Don't Stop Believing (Hold on to that Warm Fuzzy Feeling)

Jessica Isserow* Edward Elliott†

School of Philosophy, Religion and History of Science
University of Leeds

April 6, 2020

Abstract

It's natural to think that there's value to improving the accuracy of our beliefs. If beliefs are a map by which we steer our efforts to bring the world in line with our preferences, then, *ceteris paribus*, we should want a more accurate map. However, it is easy to see that this cannot be true in general. The world could be structured so as to punish learning with respect to certain topics; i.e., by coming into possession of new information, an agent's situation could be made worse than it otherwise would have been. In this paper, we investigate whether the world is structured so as to punish learning with respect to moral nihilism. We ask: If an ordinary human agent had the option to learn whether or not moral nihilism is true, then ought she to take it? We argue that, given intuitively plausible and empirically grounded assumptions about ordinary human preferences, she (probably) should not.

1 Introduction

All else being equal, it's good to have true beliefs. On a common view of human action and decision-making, we are for the most part pragmatically rational beings: we typically act in such a way as to bring about the kinds of things that we want, given the way we take the world to be. To borrow a metaphor from Ramsey (1931), our beliefs are a map by which we steer our efforts to bring the world in line with the way we'd prefer it to be. If this is so, then it stands to reason that a more accurate map is usually going to be a better guide than one that says (for example) that there's a road where no road exists, a forest where there's no trees, or a mountain where there's only a molehill.

Indeed, in the limit case, where an agent—we'll call her Alice—has *complete* knowledge of what her world is like and where she's situated within it, she will typically choose whatever action available to her will result in the best outcome by her lights. That is, if Alice had only true beliefs and were ignorant of nothing relevant to her present decision, then the subjectively *rational* choice (that which would maximise her preference satisfaction if her beliefs were fully accurate) and the objectively *correct* choice (that which actually would maximise

*Email: j.m.isserow@leeds.ac.uk
†Email:e.j.r.elliott@leeds.ac.uk

her preference satisfaction) would be one and the same. As a rough and ready generalisation: the more accurate Alice's beliefs are, the more likely it will be that the rational choice and the correct choice will coincide for any decision situation Alice might find herself in.

With these considerations in mind, it's plausible that if some theory is true, we should want to learn of its truth—especially if the truth or falsity of that theory would have far-reaching implications as to whether and how well our preferences are satisfied. And for many of us, *moral nihilism* is just such a theory. Briefly—we'll precisify in a moment—the moral nihilist says that there are no moral facts; no facts about who is morally good or bad, about what is morally right or wrong, or about what we morally ought or ought not to do; morality, in general, is bunk. If such a theory were true, then this would make a difference to how we evaluate the potential consequences of our actions. Most of us care about being morally good people and doing the morally right thing (or so we'll argue). If it turns out that there are no morally good people, or no morally right things to do, then, *ceteris paribus*, this seems the sort of thing we should want to know as soon as possible.

However, we should first check that the *ceteri* really are *paribu*.[1] It's easy to see that it can't be true *in general* that we always do better to improve the accuracy of our beliefs whenever we're given the choice to do so. Consider the case of an evil demon who hates know-it-alls: the more that Alice learns about the world around her, the more the demon limits her options to only those with the worst outcomes. In the limit case, Alice knows exactly what choices will maximise her preference satisfaction given whatever situations she finds herself in, so she always makes the best choices she can relative to her situation—but, by virtue of her now perfectly ideal epistemic state, any situation she's in will be much worse than it might have been otherwise. So, it's at least possible for the world to be structured so as to punish *general* improvements to one's epistemic state. It may also be structured so as to punish learning with respect to *specific* topics. If an overzealous moral realist credibly threatened to set off a nuclear weapon were Alice to learn any more about the truth of moral nihilism, then she may quite rationally decide to avoid any further inquiries on the topic.

The examples just given are fanciful, but they do raise an interesting question: just how plausible is it that the world the average person lives in is structured so as to punish learning specifically with respect to moral nihilism? Or to put the question in a slightly different way: if Alice, whom we'll suppose henceforth is an ordinary human being with ordinary human preferences and ordinary human beliefs, had the option to learn the truth of moral nihilism, free of charge, then ought she to take it? We will argue that Alice would be irrational to take the offer, provided her preferences and beliefs conform to (what we will argue are) common and perfectly reasonable patterns.

We do not put this conclusion forward as a necessary claim. It's obviously not *necessarily* irrational to learn about moral nihilism. Nor do we want to say that our conclusion applies to everyone alive today. People vary, some more so than others. In fact, we'll argue that philosophers in particular can have incentives to inquire after the truth of moral nihilism that plausibly outweigh the costs involved. But philosophers are unusual. Most people, if they are the way we think they are, would do better to avoid inquiry into moral nihilism.

[1] Only one of the authors knows Latin; the other is only guessing.

The remainder of the paper is as follows. In §2, we say a little more to pin down the sort of nihilistic theory we have in mind, and lay out some general background assumptions used in the ensuing discussion. Then in §3, we introduce a standard framework for thinking about the value of learning using a pair of simple hypothetical cases, and in §4 apply that same framework to the case of learning about moral nihilism, and draw our main conclusion. Finally, in §§5–8, we provide an empirical case for the key assumptions we require for that conclusion.

2 Background

It will be helpful to begin by laying out some key ideas and assumptions that we'll making with regards to moral nihilism, and how we'll be understanding beliefs, preferences, rational choice, and the relationships between them.

The moral nihilist—she may also go by *moral error theorist*—is a cognitivist with respect to moral discourse, taking ordinary moral claims to be in the market for truth and falsity. However, she parts company from other cognitivists—so-called *success theorists*—in taking such claims to be systematically false.[2] Some subtlety is called for here. The moral nihilist may very well be able to stomach the truth of a small class of moral claims. She will likely allow that some non-atomic ('Either lying is wrong or the Eiffel Tower is in Paris'), tautological ('Wrongness is wrongness'), negative ('stealing isn't wrong'), or second-order ('There are no moral facts') moral claims could still be true. But we take it that such exceptions will be of cold comfort to the opponent of moral nihilism. At the very least, the moral nihilist will want to say that all atomic, non-tautological, positive, first-order moral claims are false. This, we submit, is a sizeable portion of moral discourse—sizeable enough to render moral nihilism a *prima facie* unsettling proposal.

Different philosophers have had different grounds for endorsing moral nihilism. Perhaps our moral talk is underwritten by a problematic commitment to categorical reasons (Joyce 2001). Maybe moral facts would be unacceptably "queer", the sorts of things that could not hope to find a place in the natural world (Mackie 1977). Or perhaps such facts would be explanatorily idle, swiftly eliminated from one's ontology with an unforgiving swipe of Occam's razor (Olson 2014, pp. 123–36). Some of these claims will no doubt strike the reader as more persuasive than others. But for our purposes, there is no need to choose. Our arguments don't really stand or fall with the truth of moral nihilism—let alone with a particular variety of it.

Consequently, we'll in effect be treating 'moral nihilism' as the disjunction of the particular varieties of nihilism that these and other authors have put forward over the years. Some of those disjuncts will be what we can call *metaphysically non-contingent* theories—i.e., varieties of moral nihilism such that, if they're true, must be true as a matter of metaphysical necessity. And some of the disjuncts will be *epistemically non-contingent*—i.e., theories such that, if they're true, will be true a priori. Some might be both. But we take it that metaphysically contingent nihilist theories make sense as well, and (as far as we can tell) at least some of these cannot be ruled out a priori (cf. Miller 2010).

[2] See (Sayre-McCord 1986) for a helpful taxonomy. Nihilists who take moral discourse to fall victim to presupposition failure may prefer to characterise moral claims as neither true nor false. For discussion, see (Joyce 2001, Ch. 1) and (Kalf 2013).

It also bears mentioning here that we conceive of moral nihilism as a *local* nihilism. That is to say, the moral nihilism at issue is not merely a symptom of a more sweeping, global nihilism, according to which *no* normative claims are true (cf. Streumer 2017). We hasten to emphasise that this assumption is not idiosyncratic. Global normative nihilism is plausibly the exception rather than the rule (see Joyce and Kirchin 2010, p. xiii). In what follows, then, we will happily help ourselves to normative language, speaking of what ordinary agents like Alice ought to do, and of rational decision-making more generally.

Next, we will assume henceforth a specifically Bayesian approach to beliefs, preferences, and rational decision making.[3] With respect to Alice's beliefs, this requires the existence of a credence function, $\mathcal{C}r$, which assigns numerical strengths of belief between 0 and 1 to the various propositions regarding which Alice has opinions, and *eo ipso* represents her beliefs *in toto*. We don't have to assume that this function $\mathcal{C}r$ obeys all the axioms of the probability calculus, but we will assume that the following are all at least roughly true:

(i) If p is inconsistent, then $\mathcal{C}r(p) = 0$
(ii) If the propositions $p_1, \ldots, p_n$ are mutually exclusive and jointly exhaustive, then $\mathcal{C}r(p_1) + \cdots + \mathcal{C}r(p_n) = 1$, at least for small n
(iii) It's both possible and rationally permissible that $0 < \mathcal{C}r(\text{Nihilism}) < 1$

(One might worry that, given the possibility of metaphysically or epistemically necessary nihilist theories, these three assumptions could run up against some Bayesian models. Specifically: it's usual to define $\mathcal{C}r$ over an algebra of propositions drawn from an underlying space of worlds Ω, where 'inconsistent' propositions are modelled as the empty set; given this, the three assumptions imply that anything true at all worlds in Ω must be assigned 1. So the third assumption is inconsistent with the first two if either (i) Ω is the space of metaphysically possible worlds, and there's at least one variety of nihilism which happens to be true as a matter of metaphysical necessity; or (ii) Ω is the space of 'a priori possible' worlds, and there's at least one variety of nihilism which is true a priori. In either case, the *disjunction* of moral nihilist theories will be true at all worlds in Ω, and hence will need to be assigned a credence of 1. But we think such worries would be misplaced. In particular, if a particular choice of Ω makes rational belief (or disbelief) in nihilism impossible or irrational, then we've got a good reason to choose a different Ω. A non-ideal agent can have very good reasons for being less-than-certain of a metaphysical necessity, or even an a priori truth if it's sufficiently non-obvious, without thereby being labelled *irrational* by any standard of rationality that's reasonably applicable to ordinary human beings. Assuming moral nihilism neither follows from nor contradicts classical logic, there would be no problem of the kind being discussed here if we just let Ω be the space of classically logically possible worlds—or anything larger—and we see no reason not to think of Ω in this way.)

