# Praiseworthy Motivations

ZOË A. JOHNSON KING
*University of Michigan*

*I'm just a soul whose intentions are good;*
*Oh Lord, please don't let me be misunderstood.*
— Nina Simone

## 1. Introduction

In this paper I defend the following thesis:

> **THESIS**: If motivation by rightness *de re* is praiseworthy, then so is motivation by rightness *de dicto*.

My thesis is equivalent to the negation of a popular view, as follows:

> **FALSE VIEW**: Motivation by rightness *de re* is praiseworthy, but motivation by rightness *de dicto* is not praiseworthy.

As I have indicated, I think that this view is false. In defending my thesis, I will be arguing against it.

Let me begin by providing some context, which will help to explain what both my thesis and the false view are talking about.

We all face morally difficult decisions. Life is complicated, lots of things are morally significant, and it is frequently hard to tell precisely what is morally required of us.

Some people approach morally difficult decisions thinking something like "I just want to do the right thing in this situation, whatever it is". These people then engage in moral reflection. When they think that they have worked out what the right thing to do in their situation is – or, at least, when they have a good guess – they then do it, *because it's the right thing to do*.

Other people have more concrete concerns. Faced with morally difficult decisions, they think about what would be *honest*, or *kind*, or *fair*, or about what's in the *interests* of the people concerned, rather than thinking about what's morally right *per se*. These people then choose a course of action based on its having one of these more concrete features, rather than choosing it based on its moral rightness.

But some of the more concrete features by which these people are motivated are among the features that *make* courses of action morally right – the so-called "right-making features". So, although people moved by these concerns are not motivated by the moral rightness of their actions *per se*, they are motivated by the very features that their actions' moral rightness consists in.

Philosophers distinguish between these two types of moral concern. We say that the first type of person is motivated by rightness *de dicto.* This means that she is explicitly concerned with acting morally rightly. We say that the second type of person is motivated by rightness *de re*.[1] This means that she is concerned with the very features of actions that their moral rightness in fact consists in.

This way of drawing the distinction comes from Michael Smith's discussion of praiseworthy motivations in *The Moral Problem* (1994). Smith drew the distinction in order to disparage motivation by rightness *de dicto*. He denied that this type of moral concern can be part of what it is to be a good person, claiming that "good people care non-derivatively about honesty, the weal and woe of their children and friends, the well-being of their fellows, people getting what they deserve, justice, equality, and the like, not… doing what they believe to be right, where this is read *de dicto* and not *de re.* Indeed, commonsense tells us that being so motivated is a fetish or moral vice, not [a] moral virtue" (*ibid.,* p.75).

A lively debate ensued as to whether Smith is right about this. Some (e.g. Lillehammer 1996, Svavarsdóttir 1999, Olson 2002, Aboodi 2016) argued that Smith's so-called "commonsense" intuition is mistaken or misleading. Others (e.g. Miller 1996, Copp 1997, Dreier 2000, Zangwill 2003, Toppinen 2004, Strandberg 2007) reported sharing it. And Smith himself later clarified that what he really thinks is praiseworthy is not motivation by rightness *de re*, but rather an "executive virtue" by which agents' intrinsic motivations reliably track their beliefs about what moral rightness consists in (Smith 1996, pp.176-177).

This literature is already crowded with disputants, and I will not wade into it here.

I am interested in a different literature. While metaethicists have been discussing Smith's position, the false view – that motivation by rightness *de re* is praiseworthy, but motivation by rightness *de dicto* is not – has become popular in normative ethics.

The most developed statement of the false view is from Nomy Arpaly and Timothy Schroeder (2013). Their view is that good will is a matter of intrinsically desiring that which is in fact right or good, *de re*, and/or *not* intrinsically desiring that which is in fact wrong or bad, *de re*, and that good will in turn is both necessary and sufficient for virtue and praiseworthiness. They say that "it is the right or good conceptualized in the way preferred by the correct normative theory, and not merely via the concept RIGHT or GOOD, that motivates people moved by good will" (*ibid*., p.177). Thus they explicitly consider and reject the possibility that it might be praiseworthy to be motivated by the right or good *de dicto*. On this view, what matters for good will, virtue, and praiseworthiness is that an agent is motivated by the very features that rightness or goodness in fact consists in.

