Ergo AN OPEN ACCESS JOURNAL OF PHILOSOPHY

The Second Revolution of Moral Fictionalism

ELINE GERRITSEN
University of St Andrews
University of Groningen

If our moral beliefs rest on a mistake, as moral error theorists claim, what should we do with them? According to Richard Joyce's *revolutionary moral fictionalism*, error theorists should pretend to believe moral propositions in order to keep the benefits moral thinking has for their preference satisfaction. This, he claims, frees error theory from radical practical implications. In response, I argue that implementing fictionalism would not preserve our moral practices, but disrupt them. The change from moral belief to make-belief yields an unintended second revolution: a revolution in the content of morality. I show that fictionalism necessarily relies on a similar justification of moral practices as David Gauthier's *contractarianism*, and consequently has similar implications for moral content. Because fictionalists engage in moral thinking purely for its instrumental value, they should only accept moral obligations that are useful to them into their fiction. This restriction is important: the most useful moral fiction departs substantially from conventional moral views. Revolutionary moral fictionalism is therefore more radical than it is promised to be.

Keywords: contractarianism; David Gauthier; fictionalism; moral error theory; now-what problem; Richard Joyce

1. Introduction

What should we do if there is nothing we morally ought to do? According to *moral error theory*, moral claims ascribe moral properties that either do not exist or are never instantiated. This entails that the moral judgements we use to deliberate, evaluate and advise all rest on a mistake. If we were to discover that this is the case, how should we move on? This 'now-what problem' has received growing

Contact: Eline Gerritsen <elinegerritsen@outlook.com>

https://doi.org/10.3998/ergo.2276

attention from error theorists themselves,[1] with good reason. Error theory may be thought to imply that we should get rid of moral thinking.[2] However, our moral practices run deep in the fabric of society, and the way we behave towards each other would be drastically different without them. This threat of radical practical consequences puts a significant burden on error theory. Error theorists like Richard Joyce therefore have a clear incentive to try to remove this threat: "If a persuasive case were to be made that we could adopt the error-theoretic position and civilization would not collapse—that life would go on as before, or even go on *better*!—then the opposition to the theory might diminish, or at the very least lose some of its determination" (Joyce 2019: 151, original italics).

With this ambition, Joyce presents his *revolutionary moral fictionalism* (hereafter referred to as 'fictionalism'). He sees morality as a useful fiction, due to the special advantages he ascribes to moral thinking. Joyce recommends that we preserve these advantages after rejecting moral beliefs as mistaken by *pretending* that some moral propositions are true. The appeal of fictionalism is that it allows error theorists to continue to deliberate in moral terms yet avoid being committed to supposed falsehoods. In contrast to *hermeneutic fictionalism*, the claim of revolutionary fictionalism is not that we already merely pretend that some moral propositions are true, but that we *should* do this. Joyce suggests that everything will more or less stay the same when we make this switch from moral belief to make-belief. In this way, fictionalism is supposed to invalidate concerns about the practical impact of collectively accepting error theory (Joyce 2001: 231).

I will argue that fictionalism fails to meet this promise, because it is more revolutionary than has been assumed. The problem is that fictionalism does not merely require us to change our attitude towards morality from belief to pretence; it also requires us to change the *content* of morality. In other words, a fictionalist should endorse a different set of moral obligations than moral believers conventionally endorse.[3] This becomes clear once we recognise that Joyce's instrumental justification of moral practices is fundamentally similar to that of David Gauthier's *contractarianism*. I will argue that, just like the contractarian morality, the revolutionary moral fiction lacks a substantial range of conventionally recognised moral obligations. Although my discussion is focused on Joyce's argument for fictionalism, the implications it uncovers apply more generally: the change in morality's content is a challenge for anyone who advocates becoming a fictionalist for the

1. See, for example, Garner and Joyce (2019).

2. Throughout my discussion, 'error theory' refers specifically to moral error theory of the kind defended by Joyce, not to broader error theories about normativity (e.g., Streumer 2017; Olson 2014).

3. My focus, like Joyce's, is on fictionalism as a collective strategy for a group of error theorists; hence, my conclusions may not extend to a scenario where an isolated error theorist decides to adopt fictionalism.

sake of instrumental value (e.g., Nolan, Restall, & West 2005)—especially when, as is standard, this is combined with Joyce's prominent form of moral error theory. The fiction's revolutionary content is a significant drawback for both fictionalism and error theory. To absolve error theory from the burden of having radical practical implications, showing that error theorists should not abolish moral practices completely is not enough; it also needs to be shown that they should preserve a wide range of moral norms resembling our current commitments. This is where fictionalism fails. The goal of this paper is not to establish that fictionalism is the wrong answer to the now-what problem or that error theory is false, but rather to demonstrate that the former cannot remove the unappealing revolutionary nature of the latter.

2. The Rationale for Revolutionary Moral Fictionalism

Joyce's moral error theory and fictionalism are both motivated by his view of reasons. An important distinction here is that between categorical and hypothetical reasons. The general idea is that while hypothetical reasons are contingent on the agent's aims and preferences, categorical reasons for action obtain regardless of how the action in question relates to the agent's ends. Joyce (2001) argues that, as a matter of the meaning of our moral terms, moral obligations necessarily come with categorical reasons. At the same time, he argues that categorical reasons for action do not exist; what we have reason to do fully depends on our preferences. Together, these claims entail that moral obligations do not exist.

