that of a fictional character acknowledging that there is such an ontological divide, i.e., that there is an actual world 'out there' containing an audience and that their world is merely fictional. Fourth wall breaks can happen in variety of media (other than just theatre). Consider for instance the movie *On Her Majesty's Secret Service* which is the first (and only) movie in which George Lazenby portrayed James Bond after Bond had been portrayed by Sean Connery. In this movie, Bond is left alone on a beach holding a woman's slippers as she runs away from him. Bond says:

(5) This never happened to the other fellow

and smiles into the camera. What "other fellow" does Bond refer to? He seems to refer to Sean Connery, i.e., the actor that previously played Bond. But how can Bond possibly make reference to an actor that portrays Bond? Or consider the Deadpool comic book *Dead Presidents* which contains a panel in which the main character Deadpool states:

(6) So, I have about six pages to kill ten presidents and their henchmen. I say it's montage time. I suggest the reader crank "Five Minutes Alone" by Pantera. Perfect song for me to kick some dead president butt to.

Again, who or what is Deadpool referring to? There is no montage or reader in his world. At this point the reader may wonder whether fourth wall breaks are special to visual fictions. They are not. Although indeed a fictional character in a novel is not able to 'look the audience in the eye', they can still acknowledge their fictionality and hence break the fourth wall. For instance, consider Dickson Carr's *The Hollow Man* which is a classic example of a Locked Room Mystery novel. At some point, one of the characters, Gideon Fell, gives a lecture on the genre of Locked Room Mystery novels. When asked why this is relevant Gideon Fell answers:

(7) Because we're in a detective story, and we don't fool the reader by pretending we're not.

Before I continue, it is important to clearly distinguish the phenomenon of breaking the fourth wall (as I understand it) from a fiction in which there exists a fiction f^1 *within* the fiction f^2 and the ontological boundary between the real and fictional world is broken *within* fiction f^2. Examples of such fictions can also be found across media. For instance, in the movie *The Purple Rose of Cairo*, a woman is watching a movie when suddenly one of the fictional character walks out of the screen to interact with her. Or consider Pridelli's play *Six Characters in Search of an Author*, in which six fictional characters (abandoned by their creator) visit a theatre group to ask them to portray their story. Lastly, consider the novel *The Eyre Affair*, in which characters can enter the fictional world of literature through the use of a machine called the prose portal. Many online sources (and some theorists, e.g. Satik [2]) mention these and similar fictions as examples of (blatant) fourth wall breaks. However, I suggest that these fictions are rather simply examples of fictions in which travel between different possible worlds is possible (e.g. the *Rick and Morty* series of the movie *Everything everywhere all at once*). These may be interesting examples of impossible fictions in which it is true about one and the same entity that it *is* and *is not* a fictional entity. However, these are not examples of fictions in which a fictional character breaks the ontological barrier between the fictional world and *our* actual world.

2 Semantic analysis of BtFW-statements

Now that we have a clearer understanding of the phenomenon of breaking the fourth wall, let's consider how it relates to fictional, parafictional and metafictional discourse. I suggest that we should understand BtFW-statements as a mirror case of parafictional statements. For instance, in Bond's statement (5) "the other fellow" is taken to abbreviate something like "the other actor that portrayed the fictional character Bond". *Prima facie*, Bond is interpreted as saying something like the following:

(8) The other actor that portrayed the fictional character Bond never portrayed a Bond that women ran away from

In our world, (8) would amount to a (true) statement involving metafictional reference (i.e., reference to Bond as a fictional entity). (5) is thus a pretend assertion (made by a fictional character from a perspective internal to the fiction) that seems to involve (self-referential) metafictional (or external) reference. Or consider Dickson Carr's more explicit (7) where fictional character Gideon Fell makes metafictional reference to himself (and the people listening to his lecture) as being fictional characters by stating that they are "in a detective story".

BtFW-statements thus seem to mix internal/external perspectives in the exact opposite way that parafictional statements do. An understanding of BtFW-statements as a mirror case of parafictional statements suggests two strategies to deal with BtFW-statements that parallel the two main existing analyses of parafictional discourse, i.e., reduction to either fictional or metafictional discourse. I turn to these analyses now.

2.1 Fictional analysis

A fictional approach would analyse BtFW-statements (e.g., (5)) as a species of fictional statements. This seems consistent with our common understanding of BtFW-statements as being made by fictional characters. However, just as Bilbo does not exist in the actual world, there exists no actor that portrayed the fictional character Bond in the worlds of the Bond movies. Bond is *not* a fictional character in those worlds but a flesh and blood spy. In other words, "the other actor that portrayed the fictional character Bond" in (8) is an empty definite description in the Bond-worlds. In the fiction, (5) is thus (not even pretend-true but) a truth-valueless (or false, depending on your theory of presupposition failure) infelicitous statement made by Bond. In Bond's 'mouth', (5) is a incoherent nonsense statement. Similarly, Deadpool's "the reader of the comic book that portrays this fictional world" is an empty definite description in the *Deadpool* worlds and hence his statement (6) is similarly infelicitous.