[3] A previous referee notes that we don't *need* Bayesianism to make our argument, which could be made in a *rough* way without all of the formalism. Perhaps this is true. But rough arguments lead to rough conclusions. Our argument rests on drawing comparisons between trade-offs amongst a decision-maker's outcomes which are sensitive to the relationship between values assigned to the outcomes and the decision-maker's degrees of belief. We don't see how we could fruitfully draw such comparisons without doing so within a formal framework designed to handle exactly these kinds of trade-offs. Moreover, by making use of the Bayesian framework, we're making the structure of our argument and our assumptions as explicit as we can, so that you the reader can know *exactly* where it is you want to disagree if you so choose.

Our Bayesian approach also requires the existence of a utility function, $\mathcal{U}$, which provides a numerical measure of the extent to which Alice's preferences are satisfied under different ways the world might be. Furthermore, we assume that the *rational* choice in any decision situation is that which maximises expected utility. Where Alice has to choose among some collection of options each of which have different outcomes under different ways the world might be consistent with what she believes, Alice should pick the option (or one of the options) with the greatest $\mathcal{C}r$-weighted average utility. In more formal terms, supposing that

(i) $\{p_1, \ldots, p_n\}$ is a finite partition of propositions, where each element is causally and evidentially independent of whatever option Alice decides upon, and
(ii) option A has outcome a_i if p_i is true, and B has outcome b_i if p_i is true $(i = 1, \ldots, n)$,

then Alice should weakly prefer the option A to B if and only if

$$\sum_{i=1}^{n} \mathcal{C}r(p_i) \cdot \mathcal{U}(a_i) \;\geq\; \sum_{i=1}^{n} \mathcal{C}r(p_i) \cdot \mathcal{U}(b_i).$$

We will not defend this account of rational choice here. That's been done more than enough elsewhere. We do note, however, that alternative accounts of rational choice in which attitudes towards risk play a bigger role (e.g., Buchak 2013) will tend to favour our conclusions, given our empirical assumptions.

Finally, we take a *Humean* picture of rationality with respect to preferences over outcomes. That is: beyond basic coherence requirements like transitivity, there are no restrictions on what kinds of things an agent ought to care about when she's making her decisions. A rational agent doesn't *have* to care, *qua* rational agent, about what's morally good, or "the truth", or indeed about anything else whatsoever. This means as well that Alice's preferences need not depend essentially on matters of personal experience—i.e., our use of 'utility' is not to be interpreted as a measure of some experiential state of pleasure, happiness, or a sense of satisfaction. In terms borrowed from the economics literature, we're interested in what's often called *decision utility*, rather than *experienced* or *hedonic utility* (see Fumagalli 2013). For example, suppose that Alice cares overwhelmingly about maximising the number of puppies there are, and that world ω_1 has many more puppies than ω_2. Then ω_1 carries more utility for Alice than ω_2 does—that is, even if at ω_2 Alice has more puppies than she knows what to do with, while at ω_1 Alice sadly believes that puppies have gone extinct, and she never gets to experience the joys of having any around for herself. Preference satisfaction doesn't imply awareness of that satisfaction, and utility isn't a measure of how nice it feels when you believe your preferences to have been satisfied. Most people usually *do* care about having nice feelings, of course, but they tend to care about a lot more besides.

3 The Value of Learning

With all that out of the way, how should we think about the value of new information within the general framework we've outlined? We'll introduce this with some hypothetical cases.

Case 1: The Bet
Before Alice and Bob sits an opaque box, which contains either a red ball or a blue ball. Alice doesn't know what colour the ball is, but she's slightly more confident that it's red—specifically, $\mathcal{C}r(\text{Red}) = 0.6$. Alice knows that Bob doesn't know what colour the ball is. Bob offers Alice a bet: he'll reach into the box and pull out the ball; if it's red then Alice wins \$10, if it's blue Alice pays him \$10.

As Alice is deciding whether to take the bet, an oracle appears: she offers to tell Alice whether the ball is red or blue before she decides. The offer is free of charge, and there are no strings attached. Alice knows she can trust the oracle. Should she accept?

It's immediately obvious that Alice should accept. But we can slow things down and rationally reconstruct her reasoning using Figure 1.

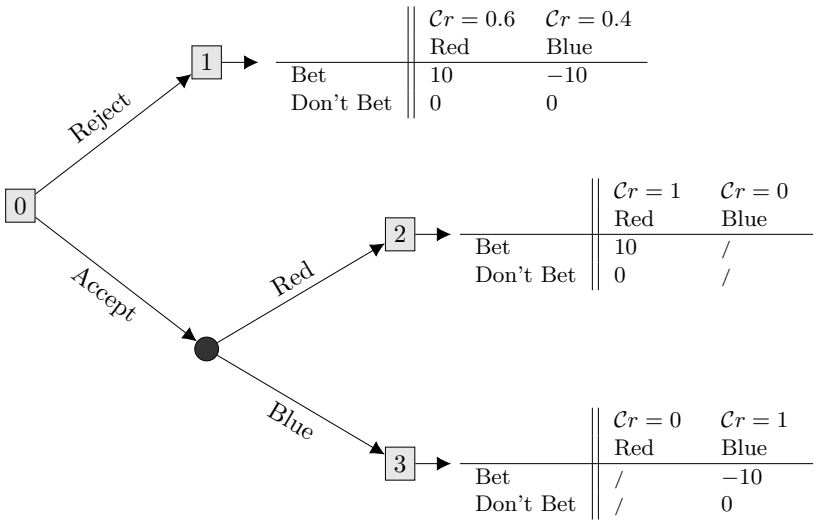


Figure 1: The Bet

At node 0, Alice is trying to decide whether to *Reject* or *Accept* the oracle's offer. If she decides to *Reject*, then she knows she'll end up in the decision situation at node 1, where she has to decide between *Bet* and *Don't Bet* given her unchanged belief state. Letting utilities equal the dollar values for simplicity, in this case the expected utility at node 1 (written $\mathcal{EU}_1$) of *Bet* is

$$\mathcal{EU}_1(Bet) \;=\; (0.6 \times 10) + (0.4 \times -10) \;=\; 2,$$

which is higher than the expected utility of *Don't Bet*. So, Alice reasons that if she were in that situation, she'd certainly choose *Bet*; hence the expected utility of *Reject* at node 0 is just the expected utility of the choice she'd end up making if she rejected—i.e., $\mathcal{EU}_0(Reject) = \mathcal{EU}_1(Bet)$.

On the other hand, if she decides to *Accept*, then she knows that the oracle is either going to tell her that the ball is *Red* or that it's *Blue*. She doesn't know what she'll be told, but she *does* have beliefs about which is more likely. She reasons that there's a 60% probability that she'll be told the ball is red, in which

case she'll be in the decision situation at node 2, where she'd choose *Bet* and will win a guaranteed $10. On the other hand, there's a 40% probability that the oracle will tell her the ball is blue, in which case she'll be in the decision situation at node 3, whereupon she'd want to avoid a sure loss and choose *Don't Bet*. Overall, then, the expected utility of *Accept* at node 0 is:

$$\big(0.6 \times \mathcal{EU}_2(\mathit{Bet})\big) + \big(0.4 \times \mathcal{EU}_3(\mathit{Don't\ Bet})\big) \;=\; (0.6 \times 10) + (0.4 \times 0) \;=\; 6,$$

which is greater than that of *Reject*. So Alice takes up the oracle's offer.

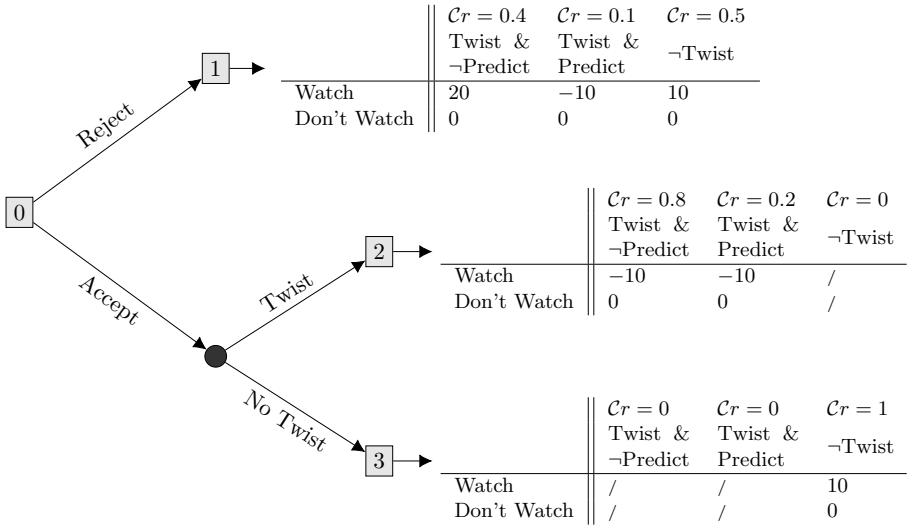
Note a special feature of the case: accepting the oracle's offer comes with no associated costs. Alice doesn't have to pay any money for the information, there's no cost in time or effort, she doesn't have to promise away her first-born child, etc. Nor is Alice required to forego any of her future options by accepting the offer. In short, if she accepts, then Alice loses no opportunities she would have had otherwise, nor does she make any of the outcomes of any later choices worse under the different states of the world she's uncertain about. In this kind of case, we can say that her learning is genuinely *cost free*. And on the basis of several quite general formal results, we have known for a long time that it is *always* rational to pursue genuinely cost free learning. (See, for example, Good 1967; Skyrms 1990; and Oddie 1997.)[4]

But truly cost free learning is rare indeed. Outside of purely fictional cases and artificial experimental situations, learning and inquiry usually involves *some* cost in effort, resources, time, or future opportunities. And often those costs can be considerable (as anyone paying back student loans will appreciate). So let us therefore consider a case of costly learning, which (we believe) is more closely analogous to the case of moral nihilism.

