# *Yes to moral fictionalism; no to religious fictionalism*

**Richard Joyce**

Penultimate draft of paper appearing in R. Joyce & S. Brock (eds.),
*Moral Fictionalism and Religious Fictionalism* (OUP, 2023) 256-76.

## 1. Introduction

Atheism and moral error theory are types of skepticism. The former is familiar and reasonably popular (in certain demographics, at least); the latter is less well known outside academic circles and, even within those circles, not a terribly widespread view. The atheist denies the existence of such things as gods, cosmic karma, the afterlife, prophesy, etc., and thus holds that many central claims of religious discourse are simply false. The moral error theorist denies the existence of such properties as moral wrongness, praiseworthiness, moral permissibility, evil, moral rights, etc., and thus holds that many central claims of moral discourse are simply false.

It seems fair to say that many people who embrace religious skepticism remain *appalled* at the prospect of moral skepticism. This strikes me as somewhat surprising, given the plausibility of the claim that most moral concepts were born in a theistic framework, outside of which it's far from clear that they can sensibly survive. An action's being *morally forbidden*, say, might make sense when there is a divine being doing the forbidding, but once that forbidder is removed from the ontological scene, it's not obvious that a secularized notion of *being morally forbidden* remains there for the taking—any more than a secularized notion of *an eternal soul* remains.[1] The last few centuries of moral philosophy can be read largely as an attempt to tether these originally-non-naturalistic moral concepts to the natural world, but the success of this project has been, to put it charitably, disputable. The moral error theorist is in all likelihood an ontological naturalist, generally speaking, but one who thinks that moral normativity has features that render it unsuitable for accommodation within the naturalistic worldview—features like a special kind of *practical authority*, a kind of *autonomy*, and/or a kind of *agency*. "So much the worse for the naturalistic worldview" say some philosophers—but then they face a nagging worry that this relaxation of naturalistic scruples might sit uncomfortably with their grounds for confidently rejecting religious ontology. "So much the worse for moral concepts" say moral error theorists.

Both the atheist and the moral error theorist face what might be called the *"what next?" question*: given that a widespread way of talking and thinking has turned out to be erroneous, what should we do with it? The most obvious answer, in both cases, is the abolitionist's response: we should just do away with the subject, much as we previously did away with talk of mermaids, phlogiston, and bodily humors. Of course, not even the abolitionist claims that we must never *utter* such non-denoting words. Even the religious abolitionist, for example, thinks it's fine to tell a joke about Satan, or assert something like "In the past, people believed in gods." But the religious abolitionist thinks that religious discourse should be relegated to a small and harmless corner of language—it should not play the central role that

[1] See Anscombe 1958; MacIntyre 1984: 2.

it currently occupies in a great many people's lives. The religious fictionalist disagrees and offers a less obvious answer. The religious fictionalist notes that engaging with religious language and thought has been quite useful in various ways—for all its error—and proposes that we could retain some of this usefulness by preserving religious language and thought in our lives. The fictionalist remains opposed to *believing* falsehoods, and so recommends the adoption of a kind of positive attitude toward religion that is not belief: let's call it "nondoxastic acceptance." Regarding religious language, the fictionalist is typically opposed to *asserting* falsehoods, and so recommends the adoption of a kind of positive speech act that is not assertion: let's call it "nonassertoric affirmation."

The moral fictionalist makes the analogous claim about morality. Even though a proposition like "Stealing is morally wrong" is, strictly speaking, false—since there's no such property as moral wrongness—it is a useful kind of falsehood, and thus we should (pragmatically speaking) carry on endorsing it but without believing or asserting it.[2]

An ambitious fictionalist will claim that the benefits of nondoxastic acceptance and nonassertoric affirmation will come close to matching the benefits of sincere belief. A less ambitious fictionalist accepts that the benefits might fall well short of those of sincere belief and assertion, but thinks that the former at least offers *some* benefit—enough to make it a preferable option to abolitionism.

My goal in this chapter is to sketch out a plausible version of ambitious moral fictionalism, but then argue that this sketch does not transfer over to lend plausibility to ambitious religious fictionalism. In other words, someone who is both an atheist and a moral error theorist might have good reason to retain *morality* as a kind of useful fiction, but should be an abolitionist about *religion*. Yes to moral fictionalism; no to religious fictionalism.[3]

## 2. The "what next?" question

There is not really a single "what next?" question; there is an open-ended plurality. Consider the atheist's version of the question as it pertains to religion. For a start, we must decide whether we're asking what should be done with religious *language* or done with religious *thought*. Then we must decide who it is that we imagine asking the question. Are we picturing the question asked by an individual atheist—and, if so, are we imagining them surrounded by theists—and, if so, are these theists inclined to burn non-believers at the stake? A sequence of "yes" answers should make it clear that it would be a good idea, practically speaking, for this individual to maintain a pretense of religion—and not just sporadically, but 24/7. Alternatively, perhaps we're picturing the question being posed by a *group* of atheists. But, if so, then what kind of practical question is the group asking when it inquires what it "should" do? What it should do to maximize preference satisfaction among its members? Or to maximize overall pleasure? Or to ensure that the worst off member couldn't be better off? Or what? There is not a single correct understanding here; rather, there are different

[2] I have in mind only revolutionary fictionalism. Hermeneutic forms of fictionalism—whether moral or religious—are not under consideration in this chapter. See Introduction to this volume for this distinction explained.

[3] Large portions of the first half of this chapter are a condensed version of some chapters of my book *Morality: From Error to Fiction* (forthcoming, OUP).

reasonable ways of taking the question, and the person asking it needs to specify. (This is not to say, though, that the decision is an inherently moral one.[4])

This indeterminacy isn't a problem for the skeptic *per se*; it permeates the way we talk about practical recommendations in general. If I steal your wallet, then presumably I harm your interests—but in what sense? By frustrating your preferences? By making you unhappy? By detracting from a list of your objective goods? Or what? It would be unreasonable to maintain that all advice-giving must be held in abeyance until we settle these thorny questions. Presumably, my stealing your wallet harms you on any reasonable account. But other cases may be less clear cut: an action may harm you in one legitimate sense of "harm" and not harm you in another legitimate sense of "harm," and that's all there is to it. Someone who wonders whether such an action harms you has to specify which sense they mean.