Arpaly and Schroeder argue for their view by comparing agents, some of whom are motivated by rightness *de dicto* but not *de re*, others of whom are motivated by rightness *de re* but not *de dicto*, and all of whom accept false and pernicious moral theories (such as a pro-slavery moral theory). Those who are motivated by rightness *de dicto* do what is in fact morally wrong, believing it to be right, since it is right according to their false theory. And those who are motivated by rightness *de re* do what is in fact morally right, believing it to be morally wrong, but being undeterred by this since they are uninterested in rightness *de dicto*. Arpaly and Schroeder suggest that the latter (*de re* morally motivated) agents seem more praiseworthy than the former. They do not use the term "fetishist" to describe the former (*de dicto* morally motivated) agents, but their intuition here is roughly the same as Smith's. I will object to this way of comparing cases in §3; for now, I simply note that the false view has received some sophisticated recent defenses.

[1] For some background on the *de dicto*/*de re* distinction see McKay and Nelson (2014).

Other authors, writing on related topics, have simply assumed that Smith is correct. For example, Brian Weatherson (2014, pp.152-154) deploys the fetishism intuition as the key move in his argument against "moral hedging", which involves taking account of one's credences in moral theories when deciding what to do.[2] Weatherson argues that someone would only engage in moral hedging if she were motivated by rightness *de dicto*. Then he suggests that this shows moral hedging to be objectionable, as it "is not possible without falling into the bad kind of moral fetishism that Smith rightly decries" (*ibid.*, p.154). Weatherson is explicit about the fact that this is his main argument against moral hedging.

Similarly, Julia Markovits (2010, p.204) deploys the fetishism intuition in her argument for the claim that someone performs an act with moral worth just in case she is motivated to do the morally right thing by the features that make it morally right. She too defers to Smith, arguing that someone who does the right thing because it is right "seems guilty of a kind of fetishism (to borrow a phrase from Michael Smith)" (*ibid*.). This is Markovits' main argument against the traditional Kantian idea that it might be sufficient for moral worth that an agent does the right thing because it is right.[3]

So the distinction between motivation by rightness *de dicto* and *de re*, and the associated idea that there is something wrong with motivation by rightness *de dicto*, is an old dog that is being put to new tricks. My aim here is to put a stop to this. I think that the widespread acceptance of the false view has been a mistake.

My own view is a form of pluralism about praiseworthy motivations. I think that it is good to be motivated by honesty, fairness, equality, and so on, *and* it is *also* good to be motivated by rightness *de dicto*. And that is not all: I also think that the traditional distinction between the right-making features and rightness itself is oversimplified. Just as there are right-making features, there are features that make it the case that the right-making features obtain – we might call them "right-making-feature-making-features". And there are further features that make it the case that the right-making-feature-making features obtain, and so on, in a hierarchy of metaphysical constitution. To preview slightly: my view is that *any* intrinsic or well-derived realizer motivation whose object is one of the moral features in this metaphysical hierarchy, including the maximally thin moral feature at the top, is a praiseworthy motivation. (I will explain this in §3.3.)

Nonetheless, some of my arguments show only that certain popular criticisms of motivation by rightness *de dicto* apply with equal force to motivation by rightness *de re.* This leaves open the possibility that neither is praiseworthy. Hence, I argue for the conditional thesis stated above: *if* motivation by rightness *de re* is praiseworthy, then so is motivation by rightness *de dicto*. My opponents already accept that motivation by rightness *de re* is praiseworthy, so I hope that they will join me in adding more praiseworthy motivations to their list. But the argument of this paper leaves open the option of throwing out both baby and bathwater and starting anew.

Here is a roadmap. After two preliminary clarifications (§2), I argue that motivation by rightness *de dicto* and *de re* have been poorly compared, and that, when we compare correctly constructed minimal pairs, it is no longer plausible that one is praiseworthy and the other not. I first discuss good cases, in which people

[2] The issues surrounding moral hedging are interesting and complex. For defenses, see Lockhart (2000), Sepielli (2009, 2013), Enoch (2014), and for criticisms see Nissan (2015), Hedden (2016), and Harman (2015). I discuss a puzzle about iterated moral hedging in my paper "Higher-Order Uncertainty", a draft of which is available on my website.