> **Case 2: The Movie**
> Alice loves movies which have a 'big twist', but only if she doesn't see the twist coming—if she were to know what twist is coming, then watching the movie would be worse than watching nothing at all. Luckily, when she's watching a movie with a twist, she only sees the twist coming about 20% of the time. (She vigorously avoids watching movies she's seen before if they have a twist.) Of those movies which don't have a twist, she usually considers them just so-so: better than nothing, but also not great.
>
> She's trying now to decide whether to watch a new movie; the alternative is to watch nothing. She knows nothing about the new movie, and she's 50/50 on whether it will have a twist. As she's making up her mind, an oracle appears again and offers to tell her the entire plot, free of charge. Should Alice accept?

Again, it's obvious what Alice should do. If Alice accepts the oracle's offer, then from her epistemic perspective there's a 50% probability that she'll learn the plot of a movie which has a twist, and hence she'll see the twist coming regardless of whether it was antecedently predictable or not. If that happens, she knows she'd dislike the movie intensely. Moreover, she'll have ruled out obtaining the best possible outcome: watching a movie with a twist she doesn't see coming. She has nothing to gain and everything to lose. The fact that the oracle's offer is *free of charge* doesn't mean that it's *cost free*.

[4] Where Alice's degrees of belief are imprecise, matters a slightly more complicated, though the basic point still holds. See (Bradley and Steele 2016) for discussion.



	$\mathcal{C}r = 0.4$ Twist & ¬Predict	$\mathcal{C}r = 0.1$ Twist & Predict	$\mathcal{C}r = 0.5$ ¬Twist
Watch	20	−10	10
Don't Watch	0	0	0

	$\mathcal{C}r = 0.8$ Twist & ¬Predict	$\mathcal{C}r = 0.2$ Twist & Predict	$\mathcal{C}r = 0$ ¬Twist
Watch	−10	−10	/
Don't Watch	0	0	/

	$\mathcal{C}r = 0$ Twist & ¬Predict	$\mathcal{C}r = 0$ Twist & Predict	$\mathcal{C}r = 1$ ¬Twist
Watch	/	/	10
Don't Watch	/	/	0

Figure 2: The Movie

The situation is represented in Figure 2 (below), where it's easy to see that the expected utility of *Reject* (12) is greater than the expected utility of *Accept* (5). (The specific values we've chosen for the utilities make no difference to the result, which in this case depends only on preferences over the outcomes.)

The reasoning we've been attributing to Alice usually goes by the name *backwards induction* in game theory, where it is used frequently. Effectively, while deciding whether to accept the offers of more information, Alice is treating her future self as a second player in a 2-player game, whose later choice:

(i) Determines Alice's outcomes now, and
(ii) Depends partly on (a) Alice's own present choice (*Accept* or *Reject*), and potentially also (b) an unknown state of the world (Twist or No Twist).

The backwards-inductive reasoning has been slightly idealised: we've been implicitly assuming that Alice knows (i.e., with certainty) exactly what her future self's options, beliefs, and utilities will be, and exactly how she will choose under the circumstances she might end up in. With some added sophistication we could remove these idealisations, but in practice they typically don't make much difference for simple cases like those we're considering here. Insofar as Alice has high confidence on these matters, the main conclusions will remain unchanged.

4 The Value of Morality

Let us now apply backwards induction to the case of learning about moral nihilism. We'll begin our discussion with a toy case, designed to bring out the basic structure of our argument and the core assumptions it rests upon.

> **Case 3: The Sofa**
> Alice is deciding whether to help Bob, who's moving a sofa up a flight of stairs. On the one hand, Alice has no intrinsic desire to carry sofas up stairs, and all else equal would prefer not to. However, there are

several strong considerations in favour of helping. First, Alice desires to help Bob because she cares about doing the right thing (whatever that may be), and she believes in this case that helping Bob is *the right thing to do*. Furthermore, whenever she does what she believes is *the right thing*, Alice gets a little warm fuzzy feeling inside. Alice enjoys this feeling, though it is by no means a primary driving force in favour of doing the right thing generally. Over and above those considerations, Alice also desires to help Bob regardless of whether it's the right thing to do, because Bob is her friend and she wants to help her friends; and she wants to avoid any social reprobation that might arise if it were to become widely believed that she is unhelpful.

As she's making up her mind, an oracle once again appears and offers to tell Alice whether moral nihilism is true, free of charge. Alice is open to the idea of nihilism—specifically, she'd assign it about 10% confidence—but the rest of her confidence resides in some form of moral realism. Should Alice accept?

Before we say anything else, it must be emphasised that we're using 'warm fuzzy feeling' (henceforth: *wff*) as a kind of placeholder. Our arguments do not rest upon the idea that agents like Alice literally experience any sort of pleasurable sensation or violent passion when they act on their moral convictions. Our use of '*wff*' may equally well denote (say) a sense of personal accomplishment, or meaningful achievement, or a disjunction of the above. You could treat 'wff' as a stipulative name, designating *something* Alice values which is specifically tied to her believing that her moral preferences have been satisfied. We'll argue that there are indeed such things in §6.

In evaluating this case, we'll make things tractable by grouping together a wide disjunction of realist views under the heading 'Realism', just as we've done for 'Nihilism'.[5] Given this, the first key point to note is that the outcomes of Alice's choices will depend not only on what state of the world is actual, but also on Alice's beliefs about which kind of world she's in. In our description of the case, we've said that Alice has a preference for doing the right thing (whatever that may be), but also numerous other preferences, e.g., the desire to help her friend and the fear of social reprobation, and at least a slight preference for the little warm fuzzy feeling.

Let's refer to the former as Alice's *moral preference*—i.e., her preference for doing the right thing as such, for being a good person, for improving the overall moral goodness of the world. And let's refer to the latter (disjunctive) kind as her *non-moral preferences*. Now whether Realism or Nihilism is true makes a difference to whether her moral preferences are satisfied, but what she *believes* about the status of Realism/Nihilism makes a difference to her warm fuzzy feeling. The other non-moral factors depend primarily on whether she chooses to help or not help.

[5] It's worth noting that the kinds of realism and nihilism that matter most here will be those that Alice herself will have in mind. We are, recall, assuming that Alice is a *ordinary* person, and we should be thinking about her beliefs, desires, and reactions to the oracle's testimony in that way. Metaethicists will want to make manifold and nuanced distinctions between various sub-classes of realist and nihilist views, and they'll have different degrees of confidence regarding each one. But we can't expect this kind of nuance *for Alice*, who probably isn't going to even be aware of most of these distinctions.

So let's turn that observation into an argument that Alice should reject the oracle's offer. We'll simplify (to begin with) by supposing that Alice doesn't yet consider how her choice *vis-à-vis* the oracle's offer might affect her outcomes in more distant future decision situations—she's focused for now just on how it will affect the immediate outcomes of her decision whether to help Bob. And one further simplification: say that Alice *believes* that p if and only if $\mathcal{C}r(p) \geq 0.9$, and then symbolise the outcomes as follows:

x = *Help* at a world where Realism is true, and Alice believes Realism
y = *Help* at a world where Nihilism is true, and Alice believes Realism
z = *Help* at a world where Nihilism is true, and Alice believes Nihilism

q = *Don't Help* at a world where Realism is true, and Alice believes Realism
r = *Don't Help* at a world where Nihilism is true, and Alice believes Realism
s = *Don't Help* at a world where Nihilism is true, and Alice believes Nihilism

The effect of the simplification here is that helping/not helping at a world where Realism holds and $\mathcal{C}r(\text{Realism}) = 0.9$ has the same utility as helping/not helping at a world where Realism holds and $\mathcal{C}r(\text{Realism}) = 1$. This probably isn't exactly true, but it's likely to be approximately true, and it will let us present a relatively easy to follow version of our argument to begin with. We promise we'll discuss de-simplifying our argument further below, in §8. For now, we use Figure 3 to represent Alice's decision situation.