If we are talking about an individual atheist, surrounded by non-atheists, then it isn't hard to imagine circumstances in which this person should definitely "play along." These circumstances needn't be so dramatic as involving stakes and burning; they may simply involve social distrust and being labeled an oddball. Let's imagine an atheist in such a situation; we'll call him *Brad*. Brad "plays along" with religion at the level of language and social practices, and we'll assume that he has good pragmatic reasons for doing so—he stands to gain the benefits of participation—or, at least, avoid the harms of non-participation. We'll assume, though, that Brad's "playing along" does not extend to his own mental life: he doesn't engage in anything deserving the name "nondoxastic acceptance"; rather, his mental attitude toward religious claims is fairly straightforward disbelief. It's reasonable to suppose, though, that Brad's securing of participatory benefits might depend on his staying quiet about his disbelief, so although he doesn't believe in God, he allows everyone else to believe that he *does* believe in God. Thus when Brad makes religious utterances he does not do so with a nudge or a wink, nor with any accompanying explanation, and therefore his audience will, naturally, take him to be making assertions. It is important to note that a speaker cannot determine unilaterally which speech acts they perform: from the fact that Brad doesn't believe that *p* it doesn't follow that, when he says "*p*," he is merely "quasi-asserting' or "pretending to assert"—this depends in part on how his audience takes his speech. Thus if all of Brad's audience take him to be asserting religious claims, then he *is* asserting them (whether he intends to or not)—his utterances are in no sense a form of nonassertoric affirmation—in which case Brad's *language* (in contrast to his mental states) remains ontologically committed to gods, etc. It is clear, then, that Brad's "playing along" with religion for pragmatic purposes is a long way from being an exemplar of what the fictionalist has in mind. The problem is not that the deception of Brad's interactions is distasteful or dishonest; it is, rather, that he practices neither nondoxastic acceptance nor nonassertoric affirmation, and, moreover, his language remains ontologically committed to the existence of problematic entities.

Ideally, the fictionalist recommendation would cover both language and thought, recommending for each a kind of assent *sans* ontological commitment. We have already seen, however, that the former is problematic if fictionalism is advice directed at an individual, since which type of speech act is performed is not up to the speaker alone—it depends in part

[4] *Pace* Lenman, this volume.

on how the audience takes the speech. If, however, the religious fictionalist recommendation is aimed at a *group* of atheists, rather than an individual, then the suggestion can be that the group should develop conventions such that, when the topic of religion is entered into, everyone knows that assertoric force, and thus ontological commitment, is lifted. (Compare, say, our conventions surrounding sarcasm—another commitment-nullifying linguistic device.)

In this chapter I will focus on the fictionalist recommendation more as it pertains to thought than language; I am, in other words, more interested here in nondoxastic acceptance than nonassertoric affirmation. Accepting something *in one's mind* suggests that the benefits of acceptance must come from within (so to speak), not just from the reactions of others—and this, I think, is the more challenging kind of case to attempt to establish. The crucial questions are, first, what kind of benefits might these be, and, second, can these benefits be gained from "mere acceptance" rather than belief? In what follows I shall offer answers to these questions for the case of *moral* fictionalism; we will return to religious fictionalism later.

## 3. The usefulness of moral beliefs

The dispute between the moral fictionalist and the abolitionist concerns whether morality is, on balance, useful. The fictionalist says yes; the abolitionist says no. But the question of the usefulness of morality is not confined to a squabble among the error theorists. We could also ask a moral *realist* whether morality is useful. Most will probably say yes, but we can certainly imagine a moral realist claiming that the practice of constantly referring to moral facts in our deliberations is, on balance, harmful and that it would be better if we stopped. So *abolitionist moral realism* is a possible (albeit sparsely populated) category (see Ingram 2015). In any case, anyone who thinks that the question "Is morality useful?" is *coherent*—even if they complain that the question requires some disambiguation before it can be addressed—is allowing two kinds of normativity into the picture: the normativity of morality and the normativity of usefulness.[5]

The error theorist thinks that the former kind of normativity is bankrupt—there's no such thing. John Mackie (1977) famously argues that moral normativity requires a kind of *objectivity*, whereas the only kinds of normativity that actually exist are "subjective." According to Mackie, we "objectify" our preferences and values: we see our moral assessments as responding (or, at least, as attempting to respond) to moral demands that are already there, in the nature of things, whereas in fact there are no demands that are "there" independently of our valuing activity. We think that there are things that we must do, or must not do, regardless of our desires, and that these rules of conduct are not simply made up by humans. According to Mackie, then, a view that might be called "Kantian realism" captures

[5] This isn't strictly true. One could have some kind of *moral* understanding of "usefulness," in which case the question would concern whether morality is self-validating (which, I guess, it is). I suggest, however, that most parties who find the question "Is morality useful?" coherent (even if indeterminate) do not have in mind the question of self-validation.

the *content* of our moral claims. But this view does not capture the *reality* of what's really going on when we make moral judgments. Hence the error at the heart of morality.

There is no obvious reason, however, for the error theorist to think that the second kind of normativity—that which pertains to *usefulness*—is also bankrupt. What makes something useful is *us*: something is useful, one might claim, if it satisfies our desires or preferences. And we know this; there is no tendency to objectify the values of usefulness, or to endorse a Kantian realist interpretation of the content of utility judgments. Rather, a view that might be called "Humean subjectivism" captures both the content of our utility judgments and the reality of what is going on when we make these judgments. Hence there is no error at the heart of what makes things useful—including the matter of what makes morality useful.

(A few quick qualifications. First, I am using the terms "Kantian" and "Humean" more as gestures of convenience than as scholarly claims about what these historical philosophers really argued for. Second, Mackie's is just one way of arguing for moral error theory; the truth of error theory does not live or die with the soundness of his argument—any more than the plausibility of atheism depends on, say, the success of the argument from evil. Third, I have no space on this occasion to attempt to support the premises of this view. It is outlined and assumed in order to devote energies to other discussions.)

What has been the practical benefit of having such a system of Kantian realist norms? Why do we have this tendency to objectify certain preferences and values?