In light of his argument for error theory, Joyce can only argue that error theorists should adopt fictionalism by pointing to hypothetical reasons to do so. In his view, the question of what to do with discredited moral thinking is settled by a calculation of what serves our ends best: "when morality is removed from the picture, what is practically called for is a matter of a cost-benefit analysis, where the costs and benefits can be understood liberally as preference satisfactions" (Joyce 2001: 177). Consequently, the only basis on which Joyce can recommend that we adopt fictionalism is its instrumental value in maximising our individual preference satisfaction. Although advocates of fictionalism can make adjustments to the exact view of hypothetical reasons used, they must understand reasons to ultimately be dependent on the agent's aims or preferences, or else Joyce's type of argument for error theory would be undermined.[4]

4. I base my discussion on the basic instrumentalist view of reasons that Joyce consistently uses in his defence of fictionalism. In his defence of error theory, he accepts an alternative instrumentalist view according to which an agent has a normative reason to do an action if and only if an idealised version of herself, who is fully informed and deliberating flawlessly, would advise her to do the action (Joyce 2001: 53–79). Joyce emphasises that, on this second view, an agent's reasons

There are two prominent alternative responses to accepting error theory that Joyce needs to dismiss as less useful than fictionalism. The first is *abolitionism*, which requires error theorists to get rid of moral thinking and stop using moral language altogether (Garner 2010). While abolitionists take moralising to cause more harm than good, Joyce argues that moral thinking is too valuable to abandon. He ascribes a special role in deliberation to moral beliefs: when we think of an action as morally required, we attach a 'must-be-doneness' to it (Joyce 2001: 181). While we may fail to act on merely prudential considerations due to weakness of will, moral thinking "functions to bolster self-control against such practical irrationality" (Joyce 2001: 184). Joyce thus takes moral beliefs to be advantageous, despite being false, in virtue of their special motivational power. Of course, this is only beneficial if moral beliefs motivate actions that are conducive to our preference satisfaction. Joyce assumes that, generally, this is the case: morality is supposed to promote sincere cooperation, which is itself instrumentally valuable (Joyce 2001: 181). I will elaborate on the instrumental value of moral behaviour below.

The second main alternative to fictionalism is *conservationism*, the recommendation that error theorists continue having moral beliefs and making moral assertions in everyday contexts, despite being disposed to believe in error theory in critical contexts, such as the seminar room (Olson 2011). Against this, Joyce (2001: 178–79) argues that making ourselves believe moral propositions while recognising evidence that they are false will have detrimental effects, since it violates the instrumentally valuable policy of aiming for true beliefs.[5] He concludes that maintaining genuine moral beliefs will not maximise error theorists' preference satisfaction.

However, Joyce does not want to give up on moral thinking as a whole: he recommends that we avoid the harms of false beliefs yet preserve the benefits of seeing the world through a moral lens by becoming *fictionalists* about moral discourse. In most contexts, a moral fictionalist sounds just like a moral believer: she will call certain actions right or wrong and employ moral terms in her deliberation. What distinguishes her from a moral believer is that the fictionalist does not actually believe moral propositions, but only pretends that they are true. Likewise, she merely pretends to assert moral propositions—her moral utterances lack assertoric force, as if she were speaking as an actor in a play. The

still fully depend on her contingent desires (2001: 80–105). This aspect of reasons being relative to individual and contingent ends is crucial for the implications I will draw for the fiction's content. Whether the agent ought to pursue her actual or idealised (but still contingent and subjective) ends is less relevant for my argument. My argument would therefore still go through if fictionalism were defended based on Joyce's alternative instrumentalist view. The same applies, arguably, to any other view of reasons that error theorists could coherently endorse.

5. For criticism of this argument, see Olson (2011: 193–95).

fictionalist's commitment to error theory will only be apparent in contexts where the status of morality is discussed, where she will be ready to admit that her engagement with morality is merely an elaborate case of make-belief. In Joyce's view, his fictionalism is revolutionary in the sense that it requires a change from moral belief to make-belief.

As long as some of the positive impact of genuine moral beliefs remains when they get turned into make-belief, fictionalism has an advantage over abolitionism. Assuming that cultivating moral beliefs known to be false is too harmful and that there are no other viable options, this would mean that error theorists should engage with morality as a fiction. For the sake of argument, I will accept Joyce's conclusion that fictionalism wins this cost-benefit analysis.[6] My focus is on what the implications are of error theorists choosing this policy for the sake of their own preference satisfaction.

After establishing that an error theorist should adopt an attitude of pretence towards morality, a substantial question remains: *which* moral propositions should she pretend to be true? Morality is not made up of a single, clear-cut set of propositions that any moral believer accepts. Since the contents of genuine moral beliefs differ, the content of the moral fiction can vary as well. As fictionalists, do we pretend that it is wrong to lie to the murderer at the door? Do we tell ourselves that we ought to donate a share of our income to those in need? There is an infinite range of specific moral fictions to choose from—which should a revolutionary fictionalist adopt? Importantly, since fictionalism is built on error theory, the answer cannot be that a fictionalist ought to endorse moral propositions that are true.

Instead, the appropriate criterion to assess moral fictions by is the same one that requires error theorists to adopt fictionalism rather than an alternative: the right option is the one with most instrumental value. The only reason an error theorist has to engage with morality as a fiction is that this is conducive to the satisfaction of her preferences. As a consequence, the specific moral fiction she should adopt is the fiction that it is most advantageous for her to engage with. A fictionalist should thus pretend that lying is wrong if and only if this is more beneficial to her than not doing so. The whole set of moral propositions that a fictionalist will pretend to be true will be determined in the same way. This follows directly from Joyce's argument for adopting fictionalism. Once error theory is accepted, the right version of morality is the most useful one. This has been recognised by Joyce (2019: 154) as well as other commentators on fictionalism (Nolan et al. 2005: 327; Olson 2011: 189). What has hardly been explored so far, however, is which moral fiction we can expect to be the most useful one (cf. Jaquet 2021).