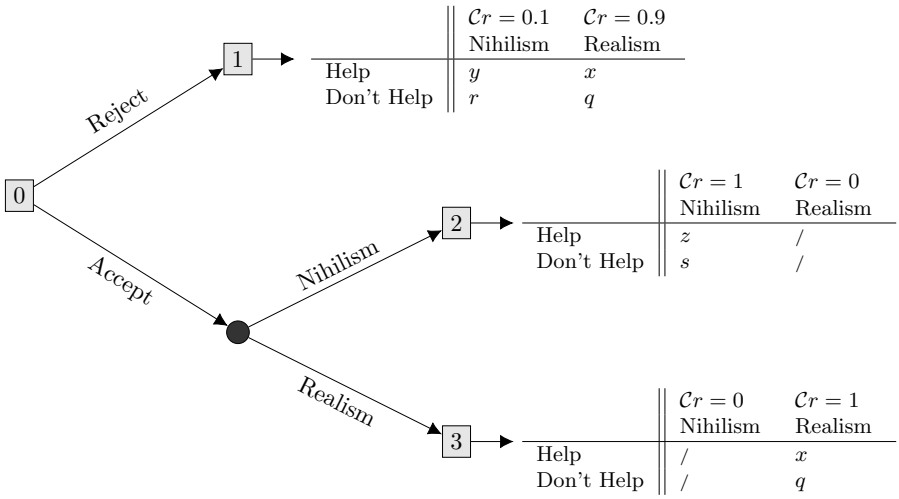


Figure 3: The Sofa

Of course, whether Alice ought to *Accept* or *Reject* comes down to how we fill out the utility values at the different nodes of the decision tree. But suppose the following claims are true:

A1. At node 1, *Help* has maximal expected utility
A2. x is at least as great as q
A3. y is greater than either z or s

Fill in the utility values any way you like consistent with A1–A3, and *Reject* will have strictly greater expected utility. The proof of this is simple. Where $\mathcal{C}r_0$ designates Alice's degrees of belief at node 0, A1 implies

$$\mathcal{EU}_0(\mathit{Reject}) \;=\; \mathcal{EU}_1(\mathit{Help}) \;=\; \big(\mathcal{C}r_0(\text{Nihilism}) \times y\big) + \big(\mathcal{C}r_0(\text{Realism}) \times x\big),$$

and from A2 and A3,

$$\mathcal{EU}_0(\mathit{Accept}) \;=\; \big(\mathcal{C}r_0(\text{Nihilism}) \times \max\{z, s\}\big) + \big(\mathcal{C}r_0(\text{Realism}) \times x\big),$$

where $y > \max\{z, s\}$; hence $\mathcal{EU}_0(\mathit{Reject}) > \mathcal{EU}_0(\mathit{Accept})$.

Great—but why should we think that A1–A3 are true? Well, our core empirical assumptions are as follows:

> Correlation. There is a correlation between Alice's moral preferences and her non-moral preferences, in the sense that she would usually prefer to do what she believes is the right thing regardless of the truth of nihilism.
>
> Cost. In worlds where nihilism is true but she *believes* it's false, Alice still gets a pleasant *wff* for having done what she *believes* is the right thing—which she would not have if she came to believe there is no right thing to do.
>
> No Compensation. The aforementioned cost of losing the *wff* is greater than any increase in utility to *Help* or *Don't Help* (whichever is the greater) that results from coming to believe in nihilism at a world where it's true.

That is, prior to the oracle's offer Alice would prefer to do what she initially takes to be the right thing regardless of whether nihilism were true, because her non-moral preferences speak in favour of acting as such. Among the many factors that speak in favour of helping is the *wff*, though she would still choose to help even if she didn't expect to have it (Correlation). Furthermore, she thinks that she'd no longer have the *wff* if she were to act the same way and did believe nihilism—if she no longer thinks there *is* a morally right thing to do, then she can't believe of herself that she has managed to do it. She'll still choose to act the same way, but she'd suffer the cost of doing so without having the *wff*, however small that cost may be (Cost). And, finally, that cost would not be compensated for by any gain to the utility of helping (or not helping) at a world where nihilism is true and she believes it. For example, if she cares about having true or more accurate beliefs, then she may get a bit of extra 'epistemic utility' for having true beliefs about nihilism—but not enough to cover the costs of losing the *wff* (No Compensation).

Correlation directly supports assumptions A1 and A2, and given Correlation, Cost and No Compensation jointly support A3. Summarising, our claim is that given her present belief state, Alice doesn't have much to gain by learning that nihilism is false—and, if it turns out that nihilism is true, then she has something to lose (the *wff*). Moreover, whatever she might gain by coming to have more true beliefs under that eventuality isn't enough to make up for what she'll lose. Hence, she should reject the oracle's offer. In the following sections, we'll provide empirical support for Correlation, Cost and No Compensation.

5 Correlation

According to Correlation, people's moral and non-moral concerns tend to correlate fairly well with one another. We should emphasise from the outset that our arguments do not require that the correlation here be perfect. It should be all too obvious that it isn't; our moral and non-moral concerns can and do sometimes come apart. Doing what is right may require sacrificing other things that we care about. Nonetheless, it is our contention that the correlation is fairly systematic and seems to hold true in general, even if it does not hold true on each and every occasion. We should also emphasise that we do not regard Correlation as a brute or necessary fact about moral agency. Our arguments only require that it amounts to a defensible generalisation as regards typical human beings. The case that we sketch in what follows is therefore intended in an abductive spirit; given what we know about human evolution and psychology, it is reasonable to suppose that Correlation applies to a great deal many human beings. Clearly, more needs to be said to tie down precisely what we have in mind. We now get to work, beginning with the distinction between moral and non-moral concerns.

As we will argue in greater detail below, people are generally invested in moral matters *qua* moral matters; they tend to care about acting in ways that they judge to be morally right, and about not acting in ways that they take to be morally egregious. But people are not *only* invested in moral matters. Most have other sorts of projects and interests as well. Some of these may be purely self-regarding (Alice may want to be popular). Others may be genuinely other-regarding (she may want to do something nice for her mother). Agents can no doubt have loftier concerns as well; Alice may enrol in university because she values knowledge for its own sake. (We attend to epistemic pursuits in greater detail in §7.) In practice, the boundaries between each of these can sometimes be fuzzy. However, we take it that non-moral concerns can be meaningfully differentiated from concerns for moral rightness as such—ethics is not first philosophy.

Insofar as Alice's moral concerns correlate with non-moral concerns such as these, the following will be true of her: whenever morality favours φ-ing, the balance of Alice's non-moral preferences will independently tend to favour φ-ing as well, and whenever morality disfavours φ-ing, the balance of Alice's non-moral preferences will independently tend to disfavour φ-ing as well, such that were morality removed from the equation, it would still be (ir)rational for Alice to φ. Correlation therefore rests upon the well-worn idea that it's generally in one's interest to be (morally) good. We are not inclined to regard this idea as especially controversial (the true battleground surely lies with the idea that it's *necessarily* in *any* agent's interest to be morally good). But we do not wish to take it for granted either. In what follows, we argue that morally good conduct generally promotes (what are reasonably regarded as) common human ends.

To begin with, humans care deeply about the welfare of others. Even young infants exhibit strong other-regarding concerns (Liszkowski et al. 2006; Warneken and Tomasello 2006, 2007; Hepach et al. 2013). This is unsurprising. The survival of our species survival has long been predicated upon successful co-operation, and there has long been biological and cultural selection for emotional responses that support it. We feel sympathy in response to others' suffering, anger in response to their selfishness, and guilt in response to our own. Each of these experiences has motivational import. Sympathy motivates helping behaviour

(Batson et al. 1981, 1988); anger fuels punitive action (Hopfensitz and Reuben 2009; Nelissen and Zeelenberg 2009; Seip et al. 2014);[6] and guilt encourages making amends (Ferguson et al. 1991; Tangney et al. 2013). It is telling that a callous disregard for others is among the key diagnostic criteria for a number of human *pathologies*—Anti-Social Personality Disorder and Conduct Disorder, for example (American Psychiatric Association 2013).

People also care about what others think of them. Social disapproval is often seen as an especially toxic form of punishment; many would prefer pain, jail time, amputation, or *death* to a heavily tarnished reputation (Vonasch et al. forthcoming). Again, this is unsurprising for an ultra-social and ultra-cooperative species. It is in our interests to secure the approval of others. It is arguably even more in our interests to avoid their disapproval. The price of being unpopular is high, ranging from lower job prospects (Western et al. 2001) to lower life expectancy (House et al. 1988). Given the great costs of being disliked, some hypothesise that natural selection may have favoured a robust concern for reputation (American Psychiatric Association 2013).

In light of evidence of this kind, we are inclined to regard Correlation as a reasonable empirical generalisation. Morality centrally concerns the claims and interests of others, and we tend to care about how other people fare in life. Likewise, morally good conduct has clear reputational benefits. Most people favour generosity. Everybody appreciates compassion. Nobody likes an arsehole.

Before moving on, we want to note that Correlation can be motivated on intuitive grounds as well. To this end, it helps to consider what it would take for it to be radically *false*. On the one hand, it may be that most people's non-moral concerns are radically out of kilter with what morality requires of them. This would be true if, for example, most people had nothing but repugnance for their fellow travellers, deriving happiness from their pain, and caring little for their good opinion. In acting rightly, such agents would represent the (somewhat caricatured) Kantian ideal of moral agency: moral automatons propelled by the sheer force of the moral law. Alternatively, it may be that most individuals subscribe to excessively demanding moralities. The radical utilitarian may struggle to live up to her moral ideals, devoting most of her resources to her beloved children—all the while believing that morally, she ought not to be doing so.

It would not necessarily be irrational for either of these agents to inquire after the truth of moral nihilism. Indeed, each stands to benefit should nihilism turn out to be true. Our Kantian would be free to act on her aversions and inclinations (however ignoble), the utilitarian no longer caught in the throes of excessive moral demands. We do not doubt that such agents exist.[7] But we do doubt their prevalence. Again, the empirical data here is telling; given what we know about human psychology, an individual completely devoid of fellow feeling is reasonably classified as anomalous. But the intuitive data is telling as well. Moral philosophers have long objected to Kantian and utilitarian ideals *precisely because* they are humanly unachievable (most famously, Stocker 1976). People care deeply for those closest to them, and it is not reasonable to expect that such concerns can be surgically removed from their maxims or moral calculus.

[6] Or, at least it does in the Western world. As (Flanagan 2017, p. 154) notes, anger seems to have a different behavioural profile in other cultures.

[7] Indeed, we were once asked (in a tone that suggested the animating thought behind the question was all too obvious) why we did not seriously consider the best part of converting to moral nihilism—*viz.* the liberty to unabashedly pursue self-interest.

We agree. Any moral theory that proposes to divorce moral action from the affective network that underwrites it is unlikely to be user-friendly.