[3] I accept a version of the traditional Kantian idea. I defend it, and criticize Markovits' position further, in my paper "Accidentally Doing the Right Thing", a draft of which is also available on my website. Another recent defense of a version of the Kantian idea can be found in Sliwa (2016).

succeed in doing what they are trying to do (§3.1). I argue that the false view is committed to implausibly harsh verdicts about agents who try to act rightly and even partially succeed, especially as compared with those who manage to act rightly without trying. I then turn to bad cases, in which people fail to do what they are trying to do due to their false moral beliefs (§3.2). I argue that these cases arise for motivation by rightness *de re* exactly as they do for motivation by rightness *de dicto*. This means that my opponents and I all need to find something plausible to say about such cases. I offer something to say: we should all attend more closely to the different ways of being praiseworthy, acknowledging that someone can have praiseworthy motivations without praiseworthy beliefs or behavior, and that someone can have some praiseworthy motivations while lacking others. It should come as no surprise that people can be criticizable in certain respects while also having some redeeming features. This, I contend, is what we should say about the well-meaning but morally mistaken.

## 2. Clarifying the phenomena

This section covers two preliminaries that are necessary for understanding my main argument. I explain the way I am thinking of motivation, and I sketch the picture of moral metaphysics that informs my view.

Here is how I am thinking of motivation. As I will construe it throughout this paper, a motivation is a type of mental state to which desires give rise, and which itself gives rise to a set of dispositions. These comprise (1) the disposition to think about what it would take to realize that which one desires, (2) the disposition to notice when one's acts seem to have some bearing on whether that which one desires will be realized, (3) the disposition to do what one thinks will realize that which one desires, doing it *because* (one thinks that) it will realize that which one desires, and (4) the disposition to refrain from doing something if one thinks that it will impede the realization of that which one desires, refraining from doing it *because* (one thinks that) it will impede the realization of that which one desires. Someone is motivated to do something to the extent that she has these four dispositions. For example, someone is motivated to eat healthily to the extent that she is disposed to think about healthy eating, notice whether her food is healthy or unhealthy, and choose to eat some foods and avoid others on the grounds that this is what it takes to eat healthily.[4]

As I am thinking of it, motivation is not quite the same thing as desire. Motivation is the part of desiring something that involves trying to bring it about.[5] Desire itself is associated with a wider set of dispositions than the four just mentioned; for example, it is associated with the disposition to be happy and satisfied when one believes that what one desires is realized, and to be unhappy and frustrated when one believes that it is not realized.[6] I think that it would be a conceptual stretch to say that these dispositions are part of motivation. But nothing hangs on this terminological point; if we spoke in terms of desire (or anything else)

[4] Like all dispositions, the dispositions associated with motivation need not always manifest. For example, someone could be motivated to eat healthily even though she sometimes eats cake, knowing full well that this will impede the coming about of that which she desires (viz., that she eats healthily). To the extent that she remains generally *disposed* to refrain from doing what she thinks will impede her eating healthily, and she also has dispositions (1—3), she still counts as motivated to eat healthily. These dispositions come in degrees, because motivation comes in degrees.

[5] The question of what it takes to try to do something is a vexed question in the philosophy of action and in law; see for instance Adams 1995, Ludwig 1995, Schroeder 2001, Yaffe 2010. For present purposes I will sidestep all the interesting issues in these literatures. I use the term "trying" for cases in which motivation non-deviantly causes action – where by "non-deviant" I mean to stipulatively rule out such cases as Davidson's climber (1973, pp.78-79).

[6] For more on desire and its relationship to motivation, see e.g. Schroeder 2006, pp.633-634; Sinhababu 2017, pp.23-28. For some relevant neuroscientific work see Morillo (1990) and Berridge (2003).

rather than motivation, it would remain the case that I am interested in the mental state that gives rise to the four dispositions just mentioned.