Daniel Dennett (1995) argues that one benefit of having moral considerations in our conceptual repertoire is that they can serve as *conversation-stoppers*: their value is to bring deliberations to a close. We are rational creatures, always able to ask for justification, and in many contexts this is a trait that has served us well. The problem is that upon receiving a perfectly good answer we can always sensibly respond "Okay, but what justifies *that*?"—and we can potentially do so *ad infinitum*, never coming to a decision, forever hesitant and doubting, undone by our own rational prowess. This is potentially as much a problem for our own private deliberations as for our public interpersonal ones. Dennett suggests that it is useful to have "consideration-generator-squelchers" (1995: 506): items that, once introduced, stop any further deliberation in its tracks. "That would be morally wrong!" would appear to work in this manner. Once this claim is accepted then there is no need or room for further consideration: the action mustn't be done, even if it is tempting, and that's all there is to it.

In saying this I am, of necessity, considering matters at a highly general level. The claim certainly isn't that using a moral judgment as a conversation-stopper is always going to be useful. If the content of the judgment were that it is morally obligatory for me to eat broken glass (regardless of whether I want to)—and if this functioned as an emphatic end-point to my deliberations, blocking me from even raising the natural objection "But why would I want to do that?"—then we could safely file this under the heading of "when good conversation-stoppers go bad."

It is worth noting that if Dennett is correct then the benefits of moral judgment need not be seen in terms of its contribution to cooperation or to social cohesion or to the solution of coordination problems—all of which are commonly assumed to be the basic "point" of human morality. Even Robinson Crusoe, alone on his island, might conceivably benefit from internal conversation-stoppers—if he is prone to over-thinking and second-guessing his daily plans, for example. That said, it is clear that conversation-stoppers are also terribly useful in

public interpersonal relations. Debates and discussions about social policy and the justification of interpersonal interactions are just as likely to spiral endlessly unless we have a shared bedrock of unquestioned values: "We won't do that because *it's morally wrong*" has a finality to it. By contrast, "We won't do that because *we don't want to*" remains open to negotiation: desires can be interrogated; they can be bargained with; they can be altered. For this reason, values with a Kantian flavor stop conversations more effectively than those with a Humean flavor. Values that are treated as objective can stop conversations more effectively than those that are taken to depend on some agent's or agents' attitudes. The latter are likely simply to throw up further calls for justification ("But why should we care about that agent's attitudes?"), while the former enjoy the realist's table-thumping conclusiveness: "That's just the way the world is!"

Conversation-stoppers that gain general currency within a group can be predicted to serve the group's collective ends. Sometimes values might do this *directly*. It should be no surprise, for instance, to find moral values requiring individuals to restrain the unbridled pursuit of their own personal gratification. But sometimes values might accomplish this *indirectly*. The value in question may, on the face of it, have little to do with cooperative projects (it might concern, e.g., dietary prohibitions), but nevertheless the fact that members of the group all share and commit to this value can contribute to social cohesion. Having a range of values that are imbued with this kind of no-questions-asked finality allows one to signal to others where one draws the line between, on the one hand, those practical matters that might be up for negotiation and for which diversity will be tolerated, and, on the other hand, those matters for which any concession or softening is emphatically off the table. The willingness of an individual to draw this line in the same place as their fellows can be a powerful indicator of group solidarity.

In many contexts it will remain in an individual's interests to cultivate a no-questions-asked approach to these prosocial values and rules. If I am constantly publicly questioning why I should follow the rules, what justifies the rules, whether I can get away with breaking the rules, etc., then this will likely go against my own practical purposes by creating mistrust and the possibility of ostracization. If various important goods are available to me only if I indicate that I embrace a range of prosocial values, then it will be in my interests to loudly advertise that I have those values—and, indeed, to *have* those values. As before, taking a Kantian and realist attitude toward those values is likely to serve me better than taking a Humean and non-realist one. If my fellows are seeking a companion who doesn't break promises (say), and they are faced with a choice between someone who values promise-keeping *come what may*, and someone who values promise-keeping because she believes that keeping promises contributes to the satisfaction of her desires, then the former looks by far the better bet. My fellows want a companion who will, upon realizing that a prospective course of action requires the breaking of a promise, immediately reject that action with no further queries about justification and no weighing of the strengths of competing desires. Richard Whately hit this nail on its head when he declared: "Honesty is the best policy; but he who acts on that principle is not an honest man" (1856: 106).

A background assumption of the hypothesis outlined in this section has been the idea that, in order to furnish such benefits, moral judgments must be *beliefs*. If one is to stop a conversation or a deliberative process with the thought "No, that would be morally wrong!",

the natural assumption is that in order to play this role effectively the thought must be a *belief*. However, in order for us to accept this hypothesis about the kinds of benefits that moral judgments bring, there's no need for us to assume that the beliefs are *true*. The hypothesis just outlined is one that the moral error theorist should have no problem endorsing.

The next question to ask is whether these practical benefits could still be secured if the attitude were not a belief but rather an act of nondoxastic acceptance. I will come at this question obliquely, which (as we shall see) is often the best route.

## 4. The suspension of disbelief

Samuel Taylor Coleridge introduced the phrase "the suspension of disbelief" (1817: 2) to describe how the enjoyment of a romantic or Gothic work will require the reader to suppress the urge to respond skeptically to the supernatural elements of the narrative. One will not enjoy *The Rime of the Ancient Mariner* if constantly thinking "Well, *that* couldn't happen!" Just as the appreciation of the work might require one to silence one's *dis*belief, so too it requires the stifling of one's beliefs. One will not enjoy any movie at the cinema if constantly thinking "I am surrounded by strangers in a large dark room." Crucially for our purposes, the "*suspension*" of belief/disbelief must not be confused with the *cessation* of belief/disbelief. One never ceases to believe that one is in a cinema, but one ceases to think about this fact; the belief goes on the back burner once the movie starts.

