Fiction and Emotion: The Puzzle of Divergent Norms

1. Introduction

Emotional responses toward fiction are puzzling. Readers of Colson Whitehead's *The Underground Railroad* (2016) pity the slave Cora when she is savagely whipped and detest the overseer Connelly for his cruelty, even though they know that neither Cora nor Connelly exists. If I fear the shadowy form in the corner, then discover that it is nothing but an old coat, I will stop being afraid; yet I remain terrified of the extra-terrestrial in Ridley Scott's *Alien* (1979) despite my certainty that there is no such creature. Why do we respond emotionally to characters we know do not exist and events we know have not occurred?

Philosophical discussion of this issue has focused on two questions: First, what is the nature of the responses described above? Should they be classified as the same kinds of states as emotions in other contexts? Second, how can emotions in the fiction context be rational, given that we know they have no objects? In this paper I focus on the second, concerning rationality.

The rationality puzzle can be traced back to Colin Radford's famous 1975 argument that emotions toward fiction are 'incoherent' and 'inconsistent' and therefore irrational.[1] It has become standard to formulate Radford's challenge in terms of the so-called Paradox of Fiction, and to take various ways of addressing the Paradox to provide answers to the rationality question. In this paper I argue that neither Radford's argument, nor any plausible argument for irrationalism, involves a paradox (§§2-3). The rationality puzzle instead concerns the *normative assessment* of emotional responses; but even philosophers who have

[1] Colin Radford, 'How Can We Be Moved by the Fate of Anna Karenina?' *Proceedings of the Aristotelian Society* 49 Supplement (1975), 67–80.

focused on the normative dimension have not appreciated the full force of Radford's original argument, and therefore have failed to address it adequately (§4). I propose a better interpretation (§5) and a new solution (§§6-7). One advantage of my proposal is that it helps to explain a mechanism by which we may learn emotionally from fiction (§8).

My purpose in making this argument is twofold: First, the interpretation of Radford's argument for irrationalism in terms of the Paradox of Fiction has become entrenched over the last several decades, continuing to influence even the most recent contributions to the debate. It is therefore important to set the record straight. Second, the persistent tendency to misread Radford's argument has prevented an adequate solution to the rationality puzzle. Understanding his argument directs us toward a more nuanced account of the rationality of emotion than typically assumed.

2. The Paradox of Fiction

Radford's 1975 paper, 'How Can We Be Moved by the Fate of Anna Karenina?', generated a sizeable debate of its own in the twenty-five years following its publication.[2] Over the last few decades, however, both his argument and the rationality puzzle have typically been presented in terms of the Paradox of Fiction (PoF).

The PoF is a set of three propositions, which are independently plausible but incompatible in combination. Here is a standard version:

(P1) We experience emotions toward fictional characters, situations and events.

[2] Radford's replies include: 'Philosophers and Their Monstrous Thoughts', *BJA* 22 (1982), 261–263; 'Stuffed Tigers: A Reply to H. O. Mounce', *Philosophy* 57 (1982), 529–532; 'Replies to Three Critics', *Philosophy* 64 (1989), 93–97; 'Fiction, Pity, Fear, and Jealousy', *Journal of Aesthetics and Art Criticism* 53 (1995), 71–75; 'Neuroscience and Anna: a Reply to Glenn Hartz', *Philosophy* 75 (2000), 437–40.

(P2) We do not believe (we disbelieve) in the existence of fictional characters, situations and events.

(P3) To have an emotion towards something, we must believe that it exists.

Radford is widely cited as offering the first formulation of the PoF.[3] The frequent assumption is that his irrationalist conclusion—that emotions toward fictions are irrational—is grounded in a commitment to the three propositions. After all, if we are committed to contradictory propositions, then there is a clear sense in which we are 'incoherent' and 'inconsistent'. Call this the *Paradoxical Interpretation of the Rationality Puzzle* (PIRP). Those who assume PIRP and defend the rationality of emotions toward fiction—the majority of theorists—take their task to be offering a solution or dissolution of the PoF. I will argue that this entire approach is misguided.

The earliest explicit statement of PIRP may be Robert Yanal's in 1994. Yanal presents the puzzle by attributing the following three claims to Radford:

(1) We feel emotions towards the characters and situations of some works of fiction.

[3] A selection: Robert Yanal, 'The Paradox of Emotion and Fiction', *Pacific Philosophical Quarterly* 75 (1994), 54–75; Steven Schneider, 'The Paradox of Fiction', *Internet Encyclopedia of Philosophy* (2002) <https://www.iep.utm.edu/fict-par/> accessed 25 August 2019; Oliver Conolly, 'Pleasure and Pain in Literature', *Philosophy and Literature* 29 (2005), 305–320; Derek Matravers, *Fiction and Narrative* (Oxford: OUP, 2014); Katherine Tullman and Wesley Buckwalter, 'Does the Paradox of Fiction Exist?' *Erkenntnis* 79 (2014), 779–796; Florian Cova and Fabrice Teroni, 'Is the Paradox of Fiction Soluble in Psychology?' *Philosophical Psychology* 29 (2016), 930–942; Marco Sperduti et al., 'The Paradox of Fiction: Emotional Response toward Fiction and the Modulatory Role of Self-Relevance', *Acta Psychologica* 165 (2016), 53–59; Fabrice Teroni, 'Emotion, Fiction and Rationality', *BJA* 59 (2019), 113–128.

(2) We feel these emotions even though we believe that such characters and situations are fictional and not real.