6 Cost

According to our second empirical assumption, there is a potential cost associated with inquiry into moral nihilism. The general idea here is that an ordinary human agent like Alice still gets *something* of value even in those worlds where moral nihilism is true, so long as she *believes* that she has done what is right—to wit, the *wff*. Buried in this general idea are two further assumptions concerning Alice's preferences:

> De dicto desires. Alice desires to do what is morally right, whatever that may be.
>
> De dicto dividends. There's an extra payoff specifically tied to Alice's believing herself to have satisfied her desire to do what is morally right.

We take the latter assumption to be the more controversial of the two, and so, it is there that we will focus the majority of our critical attention. But let us offer some brief words of support for the first assumption, so as to allay any niggling doubts.

De Dicto Desires attributes to Alice a standing desire to do the (morally) right thing, where that desire is given a *de dicto* reading: Alice desires to do the right thing, whatever that may be. Put differently, Alice desires to do the right thing as such; she wants to perform actions that are morally right under that description. She does not merely want to do the right thing *de re*—to engage in behaviour which, as it happens, is morally right. Some philosophers have questioned whether human agents generally are motivated in this way (see (Arpaly 2015, p. 149). Others still have questioned whether they ought to be (following Smith 1994, pp. 75–6). Regarding the latter complaint, it should be noted that much hostility to *de dicto* moral desires is really just hostility to the suggestion that they exhaust a moral agent's motivational resources (Brink 1997). It is therefore important to emphasise that we do not take Alice's *de dicto* moral desires to be the only force that motivates her, morally speaking. Indeed, Correlation predicts that Alice will have a number of *de re* moral desires as well; desires to help the global poor, promote peace in the Middle East, or save the whales, for example. Our arguments do not rest upon any suspicious motivational monism.

We are now in a position to defend (what we take to be) the more interesting and controversial component of Cost. This was, recall, the assumption that there is a payoff specifically tied to Alice's belief that she has done the right thing. Some care is needed in spelling this out; for, on certain natural interpretations, it is neither interesting nor controversial.

On the one hand, it seems both empirically and phenomenologically obvious that people tend to *feel good* when they *do good*. The phenomenon of 'warm glow' suggests that positive feelings often accompany prosocial behaviour. (See Andreoni 1990; Andreoni and Miller 2002; Crumpler and Grossman 2008.) However, reverting to this truism won't suffice for our purposes. It is not sufficient that whenever Alice acts rightly, she experiences a *wff*. This *wff* needs to be tied

to her belief that she has done what is morally right.[8] Otherwise, Cost won't be plausible—indeed, there won't be any *cost* to be paid at all. If Alice's *wff* is ultimately explained by her *de re* moral desires, then it is not something that she stands to lose upon coming to believe moral nihilism.

The all-important question is therefore: when feeling good accompanies doing good, is it *because* the agent believes that she has done the right thing as such? To our knowledge, there have been scarcely any philosophical expeditions into the empirical literature bearing upon this question. But suggestive evidence is there. With a little effort and determination, the *wff* tied to (believing oneself to be) doing the right thing *de dicto* can be disentangled from any *wff* that may be tied to (believing oneself to be) doing the right thing *de re*.

We attend first to important work on moral identity and self-serving biases. It is a fact now widely recognised in psychology that people care deeply about their moral selves. Moral commitments are often described as "identity-defining" (Narvaez and Lapsley 2009, p. 243); they play an important role in defining *who one is*.[9] This idea has been borne out empirically in a variety of ways. (See for example, Blasi 1983, 1984; Monroe 1994, 2001, 2003.) In what follows, we argue that there's good reason to take this moral self-conception to include desires with moral content—*viz.*, desires to be a *morally good* person, or a person who does the *morally right* thing. Thus, an agent's moral self-conception is not just a matter of her aspiring to be a helpful person, or someone who promotes happiness. If we're correct, belief in moral nihilism has the potential to effect a radical upheaval in an agent's self-understanding, a loss to her sense of self.

The importance that people attach to their moral identity is reflected in the cognitive biases that support it. Self-serving biases are surprisingly common in the moral sphere. Almost everyone thinks that they are morally above average (Messick et al. 1985; Liebrand et al. 1986; Goethals 1986; Alison et al. 1989; Van Lange 1991). It may be tempting to chalk this up to a more general human tendency; people do, after all, tend to overestimate themselves. Yet this cannot be the whole story; for self-serving biases are *selective*. They are more pronounced in some domains than others. And they turn out to be especially pronounced in the moral domain. Whereas we strongly overestimate our moral credentials, we only weakly (if at all) overestimate our intelligence.

This asymmetry has been dubbed the *Muhammad Ali effect*,[10] and the evidence seems to have converged upon the following explanation for it (see Alison et al. 1989; Van Lange 1991). On the one hand, people do well to have a high opinion of themselves. A little embellishment can be a good thing (see Taylor and Brown 1988). But there are limits. If one is to take a flattering self-portrait

[8] We're here concerned with whether there's a *wff* tied to satisfying one's moral preferences. This is not to assume that there is no *wff* tied to satisfying one's non-moral preferences. Indeed, it is precisely because the latter seems so plausible that the disentanglement problem arises.

[9] This is not to deny that non-moral qualities may also be essential to one's identity. Nor is it to deny that there is interpersonal variation in how central agents take their moral self to be (see Aquino and Reed 2002). However, one's self-conception is rarely limited to non-moral attributes, and moral attributes are rarely inconsequential to who we take ourselves to be. Indeed, and as we will argue shortly, people often take the moral element of their identity to be *especially* important.

[10] See (Alison et al. 1989). Its namesake defended his suspiciously poor performance on an army mental exam by remarking, 'I only said I was the greatest, not the smartest' (Ali 1975, p. 129).

seriously, then that portrait must be *credible* (Gilbert and Cooper 1985). Interestingly, it turns out that a misleading moral resume tends to have more staying power than a misleading picture of one's cognitive potential. Even an agent who routinely reneges on her promises can hope to preserve a saintly image—by citing the greater good that was served by her actions, say. By way of contrast, it is difficult to persist in the illusion that one is a mathematical mastermind after having struggled to add up the dinner bill. The point is often framed in terms of verifiability: judgments of intelligence tend to be more "publicly and objectively verifiable", whereas judgments of moral calibre are more subject to "interpretational or attributional ambiguity" (Van Lange, p. 692; see also Alison et al. 1989). This explanation dovetails nicely with independent evidence suggesting that there is less room for self-serving manoeuvre where objectively verifiable abilities are concerned (e.g., Felson 1981).

Now for the philosophical takeaway. What we are concerned to emphasise is this: the explanation for the Muhammed Ali effect doesn't really have legs if the moral image that self-serving biases serve to protect lacks a *moral* element; if it is merely an image of an individual who tends to help, for instance. Helpful behaviours are, after all, just as publicly and objectively verifiable as intelligent behaviours; it is hard to persist in the illusion that one is a helpful person if one never rises to the occasion when the opportunity presents itself. But the foregoing explanation *does* have legs if we suppose that the moral self-image is an image of an agent who does what is *morally right*. Given this supposition, the interpretational ambiguity of self-directed moral judgments makes sense; an agent who never seizes the opportunity to help may very well persist in believing that she is *morally good*—perhaps there is simply more important moral work to be done than attending to those in one's immediate vicinity. In prizing their moral identity, then, human agents plausibly prize being morally good people. But if moral nihilism is true, then this self-image quickly breaks down; there is no *moral goodness* for anyone to instantiate—no pride to be taken in one's moral accomplishments, nor any virtue to be cultivated throughout one's life. A belief in moral nihilism therefore poses a risk to an agent's sense of self.

We will now suggest that the costs of believing moral nihilism are greater still. Drawing upon work in media psychology, we sketch a provisional case for thinking that there is a positive experiential aspect associated with believing one's moral preferences to be satisfied.

Moral judgment is heavily implicated in the consumption of dramatic entertainment. Moral assessment determines the extent to which a character is liked or disliked; and viewers also find a drama more enjoyable when characters get their just desserts (Zillmann and Bryant 1975; Zillmann and Cantor 1977; Raney and Bryant 2002; Raney 2002, 2005). Importantly, the latter phenomenon really does appear to be mediated by *moral judgment*. The evidence for this is fairly straightforward: vary the moral standards, and the enjoyment of dramatic entertainment will vary as well.

Studies involving children are particularly telling. Zillmann and Bryant (1975) exposed four and eight-year-olds to three fairy tales, which differed only in their portrayal of the villain's fate: respectively, (i) pardoned, (ii) proportionately punished, and (iii) excessively punished. These two populations were chosen for a reason. Children around four years of age have a penchant for excessive retribution (the more suffering, the better). By eight, they typically develop a preference for proportionate retaliation, which informs their sense of justice.

In keeping with their hypothesis, Zillman and Bryant found that 4-year-olds' enjoyment of the fairy tale increased with the severity of punishment, whereas 8-year-olds enjoyed the fairy tale most in the second condition. These results suggest that enjoyment was tied to the satisfaction of subjects' moral preferences. Similar studies conducted on children support this interpretation (e.g., Zillmann and Cantor 1976), as do studies involving adults (e.g., Raney 2005).

It's been a long journey. Let us summarise the foregoing defence of Cost, and conclude with some words of caution. We have argued that (i) human agents have moral preferences, and that (ii) there are payoffs tied to the satisfaction of these preferences—payoffs that would no longer be available to them in a world where they believe the truth of moral nihilism. If this is right, then nihilistic belief comes at a price. It is a difficult question just how *high* that price is. This is likely to depend on (among other things) the extent to which the phenomena that we explore are representative of agents' moral experiences more generally. If, for example, there is a strong positive experiential aspect associated with the satisfaction of *all sorts of* moral preferences (i.e., not only preferences involving desert), then our case starts to look even stronger—likewise if cultivating a moral identity is the strongest prospect for injecting meaning into one's life.