The suspension of belief that characterizes a person's reading a poem or watching a movie is short-lived. Can there be a more "all-encompassing" suspension of belief that someone might implement in many or most everyday situations? We don't need to search anywhere terribly unfamiliar in order to locate such a case. Consider what J. S. Mill said (in his autobiography) about the pursuit of happiness:

> Those only are happy (I thought) who have their minds fixed on some object other than their own happiness; on the happiness of others, on the improvement of mankind, even on some art or pursuit, followed not as a means, but as itself an ideal end. Aiming thus at something else, they find happiness by the way. The enjoyments of life (such was now my theory) are sufficient to make it a pleasant thing, when they are taken *en passant*, without being made a principal object. … The only chance is to treat, not happiness, but some end external to it, as the purpose of life. Let your self-consciousness, your scrutiny, your self-interrogation, exhaust themselves on that; and if otherwise fortunately circumstanced you will inhale happiness with the air you breathe, without dwelling on it or thinking about it, without either forestalling it in imagination, or putting it to flight by fatal questioning. ([1873] 1924: 100)

Henry Sidgwick later called this a "paradox"—though it's really more an ironic twist of human psychology than it is a paradox. According to Sidgwick, certain pleasures are such that "in order to get them, one must forget them" (1907: 51). And this seems a fairly widespread phenomenon: we often achieve things best by aiming at them not directly, but obliquely. Someone who asks how to get the most out of a personal relationship, for example, might receive the excellent advice "Don't constantly calculate how much you are getting from the relationship." Someone who asks how to come across to others as cool might receive the excellent advice "Don't try so hard to be cool." Later we will discuss the version

of this called "the paradox of self-interest" (also not really a paradox), but for now let's stick with the happiness version.[6]

Imagine that Janet strives for happiness but finds it elusive. If she seeks advice from Mill, then the recommendation will be that she should forget about the fact that happiness is her ultimate goal—even though it *is* her ultimate goal—and instead she should focus her attention on achieving other ends: loving relationships, publishing scholarly articles, stamp-collecting, whatever. All along, Janet's ultimate goal remains her own happiness, and at no point does she cease to believe this. But she puts this belief on the back burner most of the time, since experience has taught her that if it is in the forefront of her mind then it simply gets in the way of her achieving happiness. When absorbed in her favorite hobby of stamp-collecting, for example, Janet thinks of the activity as valuable for its own sake; she does not constantly contemplate or calculate the contribution that her hobby makes to her happiness. Doing so takes all the fun out of the activity, and (ironically) makes her less happy.

This is not to say, though, that Janet cannot be brought to admit explicitly that happiness is her ultimate goal and that the only kind of value that stamp-collecting has is instrumental. When in a reflective mood—when thinking overtly about the paradox of happiness and the advice that Mill offers her—Janet frankly acknowledges exactly what's going on: namely, that in order for her to achieve her ultimate goal, most of her everyday life is an exercise in distracting her mind from what that goal really is. Indeed, it is the fact that she is prepared to admit this when she sits down in a cool hour that indicates that she believes it all along and has simply "suspended" that belief rather than ceased to believe it. All along, she *has* the disposition to acknowledge the facts, but most of the time she is unaware that she has this disposition, and even actively encourages this lack of awareness in herself in everyday situations. She may well even be reluctant to enter into this "critical mode" too frequently, since doing so dampens the effectiveness of the distraction, thus interfering with her pursuit of happiness. Still, it's not as if once she learns of the paradox of happiness then all hope of her attaining happiness is lost. Mill did not tell us about the paradox of happiness with the cruel expectation that doing so would crush our capacity to achieve happiness!

Moreover, it's not as if in moments of transparency Janet will lose all motivation to continue stamp-collecting. Even when she becomes aware that stamp-collecting lacks the inherent value with which she usually imbues it, she can still acknowledge its instrumental value for her; it's just that things go better for her (and she knows it) if she's not constantly in this state of transparency. Janet can, effectively, get the cat back into the bag by focusing her attention in the right way. Doing so might not even take much effort on her part—it may be habituated and natural, requiring nothing more complicated than perusing through her stamp albums. (Hume mentions playing backgammon as a means of driving away skeptical ruminations.)

Is Janet engaged in *make-believe*? Possibly, but describing her attitude in this manner might be misleading. If asked independently to think of episodes of make-believe, one's mind

---

[6] Under some understandings of what *happiness* is, the paradox of happiness and the paradox of self-interest might turn out to be the same thing. For a useful catalog of many different "paradoxes of happiness" that are often lumped together under that single heading, see Martin (2008). Ethical theories that have this quality (and many of them, arguably, do) are usually referred to as "self-effacing."

is likely to alight on paradigms where a person can easily and voluntarily slip in and out of the pretense, and is perfectly aware the whole time that they are engaged in make-believe. But these are not necessarily features of Janet's attitude toward the value of stamp-collecting, so I prefer to stick with the label "nondoxastic acceptance."

Is Janet *self-deceived*? She is not self-deceived in the manner of someone who sticks with a falsehood come what may—who is unable to give it up. Such a person would *not* have the disposition to admit the truth if asked in all seriousness, which indicates that this person would *believe* the falsehood. If Janet's attitude does still count as a form of "self-deception," then (if Mill is to be believed) it's a kind that we should all strive to cultivate—something on which nothing less than human happiness depends. I suggest that it is more fruitful to think of Mill (and Coleridge) as counseling "self-distraction" than self-deception.

This notion of "self-distraction" lies at the heart of the kind of moral fictionalism that I find plausible. According to this view, Humean values exist and Kantian values do not, and a smart person knows this. But a smart person also realizes that something similar in its irony to the paradox of happiness is in play: that the practice of deliberating explicitly in terms of Humean values tends to be self-defeating, whereas deliberating in terms of realist Kantian values can further important personal and social goals (goals, that is, which are understood in Humean terms). The moral fictionalist, thus, prescribes a course of self-distraction and suspension of disbelief: in your day-to-day activities, forget about the Humean truth; let your thoughts, speech, and actions be guided by Kantian normativity. In other words, when it comes to Humean values, "in order to get them, one must forget them."

This is not to say that in moments of transparency you would lose all motivation to act in accordance with your moral fiction. Even when you enter "critical mode" and realize that, say, breaking promises is not really morally wrong, you are still in a position to acknowledge that promise-breaking will likely frustrate your Humean values. (After all, if this weren't true—if your moral fiction were urging you to do something that goes against your Humean values—then clearly that is not a useful moral system: you have embraced the wrong moral fiction.) It's just that things go better for you, and you should know it, if you're not constantly in this state of transparency.