6. For objections against Joyce's argument, see Cuneo and Christy (2011), Eriksson and Olson (2019), Lutz (2014) and Olson (2011).

To be clear, the question here is which fiction we should adopt upon becoming fictionalists and continue to use thereafter. The idea is not that we should determine for each situation separately whether it is most advantageous to accept certain moral propositions; this continuous adjustment would be cognitively demanding and motivationally ineffective. As Joyce describes it, the fictionalist does not choose to engage with morality only in those situations where moral motivation would help her, but adopts a constantly present "habit of ethical thinking" (Joyce 2001: 219). For this habit to be effective, the content of one's moral thoughts should be constant as well. Thus, we are looking for a stable and continually present moral fiction that is overall the most advantageous to adopt.

Nolan, Restall and West have suggested that which moral fiction is most useful will largely depend on current moral practices: "it will be easier to institute a fiction that is a close relation of moral theories currently employed, than to construct a new one out of whole cloth" (Nolan et al. 2005: 327). Indeed, we cannot completely ignore which beliefs we hold before becoming fictionalists. If the point of keeping moral thoughts is that they motivate us, the content of the moral fiction must be restricted to what can spark moral motivation, and this will depend on which moral beliefs we are used to. Some instrumentally desirable types of actions are so far removed from traditional conceptions of moral behaviour that we could not have a sense of moral 'must-be-doneness' about them—for example, avoiding tax. Furthermore, it is possible that the specific moral thoughts that correspond to a fictionalist's previous moral beliefs will have the strongest motivational power over her.

However, this does not entail that the content of our moral fiction will be identical to the content of our prior moral beliefs; on the contrary, it is very improbable that we had already accepted precisely the most useful version of morality. It is likely that some of a fictionalist's prior moral beliefs lead to behaviour that is not ultimately advantageous to her, and it is even more likely that the set of prior beliefs does not lead to behaviour that is *maximally* advantageous to her. In choosing the fiction, we need to consider which moral thoughts can give us a sense of 'must-be-doneness', but it is also crucial whether the actions these thoughts would motivate us to do actually maximise our preference satisfaction. Moral beliefs that are not to the believer's advantage should not be preserved in the fiction. Surely, if we give up on moral truth and engage with morality purely for the instrumental value of doing so, improvements can be made.

It may well be true that adopting a conservative fiction is more useful than not adopting a moral fiction at all; yet, for Joyce's purposes, this is not enough. His case for fictionalism is not that it is sufficiently advantageous, but that it is the most advantageous policy available to the error theorist. Consequently, he cannot settle for a fiction that is merely good enough. In Joyce's framework, maximising preference satisfaction is all that matters. A fiction should therefore

be chosen if and only if it is optimally useful in this sense; there is no room for other considerations.[7] A recommendation to adopt a fiction that corresponds to our prior moral beliefs rests on a failure to include other possible moral fictions in the calculation of which policy is best.

Joyce does not explicitly defend adopting a fiction that preserves the content of our moral beliefs; his recommendation is rather that error theorists accept the "conceptual framework" of morality (Joyce 2001: 195). In his words, this merely involves accepting general claims such as "There are obligations and prohibitions" and "Wrong-doers deserve punishment", as well as minimal constrains on the content of moral norms—for example, "Torturing babies to pass the time is always wrong" (Joyce 2001: 195). However, as I have argued, fictionalism comes with a criterion for settling the specific content of the fiction. Recommending a moral fiction in the form of an open-ended conceptual framework is not in line with this. More implicitly, Joyce does suggest that the fiction would have conservative content. Only if fictionalism preserves the traditional content of moral discourse can it play the role, which Joyce ascribes to it, of undermining worries about the practical impact of accepting error theory. Furthermore, Joyce typically describes fictionalism as the recommendation to *continue* with moral discourse, but as a fiction (e.g., 2001: 221). This implies that adopting fictionalism is merely a matter of transforming an attitude of belief in moral propositions to one of pretence. As an illustration of the transition to fictionalism, Joyce describes a hypothetical person who is raised to think of actions as right or wrong but eventually becomes an error theorist, at which point "these patterns of thought are now so deeply embedded that in everyday life she carries on employing them, and is happy to do so—she becomes a moral fictionalist" (Joyce 2001: 224). A natural interpretation of this is that the content of her moral thoughts is not affected by her transforming from a moral believer into an error theorist and fictionalist.

I believe this is mistaken: what is missing from this story is that once the error theorist decides to carry on with moral thinking for the sake of her preference satisfaction, she should wonder if she can adjust her patterns of thought to be more advantageous. Unless her moral views were already precisely the most beneficial ones for her, she should make changes. This shows that there are two steps involved in becoming a revolutionary moral fictionalist. What is normally emphasised is the revolution in one's *attitude* towards morality: adopting fictionalism means switching from believing moral propositions to pretending that they are true. However, it follows from the argument for fictionalism that a second revolution is involved: a revolution in the *content* of morality. When the

7. A proponent of fictionalism may want to reject the commitment to maximisation and defend a satisficing conception of rationality instead, which might allow the choice of a suboptimal fiction to be rational. However, it is doubtful that a true satisficing conception of rationality is compatible with a strictly instrumental and preference-based view of reasons (Byron 1998).

fictionalist switches from belief to pretence, she should also change the content of the moral propositions she accepts in this way.