We see no clear path to adjudicating the latter issue at present. Psychologists haven't tested the potential consequences of belief in moral nihilism directly—at least not to our knowledge. To some extent, then, the evidence for our empirical assumptions must be mined rather than hand-picked. There is certainly room to debate the degree to which these results can be generalised, and how high the relevant costs would be. Nonetheless—and we emphasise—there would be *costs*. Human beings do not merely attach credence to the idea that they inhabit a moral world. They also attach (at least some) utility to being in one.

Before moving on to a defence of No Compensation, it's worth addressing an important concern with what we've been saying. Some may worry that we've over-simplified things by assuming that recognising the truth of moral nihilism merely involves believing it. This overlooks the sophisticated cognitive strategies that often accompany defences of moral nihilism. Some nihilists adopt a conservationist strategy, which recommends outright belief in nihilism only in particular contexts (for example, a philosophy classroom); in everyday life, nihilists are encouraged to hold onto their ordinary moral beliefs (Olson 2014). Others advocate a fictionalist approach, which advises nihilists to make-believe first order moral propositions (Joyce 2001).

It may be supposed that insofar as Alice adopts either of these strategies, there would be no threat to her *wff*, and thus, no cost to her coming to discover the truth of moral nihilism. If she is a conservationist, then she can continue to believe that she acts rightly in helping Bob, and to experience a *wff* when she does so. Matters are more complicated as far as the fictionalist strategy is concerned, but there is evidence to suggest that moral make-belief is capable of engaging similar emotions to moral belief, including (perhaps) the *wff* to which our arguments appeal (see Joyce 2001, p. 197).

Each of these are well-developed responses to the 'what next?' question for moral nihilists, and we cannot hope to attempt a thorough assessment of them here. Still, it's worth noting why we do not take these proposals to be devastating for Cost. We'll begin with the fictionalist option. It is not implausible that make-belief can produce similar emotional experiences to belief. (Few feel warm and fuzzy inside when watching man-eating spiders on their television.) But

whatever affective responses make-beliefs are capable of eliciting, these seem importantly different from the affective responses triggered by belief (see Nichols and Stich 2000; Nichols 2006). One who believes that there is a poisonous spider lurking somewhere in their bedroom is apt to feel a very real kind of fear—a fear that someone who make-believes that a rock is a spider is unlikely to experience. Thus, it seems that the fictionalist option would at best soften the blow for Alice, in virtue of preserving something like a *wff*—but there is still a cost.

Now to the conservationist. Even assuming that it were possible for Alice to hold onto beliefs that she knows to be false—a claim with which many would take issue—her belief that moral nihilism is true may very well cross her mind in everyday contexts. She may, for example, find herself thinking that helping Bob with his sofa would not really be the right thing to do at all (since nothing is right). This is, after all, something that she believes, and it seems relevant to deciding whether or not she ought to put a strain on her back. Attending to this belief, would, however, likely prevent Alice from experiencing the *wff* that she usually experiences when she does what she believes to be right. Importantly, our suggestion is not that moral nihilism would always be on Alice's mind. The point is simply that there's no reason to think that it wouldn't cross her mind on at least some occasions (outside the philosophy classroom). Even if Alice (qua conservationist) does sometimes experience a *wff*, then, these experiences are likely to be far less reliable than those of her realist analogue. For these reasons, we are inclined to think that adopting a post-nihilist strategy would at best reduce the costs of nihilist belief for Alice; it would not remove those costs completely.

7 No Compensation

We have argued that there is something that Alice stands *to lose* should she inquire after the truth of moral nihilism. This does not itself establish that it would be irrational for Alice to accept the oracle's offer; for there may also be something that she stands *to gain*. Nihilistic belief may come at a price. But perhaps that price is worth paying. It is our contention that this is unlikely to be true for an ordinary human agent like Alice. Whatever compensation (if any) she receives, it won't suffice to offset the costs of nihilistic belief. We'll now consider some challenges to this claim: objections from *goody-two-shoes*, *philosophers*, and *the value of true belief*.

(We should like to remind the reader that we have already dispatched one potential line of resistance: the idea that Alice would be *liberated from the shackles of morality*. Insofar as Correlation is true, morally recommended actions are, for the most part, actions that Alice has good independent reason to pursue. Thus, it's not as though, upon coming to believe moral nihilism, Alice will finally be free to do what she *really* wants.)

The first challenge to No Compensation comes in the form of an overzealous goody-two-shoes. It remains possible that Alice is excessively scrupulous. Perhaps she is extremely guilt-prone, carrying the weight of the world on her shoulders following even the slightest misdemeanour. Or maybe she is extremely anxious about doing what is right, planning her schedule well in advance to minimise the potential for moral mishaps. To be sure, Correlation may still apply to Alice—her non-moral preferences may tend to favour acting in ways that are morally desirable. But if moral nihilism is true, then she can continue

to do so *without* the associated guilt and anxiety. Presumably, these unpleasant experiences are something she can do without. And if they're unpleasant enough, then it may be worth her while to forego any *wff* in order that she may finally be rid of them.

We doubt that the moral inner-life of Alice's goody-goody counterpart is representative of human agents more generally. Guilt and anxiety may very well be features of our moral experience. But they are unlikely to be anywhere near as pervasive for an ordinary human being as they are for a relentless goody-two-shoes. Here, it is instructive to consider real people who *do* satisfy the above description: those who suffer from *scrupulosity*. Scrupulosity patients exaggerate the moral gravity of their behaviour, are often paralysed with moral indecision, and regularly revisit and scrutinise their moral past.[11] For a very select group of individuals, then, the goody-two shoes challenge applies; scrupulosity patients would be well-advised to inquire after the truth of moral nihilism. But insofar as such persons form a small pathological population, we are not inclined to regard them as a threat to our generalisation.

On a different note, it has been put to us that we ourselves are walking counterexamples to No Compensation. As we hinted earlier, *philosophers* are likely to have strong incentives to inquire after the truth of moral nihilism. They may want to attend an upcoming conference, or publish a paper on the topic. Or (being philosophers) they may simply enjoy pondering life's great questions. For a philosopher, even a lifetime's worth of warm fuzzy feelings may be a small price to pay for news on the nihilist front. We are open to this possibility. However, it must be admitted that philosophers aren't representative of the general human population. Indeed, they are grossly *unrepresentative*. So it is no threat to our arguments if philosophers ought to learn more about moral nihilism. Insofar as there is an exception here, it is surely a principled one.

A final challenge alleges that Alice may simply value having true beliefs *for its own sake*.[12] Humans are curious creatures. Sometimes, we just want to know the truth about things, independently of any ends this may serve (see Goldman 1986, p. 98; Kvanvig 2003, p. 41). Are there more than ten thousand lightbulbs in the Sydney Opera House? How many blades of grass were in the Hanging Gardens of Babylon? What is the exact number of wildebeest presently sweeping majestically across the African plane? These seem like pointless questions; their answers would be of no obvious practical benefit to Alice. Nonetheless, she may still prefer to learn them. The same may be true of moral nihilism; whatever Alice loses in warm fuzzy feelings may be compensated for in the currency of true beliefs.

We are not persuaded. The transition from human curiosity to an intrinsic concern for truth ought to be viewed with a healthy suspicion. The answers to seemingly pointless questions may very well have *non-obvious* benefits for Alice. She may want closure (think about having to miss the end of an exciting football match). She may enjoy entertaining her friends with surprising titbits over dinner. Alternatively, she may subscribe to a better-safe-than-sorry policy. (What if she one day had to bet on the number of lightbulbs in the Sydney

[11] For an excellent overview, see (Miller and Hedges 2008), and the references therein. For a fictional example, see Chidi from 'The Good Place'.

[12] This is importantly distinct from the claim that truth *has* value for its own sake. What's important give how we're understanding utility is how truth figures in an agent's preferences, whether the agent herself cares about having true or more accurate beliefs.

Opera House?). Following (Brady 2009, p. 270), an agent's interest in the truth may be exhausted by her interest in some "unacknowledged practical goal." Curiosity need not reflect an interest in the truth as such.

Of course, these confounding factors can be stipulated away. Suppose now that Alice must choose between two worlds, ω_1 and ω_2, which differ only in the following respect: at ω_2, Alice believes some pointless truth T. Suppose further that T really is pointless: *none* of the outcomes of Alice's choices will ever hang upon it. For our part, we see nothing to clearly recommend one world over the other. It certainly wouldn't strike us as *bizarre* if someone were indifferent between the two. That being said, it wouldn't strike us as bizarre if an agent preferred ω_2 to ω_1 either. After all, preference is cheap; even very small factors can make a difference when nothing else is at stake. What we do want to emphasise, however, is this: inasmuch as there is a preference here, it is at best an *extremely miniscule* one. Insofar as ordinary human agents like Alice do value truth, it's not clear that they value it very much.[13] Alice may prefer to pursue the truth when *literally nothing else is at stake*. But if this is truly the best that truth can do, then No Compensation remains in good stead; for there clearly is something at stake when it comes to belief in moral nihilism.

8 Complications

We promised we'd say a bit more about the simplifications made in the argument of §4, so we'll do that before concluding. There are two main points to discuss.

First, we've been supposing throughout that Alice considers only the immediate effects that accepting or rejecting the oracle's offer will have with respect to her decision whether to help Bob. This greatly over-simplifies her real decision situation: any minimally rational agent like Alice shouldn't be considering only how a change in her information state might affect her outcomes at 10am on Tuesday, March 15th when Bob is attempting to enlist her help. If she learns that moral nihilism is true now, then that change of belief is liable to have far-reaching implications for how well her preferences are satisfied in many of her future 'moral' choices, and Alice ought to take these implications under consideration to the extent she's able.