So long as you retain the disposition to sincerely deny the Kantian foundations if asked about them in all seriousness, then you do not really believe in them, and we can classify your attitude toward morality as one of "nondoxastic acceptance." If you follow this advice studiously, then in your day-to-day life you won't even be particularly aware that your attitude toward morality is not one of sincere belief—indeed, you won't be conscious that you are "following advice studiously" at all. In everyday contexts, your attitude toward morality will have the phenomenology of belief—all the emotional, motivational, and practical advantages of moral belief—without being moral belief.

## 5. A fictionalized morality is *better* than a believed morality

Ambitious moral fictionalism attempts to show that the practical benefits of taking a nondoxastic attitude toward morality are not much different from those of having sincere moral beliefs. An even more ambitious project is to attempt to show that the benefits of the

former are *greater* than those of the latter. In order to see that the latter is plausible, let's return to Janet and the paradox of happiness.

Janet seeks happiness and has taken on board Mill's advice that the best way (perhaps the only way) to achieve it is obliquely, by seeking something else. For Janet I've imagined that this activity is stamp-collecting. Janet does not really *believe* that stamp-collecting is a worthwhile activity in itself, but she thinks and acts and talks as if it is—she nondoxastically accepts that stamp-collecting has a kind of value that it does not have (and that she knows it does not have). Suppose, though, that over time the pleasures of stamp-collecting begin to fade for Janet. She is able to step back and acknowledge that stamp-collecting is not inherently valuable, that it never really was (even though it was useful for her to think that it was), and that the only kind of value that stamp-collecting ever had was as an instrument to her happiness. Nondoxastic acceptance, unlike belief, does not have to wait upon evidence in order for rejection to be rational. When, therefore, Janet realizes that her hobby no longer makes her happy, there is nothing to stand in the way of her simply stopping this activity; she can instead throw herself into a new hobby that brings her renewed happiness. If, by contrast, Janet had sincerely *believed* stamp-collecting to be inherently valuable, then she might have felt the need to persist with the activity, albeit stoically and unhappily.

The moral fictionalist is in a comparable position. It may well be useful to have moral values as conversation-stoppers—both personally and publicly—but we don't want to be so committed to moral values that we refuse to recognize any context wherein they can be critically examined. Our situation may change over time: what was once a useful conversation-stopper might become a destructive conversation-stopper. The individual or group who can periodically reflect on how well their conversation-stoppers are serving their ends will (*ceteris paribus*) do better than an individual or group who cannot. If one of the benefits of moral judgment is that it allows an individual to signal their normative commitments to others in the community as a way of marking group solidarity, then if the community's values shift (for whatever reason), then the individual's commitments had better shift with them if they are going to continue to play this role. That's not to say that the individual was never really *committed* in the first place. There is a great deal of space between not being committed in the slightest to something and being willing to die for it.

The benefits of morality may require a kind of steadfast and inflexible commitment, but a moral system that is *utterly* steadfast and inflexible—that will crash and burn rather than adapt—is very probably less beneficial overall than a moral system that is able, *in extremis*, to adjust and evolve. This holds whether considering matters at the level of the group or the individual. It is here, then, that an attitude of nondoxastic acceptance toward morality may serve us better than sincere moral belief.

## 6. Kant's moral hazard argument

It is time to return our attention to religious fictionalism. Might it also be that an attitude of nondoxastic acceptance of religion is better than belief? My point of departure will be Kant's argument for a positive answer to this question (with my views on this matter being heavily influenced by Christopher Jay's 2014 article).

Kant certainly seems to be a religious fictionalist. In the *Critique of Pure Reason*, he argues that we have good reason for postulating God and an immortal soul—but these reasons are practical rather than epistemic. It is reasonable (Kant thinks) for us to hope that our actions will be appropriately rewarded or punished, and in order for this hope to be rational we must allow the possibility of its being satisfied. Kant rejects the proposal that there is an analytic connection between being moral and the attainment of happiness (a view he associates with Stoicism and Epicureanism), and so concludes that this connection could only be synthetic.[7] But since we pretty clearly cannot rely on the natural world to satisfy the expectation that acting morally will be rewarded with happiness, we must postulate an eternal afterlife and a "wise Author" who can act as guarantor of the connection. "Such a Ruler, together with life in such a world [i.e., an afterlife], which we must regard as a future world, reason finds itself constrained to assume; otherwise it would have to regard the moral laws as empty figments of the brain."[8]

Kant's "moral hazard argument" is that taking an attitude of (what Jay calls) "nondoxastic acceptance" toward the existence of God is actually *better* than belief, since belief would encourage the wrong kind of motivation for our actions. If we really believed that virtue will be rewarded in heaven, then this would risk becoming our motivation for acting, in which case we'd become like Kant's shopkeeper who's honest with customers only because it's profitable to be so: acting in accordance with duty but not acting *from* duty. In other words, though it's fine and necessary to expect that good actions will be rewarded, in order for those actions to *be* good (i.e., morally praiseworthy) they'd better not be performed for the sake of that reward. According to Jay's interpretation of Kant, this risk is absent (or at least reduced) if one's attitude is one of nondoxastic acceptance rather than belief.

I don't propose to go any further into the intricacies of Kant's views than this, and it is certainly not an argument I intend to advocate. I want to point out one plausible thing about the argument—which will be the basis of further discussion—and one implausible thing which, I think, sinks the moral hazard argument.

The component of the argument that has an intuitive ring is Kant's concern about the tension between the undesirability of selfish motivations and the belief in postmortem rewards. If we consider someone deeply religious and also morally admired—someone like Mother Teresa, perhaps—although we presume both (1) that she believes that she is acting in a morally right way, and (2) that she believes that acting in a morally right way will receive some kind of divine reward, we would surely be troubled to also discover (3) that she is *motivated by the anticipation* of that reward. We would think less of her, morally, if we knew that she was always dreaming of her own future eternal bliss as she ministered to the needs of the poor.