3. Fictionalism and Contractarianism

Fictionalism is strongest as a policy that is accepted collectively rather than independently. Since the instrumental value of acting in accordance with moral norms largely depends on others reciprocating, the habit of moral thinking is advantageous to an individual under the condition that this practice is widespread in her society. Joyce presents his fictionalism as a way forward for a group of error theorists: "By asking what *we* ought to do I am asking how a *group* of persons, who share a variety of broad interests, projects, ends—and who have come to the realization that morality is a bankrupt theory—might best carry on" (Joyce 2001: 177, original italics).

Arguably, such a group will be best advised to adopt a single moral fiction together: if fictionalists do not coordinate their moral thoughts and resulting behaviour, the cooperation their engagement with morality leads to will be suboptimal. For example, pretending that it is wrong to break promises is advantageous for an agent if done collectively, but puts her at risk of exploitation if some members of the group do not accept this prohibition. To see which moral fiction a fictionalist should adopt, then, we need to investigate which moral fiction is best for the members of her society to adopt together. However, there is no room for considerations about the collective good in Joyce's framework; individual preference satisfaction is all that matters. Thus, the content of the moral fiction of a society of fictionalists will be restricted by what is advantageous to its members as individuals.

With this in mind, the moral fiction can be understood as a social contract that fictionalists endorse for their own benefit. Interestingly, this means that there is an important resemblance between fictionalism and accounts of morality as a rational agreement between individuals. In particular, Joyce's fictionalism is closely related to David Gauthier's contractarian theory of morality, which aims to vindicate morality as a set of constraints which it is rational for actual persons to agree to and comply with (Gauthier 1986).[8] Note that Joyce and Gauthier are nevertheless opponents: Gauthier is not an error theorist. He assumes that genuine moral requirements exist and provides a theory of why we should comply with them, rather than a theory of what to do in absence of moral facts. Still, we can learn more about fictionalism by looking at contractarianism. As I will show,

8. The term 'contractarianism' can be used for both Hobbesian social contract theories and Kantian contractualist theories. In this paper, it refers specifically to Gauthier's Hobbesian theory.

Gauthier and Joyce justify moral practices along the same lines. For contractarianism, it has long been established what implications this justification has for the content of the vindicated morality. Due to the parallels between Gauthier's and Joyce's arguments, these implications are relevant for fictionalism too.

There are three main elements that make Joyce's fictionalism and Gauthier's contractarianism significantly similar. The first is that both use strictly non-moral premises to argue in favour of adopting moral practices. It is clear that Joyce, as an error theorist, must argue for embracing moral thinking without referring to moral facts. Gauthier likewise justifies morality on terms that are acceptable to someone who does not already recognise moral demands: "We are committed to showing why an individual, reasoning from non-moral premisses, would accept the constraints of morality on his choices" (Gauthier 1986: 5). This agent Gauthier has in mind does not yet distinguish "between what he may and may not do" or "recognize a moral dimension to choice" (Gauthier 1986: 9). Therefore, just like Joyce's, Gauthier's justification of morality is characterised by a complete absence of moral considerations.

Related to this is the second important shared element, the strictly instrumental view of reasons. Gauthier dismisses attempts to vindicate morality with a 'moralised' conception of rationality, such as notions of rationality that presuppose an impartial viewpoint (Gauthier 1986: 4–8). Instead, in Gauthier's view, a person acts rationally if and only if she "seeks the greatest satisfaction of her own interests" (Gauthier 1986: 7). Here, a person's interests are understood simply in terms of her preferences. Thus, like Joyce, Gauthier believes that an agent has a reason to perform an action if and only if it maximises her own preference satisfaction. This gives them the same starting point for a justification of moral practices: for both Joyce and Gauthier, the task is to show that engaging with morality maximises an individual's preference satisfaction.[9] Any other consideration in favour of morality will not be of the right kind.

Finally, Joyce and Gauthier make a similar case for why morality is justified in this way: central to both accounts is that being moral entails being cooperative, and that cooperation is beneficial to the individual. Where their arguments for these two claims diverge, this is mainly because Gauthier's account is much more elaborate in this respect than Joyce's. Since Joyce mostly leaves open why it is advantageous to be cooperative in interaction with others, Gauthier's account of this can be plugged into Joyce's argument to make the latter more complete. In my view, then, a more sophisticated version of Joyce's justification of moral practices would be even more similar to Gauthier's. In any case, the contractarian and fictionalist justifications of morality must run largely parallel, due to

9. For Joyce, this can include the satisfaction of other-regarding preferences. I address how this affects the content of the fiction in Section 4.2.

their shared starting point. With this in mind, I will now discuss Gauthier's justification of morality and its implications for morality's content.