Such an analysis seems to get part of our interpretative intuitions right. Indeed, there is a strong sense in which Bond said something nonsensical on the beach that made no sense for him to say. However, this analysis does not account for our intuition that what is expressed (i.e., that Connery never portrayed a Bond that women ran away from) is also somehow *really* true. It's not *just* nonsense.

2.2 Metafictional analysis

A metafictional approach to BtFW-statements would deem (5) a type of metafictional statement. In other words, BtFW-statements would be analysed as a special case of what Gendler [5] calls 'pop-out', i.e., an assertion by the author that 'interrupts' the fictional discourse. We are thus dealing with a temporary interruption of the pretense to make an assertion, i.e. a temporary lifting of all four walls.

In the case of (graphic) novels it is clear who takes the role of 'the author'. Similarly, in cases of pretend play or improvisational theatre, it is clear who we can ascribe individual fictional statements (and hence assertions in case of pop-out) to, i.e., the person who came up with and said the statement. Movies and scripted theatre pose an interesting intermediate case. Who is 'the author' in the case of (5)? We are dealing with an actor who (most probably) based his fictional statements on a script written by others. We can ascribe the speech act to the actor Lazenby (who was instructed to make certain fictional statements by the script writers); we can ascribe the speech act to the script writers (who make fictional statements 'through' Lazenby); or ascribe the speech act to a collective of agents including Lazenby, the writers and possibly others. For simplicitly, I am going to ascribe the relevant fictional statements (and the metafictional pop-out assertion) to Lazenby. Thus in the case of (5), Lazenby makes the (true) metafictional assertion that Connery never portrayed a Bond that women ran away from.

Again, the analysis seems to get only part of our interpretative intuitions right. Lazenby can felicitously refer to "the other actor that portrayed Bond" and hence we account for the intuition that what has been said, is really true: (5) is not nonsense in the mouth of Lazenby but a true assertion about Connery. Also, a pop-out analysis does justice to part of the aesthetic experience often associated with breaking the fourth wall, i.e., a feeling of 'interruption'. Fourth wall breaks interrupt the suspension of belief and seem to throw you out of the movie/book/fiction for a moment. However, just like the fictional analysis, the metafictional analysis also doesn't seem to do complete justice to all our interpretative intuitions. Isn't it true in the fiction that Bond says something funny/weird, i.e., that Bond says (5)? Especially in the case of the *Deadpool* comics it is clearly true in the fiction that Deadpool frequently says and does weird things. Other characters react in confusion when Deadpool breaks the fourth wall in their presence.

2.3 Definition of statements that Break the Fourth Wall

Both discussed analyses seem to get only a part of our interpretative intuitions concerning BtFW-statements right. I suggest that a synthesis of these two approaches captures the phenomenon best. I thus propose the following definition of BtFW-statements:

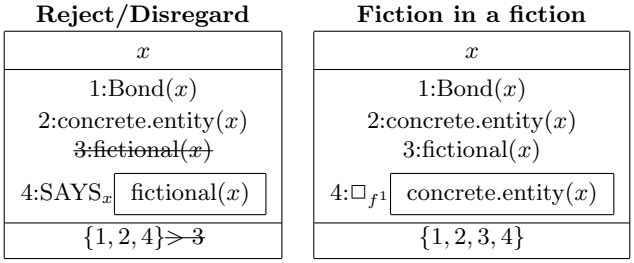
(9) Statement s in fiction f breaks the fourth wall iff s is [1] a fictional statement that makes it true in f that an infelicitous/incoherent and truth-valueless statement is made, and [2] a true and felicitous metafictional statement about f in the actual world simultaneously

For example, a BtFW-statement such as (5) instantiates two speech acts simultaneously: (5) is a fictional statement made by Lazenby that makes it true in the fiction that Bond makes a nonsensical and truth-valueless statement *and* (5) is a true metafictional assertion made by Lazenby. Or consider Dickson Carr's (7) which instantiates a fictional statement (that makes it true in *The Hollow Man* that Gideon Fell makes a incoherent statement that he and the people surrounding him are characters "in a detective story") *and* a true metafictional assertion by Dickson Carr that the relevant characters are in a detective story.

On such an analysis, there is no actual breaking of the ontological barrier between the fictional and actual world (because the rules of fiction don't allow this), but there are two parallel speech acts taking place at the same time. Apart from addressing all discussed intuitions, the proposed synthesis has the added benefit of offering an explanation for the confusing and/or funny aesthetic experience often associated with fourth wall breaks: A feeling of unexpected and awkward uncertainty on how to interpret the fiction. What's funny about BtFW-statements is that they are ambiguous between two interpretations that force a switching between internal and external perspectives. They thus put the audience in a somewhat awkward position of oscillating between two conflicting interpretations.