This is reminiscent of the paradox of happiness. A religious person might much prefer the prospect of receiving posthumous rewards over posthumous punishments, but in order to secure those rewards they must take an oblique approach: they must cultivate concerns for, and interests in, other things (e.g., genuine sympathy for others' suffering). What is

[7] This rejection is found in the *Critique of Practical Reason* (5.111 ff.).

[8] *Critique of Pure Reason* A811 (trans. Kemp Smith). A moral error theorist, reading this passage, is likely to be reminded of the adage that one person's *ponens* is another's *tollens*.

surprising, however, is that this intuitively acceptable component of Kant's argument speaks in favor of a kind of fictionalism that is directly opposed to that which he actually supports. Simplifying things dramatically: the Kantian religious fictionalist might offer the recommendation "Nondoxastically accept that the afterlife exists, but do not really believe it"; but the viewpoint we have just been discussing—the advice for Mother Teresa et al.—is "Believe that the afterlife exists, but distract yourself from that belief in everyday contexts—that is, nondoxastically accept that the afterlife does *not* exist." (More on this in a moment.)

The less intuitively acceptable component of Kant's argument is the assumption that taking a nondoxastic attitude toward God and an afterlife will make one immune to the corrupting effect upon motivation that sincere belief in God and the afterlife would (allegedly) have. This is an instance of a generic challenge for fictionalism—whether moral or religious. The fictionalist might hope to recoup the *benefits* of belief by recommending a nondoxastic attitude that "feels" a lot like belief, but this attitude is likely also to bring the *costs* of belief. Finding an attitude that allows one to cherry-pick the benefits while avoiding the costs is tricky. And so it is with Kant's moral hazard argument. If *believing* in the promise of posthumous rewards risks corrupting one's motivation by fostering selfishness, then *nondoxastically accepting* the promise of posthumous rewards is likely to have exactly the same corrupting influence—that is, so long as this attitude of acceptance has enough of an impact on one's other psychological states to supply the touted practical benefits. There are, then, serious doubts to be raised about the capacity of the Kantian "moral hazard argument" to show that an attitude of nondoxastic acceptance of the existence of God and an afterlife is better than one of sincere belief.[9]

## 7. Postmortem rewards and the paradox of self-interest

The discussion of Kant has, however, succeeded in bringing into the light a general complication for religious belief systems, centered on the issue of selfish versus altruistic motivation. Most major religions endorse values or requirements that are opposed to flat-out selfishness. All the major religions, for example, prize love, kindness, and generosity—and not just in the realm of *action*, but motivation. (The first hadith of Islam, for instance, could hardly be more explicit on this last point: "Actions are judged according to intentions.") But, at the same time, all major moralizing religions include an elaborate ontology of postmortem rewards and punishments.[10] So before the possibility of atheism even raises its skeptical head, there is already something psychologically complicated going on with most religious belief systems. "If you're good you'll get into paradise, but if your motivation is to get into paradise then you're not good!" There is a straightforward solution to this complication, but, I will argue, it is not a solution for the religious fictionalist.

The challenge here is an instance of the paradox of self-interest—a close cousin to the paradox of happiness. Robert Frank presents the former as "a simple paradox, namely, that in many situations the conscious pursuit of self-interest is incompatible with its attainment" (1988: ix). The key word here is "conscious." How do you advance your long-term best

[9] See also Leng, this volume.

[10] See Johnson & Krüger 2004; Bernstein & Katz 2010; Baumard & Boyer 2013.

interests? Answer: By not constantly deliberating about those interests as you act; by not always taking those ends as your motivating reasons. Similarly: How can you get into heaven? Answer: By not thinking about heavenly rewards as you minister to the needs of the poor; by being genuinely loving, kind, and generous. There is no incoherence to any of this, just sensible advice for complicated creatures. One can see such advice to practice "self-distraction" as a kind of fictionalist recommendation; ultimately what you are being counseled to do is imbue things with a kind of value that they do not have, and that you know they do not have. (Think back to Janet and her stamp-collecting.) The possibility of following such advice is what allows one to navigate the obvious tension inherent in those religions that promise glorious rewards (or threaten terrible punishments) while also decrying ubiquitous selfishness.

It's important to note here that the undesirability of egoism need not come from values of the religion itself. Even if the religion is silent on what our motivations should be, one might independently deem it morally undesirable to be always motivated by thoughts of self-gain. Indeed, the undesirability of egoism need not be prompted by *moral* considerations at all. Even a moral error theorist, if asked a straightforward practical question about, say, what will generally best advance a person's interests, will likely give an answer that encourages one to cultivate genuine friendships, to sympathize with others, to fall in love, to be generous with no thought of compensation, and so on. Basically, if someone is wondering how to advance their own interests, then even the moral error theorist should recognize that any response that includes "always keep your own interests in the forefront of your mind" is almost certainly *terrible* advice; it amounts to counseling the cultivation of sociopathy, which is not a good recipe for happy humans.

Once we distinguish between what ends a person is pursuing, on the one hand, and what ends the person is *thinking about* while engaged in the pursuit, on the other, then the solutions to a lot of these seemingly-tense situations fall into place. We should seek happiness, sure, but we should not *think about* the pursuit of happiness while acting—doing so actually makes us less happy. We should seek our Humean ends, sure, but we should not *think about* those ends while acting—doing so actually prevents our attaining those ends. We should seek heavenly rewards, sure, but we should not *think about* those rewards while acting—doing so makes us unworthy to receive those rewards.

The point to which I'm keen to draw attention is that the last solution, pertaining to the attitude one might take toward a religion's promise of postmortem reward, is a solution for those who *believe* in those rewards; it's sensible advice for a theist. If, however, we picture a religious *fictionalist* trying to follow this advice, then things get weird.

Consider an atheist, Brad, who resolves to embrace a certain religion with a kind of nondoxastic acceptance: he's going to accept that God exists, that God created all life, that the virtuous are rewarded in the afterlife, etc.—all the while believing none of these things. Remember that we're assuming that Brad's "acceptance" is not just a matter of his playing along with religious language and practices (attending church, etc.); we're assuming that he benefits from adopting a positive psychological attitude toward religion even in private. We might describe this as Brad's cultivating a habit of holding religious thoughts in his mind, of being guided by them, of allowing them a role in his deliberations—all the while remaining disposed to deny these religious propositions if probed in a sufficiently critical manner.