Gauthier claims that morality promotes cooperation by demanding impartial constraints on an agent's utility maximisation. In his view, adopting morality consists of letting go of a policy of directly pursuing what is best for you without restrictions. He argues that restricting yourself is ultimately in your best interest. In the first place, it is beneficial for you if *others* adopt impartial constraints on their behaviour, since they may otherwise pursue their own preference satisfaction at great cost to you. However, others cannot be expected to constrain their behaviour towards you unless you accept these constraints as well. To access the benefits of cooperation, you must play by the rules. It is therefore in one's best interest to forego a strategy of directly pursuing what is in one's best interest, on the condition that others do the same (Gauthier 1991: 23). Gauthier claims that others will notice if you try to fool them into cooperating with you by merely superficially accepting moral constraints. Therefore, to benefit from moral practices, you must fully endorse them by developing a deeply ingrained disposition to constrain your direct utility maximisation, including in specific situations in which this is disadvantageous to you (Gauthier 1986: 172–77).[10]

The nature of this justification has important consequences for the scope of moral obligations it supports. In contractarian thought, you have reason to behave in a considerate way towards others not because they have intrinsic value, but because they are instrumentally valuable to you (Hampton 1991: 48). However, crucially, not every person provides instrumental value for us in the sense Gauthier is interested in. Not everyone is in a position to engage in mutually beneficial cooperation with you; some relations are unavoidably asymmetrical, and here restrained cooperation can bring one party more costs than benefits. Yet, it follows from Gauthier's argument that if others cannot offer beneficial cooperation in return for you constraining your behaviour towards them, then you do not have a reason for adopting such constraints. He acknowledges that this leads contractarianism away from common conceptions of the content

10. I will take for granted that this argument succeeds. In reality, there are several major problems that Gauthier is known to face and seems unable to overcome. For one, there are many conceivable policies to adopt towards utility maximisation, and Gauthier has not established that fully committing yourself to constrained utility maximisation is the best option of all (Copp 1991: 220–21; Smith 1991: 238–43; Sayre-McCord 1991). Moreover, even if it is true that it is rational to adopt a disposition of constraining your behaviour, this fails to show that it is rational to *act on* such a disposition (Parfit 2011: 433–47; Smith 1991: 244–49; Copp 1991: 207). An interesting upshot of the similarities with contractarianism is that fictionalism faces these same challenges. Perhaps Joyce is able to provide an answer to Gauthier's problems—for example, he may argue that committing to constrained utility maximisation is the best available option due to our weakness of will. However, it is important that we are as critical of these steps in the fictionalist argument as we are in the case of the contractarian argument.

of morality: "we may agree that the moral constraints arising from what are, in the fullest sense, conditions of mutual advantage, do not correspond in every respect to the 'plain duties' of conventional morality" (Gauthier 1986: 268).

In particular, Gauthier cannot recognise moral obligations towards several categories of persons that we do normally take ourselves to owe moral consideration to. The first is persons outside of the society that our social interactions take place in. Interactions with outsiders are rare—when you do have an opportunity to benefit them, this is unlikely to result in a benefit to you. Gauthier thus argues that the needs of a different society that one does not cooperate with are morally irrelevant (1986: 282–88). Non-human animals are a major subcategory of outsiders that do not warrant moral consideration on the contractarian picture (Gauthier 1986: 268).

A second category excluded from the contractarian morality is future persons whose lives do not overlap with yours. The asymmetric relation here is obvious: any constraints you put on your behaviour for the sake of future persons cannot be reciprocated. If no cooperation is possible with our descendants, no injustice is committed when you do not consider their interests, even if this leaves them with an uninhabitable world (Gauthier 1986: 298).[11]

Another striking omission from the contractarian morality are moral obligations towards severely disabled or chronically ill persons: "Only beings whose physical and mental capacities are either roughly equal or mutually complementary can expect to find cooperation beneficial to all. [. . .] Among unequals, one party may benefit most by coercing the other, and on our theory would have no reason to refrain" (Gauthier 1986: 17). Some disabled persons permanently cannot engage in mutually beneficial cooperation, either because of the nature of their disability itself or because public space is not made accessible to them (Nussbaum 2007: 117–18). These persons therefore have no grounds for demanding any kind of moral respect in the contractarian picture. Clearly, Gauthier's instrumental justification of moral practices leaves him with a very limited form of morality.

Because fictionalism has the same starting point as contractarianism, Joyce unavoidably faces very similar implications for the content of morality. Revolutionary fictionalists should accept the specific moral fiction that is maximally useful for their individual preference satisfaction. Just like Gauthier, Joyce makes the usefulness of moral thoughts and behaviour contingent on the usefulness of cooperation. However, as we learned from contractarianism, a morality purely based on the rewards of cooperation cannot match the extension of ordinary conceptions of morality. Many persons we normally ascribe moral status

11. Gauthier (1986: 299) argues that there are cooperative links between all generations due to generational overlap. This argument is shown to rest on highly implausible assumptions in Arrhenius (1999).

to do not qualify for mutually beneficial cooperation. If the moral fiction is to be chosen on this basis, then a fictionalist should not pretend that it is morally wrong to harm outsiders, non-human animals, future persons or severely disabled persons.

My claim that fictionalism fails to account for some conventional moral obligations is not meant to imply that there is a single conventional morality that all moral believers endorse. Moral beliefs are not uniform. Nonetheless, there appears to be agreement among most people about some aspects of morality—and fictionalism turns out to fall short here. It seems highly unusual to recognise no moral obligations whatsoever towards other societies, non-human animals and future generations. Moreover, the requirement to take care of the most vulnerable can be seen as one of the basic elements of our moral code. The revolutionary fiction goes directly against this if it undermines the moral status of severely disabled members of society.