Some motivations are related to one another, because there are structural relationships between the desires that give rise to them. There are three types of desire. A desire to φ is *intrinsic* if it serves no further end; philosophers sometimes express this by saying that the agent wants to φ "for its own sake".[7] A desire to φ is *instrumental* if it is generated by a desire to ψ plus a belief that φ-ing is a causal means to ψ-ing. And a desire to φ is a *realizer* desire if it is generated by a desire to ψ plus a belief that φ-ing constitutes ("realizes") ψ-ing. Thus, both instrumental and realizer desires depend on prior desires and beliefs about relationships between their objects and the objects of these prior desires. But they are different, since causal relationships are different from relationships of metaphysical constitution. For example, suppose that a track-and-field athlete intrinsically desires to win an upcoming race. She may desire to train regularly, but only insofar as she believes that training regularly will help her to win; if she ceased to believe that it will help her to win, then (*ceteris paribus*) she would no longer want to train regularly. This makes her desire to train regularly an instrumental desire. The athlete may also desire to get her torso across the finish line faster than all her competitors, but only insofar as she believes that this is *what it is* to win the race; if she ceased to believe that it constitutes winning the race – say, if she heard that the rules had changed and that getting one's foot across the finish line first now constitutes winning – then (*ceteris paribus*) she would cease to care about the position of her torso. This makes her desire to get her torso across the finish line first a realizer desire.

We can now clarify the nature of intrinsic motivation by rightness *de dicto*. This is a mental state that arises when an agent desires that she act morally rightly, and that gives rise to dispositions to think about what it takes to act rightly, to notice the moral quality of her acts, to do things she thinks are right, *because* they are right, and to refrain from doing things she thinks are wrong, *because* they would be wrong. Importantly, for a motivation to act rightly to be *intrinsic*, it must not depend on the agent's beliefs about what acting rightly would cause or constitute. For example, the agent is not intrinsically motivated to act rightly if she has these dispositions only because she believes that a person she finds attractive will go on a date with her if she acts rightly. And she is not intrinsically motivated to act rightly if she has these dispositions only because she believes that acting rightly constitutes earning good karma, and she wants to earn good karma.

We can similarly see what it is to be intrinsically motivated by a right-making feature. For example, suppose that fairness is a right-making feature. To be intrinsically motivated by this feature is to be in a mental state that arises when the agent desires that she act fairly, and that gives rise to dispositions to think about what it takes to act fairly, to notice the fairness or unfairness of her acts, to do the things she thinks are fair, *because* they are fair, and to refrain from doing the things she thinks are unfair, *because* they are unfair. For an agent's motivation to act fairly to be intrinsic, it must not depend on her prior beliefs about what acting fairly would cause or constitute. For example, she is not intrinsically motivated to act fairly if she has these dispositions only because she believes that she will be financially rewarded for her fairness and wants some financial reward. And she is not *intrinsically* motivated to act fairly if she has the dispositions only because she believes that acting fairly constitutes acting rightly, and she wants to act rightly. (This last point will be important for my argument in §3.1.) The same applies, *mutatis mutandis*, to all other right-making features.

That was the first preliminary. The second is a brief sketch of the metaphysical picture that informs my thinking on this topic. Lots of the details of this picture are unimportant for present purposes, and could be filled out in many ways. What is important is this: *the right-making features are not fundamental*. This means

[7] This muddies the waters somewhat by ignoring the distinction between intrinsic and final desires, which does not matter for present purposes. For the distinction see Korsgaard (1983); Rabinowicz and Rønnow-Rasmussen (2000).

that the very same metaphysical relationship – the "makes it the case" relationship, however its details are construed – that moral rightness bears to the right-making features is in turn borne by the right-making features to various other features of acts. For example, the fact that an act is fair is not a brute fact. This fact obtains in virtue of further facts about the act; perhaps the fact that the act distributes benefits and burdens on reasonable, non-arbitrary grounds. That is also not a brute fact. It obtains in virtue of further facts about the act; perhaps that it is meritocratic, or that it makes reparations for past injustice, or that it distributes resources based on need. This yields a metaphysical hierarchy of features of acts that continues down to the fundamental level.[8] Here I will focus on the first few levels, *qua* possible objects of people's motivations.