Now, we cannot feasibly recreate in a simple decision tree all of the temporally downstream factors that might matter to Alice's decision that she's aware of. But we see Alice's decision whether to help Bob as *representative* of the kinds of moral choice situations that an agent is likely to face over the course of a life-time. If so, then Alice needs to weigh up the *average* cost of losing the warm fuzzy feeling *over the course of a lifetime*, versus whatever benefits she may receive now or in the long run from learning more about the truth of nihilism. We take it that our arguments in support of Correlation support the claim that the case of the sofa is representative; and that our arguments in support of Cost and No Compensation can be naturally adduced in favour of the claim that the (potential) loss of a *wff* over the course of a lifetime isn't worth whatever benefits come with having more true beliefs about nihilism.

The second complicating factor is not so easy to deal with. Alice's decision situation in the sofa case is somewhat more complicated if we relax the assumption (implicit in our use of the values x and q at both nodes 1 and 3 in Figure 3)

[13] Wren (2017) uses these considerations to support a parallel claim about truth's intrinsic value; at best, truth is the *least* valuable intrinsic good.

that there's no difference in the utility that attaches to *Help*/*Don't Help* when $\mathcal{C}r(\text{Realism}) = 0.9$ and when $\mathcal{C}r(\text{Realism}) = 1$, at those worlds where Realism holds. If this is false then Figure 3 is a misrepresentation, and our earlier formal argument needs to be generalised to accommodate the possibility that the utilities of those outcomes might be different.

Hence, let $x^\star$ designate the value of *Help* at worlds where Realism is true and Alice is *certain* that it's true, and let $q^\star$ designate the value of *Don't Help* at the same kind of worlds. Given this, we'll get the same result in favour of Alice choosing *Reject* if we keep assumption A1 as is, and replace assumptions A2 and A3 with:[14]

A2*. $x^\star \geq q^\star$
A3*. $\big((x^\star - x) \times \mathcal{C}r_0(\text{Realism})\big) \; < \; \big((y - \max\{z, s\}) \times \mathcal{C}r_0(\text{Nihilism})\big)$

A2* should be more or less just as plausible as A2: if *Help* is more valuable than *Don't Help* at worlds where Realism is true and $\mathcal{C}r(\text{Realism}) = 0.9$, then it should also be more valuable when $\mathcal{C}r(\text{Realism}) = 1$. So we take it that A2* is supported already by Correlation. The harder one to justify is A3*.

What A3* says is not at all easy to put into plain English. Nevertheless (take a deep breath): if there's any increase/decrease to the utility of *Help* at Realism-worlds that would result from a shift from $\mathcal{C}r(\text{Realism}) = 0.9$ to $\mathcal{C}r(\text{Realism}) = 1$ (weighted by Alice's original degree of belief in Realism), then that increase/decrease is less/greater than any decrease/increase in the utility of the optimal choice at Nihilism-worlds that would result from a shift from $\mathcal{C}r(\text{Nihilism}) = 0.1$ to $\mathcal{C}r(\text{Nihilism}) = 1$ (weighted by Alice's original degree of belief in Nihilism). If we assume that $y > z > s$ and $x^\star > x$, then there's a straightforward consequence of A3* for the purposes of our argument: the higher $\mathcal{C}r_0(\text{Realism})$ is, the less any *cost* (the difference between y and z) matters, and the more any *gain* (the difference between $x^\star$ and x) matters. So, for example, to get the conclusion that Alice should choose *Reject*,

- If $\mathcal{C}r_0(\textit{Realism}) = 1/2$, the cost must be more than the gain
- If $\mathcal{C}r_0(\textit{Realism}) = 2/3$, the cost must be more than 2 times the gain
- If $\mathcal{C}r_0(\textit{Realism}) = 9/10$, the cost must be more than 9 times the gain
- If $\mathcal{C}r_0(\textit{Realism}) = 99/100$, the cost must be more than 99 times the gain

The upshot here is that there's a further dimension that needs to be considered before we can draw any conclusions about Alice's decision: *if* there's a difference between $x^\star$ and x, then Alice's initial degree of belief in Realism matters.

So what reason would we have for thinking that there's a difference between $x^\star$ and x? We've already argued that having slightly more accurate beliefs isn't (very) valuable for its own sake, so at most there's only a tiny gain in this case that might come from 'epistemic utility'. But perhaps Alice also gets *more* of a *wff* (or a more valuable *wff*) for doing what she thinks is the right thing, the more confident she is that there *is* a right thing to do. This seems plausible

[14] *Proof.* Let $\mathcal{C}r_0(\text{Nihilism}) = a$, and $\mathcal{C}r_0(\text{Realism}) = b$. From A1, $\mathcal{EU}_0(\textit{Reject}) = ay + bx$; and from A2*, $\mathcal{EU}_0(\textit{Accept}) = au + bx^\star$, where u designates whichever is the largest of z or s. Now $bx^\star - bx = b(x^\star - x)$, which rearranged is $bx^\star = bx + b(x^\star - x)$. Likewise, $ay - au = a(y - u)$; so, $au = ay - a(y - u)$. Thus, $\mathcal{EU}_0(\textit{Accept}) = ay - a(y - u) + bx + b(x^\star - x)$. Or in other words, $\mathcal{EU}_0(\textit{Accept}) = \mathcal{EU}_0(\textit{Reject}) - a(y - u) + b(x^\star - x)$. By A3*, $a(y - u)$ is more than $b(x^\star - x)$, so $\mathcal{EU}_0(\textit{Accept}) < \mathcal{EU}_0(\textit{Reject})$.

enough, so we're happy to take the suggestion on board. The question now is how it impacts on our conclusion.

We're inclined to think that it doesn't make too much difference. In particular, it's likely that the higher $\mathcal{C}r_0$(Realism) is,

(a) the smaller the difference between $x^\star$ and x, and
(b) the greater the difference between y and z.

That is, if there *is* a positive correlation between the amount of *wff* and one's confidence in Realism, then we'd expect very small increases in one's confidence to generate correspondingly small changes in the amount of *wff*. Moreover, we'd expect that the correlation isn't perfectly linear—in particular, we think it's plausible that after a certain point, increasing one's confidence in Realism comes with diminishing returns in increases in amount of *wff*. As one's confidence in Realism gets very high, there won't be much difference to the amount of *wff* that arises from becoming certain of Realism. And on the flip side, the higher $\mathcal{C}r_0$(Realism) is, the more *wff* there is to lose from becoming certain in Nihilism. Thus, if $\mathcal{C}r_0$(Realism) is quite high indeed—say, $^{99}/_{100}$—then we'd expect the difference between $x^\star$ and x to be correspondingly *tiny*, and the difference between y and z to be much larger. And a cost that's 99 times larger than a very small amount need not be very large at all.

We've represented Alice's initial belief in Realism at 90% confidence. We don't have any strong empirical evidence to back that up—it's a rough estimate supported by tutorial discussions with undergraduate philosophy students in ethics and metaethics classes in English-speaking universities. Make of that what you will. But the point just raised is quite general. While it's true that the higher $\mathcal{C}r_0$(Realism) is, the greater the difference between the cost and the gain must be for our argument to go through, there's also a trade-off given that the $\mathcal{C}r_0$(Realism) is, the greater the cost and the less the gain. In light of that trade-off, and in conjunction with the points raised in §§5–§7, we take it that there's a strong case that Alice (probably) ought to reject.

9 Conclusion

We have argued that nihilistic belief is not cost free for an ordinary human agent like Alice. Moreover, this cost is unlikely to be offset by anything that an agent stands to gain in learning more about moral nihilism. True beliefs may count for something. But it is doubtful that any modicum of value they have would compensate for the costs of nihilistic belief over a lifetime. We have defended these assumptions about Alice's preferences on both intuitive and empirical grounds. If they are roughly correct, then the average person would be well-advised not to inquire after the truth of moral nihilism.

It's easy to mistake our conclusion for a Pascalian one. As a previous referee lamented, we seem to follow Pascal in recommending "pleasing lies over truth". However, nothing that we say supports this conclusion more generally. Presumably, there are many cases where opting for convenient falsehoods would not maximise an agent's expected utility; in such cases, our reasoning would counsel pursuit of the truth. Some among our readership may have hoped for the result that it is never rational or advisable to pursue lies over truth. But if that's so, we suspect that their real beef is with the general theory of rationality

that we assume; and a defence of that theory is not a burden that this paper can reasonably be expected to bear.

Though our conclusions apply in the first instance to ordinary human agents—a population from which philosophers have been swiftly and unceremoniously evicted—they do of course have implications for philosophers as well. According to a well-known tradition of thought, philosophers would sometimes do best to keep uncomfortable truths hidden from non-philosophers or (as they are sometimes more affectionately known) 'the folk'.[15] And our arguments suggest that moral nihilism is apt to be a very uncomfortable truth indeed. Philosophers, then, may very well do best to keep their nihilistic opinions to themselves.[16]

References

Ali, M. (1975). *The Greatest: My own story*. New York: Random House.

Alison, S. T., D. M. Messick, and G. R. Goethals (1989). On being better but not smarter than others: The Muhammad Ali effect. *Social Cognition 7*, 275–296.

American Psychiatric Association, . (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5 ed.). Arlington, VA: American Psychiatric Publishing.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal 100*, 464–477.

Andreoni, J. and J. Miller (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica 70*, 737–753.

Aquino, K. and A. Reed (2002). The Self-Importance of Moral Identity. *Journal of Personality and Social Psychology 86*, 1423–1440.

Arpaly, N. (2015). Huckleberry Finn Revisited: Inverse Akrasia and Moral Ignorance. In R. Clarke, M. Mckenna, and A. Smith (Eds.), *The Nature of Moral Responsibility: New Essays*, pp. 141–156. New York: Oxford University Press.

Batson, C., B. Duncan, P. Ackerman, T. Buckley, and K. Birch (1981). Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology 40*, 290–302.

Batson, C., J. B. Dyck, J. R. Batson, and A. Powell (1988). Five Studies Testing Two New Egoistic Alternatives to the Empathy-Altruism Hypothesis. *Journal of Personality and Social Psychology 55*, 52–77.