…

(3) We feel emotions towards characters or situations only when we believe them to be real and not fictional.[4]

Yanal asserts: 'It is the conjunction of (3) with (1) and (2) that leads Radford to' his irrationalist conclusion.[5]

Steven Schneider's *Internet Encyclopedia of Philosophy* entry, 'The Paradox of Fiction', similarly introduces the PoF as originating in Radford's argument. He claims that Radford draws his conclusion on the grounds that '*existence beliefs* concerning the objects of our emotions … are necessary for us to be moved by them' (P3), but 'such beliefs are lacking when we knowingly partake of works of fiction' (P2); nonetheless, the 'works *do* in fact move us at times' (P1).[6] Schneider goes on to describe various solutions to the PoF. Each denies one of the three claims, a strategy makes sense on the assumption of PIRP that the irrationality is due to our commitment to incompatible propositions.

The solutions outlined by Schneider are familiar. Some argue that emotions in response to fiction are not genuine emotions, or at least not of the same kind as ordinary emotions, thereby denying (P1); this is the well-known proposal of Kendall Walton.[7] Others

---

[4] Yanal, 'The Paradox of Emotion and Fiction', 54–55. Yanal no longer accepts PIRP in *Paradoxes of Emotion and Fiction* (Philadelphia: Penn State University Press, 1999).

[5] Ibid., 55.

[6] Schneider, 'The Paradox of Fiction'.

[7] Kendall Walton, 'Fearing Fictions', *Journal of Philosophy* 75 (1978), 5–27; *Mimesis as Make-Believe* (Cambridge, MA: Harvard University Press, 1990). See also Jerrold Levinson, 'Emotion in Response to Art: A

maintain that we *do* temporarily believe in the existence of fictional characters, thereby denying (P2).[8] The most popular approach is to deny (P3), maintaining that belief in existence is not required for emotion.[9] A more recent strategy is to *dissolve* the paradox by showing that the three propositions are actually compatible, once we understand the meanings of the terms involved.[10] The idea in common is that if we are not committed to incompatible propositions, we cannot be irrational.

Since the source of PIRP is widely assumed to be Radford, it is instructive to consider what he says. The attribution rests on his apparent defence of all three claims. For example, in reply to the proposal that we are not *really* moved by fictions—the denial of (P1)—Radford insists that 'we cannot say that we do not feel for fictional characters …. We shed real tears for Mercutio'.[11] Later he considers the possibility that 'there are two sorts of being moved', one that applies to fiction and one that applies in ordinary cases.[12] He concedes that there are differences between emotions in these contexts, for instance in intensity or motivational capacity. Nonetheless, he argues, the emotions are similar in the only respect that matters: 'the essential similarity seems to be that we are saddened'.[13] He is committed to (P1).

Survey of the Terrain', in Mette Hjort and Sue Laver (eds), *Emotion and the Arts* (New York: OUP, 1997), 20–34.

[8] E.g., Glenn Hartz, 'How We Can Be Moved by Anna Karenina, Green Slime, and a Red Pony', *Philosophy* 74 (1999), 557–578; David B. Suits, 'Really Believing in Fiction', *Pacific Philosophical Quarterly* (87), 369–386.

[9] The advocates of this position are too numerous to list; they include everyone cited in this paper who doesn't deny (P1) or (P2).

[10] Proposed by Tullman and Buckwalter, 'Does the Paradox of Fiction Exist?'.

[11] Radford, 'How Can We Be Moved by the Fate of Anna Karenina?', 70.

[12] Ibid., 75.

[13] Ibid., 77.

Radford is equally committed to (P2). He considers two ways of denying this claim: that we temporarily 'forget' that we are engaged with fiction; or that we actively 'suspend our disbelief' in the reality of the fictional events.[14] Against the first he points out that if it were right, we would 'shout and try to get on the stage when, watching *Romeo and Juliet*, we see that Tybalt is going to kill Mercutio'.[15] Against the second he replies that while we do not typically focus on the fact that we are watching a play, 'we are never unaware that we are watching a play, and one about fictional characters'.[16]

Radford *seems* to defend (P3) as well. In an example that has been cited repeatedly in the literature on emotions and fiction, Radford asks the reader to suppose that a man tells 'you a harrowing story about his sister and you are harrowed'. Then the man confesses that he has no sister and has made up the entire story. Radford suggests that you would no longer be harrowed: 'the possibility of your being harrowed … seems to require that you believe someone suffered'. Considering this and other examples, he writes: 'It would seem then that I can only be moved by someone's plight if I believe that something terrible has happened to him. If I do not believe that he has not and is not suffering or whatever, I cannot grieve or be moved to tears.'[17]

Radford's famous conclusion is 'that our being moved in certain ways by works of art, though very "natural" to us and in that way only too intelligible, involves us in inconsistency and so incoherence'.[18] If PIRP is correct, the inconsistency and incoherence are due to our commitment to (P1), (P2) and (P3). But PIRP is unfounded, not only as an

---

[14] Radford also argues against the second view in 'Neuroscience and Anna'.

[15] Radford, 'How Can We Be Moved by the Fate of Anna Karenina?', 71.

[16] Ibid., 72.

[17] Ibid., 68.

[18] Ibid., 78.

interpretation of Radford, but as a plausible account of what is at issue in the irrationalist challenge.

3. Against the Paradoxical Interpretation

There are two difficulties with PIRP. The first is that it renders any argument for the irrationality of emotions toward fiction self-refuting. The second is that it mislocates the irrationality itself. I take each in turn.

Advocates of PIRP trace the PoF to Radford because, they say, he argues for (P1), (P2) and (P3). They further hold that the irrationality identified by Radford is due to the inconsistency among (P1), (P2) and (P3). It follows that Radford is arguing for the truth of propositions that he himself maintains are incompatible. Yanal makes this clear, suggesting that anyone who advocates the three propositions must hold that 'it is not only *we* who would be involved in 'incoherence' but the *universe.* (1)-(2)-(3), if all true, entail a contradiction.'[19] As Derek Matravers, who attributes the same contradiction to Radford, objects: 'It does not help to say that we are inconsistent; that is akin to saying that, as we cannot be in two places at once, we are being inconsistent when we do.'[20] In other words, no one can coherently argue for the truth of the three propositions that constitute the PoF, as Radford is typically taken to do.