3 Further steps

I suggested that BtFW-statements instantiate (apart from pop-out) fictional statements that make it true in the fiction that a fictional character says something infelicitous. Put differently, the character says something that leads to an inconsistency. For instance, (8) implies that Bond is a fictional character while it is a well established fictional fact in *On Her Majesty's Secret Service* that Bond is *not* a fictional character but a flesh and blood individual. I would like to explore to what extent existing literature on interpretative strategies for inconsistency in fiction (see e.g. Matravers [9]) are applicable to fourth wall breaks. We can represent the available strategies using Maier and Semeijn's [8] ordered DRS's (Discourse Representation Structures (see Kamp [6]) that include an epistemic entrenchment ordering on the conditions):



One interpretative strategy is to reject the inconsistency by removing one of the problematic conditions. The fictional truth that is epistemically least entrenched (easiest to disregard) seems to be the 'fictional(x)' condition. This move can be accompanied by what Eckardt [3] calls a 'cautious update', i.e., I do not accept that p, but I do accept that a character believes/accepts/states that p (e.g., In *On Her Majesty's Secret Service*, Bond is not a fictional character but it *is* true that Bond states (5)). For fictions in which the fourth wall is broken frequently, it may be tempting to accept the inconsistency instead, i.e. accept that there is some entity that *can* break the ontological barrier between fiction and non-fiction and hence somehow functions on both planes of existence. Since this cannot happen in the actual world this strategy amounts to (temporarily) accommodating a fiction f^1 in the fiction. Another fourth wall is placed 'behind the audience' (e.g., the audience pretend-plays that Deadpool can actually see them). I hypothesize that excessive fourth wall breaking leads to permanent accommodation of a fiction in a fiction interpretation and hence at some point disables further fourth wall breaks.

Second, because this paper focuses on the mirror case of parafictional statements, the provided definition (9) deals only with *assertions* through which characters break the fourth wall. However, just as other kinds of parafictional speech acts are possible (e.g., "Is it true in *The Lord of the Rings*, that Bilbo is Frodo's uncle?"), characters can also break the fourth wall with other speech acts than assertions (e.g., "Would this ever have happened to the other fellow"). A general definition of BtFW-discourse could be formulated along the following lines:

(10) Discourse s in fiction f breaks the fourth wall iff s [1] is fictional discourse that makes it true in f that someone engages in infelicitous discourse, and [2] felicitous metafictional discourse about f in the actual world simultaneously

Even broader, characters can also break the fourth wall without using any language (e.g., in the videogame *Sonic the Hedgehog*, the avatar will look at the player impatiently in case they don't press any buttons for too long). In order to do justice to all varieties of fourth wall breaks in terms of (meta)fictional reference, we would need an account of reference that can for instance analyse a meaningful look into a camera as a kind of pointing or demonstrative reference.

Last, it would be interesting to see how the provided analysis relates to other interesting phenomena such as 'leaning against the fourth wall', 'fourth wall psychs' and 'asides'.

References

[1] Diderot Denis. On dramatic poetry. In B. H. Clark, editor, *European Theories of the Drama*. Crown Publishers, New York, 1947 [1758].

[2] Satik Deniz. The fourth wall against possibilism on truth-in-fiction (manuscript). 2021.

[3] Regine Eckardt. *The Semantics of Free Indirect Discourse*. Brill Publishers, Leiden, 2014.

[4] Gareth Evans. *The Varieties of Reference*, volume 10. Oxford University Press, Oxford, 1982.

[5] Tamar Gendler. Imaginative resistance revisited. In Shaun Nichols, editor, *The Architecture of the Imagination: New Essays on Pretence, Possibility, and Fiction*, pages 149–173. Oxford University Press, Oxford, 2006.

[6] Hans Kamp. A theory of truth and semantic representation. In Jeroen A. G. Groenendijk, Theo M. V. Janssen, and Martin B. J. Stokhof, editors, *Formal Methods in the Study of Language, Part 1*, pages 277–322. Blackwell Publishers Ltd, Oxford, 1981.

[7] Tobias Klauk. Zalta on encoding fictional properties. *Journal of Literary Theory*, 8(2):234–256, 2014.

[8] Emar Maier and Merel Semeijn. Extracting fictional truth from unreliable sources. In Emar Maier and Andreas Stokke, editors, *The Language of Fiction*, pages 186–220. Oxford University Press, Oxford, 2021.

[9] Derek Matravers. *Fiction and Narrative*. Oxford University Press, Oxford, 2014.

[10] Semeijn Merel. The 'in' and 'according to' operators. In *Proceedings of the ESSLLI & WeSSLLI Student Session 2020*. 2020.

[11] François Recanati. Fictional, metafictional, parafictional. *Proceedings of the Aristotelian Society*, 118(1):25–54, 2018.

[12] Merel Semeijn and Edward N. Zalta. Revisiting the 'wrong kind of object' problem. *Organon F*, 28(1):168–197, 2021.

[13] Nele Van de Mosselaer. Breaking the fourth wall in videogames. In Enrico Terrone and V. Tripodi, editors, *Being and Value in Technology*. Palgrave Macmillan, Cham, 2022.

[14] Peter van Inwagen. Quantification and fictional discourse. In Anthony Everett and Thomas Hofweber, editors, *Empty Names, Fiction and the Puzzles of Non-Existence*, pages 235–47. CSLI, Stanford, 2000.

[15] Edward N. Zalta. *Abstract Objects: An Introduction to Axiomatic Metaphysics*. Springer, Berlin, 1983.