Blasi, A. (1983). Moral Cognition and Moral Action: A Theoretical Perspective. *Developmental Review 3*, 178–210.

Blasi, A. (1984). Moral identity: Its role in moral functioning. In W. M. Kurtines and J. J. Gewirtz (Eds.), *Morality, Moral Behavior and Moral Development*, pp. 128–139. New York: John Wiley and Sons.

[15] Sidgwick's (1984, Ch. 5) so-called 'government house utilitarianism' is a nice illustration. For an application of this idea to moral nihilism, see Cuneo and Christy (2011).

[16] Thanks are due to Heather Browning, Michael Huemer, Norbit Paulo, Kim Sterelny, Jack Woods, Pekka Vayrynen, and audiences at the University of Sydney, the 2018 Evolution & Epistemology conference (Utrecht), Leeds CMM workshop, ACU 2018 international moral epistemology conference (Melbourne), and the 2018 AAP/AAPNZ (Wellington). We apologise for any omissions. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 703959.

Bradley, S. and K. Steele (2016). Can Free Evidence Be Bad? Value of Information for the Imprecise Probabilist. *Philosophy of Science 83*, 1–28.

Brady, M. (2009). Curiosity and the value of truth. In A. Haddock, A. Millar, and D. Pritchard (Eds.), *Epistemic value*, pp. 265–283. New York: Oxford University Press.

Brink, D. O. (1997). Moral Motivation. *Ethics 108*(1), 4–32.

Buchak, L. (2013). *Risk and Rationality*. Oxford: Oxford University Press.

Crumpler, H. and P. J. Grossman (2008). An experimental test of warm glow giving. *Journal of Public Economics 92*, 1011–1021.

Cuneo, T. and S. Christy (2011). The Myth of Moral Fictionalism. In M. Brady (Ed.), *New Waves in Metaethics*, pp. 85–102. Basingstoke: Palgrave MacMillan.

Felson, R. B. (1981). Ambiguity and Bias in the Self-Concept. *Social Psychological Quarterly 44*, 64–69.

Ferguson, T. J., H. Stegge, and I. Damhuis (1991). Children's understanding of guilt and shame. *Child Development 62*, 827–839.

Flanagan, O. J. (2017). *The Geography of Morals: Varieties of Moral Possibility*. Oxford University Press.

Fumagalli, R. (2013). The Futile Search for True Utility. *Economics and Philosophy 29*, 325–347.

Gilbert, D. T. and J. Cooper (1985). Social psychological strategies of self-deception. In M. Martin (Ed.), *Self-deception and self-understanding*, pp. 75–94. Lawrence, KS: University of Kansas Press.

Goethals, G. R. (1986). Fabricating and ignoring social reality: Self-serving estimates of consensus. In J. M. Olson, C. P. Herman, and M. P. Zanna (Eds.), *Social comparison and relative deprivation: The Ontario Symposium*, pp. 147–170. Hillsdale, NJ: Lawrence Erlbaum.

Goldman, A. (1986). *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

Good, I. J. (1967). On the Principle of Total Evidence. *British Journal for the Philosophy of Science 17*, 319–321.

Hepach, R., A. Vaish, and M. Tomasello (2013). A New Look at Children's Prosocial Motivation. *Infancy 18*, 67–90.

Hopfensitz, A. and E. Reuben (2009). The importance of emotions for the effectiveness of social punishment. *The Economic Journal 119*, 1534–1559.

House, J. S., K. R. Landis, and D. Umberson (1988). Social relationships and health. *v 241*, 540–545.

Joyce, R. (2001). *The Myth of Morality*. Cambridge, MA: Cambridge University Press.

Joyce, R. and S. Kirchin (2010). Introduction. In R. Joyce and S. Kirchin (Eds.), *A World Without Values: Essays on John Mackie's Moral Error Theory*, pp. ix–xxiv. Dordrecht: Springer.

Kalf, W. F. (2013). Moral Error Theory, Entailment and Presupposition. *Ethical Theory and Moral Practice 16*(5), 923–937.

Kvanvig, J. (2003). *The Value of Knowledge and the Value of Understanding*. Cambridge: Cambridge University Press.

Liebrand, W. B. G., D. M. Messick, and F. J. M. Wolters (1986). Why we are fairer than others: A cross-cultural replication and extension. *Journal of Experimental Social Psychology 22*, 590–604.

Liszkowski, U., M. Carpenter, T. Striano, and M. Tomasello (2006). Twelve- and 18-month olds point to provide information for others. *Journal of Cognition and Development 7*, 173–187.

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. New York: Penguin.

Messick, D. M., S. Bloom, J. P. Boldizar, and C. D. Samuelson (1985). Why we are fairer than others. *Journal of Experimental Social Psychology 21*, 480–500.

Miller, C. H. and D. W. Hedges (2008). Scrupulosity disorder: An overview and introductory analysis. *Journal of Anxiety Disorders 22*, 1042–1058.

Miller, K. (2010). On contingently error-theoretic concepts. *Americal Philosophical Quarterly 47*(2), 181–190.

Monroe, L. (1994). But What Else Could I do? Choice, Identity, and a Cognitive-perceptual Theory of Ethical Political Behaviour. *Political Psychology 15*, 201–226.

Monroe, L. (2001). Morality and the Sense of Self: The Importance of Identity and Categorization for Moral Action. *American Journal of Political Science 45*, 491–507.

Monroe, L. (2003). How Identity and Perspective Constrain Moral Choice. *International Political Science Review 24*, 405–424.

Narvaez, D. and D. K. Lapsley (2009). Moral identity, moral functioning, and the development of moral character. In D. M. Bartels, C. W. Bauman, L. J. Skitka, and D. L. Medin (Eds.), *Psychology of Learning and Motivation: Moral Judgment and Decision Making*, pp. 237–274. Elsevier.

Nelissen, R. M. A. and M. Zeelenberg (2009). Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions. *Judgment and Decision Making 4*, 543–553.

Nichols, S. (2006). Just the Imagination: Why Imagining Doesn't Behave Like Believing. *Mind & Language 21*(4), 459–474.

Nichols, S. and S. Stich (2000). A cognitive theory of pretense. *Cognition 74*, 115–147.

Oddie, G. (1997). Conditionalization, Cogency, and Cognitive Value. *British Journal for the Philosophy of Science 48*(4), 533–541.

Olson, J. (2014). *Moral Error Theory: History, Critique, Defence*. Oxford: Oxford University Press.

Ramsey, F. P. (1931). General Propositions and Causality. In R. B. Braithwaite (Ed.), *The Foundations of Mathematics and other Logical Essays*, pp. 237–255. London: Stephen Austin and Sons.

Raney, A. A. (2002). Moral judgment as a predictor of enjoyment of crime drama. *Media Psychology 4*, 305–322.

Raney, A. A. (2005). Punishing Media Criminals and Moral Judgment: The Impact on Enjoyment. *Media Psychology 7*, 145–163.

Raney, A. A. and J. Bryant (2002). Moral judgment and crime drama: An integrated theory of enjoyment. *Journal of Communication 52*, 402–415.

Sayre-McCord, G. (1986). The Many Moral Realisms. *The Southern Journal of Philosophy 24*, 1–22.

Seip, E. C., W. W. Van Dijk, and M. Rotteveel (2014). Anger motivates costly punishment of unfair behaviour. *Motivation and Emotion 38*, 578–588.

Sidgwick, H. (1984). *The Methods of Ethics*. London: Macmillan and Co.

Skyrms, B. (1990). *The Dynamics of Rational Deliberation*. Cambridge, MA: Harvard University Press.

Smith, M. (1994). *The Moral Problem*. Oxford: Blackwell.

Stocker, M. (1976). The Schizophrenia of Modern Ethical Theories. *Journal of Philosophy 73*, 453–66.

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

Tangney, J. P., J. Stuewig, E. Malouf, and K. Youman (2013). Communicative Functions of Shame and Guilt. In K. Sterelny, R. Joyce, B. Calcott, and B. Fraser (Eds.), *Cooperation and its Evolution*, pp. 485–502. Cambridge: MIT Press.

Taylor, S. E. and J. D. Brown (1988). Illusion and Wellbeing: A Social Psychological Perspective on Mental Health. *Psychological Bulletin 103*, 193–210.

Van Lange, P. A. (1991). Being Better but Not Smarter than Others: The Muhammad Ali Effect at Work in Interpersonal Situations. *Personality and Social Psychology Bulletin 17*, 689–693.

Vonasch, A. J., T. Reynolds, B. M. Winegard, and R. F. Baumeister (Forthcoming). Death Before Dishonor: Incurring Costs to Protect Moral Reputation. *Social Psychological and Personality Science Online First*, 1–10.

Warneken, F. and M. Tomasello (2006). Altruistic helping in human infants and young chimpanzees. *Science 311*, 1301–1303.

Warneken, F. and M. Tomasello (2007). Helping and cooperation at 14 months of age. *Infancy 11*, 271–294.

Western, B., J. R. Kling, and D. F. Weiman (2001). The labor market consequences of incarceration. *Crime and Delinquency 47*, 410–427.

Wren, C. (2017). Truth is not (very) intrinsically valuable. *Pacific Philosophical Quarterly 98*, 108–128.

Zillmann, D. and J. Bryant (1975). Viewer's moral sanction of retribution in the appreciation of dramatic presentations. *Journal of Experimental Social Psychology 11*, 572–582.

Zillmann, D. and J. Cantor (1976). A disposition theory of humor and mirth. In T. Chapman and H. Foot (Eds.), *Humor and laughter: Theory, research, and application*, pp. 93–115. London: Wiley.

Zillmann, D. and J. Cantor (1977). Affective responses to the emotions of a protagonist. *Journal of Experimental Social Psychology 13*, 155–165.