In any case, there is little reason to take the incompatibility of (P1), (P2) and (P3) to be the source of irrationality. This is the second difficulty with PIRP. Radford's conclusion is that it is 'our being moved'—that is, our *responding emotionally* to fiction—that is irrational. The PoF, by contrast, represents an inconsistency in *beliefs about* those responses. The irrationality is attributed to the theorist of emotion, rather than to the experiencer.

---

[19] Yanal, 'The Paradox of Emotion and Fiction', 55.

[20] Matravers, *Fiction and Narrative*, 104.

One might reply that anyone who argues from the PoF to the irrationality of emotional responses to fiction is presupposing *cognitivism* about the emotions, the position that emotions are at least partly constituted by judgements or beliefs.[21] For cognitivists, to pity someone involves the belief that they have suffered. If we also believe that no one has really suffered, we have contradictory beliefs. It is frequently assumed that Radford is a cognitivist about emotions.[22] For example, Katherine Tullman and Wesley Buckwalter, in presenting the PoF, attribute the following interpretation of (P3) to Radford: 'having a genuine emotional response implies a belief in the existence of a fictional thing'.[23] That is, the emotions are irrational because they entail inconsistent beliefs: On the one hand, we do not believe that fictional characters exist; on the other, our emotions commit us to the belief that they do.

Unfortunately for this interpretation, Radford is not a cognitivist. First, Radford's insistence that we have genuine emotions in response to fictional characters without believing in their existence is incompatible with cognitivism.[24] Second, Radford explicitly rejects a belief requirement on emotion. He argues that phobias constitute a counterexample; a phobic about spiders is afraid even while she knows that they cannot harm her.[25] He goes on to say

---

[21] On cognitivism see Andrea Scarantino and Ronald de Sousa, 'Emotion', in Edward Zalta (ed.), *Stanford Encyclopedia of Philosophy* (2018) <https://plato.stanford.edu/archives/win2018/entries/emotion/> accessed 3 September 2019; Bennett W. Helm, 'Cognitivist Theories of Emotion', in Andrea Scarantino (ed.), *Handbook of Emotion Theory* (New York: Routledge, forthcoming).

[22] By, e.g. Derek Matravers, *Art and Emotion* (Oxford: OUP, 1998), 68; Hartz, 'How We Can Be Moved by Anna Karenina, Green Slime, and a Red Pony', 559; Elisa Galgut, 'Poetic Faith and Prosaic Concerns: A Defense of "Suspension of Disbelief"', *South African Journal of Philosophy* 21 (2002), 190; Catherine Wilson, 'Grief and the Poet', *BJA* 53 (2013), 78.

[23] Tullman and Buckwalter, 'Does the Paradox of Fiction Exist?', 780.

[24] Richard Joyce, 'Rational Fear of Monsters', *BJA* 40 (2000); Yanal, *Paradoxes of Emotion and Fiction*.

[25] Radford, 'Fiction, Pity, Fear and Jealousy', 72.

that the emotion can occur despite the belief's being absent. If the belief is absent, there is no pair of contradictory beliefs to be the source of irrationality.

In other words, one can maintain that our emotional responses to fiction are irrational while rejecting cognitivism. Nor is this combination of positions exclusive to Radford. Jenefer Robinson defends an account of emotions as non-cognitive 'affective appraisals', which turn in part on the ways in which our interests and values are at stake in a situation. Though she rejects cognitivism, she agrees that Radford proves his point: 'strictly speaking, Radford is right. It *is* irrational to … to feel one's own interests at stake when Little Grey Rabbit is kidnapped by the weasels … [or] to have wants and goals with respect to Anna Karenina and her ilk'.[26] (She will go on to claim that despite this irrationality 'from a strictly cognitive point of view', we are nonetheless 'emotionally rational'; see §4 below.)[27] Since Robinson rejects the claim that emotions involve beliefs or judgements, the cognitive irrationality she ascribes cannot be due to any commitment to incompatible propositions such as (P1), (P2) and (P3). Rather than discussing a logical contradiction, Robinson, like Radford, is offering a *normative assessment* of our emotional responses.

Before considering the normative issue, it is worth pausing to ask why PIRP has been so persistent. I suspect that its popularity is due to conflating Radford's argument for irrationalism with Walton's argument that emotions in response to fiction are of a different kind from emotions in other contexts. It is plausible to construe Walton as addressing the PoF by denying (P1); but this is an argument concerning the *nature* or *possibility* of emotion rather than its rationality.[28] Still, the conflation is understandable: Both Radford and Walton

---

[26] Jenefer Robinson, *Deeper than Reason* (Oxford: OUP, 2005), 146-147.

[27] Ibid., 147.

[28] This point is made by Paisley Livingston and Alfred R. Mele, 'Evaluating Emotional Responses to Fiction', in Mette Hjort and Sue Laver (eds), *Emotion and the Arts* (New York: OUP, 1997), 158.

highlight a puzzle about emotional responses without belief, and each draws a counterintuitive conclusion. Once we have the PoF in mind, it is easy to read Radford as espousing versions of each of the three propositions and to understand Walton as replying to Radford. There is no evidence, however, that Walton was addressing Radford; he never mentions Radford in 'Fearing Fictions' or *Mimesis as Make-Believe*. Radford does occasionally mention Walton, but only in passing.[29] So however understandable PIRP may be, there is just no reason to suppose it correct.

4. Normative Approaches

I turn now to philosophers who recognize that the rationality puzzle concerns the normative assessment of experienced emotions, rather than a logical contradiction. An early example is Richard Joyce, who points out, as against PIRP, that Radford rejects (P3), the claim that emotions require existence beliefs. According to Joyce, Radford still

> maintains that there is *some* intimate connection between belief and emotion. For him, the dependence is not the existential one stated in (P3), but a normative one: we do not *rationally* feel fear unless we believe ourselves (or someone actual) to be in danger.[30]

A number of other philosophers agree with Joyce that the argument for irrationalism hinges, not on (P3), but on a normative claim that I will formulate as follows:

---

[29] E.g., in 'Philosophers and their Monstrous Thoughts'.