The fact that the right-making features are not fundamental means that there are *de re*/*de dicto* distinctions to be drawn with respect to motivation by any of these features, just as there is for motivation by rightness. For example, someone could have an explicit concern with fairness as such: a concern with doing the fair thing in a certain situation, whatever it may be. This is motivation by fairness *de dicto*. Someone could also care directly about whatever it is that fairness in fact consists in – i.e., whatever comes immediately below fairness in the metaphysical hierarchy of features of acts (perhaps distributing social benefits and burdens on reasonable, non-arbitrary grounds). This is motivation by fairness *de re*. The same holds for all other right-making features. Someone can care about performing whichever acts have those features, which is caring about them *de dicto*. Or she can care about that which they in fact consist in, which is caring about them *de re*. Or both.

This matters, because it means that *motivation by rightness* de re *just is motivation by one of the right-making features* de dicto. For example, assuming that fairness is a right-making feature, being motivated by fairness *de dicto* is one way of being motivated by rightness *de re*. We have two ways to refer to a single motivation: someone's explicit concern for acting fairly is accurately described either as motivation by fairness *de dicto* or as motivation by rightness *de re*. The same holds for all other right-making features. Being motivated by rightness *de re* might be a matter of being motivated to treat people with respect *de dicto*, or being motivated to promote well-being *de dicto*, or being motivated by the thought of people getting what they deserve *de dicto*. And so on, for whichever features rightness in fact consists in.

With these preliminaries in mind, we can now clarify the false view. This is what the false view says:

> **FALSE VIEW**: Intrinsic motivation by one of the right-making features *de dicto* (rightness *de re*) is praiseworthy. But intrinsic motivation by rightness *de dicto* is not praiseworthy.

I think that this is a faithful interpretation of what defenders of the false view have in mind. Defenders of this view typically explicitly restrict their focus to intrinsic motivations. For instance, in the introduction to their book, Arpaly and Schroeder say that "in this work the focus will be on intrinsic desires" and that they hold that instrumental and realizer desires have "little or [no] moral significance" (*op. cit.*, p.6). There is a rationale for this restriction, which I will discuss (and criticize) in §3.1.

Defenders of the false view are also fairly explicit about the fact that it is the right-making features, rather than the right-making-feature-making-features (or any other lower-order features), that they take to be the objects of praiseworthy motivations. When Arpaly and Schroeder say that the object of a virtuous agent's motivation is the right or good "correctly conceptualized", and that this amounts to being "conceptualized in the way preferred by the correct normative theory", their examples are all mid-level moral properties

[8] I discuss this picture in greater detail, and draw out some metaethical implications, in my "We Can Have Our Buck and Pass It, Too" (provisionally forthcoming in *Oxford Studies in Metaethics*). A draft is available on my website.

that one may care about either *de dicto* or *de re* – including "respecting persons", "happiness maximized", "welfare", and "justice" (2013, p.164). Smith's examples are also mid-level properties, like "honesty", "equality", and "people getting what they deserve" (*op. cit*.). And Arpaly and Schroeder make it clear that motivation by right-making features *de re* is not praiseworthy, on their view. They consider the case of an alien scientist who is motivated to produce high levels of activity in the perigenual anterior cingulate cortex of healthy humans, which is, in fact, what pleasure consists in. This alien is motivated to produce pleasure *de re*. But Arpaly and Schroeder say that "one would not want to credit the alien with even partial good will" (2013, p.167), even if it turns out that pleasure-production is a right-making feature. So the false view favors intrinsic motivation by the right-making features *de dicto*, not *de re*: this view holds that it is intrinsic motivations whose objects are right-making features, rather than right-making-feature-making-features (or any other lower-order features), that is praiseworthy.

## 3. Main argument

As a reminder, here's my thesis again:

> **THESIS**: If motivation by rightness *de re* is praiseworthy, then so is motivation by rightness *de dicto*.

I will now give my main argument for this thesis.