4. Objections and Replies

4.1. The Feasibility of Changing Moral Thinking

A potential objection to my account of fictionalism is that we are psychologically incapable of fully adopting the moral fiction with revolutionary content. The worry is that our moral judgements, emotions and dispositions are so deeply ingrained in us that swapping them for a more useful set is not an option. Alternatively, one could object that radically changing our moral thinking would not be most advantageous, even if it is possible. It may be thought that a newly adopted set of moral thoughts would lack the motivational power that is supposed to make them useful. In addition, accepting a revisionary fiction would come with psychological resistance, while a conservative fiction is more user-friendly in comparison.[12] Due to the mental adjustment required, the thought goes, it would be either impossible or disadvantageous to adopt the revised moral fiction I have sketched. How much change is actually possible in the moral lives of a hypothetical group of committed error theorists is an empirical question which we cannot settle here. Still, there are some reasons to expect that a revolution is both possible and most conducive to fictionalists' preference satisfaction.

Firstly, my argument does not presuppose that fictionalists are capable of endorsing and being motivated by fictional moral obligations they did not previously believe in; my suggestion is merely that a range of traditional obligations would be *left out* of the fiction, without new obligations or rights replacing them.

12. Thanks to an anonymous referee for raising this point.

That a fictionalist should not pretend that she has obligations towards future persons does not mean that she should pretend that it is morally permissible to ruin future lives; instead, she should not think about future persons in moral terms at all. Her moral thinking should stay 'turned off' when she is considering actions that affect them, just like she does not engage in moral thinking when considering whether to go for a run. This change will still take some mental effort. However, it is an effort that will be worth it: since the obligations that are removed from the fiction are precisely those that constitute an uncompensated burden to many, the fictionalist has a strong motive to do what she can to remove any traces of moral beliefs about them. It seems plausible that many persons would happily stop seeing themselves as obligated to donate to an overseas disaster relief fund, for example. When the fictionalist does revert to traditional moral thoughts, she should try to correct herself and refrain from acting on them, with an eye to the substantial future benefits of being unburdened by these obligations.

Secondly, fictionalism is a long-term policy. Presumably, the ideal is that groups of error theorists who adopt fictionalism will go on to be immersed in the moral fiction for good. While they may struggle to fully internalise the revised moral content at first, they will have the rest of their lives to get used to the more advantageous moral practices. When the erosion of moral thinking is gradual, the psychological resistance to it will be less strong. New generations, moreover, can be brought up to be committed to the adjusted moral values from the start. Therefore, if the revisionary nature of the fiction I sketched makes it difficult to internalise effectively, or if the mental effort involved in internalising it counterbalances some of its advantage over a conservative fiction, this will plausibly be minimised over time. Nonetheless, this does not make the moral practices that the society of fictionalists would end up with—and which fictionalists should aspire to move towards now—any less radical from our own perspective.

4.2. Other-Regarding Preferences

Another possible worry about my account of the revolutionary fiction is that it seems to ignore our other-regarding preferences, which might justify more extensive moral practices. Gauthier dismisses social practices that are only beneficial for agents due to their sympathy for others as a form of exploitation (Gauthier 1986: 11). Consequently, his goal is to show that endorsing moral obligations gives a net benefit to an agent even if we do not take her other-regarding preferences into account. Joyce need not limit himself in this way: he can defend fictionalism as the best way forward for error theorists given their full range of preferences, not just their selfish ones.

However, it is unlikely that fictionalists' other-regarding preferences can support all traditionally recognised moral obligations; the average agent's desires to help others are too limited. Empirical research has shown that purely altruistic behaviour is normally motivated by feelings of empathy (Batson, Ahmad, & Stocks 2011: 111).[13] Here, empathy is understood as an "another-oriented emotional response elicited by and congruent with the perception of another person in need" (Batson et al. 2011: 110). When we do accommodate our capacity for empathy, we must also acknowledge its limits. Firstly, empathy is subject to a 'familiarity bias': we are more likely to empathise with persons who are similar or in some way personally connected to us. This bias applies to friends and family as well as to members of one's own ethnic or racial group (Hoffman 2000: 206–9).[14] Secondly, there is evidence of a 'here-and-now' bias: we are more likely to have an emphatic response towards a person who is physically present or salient in some way that draws our immediate attention (Hoffman 2000: 209–13). To illustrate, seeing the photo of a specific drowned refugee all over the news will evoke feelings of empathy and an accompanying willingness to help, which is absent when large numbers of similar victims are only presented as anonymous statistics. Relatedly, Jesse Prinz (2011: 224) points out that "empathy is hard to evoke for foreign masses". Empathy focuses on immediate, local needs, and ignores more widespread or systematic problems (Prinz 2011: 228). Strikingly, these limits to our capacity for empathy largely map onto the gaps in the contractarian morality. As a result of both the familiarity bias and the here-and-now bias, we normally do not feel a substantial level of empathy towards future persons or strangers with no relation to us, including non-human animals. Consequently, we are unlikely to have a preference to help them for their own sake.

Furthermore, while selfish behaviour is normally kept in check by moral beliefs, fictionalists will lack such a constraint. A moral believer who does not feel empathy for a group of strangers may nonetheless prefer to help them because she believes that is the right thing to do. This type of motivation is not available to a fictionalist, and is therefore irrelevant for determining the content of the moral fiction. The same is true for our preferences to help others that stem from a preference to comply with moralised social norms: a society of fictionalists should get rid of such norms when they are at odds with the most useful version of morality. Again, even if these norms remain internalised by fictionalists at first, the long duration of the fictionalist project ensures that they can eventually be phased out. Then, there will be no social punishment to fear when you fail to help people who cannot reciprocate. Hence, some of the motives people currently have to take the needs of others into consideration do not apply to

13. See also Batson, Duncan, Ackerman, Buckley, and Birch (1981) and Krebs (1975).
14. See also Prinz (2011: 227–28).

fictionalists. We can conclude that the most useful moral fiction will not include all traditional moral obligations even if we take fictionalists' full range of preferences into account.