[30] Joyce, 'Rational Fear of Monsters', 210 (numbering changed, emphasis added).

(CR) To have a *rational* emotion towards something, we must believe (not disbelieve) that it exists.[31]

Both Radford and his opponents accept that (i) we experience emotions toward fictional characters, situations and events, and (ii) we do not believe that they exist. These premises correspond to P1 and P2 in the PoF, but the combination of (i), (ii) and (CR) does not constitute a logical paradox in the sense of a set of incompatible propositions.[32] Accepting (CR), as Radford does, entails that the emotions are irrational. Radford's opponents reject both (CR) and his conclusion.

I have articulated (CR) as a premise, alongside (i) and (ii), in a valid deductive argument. This is not, however, the way that several of Radford's opponents present it. Illustrating yet again the tenacity with which PIRP influences the debate, these philosophers take (CR) to be one of three incompatible propositions that constitute a *normative* version of the PoF.[33] Setting aside differences in their formulations, they attribute a concern with the following paradox to Radford:

---

[31] E.g., Yanal, *Paradoxes of Emotion and Fiction*; Tamar Szabó Gendler and Karson Kovakovich, 'Genuine Rational Fictional Emotions', in Matthew Kieran (ed.), *Contemporary Debates in Aesthetics and the Philosophy of Art* (Oxford: Blackwell, 2006), 241-253; Berys Gaut, *Art, Emotion and Ethics* (Oxford: OUP, 2007); Cova and Teroni, 'Is the Paradox of Fiction Soluble in Psychology?'; Shen-yi Liao and Tamar Gendler, 'Puzzles and Paradoxes of Imagination and the Arts', in Zalta (ed.), *Stanford Encyclopedia of Philosophy* (2009), accessed 2 September 2019; Teroni, 'Emotion, Fiction and Rationality'.

[32] Conolly (ibid.) and Liao and Gendler (ibid.) recognize that this is Radford's argument but still call it a 'paradox'.

[33] Joyce, 'Rational Fear of Monsters'; Gendler and Kovakovich, Genuine Rational Fictional Emotions'; Cova and Teroni, 'Is the Paradox of Fiction Soluble in Psychology?; Liao and Gendler, 'Puzzles and Paradoxes of Imagination and the Arts'; Teroni, 'Emotion, Fiction and Reality'.

(P1*) We experience *rational* emotions toward fictional characters, situations and events.

(P2) We do not believe (we disbelieve) in the existence of fictional characters, situations and events.

(CR) To have a *rational* emotion towards something, we must believe that it exists.

By contrast with the traditional PoF, this version of the paradox has the virtue of distinguishing between the theorist and the experiencer. A theorist who believed all three propositions would be irrational, but the formulation does not indicate that emotions toward fiction are irrational on this ground.

For precisely this reason, however, formulating the rationality puzzle in terms of this new paradox contributes nothing to the debate. Radford denies (P1*) because he accepts (CR); his opponents accept (P1*) and reject (CR). The philosophically interesting question is whether to agree with (CR), not how to solve or dissolve a theoretical paradox.

One way to refute (CR) is to offer criteria of rationality for emotions in other contexts, and then to argue that emotions in response to fiction meet those criteria. For instance, Yanal argues that emotions toward fictional characters and situations are *behaviourally* rational because they do not thwart the achievement of our ends.[34] Others maintain that such emotions are *instrumentally* rational. For Joyce, works of fiction prompt emotions in the valuable process of giving us 'life experience on the cheap'.[35] Tamar Gendler and Karson Kovakovich point to empirical results that appear to show that affective responses to merely imagined scenarios are essential to rational decision-making, and speculate that emotions in response to

[34] Yanal, *Paradoxes of Emotion and Fiction*, 22–30.

[35] Joyce, 'Rational Fear of Monsters', 223.

fiction similarly 'contribute to our capacity for rational action through the role they play in educating our sensibilities'.[36]

However, a strategy that invokes behavioural or instrumental rationality cannot meet Radford's challenge.[37] Radford himself maintains that the responses are instrumentally valuable in offering pleasure.[38] And Robinson argues that responding to fiction affectively is *emotionally* rational, enhancing social adaptation by getting us to identify and sympathize with other people. Nonetheless, as we have seen, she agrees with Radford that they are irrational from a 'cognitive point of view'. The relevant contrast is between *practical* or *strategic* rationality on the one hand, and *cognitive* or *epistemic* rationality on the other.[39] What exactly cognitive irrationality amounts to once cognitivism is rejected is a question I consider in the next section.

5. The Real Challenge

No matter how we interpret cognitive rationality, there is a serious obstacle to arguing that (CR) is false: in many contexts it is obviously true. If I believe that there are no vampires, fear of vampires is irrational. If I know that Santa Claus does not exist, pinning my hopes on his largesse is equally so. This is also the lesson of the story about the harrowing sister.

---

[36] Gendler and Kovakovich, 'Genuine Rational Fictional Emotions', 252.

[37] See Derek Matravers, 'The Challenge of Irrationalism and How Not to Meet It', in Matthew Kieran (ed.), *Contemporary Debates in Aesthetics and the Philosophy of Art* (Oxford: Blackwell, 2005), 254–264; Kim, 'The Rationality of Emotion toward Fiction'.

[38] Radford, 'Replies to Three Critics', 97.

[39] See Scarantino and de Sousa, 'Emotion'.