Now let's introduce into the picture another piece of advice for Brad: that it goes against his interests to be *motivated* by thoughts of self-gain in many interpersonal contexts. This is not to deny that self-gain should be Brad's ultimate end; it's a claim about what kind of thoughts and concerns he should have in mind when deliberating and acting, if he is to pursue that end successfully.

Brad, then, appears to be subject to competing pieces of advice about what thoughts he should attend to and which he should suspend when acting in everyday contexts: that he should hold religious thoughts in mind—including, presumably, thoughts of the postmortem rewards—and that he should at the same time endeavor to banish thoughts of postmortem rewards from his mind. Combining these pieces of advice, we seem to have the recommendation that Brad should keep in his thoughts some thoughts that he should keep out of his thoughts. This does not sound like a coherent recommendation.

One might respond that this is simply a case of a fiction within a fiction, and there's nothing incoherent about *that*. The play *Hamlet*, for example, contains the play *The Murder of Gonzago*. But upon consideration the comparison is not an apt one.

Let's consider a contemporary actor whose part (in a performance of *Hamlet*) is to be one of the players whose part (in *The Murder of Gonzago*) is to be the King. The actor pretends to be someone who is pretending to be someone. More specifically, the actor pretends to be not a contemporary actor but rather a Medieval player; and the player (the character in *Hamlet*) pretends to be not a player but rather a King (and the King doesn't pretend to be anyone; he just dies). At first glance, one might think that the actor is expected to both pretend to be a player and pretend not to be a player. But if we keep our eye on the ball then we can see that this is mistaken. There *are* two acts of pretense here—one nested in the other—but only one of them is real. The actor pretends to be someone who is pretending something, but the second pretense is no more a real act of pretense for the actor, psychologically speaking, than the pretended death is a real death.

An actor can certainly pretend to be someone who believes in postmortem rewards and who is distracting themselves from their belief. But the actor would not thereby be *actually* distracting themselves from thoughts of postmortem rewards, only pretending to. However, the practical advice that one should not be always motivated by thoughts of self-gain is that one should *really* not be always motivated by thoughts of self-gain, not that one should pretend to not be always motivated by thoughts of self-gain. We would be no less disappointed in our discovery that Mother Teresa was always gleefully anticipating heavenly rewards if we learned that she was convincingly pretending otherwise.

## 8. Religious fictionalism versus moral fictionalism

I have raised an awkward problem for religious fictionalism. Moral fictionalism does not face this problem.

The obvious escape route for the religious fictionalist is to protest that I have focused far too much on postmortem rewards and punishments—a focus that, it might be complained, reflects an old-fashioned view of religion. The contemporary religious fictionalist might prefer to recommend the nondoxastic acceptance of a much more contemporary vision of religion—one that's indiscriminately loving and forgiving, for example, and where nobody's

eternal soul is treated differently on the basis of their worldly behavior. After all, if (1) nondoxastic acceptance of religion is recommended on pragmatic grounds, and (2) the religion that one has embraced involves postmortem rewards and punishments, and (3) reflecting on these rewards and punishments encourages selfishness in one's attitudes, but (4) selfish attitudes are ultimately a poor strategy, pragmatically speaking—then the religious fictionalist can simply retort that this shows that one has chosen the wrong religion. If you'd be better off nondoxastically accepting a religion that does not include postmortem rewards and punishments, then that's the religion for you!

While I would not argue that having a system of postmortem rewards and punishments is a necessary feature of religion, I'm doubtful that it's a feature that is as easily separated from religion as one might think. A system of rewards and punishments that involves beatific angels among the clouds and gruesome Boschian hellscapes is one that, I'm sure, many modern theists will dismiss as florid excess; but, nevertheless, the general idea that there is a divine being who *cares* about how we act, and that this will somehow be differentially reflected in one's ultimate fate, may be, though vaguer and less tangible, a fairly important component of many theists' worldview. This is an empirical claim upon which I won't speculate further.

Even if we countenance embracing a religion that lacks any system of postmortem rewards and punishments, one thing that seems safe to assume is that it will still involve a fairly elaborate ontology. The ontological commitments of morality, by contrast, are less extravagant. Indeed, the ontological commitments of morality are so modest that the error theorist is sometimes criticized for thinking that there are any at all (see Lenman, this volume). As a moral error theorist sympathetic to Mackie-style arguments (of the kind touched on earlier), I think that this criticism is mistaken—but, still, if a sensible philosopher can reasonably suspect that a discourse carries no ontological burden, then the ontological burden that it does in fact carry must be fairly unassuming.

When I talk here about ontological commitments being "extravagant" or "unassuming," I am not talking so much about the modality of the ontology as about the difficulty of the psychological task of entertaining the thought. For example, pretending that my dog understands most of what I say to her (when in fact I know that she understands very little) is an easy act of pretense: it comes naturally and I'm not constantly encountering counter-evidence. (When she fails to respond to my reasonable requests, I just roll my eyes and interpret her as willful.) By contrast, pretending that I have a leopard for a pet would require a constant cognitive effort on my part. Nevertheless, the possible world at which the former is true is a lot further from actuality than that at which the latter is true—and so in that sense the former involves "more ontology" than the latter.

The moral fictionalist recommendation outlined earlier sees our actual Humean reasons and values painted with a Kantian veneer. The Kantian overlay renders the reasons and values *false*, but also makes them more attainable. Some critics of moral fictionalism argue that it would require a strenuous level of self-surveillance to maintain this kind of fiction.[11] If this were true, then we should accept it as a mark against moral fictionalism. After all, if the fictionalist stance is recommended on pragmatic grounds, then, when we perform the cost-

[11] See Cuneo & Christy 2011; Eriksson & Olson 2019.

benefit analysis of comparing it with other options, one of the things that must be taken into account is whether the adoption of the type of psychological attitude has generic running costs. If nondoxastic acceptance requires cognitive effort that is exhausting and distracts from other mental activities, then this would need to be taken into consideration.