With respect to moral obligations towards severely disabled persons, taking empathy-based preferences into account may have a better effect. The average abled person may be close enough to at least some disabled persons for her empathy to be triggered by their needs. If so, moral obligations towards disabled persons can be justified with reference to fictionalists' other-regarding preferences. Still, the foundation of these moral obligations would be very fragile: if it turns out that not enough fictionalists care about persons who cannot offer beneficial cooperation, there is no ground for extending moral consideration to them.

Of course, there is no doubt that the preferences of fictionalists who are disabled themselves will give them good reason to endorse obligations towards disabled persons. These preferences are not to be ignored. However, as I have argued, implementing a moral fiction should be treated as a collective enterprise, and its content will depend on what other fictionalists have reason to accept. The needs of vulnerable persons carry no extra weight here and, when they are a minority, can easily fail to be reflected in the moral obligations that fictionalists are collectively willing to accept. Preference-based cost-benefit analyses of the kind Gauthier and Joyce use just cannot produce a conventional morality.

4.3. So What?

Faced with my sketch of the revolutionary content of the moral fiction, the advocates of fictionalism may respond: 'So what?'. In the case of Gauthier, the result that his theory breaks with conventional moral views directly suggests that it is false: unless our moral beliefs are strongly mistaken, the contractarian picture of morality is inaccurate (Sinnott-Armstrong 2006: 170–71). In contrast, because revolutionary fictionalism is not a theory of the moral facts, its revisionary implications do not show it to be false. Those who are already fully committed error theorists may therefore see the discrepancies between the moral fiction and conventional moral beliefs as an interesting yet unproblematic result. Indeed, a Nietzschean nihilist may even embrace the 'death of morality' as an opportunity for us to realise that "many other, above all higher, moralities are possible or ought to be possible" (Nietzsche 1989: 202)—now, in the form of a fiction.

I grant that, for all I have said, error theory could still be true and fictionalism could be the best way to move forward. Even so, fictionalism's revolution in moral content does constitute a problem for both. Joyce presents fictionalism as a strategy that can invalidate concerns about the disruption that col-

lectively accepting error theory might bring (2001: 231). If it were to succeed in fulfilling this promise, that would significantly improve the status of error theory. Having radical implications is a disadvantage of a theory, even if those who are already on board are willing to accept them. Error theory is an unattractive position if it undermines both our common-sense beliefs and our way of life. Hence, some argue that a good proposal for what to do with morality after embracing error theory must let us keep practices that seem important to us, thereby making error theory more palatable (Lutz 2014: 352–53). Achieving this stability is supposed to be one of the main strengths of fictionalism. Therefore, it is highly problematic that it turns out to fail in exactly this respect. The moral fiction that should be chosen, based on its value defined in terms of individual preference satisfaction, would significantly restrict our moral practices. This amplifies the familiar counter-intuitiveness objection to error theory: combined with fictionalism, error theory not only entails that all moral beliefs are mistaken, but also contradicts our social practices and judgements on how to behave towards others. It is not the case that adopting fictionalism allows a society of error theorists to go on as before. Consequently, it cannot be employed to remove the burden of having revolutionary practical implications from error theory.

While fictionalism is only one possible way forward after giving up on moral facts, error theorists cannot simply avoid the problem I have raised by opting for one of the alternatives. Abolitionism clearly reinforces practical concerns about error theory, because it involves a complete disruption of our moral practices by definition. To alleviate these concerns, the abolitionist would need to provide a very convincing story about why, on the whole, we should not fear but welcome this disruption. That conservationism may also imply a revision of thought and behaviour is perhaps less obvious. However, it is important that, after accepting error theory, conservationists continue with having moral beliefs purely for their instrumental value. Given the assumption that there are no true moral beliefs, the content of these beliefs is not to be determined in the normal way, by the aim of appropriately responding to the evidence and representing the facts. Instead, conservationists should arguably make themselves have the moral beliefs that are most useful to them. If so, the implications I have drawn for fictionalism can be repeated for conservationism. While there may be relevant differences between believing and make-believing, I suspect that the most useful moral beliefs will be very similar in content to the most useful moral fiction. Therefore, my conclusion that fictionalism has revolutionary practical implications uncovers a more general problem for Joyce's form of moral error theory: we cannot consistently go on as before when instrumental value is all we have left. This does not entail that error theory is false, but does make it an even more radical and unattractive position than it would otherwise be.

5. Conclusion

Revolutionary fictionalism is more radical than has been assumed, due to its overlooked implications for the content of morality. Because error theorists are supposed to adopt a fictional morality purely for its instrumental value, they should choose the specific fiction that has most instrumental value for them, rather than the fiction that corresponds to their prior moral beliefs. This leads to the second revolution of fictionalism: although fictionalism is presented as merely requiring a change in attitude, it also requires a change in the set of moral obligations we accept. I have argued that the most useful moral fiction can be expected to exclude obligations to persons outside our society, non-human animals and future persons, and to make the moral status of severely disabled persons contingent on our empathy. This substantially limits the scope of moral thinking in comparison to ordinary moral beliefs, leaving fictionalists with impoverished moral practices. As a result, fictionalism fails to save error theory from the threat of having radical practical consequences. In the end, the comforting idea that we could coherently respond to moral error theory by simply transforming our moral beliefs into make-belief appears to be a mere fiction.