Radford's point is that *in ordinary circumstances*, an emotion directed at someone you are firmly convinced does not exist is problematic.[40]

The problem in these cases is that the person knowingly violates a norm widely agreed to govern emotions: the norm of *fittingness* or *correctness*.[41] As Justin D'Arms and Daniel Jacobson put it, an emotion is fitting when it 'accurately presents its object as having certain evaluative features'.[42] The features are the evaluative properties attributed in typical instances of an emotion type, such as *danger* for fear or *undeserved suffering* for pity.[43] To fear something is (at least in part) to evaluate it as dangerous; to pity someone is (at least in part) to evaluate them as suffering undeservedly. Suppose that I am angry at my friend Andrew for stealing a precious heirloom. If Andrew is innocent, my anger is incorrect. Similarly, if I respond with fear to a coiled shape in the shadows, fear will be unfitting if its target is no danger (e.g., a rope).

Now, misdirected emotions are not necessarily irrational. If I have every reason to think that Andrew is guilty of theft, anger makes sense. If, however, I am firmly convinced of Andrew's innocence, I *knowingly* violate the norm. Such a knowing violation is cognitively irrational. The incoherence is between the evaluative features attributed by my anger on the one hand, and my beliefs about Andrew on the other: more simply, between my emotion and

---

[40] Cova and Teroni, 'Is the Paradox of Fiction Soluble in Psychology?', 934.

[41] Justin D'Arms and Daniel Jacobson, 'The Moralistic Fallacy: On the "Appropriateness" of Emotions', *Philosophy and Phenomenological Research* 61 (2000), 65–90; see also Julien Deonna and Fabrice Teroni, *The Emotions: A Philosophical Introduction*, 1st edn (London: Routledge, 2012).

[42] D'Arms and Jacobson, 'The Moralistic Fallacy', 65.

[43] Such properties are the *formal objects* of emotions, in the terminology introduced in Anthony Kenny, *Action, Emotion and Will* (Routledge & K. Paul, 1963). For a more complex account of formal objects, see Kris Goffin, *Emotional and Affective Representation: Reliability, Complexity and Aesthetics* (Ghent, 2018).

my beliefs.[44] This is the kind of incoherence at issue in so-called *recalcitrant* emotions, such as fear of flying when one believes that flying is safe.[45] Those who reject cognitivism about emotion, including D'Arms and Jacobson, differ on how exactly to characterize the conflict between emotions and beliefs at the heart of recalcitrance. However, they agree that one need not (and arguably should not) be a cognitivist to recognize it.[46]

The examples Radford takes to be analogous to fiction are even more egregious violations of correctness than standard instances of recalcitrance. In these examples, our emotions do not merely conflict with our beliefs about their object; we do not believe that there *is* an object. If I know that there is no sister, continuing to pity 'the sister' is surely problematic. In such cases, the irrationality is due to disbelief in the existence of the objects of emotion, as (CR) claims.

Radford assumes that if these emotions are irrational, the same must hold for pity of Anna Karenina or grief for Mercutio. This conclusion turns on (i), that our responses toward fictional characters are genuine emotions that do not differ in kind from other emotions. Nearly everyone agrees with (i) (I return to this assumption in §7). But if the emotions are of the same kinds, presumably the same normative constraints should apply.[47] Should we therefore conclude that the emotions are cognitively irrational? After all, human beings are irrational in myriad ways. Why balk at recognizing one more?

I resist this conclusion. Emotional responses to fictional characters and events are not

---

[44] Scarantino and de Sousa in 'Emotion' call this 'incoherence' and associate it with recalcitrance.

[45] The term is introduced in Justin D'arms and Daniel Jacobson, 'The Significance of Recalcitrant Emotion', *Royal Institute of Philosophy Supplement* 52 (2003), 127–145.

[46] See Alex Grzankowski, 'Navigating Recalcitrant Emotions' (n.d.), for an overview of this debate.

[47] Gaut, *Art, Emotion and Ethics*, 222. Gendler and Kovakovich suggest that this might be Radford's complaint ('Genuine Rational Fictional Emotions', 252). Their reply is to give up (CR) entirely.

plausibly construed as irrational *simply* in virtue of being responses to fictional characters and events. In other instances of irrational emotion, the subject is apt to acknowledge something wrong: 'I know it doesn't make sense, but I'm still afraid of Fido.' By contrast, our pity of Anna or grief for Mercutio still appear unproblematic upon reflection. As Eva Schaper puts it, responding emotionally to fiction 'is not … a deplorable state of affairs calling for elimination or cure'.[48]

These observations motivate answering Radford's argument, but they do not answer it themselves. If some emotions count as irrational in the absence of beliefs and others do not, there must be a reason. What is the normative difference? This question is the core of Radford's challenge.

6. A Normative Difference

The challenge cannot be answered simply by denying (CR); (CR) is clearly true in many contexts. Instead, one must explain why (CR) does not apply across the board. That is, one must specify a systematic difference between (for instance) pitying Anna Karenina on the one hand and pitying the sister in the harrowing story on the other. In the latter case the emotion is irrational because it involves a knowing violation of the correctness norm; we know that no one has suffered. If the same does not hold of our pity for Anna, why is that?

Perhaps it is the fact that we are responding to *fiction* which matters. That is, when we are engaged with fiction, our emotions are no longer sensitive to what is *true* but only to what is *true-in-the-story*, or as I prefer, *storified*. Confronted with fiction, the norm of correctness is suspended, replaced with what we might call *story-correctness*. What matters to the story-

[48] Eva Schaper, 'Fiction and the Suspension of Disbelief', *BJA* 18 (1978), 33.

correctness of pity is just that *in Tolstoy's novel*, Anna suffers.[49] If this is right, then (CR) does not apply when we are engaged with fiction.