However, I find the claim that the nondoxastic acceptance of morality would have heavy "upkeep costs" quite unconvincing. Such critics, I suspect, simply have the wrong paradigm of nondoxastic acceptance in mind: they're thinking of an act of make-believe that takes effort to maintain and of which one is constantly transparently aware. I have endeavored in this chapter to focus on a different kind of paradigm. With this correction made, we should see that adding the Kantian veneer to one's practical deliberations is, far from requiring strenuous effort, a very natural and easy act of self-distraction to undertake; there is very little to it. Instead of thinking "Promise-keeping will serve my long-term preferences" (say), one thinks "I must keep promises regardless of my preferences." This is no more taxing than Janet's thinking of stamp-collecting as being a worthwhile goal in its own right rather than something she pursues only because it is an instrument to her happiness. Despite neither example's requiring any great feat of psychological maneuvering, in both cases the shift from deliberating in terms of true beliefs to nondoxastically accepting something false may be of the utmost practical significance. One's happiness may depend on it.

Here, I think, is another place where religious fictionalism suffers in comparison to moral fictionalism. The ontology of a religion is invariably a more elaborate affair than the ontology of morality. First of all, the former typically has a much broader range: covering all that morality covers, but in addition purporting to explain what happens when we die, where humans came from, the origin of the universe, etc. Second, the ontology of religion is more "substantial" than that of morality alone. Morality involves positing properties and relations (e.g., values and reasons), while religion tends, in addition, to be centrally oriented toward objects (e.g., gods), places (e.g., paradise), and events (e.g., the Last Judgment). Morality still *has* ontological commitments, but they seem nebulous in comparison with the ontology that religion wears boldly on its sleeve.

For this reason, the falsity of religion is going to intrude more frequently on the fictionalist's experience. (See Scott, this volume.) Imagine, for example, a religious fictionalist attempting to nondoxastically accept a religious account of the origin of life. The fictionalist doesn't *believe* the account—they know very well that life evolved through Darwinian selection—but they attempt to distract themselves from their scientific beliefs by encouraging in themselves thoughts of life being produced in a single act of divine creation. The problem is that this project of self-distraction looks vulnerable. Every time this would-be fictionalist reads an article on biology, every time they watch a David Attenborough nature documentary, every time they encounter a fossil in the museum, the falsity of their religious fiction is going to slap them in the face. It will not be viable to maintain a fictionalist attitude if the fact that one has the disposition to deny the content of the fiction is something of which one is made constantly aware. It's like trying to enjoy a movie but the phone keeps ringing, jolting one from the fictional world and back to reality. For the *moral* fictionalist, by contrast, the phone rarely rings.

## 9. Conclusion

I have outlined the case for an ambitious yet plausible kind of moral fictionalism, modeled roughly on a solution to the paradox of happiness that is focused on our capacity to distract ourselves from what we really believe, without ceasing to believe it. I have argued, however, that this model is unlikely to provide any matching plausibility for religious fictionalism.

This is not to deny that an alternative kind of religious fictionalism might be reasonable—one that understands "nondoxastic acceptance" in a different fashion, or one that focuses on the benefits of "playing along" with religion in a manner that doesn't require any substantial acceptance at the psychological level. I have also aimed to establish an ambitious form of fictionalism for which the net benefits of acceptance aspire to rival, or even surpass, those of sincere belief. A less ambitious form would require the net benefits of acceptance to outstrip only those offered by abolitionism (and perhaps only marginally so), while allowing that they may fall well short of the net benefits secured by sincere belief.

The proposal that it might be a good idea, pragmatically speaking, for an atheist to "play along" with religion at the level of language and social practices is something I am quite happy to label a minimal form of fictionalism; and I think that it is so easily established as a sensible recommendation for some people in certain circumstances that it probably doesn't warrant a great deal of discussion. But this just shows, in my opinion, that there are easy wins available for the religious fictionalist (and, indeed, the moral fictionalist) when ambitions are kept modest. As for more interesting and ambitious forms of religious fictionalism, however—especially when the focus is on the cultivation of a kind of psychological nondoxastic acceptance—I have argued that it's a very different story.

## References


Anscombe, G.E.M. 1958. "Modern moral philosophy." *Philosophy* 33: 1-19.

Baumard, N. & Boyer, P. 2013. "Explaining moral religions." *Trends in Cognitive Sciences* 17: 272-280.

Bernstein, A. & Katz, P. 2010. "The rise of postmortem retribution in China and the West." *Medieval History Journal* 13: 199-215.

Coleridge, S. T. 1817. *Biographia Literaria*, Vol 2. R. Fenner.

Cuneo, T. & Christy, S. 2011. "The myth of moral fictionalism." In M. Brady (ed.), *New Waves in Metaethics*. Palgrave Macmillan. 85-102.

Dennett, D. 1995. *Darwin's Dangerous Idea*. Penguin.

Eriksson, B. & Olson, J. 2019. "Moral practice after error theory: Negotiationism." In R. Garner & R. Joyce (eds.), *The End of Morality: Taking Moral Abolitionism Seriously*. Routledge. 113-130.

Frank, R. 1988. *Passions within Reason: The Strategic Role of the Emotions*. W.W. Norton & Company.

Ingram, S. 2015. "After moral error theory, after moral realism." *Southern Journal of Philosophy* 53: 227-248.

Jay, C. 2014. "The Kantian Moral Hazard Argument for religious fictionalism." *International Journal for Philosophy of Religion* 75: 207-232.

Johnson, D. & Krüger, O. 2004. “The good of wrath: Supernatural punishment and the evolution of cooperation.” *Political Theology* 5: 159-176.

Joyce, R. 2006. *The Evolution of Morality*. MIT Press.

MacIntyre, A. 1984. *After Virtue*. University of Notre Dame Press.

Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. Penguin.

Martin, M. 2008. “Paradoxes of happiness.” *Journal of Happiness Studies* 9: 171-184.

Mill, J. S. [1873] 1924. *Autobiography of John Stuart Mill*. Columbia University Press.

Sidgwick, H. 1907. *The Methods of Ethics* (7th edition). Macmillan.

Whately, R. 1856. *Thoughts and Apophthegms*. Lindsay & Blakiston.