Acknowledgments

I am grateful to Adam Etinson, Jakob Hinze, Kent Hurtig, Erik Kassenberg, Merel Semeijn, Justin Snedegar, Bart Streumer and participants of the Ethics, Social and Political Philosophy research seminar at Groningen for feedback and discussion of earlier versions of this paper. I also thank two anonymous referees and an editor of this journal for useful comments.

References

Arrhenius, Gustaf (1999). Mutual Advantage Contractarianism and Future Generations. *Theoria*, *65*(1), 25–35. https://doi.org/10.1111/j.1755-2567.1999.tb00112.x

Batson, C. Daniel, Nadia Ahmad, and E. L. Stocks (2011). Four Forms of Prosocial Motivation: Egoism, Altruism, Collectivism, and Principlism. In D. Dunning (Ed.), *Social Motivation* (103–26). Psychology Press.

Batson, C. Daniel, Bruce D. Duncan, Paula M. Ackerman, Terese Buckley, and Kimberly Birch (1981). Is Empathic Emotion a Source of Altruistic Motivation? *Journal of Personality and Social Psychology*, *40*(2), 290–302. https://doi.org/10.1037/0022-3514.40.2.290

Byron, Michael (1998). Satisficing and Optimality. *Ethics*, *109*(1), 67–93. https://doi.org/10.1086/233874

Copp, David (1991). Contractarianism and Moral Skepticism. In Peter Vallentune (Ed.), *Contractarianism and Rational Choice: Essays on David Gauthier's* Morals by Agreement (196–228). Cambridge University Press.

Cuneo, Terence and Sean Christy (2011). The Myth of Moral Fictionalism. In Michael Brady (Ed.), *New Waves in Metaethics* (85–102). Palgrave-Macmillan.

Eriksson, Björn and Jonas Olson (2019). Moral Practice after Error Theory: Negotiationism. In Richard Garner and Richard Joyce (Eds.), *The End of Morality: Taking Moral Abolitionism Seriously* (113–30). Routledge.

Garner, Richard (2010). Abolishing Morality. In Richard Joyce and Simon Kirchin (Eds.), *A World Without Values: Essays on John Mackie's Moral Error Theory* (217–33). Springer. https://doi.org/10.1007/978-90-481-3339-0_13

Garner, Richard and Richard Joyce (Eds.) (2019). *The End of Morality: Taking Moral Abolitionism Seriously*. Routledge.

Gauthier, David (1986). *Morals by Agreement*. Clarendon Press.

Gauthier, David (1991). Why Contractarianism? In Peter Vallentyne (Ed.), *Contractarianism and Rational Choice: Essays on David Gauthier's* Morals by Agreement (15–30). Cambridge University Press.

Hampton, Jean (1991). Two Faces of Contractarian Thought. In Peter Vallentyne (Ed.), *Contractarianism and Rational Choice: Essays on David Gauthier's* Morals by Agreement (31–55). Cambridge University Press.

Hoffman, Martin L. (2000). *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge University Press. https://doi.org/10.1017/CBO9780511805851

Jaquet, François (2021). Utilitarianism for the Error Theorist. *The Journal of Ethics*, *25*(1), 39–55. https://doi.org/10.1007/s10892-020-09339-x

Joyce, Richard (2001). *The Myth of Morality*. Cambridge University Press.

Joyce, Richard (2019). Moral Fictionalism: How to Have Your Cake and Eat It Too. In Richard Garner and Richard Joyce (Eds.), *The End of Morality: Taking Moral Abolitionism Seriously* (150–65). Routledge.

Krebs, Dennis (1975). Empathy and Altruism. *Journal of Personality and Social Psychology*, *32*(6), 1134–46. https://doi.org/10.1037/0022-3514.32.6.1134

Lutz, Matt (2014). The "Now What" Problem for Error Theory. *Philosophical Studies*, *171*(2), 351–71. https://doi.org/10.1007/s11098-013-0275-7

Nietzsche, Friedrich (1989). *Beyond Good & Evil: Prelude to a Philosophy of the Future* (Walter Kaufmann, Trans.). Vintage Books.

Nolan, Daniel, Greg Restall, and Caroline West (2005). Moral Fictionalism versus the Rest. *Australasian Journal of Philosophy*, *83*(3), 307–30. https://doi.org/10.1080/00048400500191917

Nussbaum, Martha C. (2007). *Frontiers of Justice: Disability, Nationality, Species Membership*. The Tanner Lectures on Human Values. Belknap Press.

Olson, Jonas (2011). Getting Real about Moral Fictionalism. In Russ Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol. 6, 181–204). Oxford University Press.

Olson, Jonas (2014). *Moral Error Theory: History, Critique, Defence*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198701934.001.0001

Parfit, Derek (2011). *On What Matters: Volume One*. Oxford University Press.

Prinz, Jesse (2011). Against Empathy. *The Southern Journal of Philosophy*, *49*, 214–33. https://doi.org/10.1111/j.2041-6962.2011.00069.x

Sayre-McCord, Geoffrey (1991). Deception and Reasons to Be Moral. In Peter Vallentyne (Ed.), *Contractarianism and Rational Choice: Essays on David Gauthier's* Morals by Agreement (181–95). Cambridge University Press.

Sinnott-Armstrong, Walter (2006). *Moral Skepticisms*. Oxford University Press.

Smith, Holly (1991). Deriving Morality from Rationality. In Peter Vallentyne (Ed.), *Contractarianism and Rational Choice: Essays on David Gauthier's* Morals by Agreement (229–53). Cambridge University Press.

Streumer, Bart (2017). *Unbelievable Errors: An Error Theory about All Normative Judgements*. Oxford University Press.