However, the correctness norm is not *always* suspended in response to fiction. Suppose that from Dickens's *Oliver Twist* I infer that real orphans in Victorian England suffered in miserable conditions, or from NoViolet Bulawayo's *We Need New Names* (2013) that real immigrants face many obstacles. If I therefore pity real Victorian orphans or sympathize with real immigrants, these emotions are subject to the standard norms that govern emotions in other contexts. Similarly, if I believe Shakespeare's portrayal of Richard III and come to despise Richard in reality, this emotion will be fitting only if Shakespeare portrays Richard accurately.

A different proposal is that the ordinary normative constraints are suspended when the emotions arise in the context of *imagining* rather than belief.[50] Because our pity of Anna is generated by imagining her plight, it cannot conflict with our belief that no one has suffered. Philosophers often assume that emotions involving imagining do not generate recalcitrance. For instance, Michael Brady claims that 'emotions must involve more than mere evaluative construals [e.g. imaginings] if they are to come into rational conflict with evaluative beliefs'.[51] If this is right, then (CR) does not apply when we are engaged in imagining.

However, the fact that an emotion originates in the context of imagining does not, *by itself*, suspend ordinary normative constraints. Imagining Barack Obama gleefully torturing kittens does not make it correct to despise him, just as imagining one's partner being unfaithful offers no justification for anger. If any emotion counts as irrational, it is

---

[49] See Paisley Livingston and Alfred L. Mele, 'Evaluating Emotional Responses to Fiction', in Mette Hjort and Sue Laver (eds), *Emotion and the Arts* (New York: OUP, 1997), 157–176.

[50] Gaut, *Art, Emotion and Ethics*, 216–225.

[51] Michael S. Brady, 'The Irrationality of Recalcitrant Emotions', *Philosophical Studies* 145 (2009), 420.

condemning someone while knowing that their actions were nothing but figments of one's imagination. The point applies equally to the view that normative constraints are different when they are based on beliefs about what is storified in a fiction.[52] Believing that Obama tortured kittens *in a fiction* does not justify despising him *in reality*.

The objections to the proposals so far turn on emotions directed at real individuals, whether prompted by fictions or in the context of imaginings. Such emotions are typically at issue in recalcitrance as well. A natural third proposal is therefore that the ordinary correctness norm is suspended when the emotions are directed at fictional characters, individuals who exist only within an imaginative project or fictional world. Berys Gaut suggests something along these lines in his condition for the cognitive rationality of fear: 'The rationality of fear of objects believed to exist requires one to believe that they are dangerous; and the rationality of fear of objects merely imagined to exist requires one (correctly) to imagine that they are dangerous.'[53]

The difficulty with this proposal is familiar. Suppose that I read a gripping vampire novel and experience intense fear. This fear is story-correct, since the vampires are portrayed as dangerous. Yet I cannot shake this fear after I close the book; though I know that there are no vampires, I am terrified of them. By this I do not mean merely that I am in a heightened state that prompts me to start at sudden noises and the like. I mean rather that I am afraid *of vampires*—in my ordinary life, outside the imaginative project—despite my firm conviction that they do not exist. It is difficult to see this emotion as anything other than irrational. (CR) appears true even when we are dealing with invented characters.

---

[52] See Cova and Teroni, 'Is the Paradox of Fiction Soluble in Psychology?'; Teroni, 'Emotion, Fiction and Rationality'.

[53] Gaut, *Art, Emotion and Ethics*, 220.

The objections I have raised share a common theme: that emotions, even if they are prompted by fiction or imagining, may be irrational if they are *carried over* to the real-world context (henceforth, RWC).[54] If I detest Obama in the RWC as a result of my feelings about his imaginary kitten-torture; sympathize with real Victorian orphans as a result of feeling for Oliver Twist; or fear non-existent vampires who might attack me in my home, the emotions have carried over. When emotions are directed at the RWC, they are subject to ordinary norms. This suggests a new proposal: that the ordinary norm of correctness is suspended when our emotions are *compartmentalized*, that is, when they are prevented from being carried over to the RWC. In such cases (CR) does not apply. I develop this proposal in the next section.

7. Compartmentalized Emotions

Compartmentalized emotions remain confined within a non-committal project, whether an episode of imagining, engagement with a story, game of pretence, counterfactual supposition, or what have you. The idea is familiar from discussions of *quarantine* as a feature of imagining as opposed to belief.[55] Imaginings are said to be quarantined insofar as their effects are circumscribed within a limited domain, so that they do not interact freely with other mental states or motivate action. Thus imagining myself to be deathly ill does not result in a hospital visit, nor do I call the police after pretending someone has been murdered.

---

[54] Stacie Friend, 'Getting Carried Away', *Midwest Studies in Philosophy* 34 (2010), 77–105.

[55] Tamar Szabó Gendler, 'On the Relation Between Pretense and Belief', in Matthew Kieran and Dominic Lopes (eds), *Imagination, Philosophy, and the Arts* (London: Routledge, 2003), 125–141.The term *compartmentalization* originates in George Potts and Sharyl Peterson, 'Incorporation versus Compartmentalization in Memory for Discourse', *Journal of Memory and Language* 24 (1985): 107–18. See Matravers, *Fiction and Narrative*, 79–80, for discussion.

There are different explanations of this phenomenon. Within philosophy, simulationists say that when we imagine we take our ordinary inferential systems 'offline',[56] while others maintain that imaginings involve a separate cognitive system, a *possible worlds box*[57] or *imagination box*.[58] There are still different accounts in psychology and linguistics.[59]

The choice of cognitive architecture does not matter here. What is important is that many emotions prompted by fiction exemplify the same characteristics. No matter how powerful my fear in reading the vampire novel, for example, I do not fashion wooden stakes or string garlic around the house. Such emotions are plausibly compartmentalized.

Uncompartmentalized emotions, by contrast, are not restricted in these ways. They can be carried over to the RWC, interact with ordinary beliefs and desires, motivate action, and so on. Suppose that reading about the sufferings of the fictional Juana in Reyna Grande's 2007 novel *Across a Hundred Mountains* prompts me to sympathize with the plight of Mexican migrants to the U.S. The emotion might reconfigure my understanding of debates over immigration or motivate me to help. Even if uncompartmentalized emotions do not cause such changes, they are *available* for interaction with other mental states and for motivating action. Consequently, they are subject to the correctness norm.

Now, this contrast might suggest a more straightforward reply to Radford: Rather than say that (CR) is limited in scope, we simply reject (i), holding that (CR) fails to apply to

---

[56] E.g., Gregory Currie and Ian Ravenscroft, *Recreative Minds* (Oxford: OUP, 2002).

[57] Shaun Nichols and Stephen Stich, *Mindreading* (Oxford: OUP, 2003).

[58] Jonathan M. Weinberg and Aaron Meskin, 'Puzzling over the Imagination: Philosophical Problems, Architectural Solutions', in Shaun Nichols (ed.), *The Architecture of the Imagination* (Oxford: OUP, 2006), 175–202.

[59] See, e.g., Anthony Sanford and Catherine Emmott, *Mind, Brain and Narrative* (Cambridge: Cambridge University Press, 2012).

compartmentalized emotions just because these are not genuine emotions of the same kind as their uncompartmentalized brethren. This would be to follow Walton's lead, though with a new explanation of the difference between emotion-kinds.

There are two reasons I do not take this route here. First, given that (i) is so widely accepted, it is strategic to answer the irrationalist without relying on such a controversial assumption. Second, and more importantly, it is unlikely that compartmentalization by itself justifies a distinction between emotion-kinds. For instance, both simulationists and philosophers who postulate a separate imagination system recognize that emotions in imaginative contexts are shorn of their typical motivational role. But this is not because the affective states are offline or contained within a distinct cognitive 'box'; it is because, though entirely ordinary emotions, they are disconnected from (online) beliefs and desires.[60]

An additional reason to doubt a sharp distinction becomes apparent in considering *when* we compartmentalize our emotions. The prompts are likely to be many and complex, but I suggest one explanation that will be directly relevant to answering Radford's challenge. Consider: Why do we carry over our feelings about Oliver Twist to real orphans or our feelings about Juana to real migrants, but not our feelings about Obama-the-kitten-torturer? A plausible answer is that in the former cases we treat the fictional content as realistic about real orphans and migrants, whereas in the latter we know that Obama is misrepresented. That is, we compartmentalize our emotions to the extent that we take the content to be unrealistic or inaccurate in respects relevant to the emotion.

With the Obama imagining our negative feelings are likely to be completely compartmentalized, because the content of the episode is entirely contrary to our beliefs. In most cases, though, compartmentalization and carryover are matters of degree along different

[60] Currie and Ravenscroft, *Recreative Minds*, 190–191; Meskin and Weinberg, 'Puzzling over the Imagination', 184.

dimensions. We do not carry over *every* feeling about Juana or Oliver to real people, recognizing that the characters are fictionalized in some respects. Such differences of degree are difficult to square with a sharp distinction between emotion-kinds.

Moreover, we do not always compartmentalize our emotions when we should. If I imagine my partner's infidelity sufficiently vividly, I may find myself angry with him though I recognize that this is for no good reason.[61] If I find Hilary Mantel's (2009) portrait of Thomas Cromwell in *Wolf Hall* compelling, I might be more sympathetic to the historical figure than can be justified by my beliefs. Seeing Steven Spielberg's *Jaws* (1975) can provoke fear of swimming in the ocean even among those who know that the chances of a shark attack are miniscule. In each case, it looks like one and the same emotion 'escaping' the confines of a circumscribed imaginative project. When such escapes occur, we knowingly violate the correctness norm; to this extent, the emotions appear irrational.

Some may disagree. There is substantial debate about whether recalcitrant emotions count as *irrational*, as opposed to merely inappropriate.[62] After all, there are great white sharks and they do sometimes attack humans; the fear is not totally without foundation. Suppose, however, that someone who sees Steven Spielberg's *Jaws* (1975) develops a real-world fear, not of great white sharks, but of *giant, man-eating* great white sharks like the one in the film. And suppose that this person is fully aware, with not a shred of doubt, that there are no such sharks in reality. *This* fear is indisputably irrational.

Radford would retort that exactly the same combination of fear and disbelief characterizes ordinary viewers of *Jaws*. They disbelieve in the existence of giant man-eating

---

[61] This example is owed to [suppressed].

[62] See, e.g., Brady, 'The Irrationality of Recalcitrant Emotions'; Sabine A. Döring, 'What's Wrong with Recalcitrant Emotions? From Irrationality to Challenge of Agential Identity', *Dialectica* 69 (2015); 381–402; Alex Grzankowski, 'The Real Trouble with Recalcitrant Emotions', *Erkenntnis* 82 (2016), 641–651.

sharks, yet they experience fear in watching Spielberg's film. Why are they not irrational? Why doesn't (CR) apply to them as much as it does to the person who fears non-existent giant sharks in actual oceans?

My answer is that their fear is appropriately compartmentalized, restricted to the context of the film and prevented from interacting with other mental states or producing action. This compartmentalization is not all-or-nothing. A mild fear of great white sharks after seeing *Jaws* could be perfectly rational, insofar as the film reminds us that such sharks exist in the oceans. But fear based on what we recognize to be unrealistic or inaccurate is irrational. In the next section I argue that our capacity to make such discriminations is essential to certain claims about learning emotionally from fiction.

8. Answering the Challenge

The idea that fiction is a source of emotional learning is familiar. Some philosophers and psychologists argue that feeling for fictional characters renders us more empathetic,[63] while others maintain that it enhances understanding or insight into the emotions.[64] A number adopt the Aristotelian idea that responses to fictional characters can train us to 'have the right emotional reactions at the right time, in the right way directed to the right objects'.[65] Such

---

[63] E.g., Martha C. Nussbaum, *Poetic Justice: The Literary Imagination and Public Life* (Boston, MA: Beacon, 1997); Raymond A. Mar et al., 'Bookworms versus Nerds: Exposure to Fiction versus Non-Fiction, Divergent Associations with Social Ability, and the Simulation of Fictional Social Worlds', *Journal of Research in Personality* 40 (2006), 694–712; Maja Djikic, Keith Oatley, and Mihnea C. Moldoveanu, 'Reading Other Minds: Effects of Literature on Empathy', *Scientific Study of Literature* 3 (2013), 28–47.

[64] E.g., Robinson, *Deeper than Reason*, 154-194; Gaut, *Art, Emotion and Ethics*, 134-164.

[65] Amélie Rorty, 'The Psychology of Aristotelian Tragedy', in *Essays on Aristotle's Poetics* (Princeton: Princeton University Press, 1992), 12. See also Ronald de Sousa, *The Rationality of Emotion* (Cambridge, MA: MIT Press, 1990); Martha C. Nussbaum, *Love's Knowledge: Essays on Philosophy and Literature* (Oxford:

claims take for granted that we abstract away appropriately, so that we carry over emotions to 'the right objects' in the real world. These claims only make sense if we recognize the role of compartmentalization.

Suppose that in reading George Orwell's *Animal Farm* (1945), I initially sympathize with the animals who have been mistreated by the farmer, Mr Jones. I root for them in the Battle of the Cowshed, hope that they will live a peaceful life in an equal society, and find myself disappointed when the pigs become ruthlessly dictatorial. I particularly detest and fear Napoleon and pity his victims, such as Boxer. Needless to say, Orwell's purpose is not to elicit sympathy for revolutionary farm animals, suspense about the outcome of human-animal farm battles or fear of autocratic pigs. Rather, it is to generate emotional responses toward the real-world models for the satire—fear and detestation of Stalin and his henchmen, and sympathy for their victims—so as to educate people about the realities of Stalinism. Only if readers compartmentalize those dimensions of their emotional responses that are sensitive to the fictionalizations, while carrying over only those aspects relevant to the RWC, can we take them to have learned from the fiction.

Other accounts cannot accommodate these discriminations. If we simply compartmentalized our emotions whenever we were engaged with fiction, we would fail to carry over our emotional responses to Orwell's real-world targets. That would be to miss the point of the novel. Nor does the fact that we merely imagine the characters and events portrayed preclude our carrying over the emotions. Oliver Twist does not exist, and we merely imagine his adventures in response to a work of fiction; but we do not doubt that our pity of his circumstances should be carried over to real Victorian orphans. At the same time, we keep some responses compartmentalized. For instance, we do not carry over our feelings

---

OUP, 1992); Noël Carroll, 'Art, Narrative, and Moral Understanding', in Jerrold Levinson (ed.), *Aesthetics and Ethics: Essays at the Intersection* (Cambridge: Cambridge University Press, 1998), 126–160.

about the pigs in *Animal Farm* to actual pigs. It is because we recognize that this aspect of the story is unrealistic, rather than because we are dealing with fiction or imagining, that prompts the compartmentalization.

It is this connection between compartmentalization and unreality that explains why we do not treat emotional responses to fictional characters as violating the correctness norm *simply* in virtue of the non-existence of their objects. We know that works of fiction are inaccurate with respect to ontology; most of their characters are invented. As a result, our sympathy for Oliver Twist is compartmentalized insofar as it is directed *at Oliver*; and similarly for pity of Anna Karenina, grief for Mercutio, fear of giant sharks, and so on. The departure from real-world ontology is also a premise of many imaginative projects. If I imagine that my partner has been unfaithful, I might also imagine the lover who (I know) does not exist; my feelings of jealousy will be compartmentalized insofar as they are directed *at the lover*. These emotions remain compartmentalized with respect to their specific intentional objects. This is why (CR) does not apply to them.

Let us take stock. I have argued that the paradoxical interpretation of the rationality puzzle is fundamentally misguided. As Radford made clear already in 1975, the puzzle does not concern the theorist who might accept incompatible propositions *about* emotional responses to fiction (as articulated in the Paradox of Fiction), but rather the person who *experiences* the emotions. Why are we not irrational in responding emotionally to fictional characters? We cannot answer this question with a wholesale rejection of (CR); it is clear that in many contexts, rational emotions *do* require belief in the existence of their objects. In such cases, the emotions constitute knowing violations of the correctness norm. The real challenge is thus to explain the normative difference between fearing the monster under the bed—a knowing violation of the norm—and fearing a movie monster—which does not seem subject

to the norm at all. I proposed that the correctness norm is suspended when our emotions are *compartmentalized*, typically in response to unrealistic or inaccurate content.

Importantly, the emotions need not be compartmentalized in all respects; if they were, we could not make sense of the claim that we learn by carrying over emotions to similar real-world people and situations. Because Oliver Twist (closely) resembles real orphans, and Napoleon the pig (more distantly) resembles Stalin, emotions sensitive to the realistic aspects of their stories can be uncompartmentalized, available to be carried over to the RWC. Compartmentalization is a matter of degree. This is one reason to doubt that compartmentalization underwrites a contrast between emotion-kinds.

It is also a reason why we cannot answer the challenge of irrationality *just* by pointing out that ordinary norms of correctness do not apply when we are responding to fiction or fictional characters, to what is storified, or to what we imagine. Instead, what matters is the extent to which our emotional responses remain compartmentalized within the fictional or imaginative context. Emotions that are compartmentalized are not subject to the same norms. It is because we compartmentalize our emotional responses toward fictional characters with respect to their intentional objects that these emotions can be rational even when we believe their objects do not exist. This is, finally, the answer to Radford's challenge.[66]

[66][Acknowledgements suppressed].