To assess this thesis, we should compare pairs of cases: one in which the agent is motivated by rightness *de dicto* and another in which she is motivated by rightness *de re*. But, in constructing these cases, we should tread carefully. We should ensure that we compare minimal pairs – pairs of cases in which one agent is motivated by rightness *de dicto* and the other motivated by rightness *de re*, *with all else held fixed*. We should avoid varying other potentially relevant factors, so as not to create noise. In particular, we should not compare one agent who tries to act rightly but has *false* beliefs about what rightness consists in, and thus ends up acting *wrongly*, to another agent who tries to perform acts with a certain right-making feature and has *true* beliefs about what it consists in, and thus ends up acting *rightly*. This comparison is unhelpful, because our judgement about the cases does not necessarily reflect our intuitive assessment of the relative praiseworthiness of motivation by rightness *de dicto* and *de re*. It could be a response to another difference between the cases: the fact that one agent succeeds in what she is trying to do while the other fails, or the fact that one agent has true beliefs about the object of her motivation while the other has false beliefs about the object of her motivation, or the fact that one agent acts wrongly while the other acts rightly.

### *3.1. Good cases*

For a genuine minimal pair, both agents – the one motivated by rightness *de dicto* and the one motivated by rightness *de re* – should succeed in doing what they are motivated to do. They should also perform the same act under the same circumstances. I will offer one example of such a pair, and then a recipe for how to construct further examples.

Here is the example:

> CHAIRING 1: Maryam is chairing a session at a prestigious Philosophy conference, which is notorious for getting nasty during Q&A. Maryam wants to act rightly – that is, she wants to conduct Q&A in such a manner as to meet all of her obligations not only *qua* chair but

> also *qua* moral agent. So she thinks carefully about what her obligations might be, planning to modify her behavior in light of her conclusions. After much soul-searching and careful thought, Maryam decides that four things matter morally in her case: prioritizing junior scholars over senior scholars, prioritizing those who have asked fewer questions at the conference over those who have asked lots already, discouraging audience members from asking repeated versions of the same question, and discouraging them from battering the speaker with multiple lengthy follow-ups. Maryam devises a set of principles that allows her to promote these four ends in a manner that reflects her estimation of their relative importance. She then conducts Q&A in perfect accordance with her principles. Moreover, *Maryam is completely right about all of this*. She has exhaustively identified the considerations that matter morally in her case, and has chosen principles that precisely reflect their relative importance. Maryam has perfected the principles of conference ethics. Since she guides her behavior in accordance with her conclusions, she also acts perfectly.
>
> CHAIRING 2: Mario is chairing a session at a prestigious Philosophy conference, which is notorious for getting nasty during Q&A. Mario introspects and finds that he has four intrinsic motivations relevant to his circumstances: to prioritize junior scholars over senior scholars, to prioritize those who have asked fewer questions over those who have asked lots already, to discourage audience members from asking repeated versions of the same question, and to discourage them from battering the speaker with multiple lengthy follow-ups. So Mario devises a set of principles that allows him to promote these four ends in a manner that reflects the relative degrees to which he cares about each of them. Mario also comes to believe that the objects of his motivations are the right-making features in his situation, and that it is morally right to conduct Q&A in accord with his principles, since these beliefs fit well with his pre-theoretical intuitions. But these beliefs are motivationally otiose. Mario conducts Q&A in perfect accord with his principles just because his intrinsic motivations already incline him in this direction. He could change his beliefs about how it is morally right to conduct Q&A without his behavior changing at all. Happily, though, *Mario's intrinsic motivations are directed toward all and only the things in his situation that are in fact morally significant, and their relative strength corresponds precisely to these things' relative importance*. So, since these motivations guide his behavior, Mario also acts perfectly.

Let's assume that CHAIRING 1 and CHAIRING 2 are part of a broader pattern, as follows. Maryam has one intrinsic motivation operative in her decisions: the motivation to act rightly.[9] Mario, on the other hand, has a hodge-podge of various intrinsic motivations. But Maryam has all and only the true moral beliefs, and all and only true beliefs about morally relevant non-moral matters. So she has realizer motivations directed toward all of the right-making features, the right-making-feature-making-features, and so on. Meanwhile, Mario's intrinsic motivations are directed toward all and only the right-making features in every situation. He too believes these features to be right-making, having undergone reflective equilibrium based on his pre-theoretical intuitions. Mario also has true beliefs about what each right-making feature consists in, and has developed the appropriate realizer motivations. In short, for every intrinsic motivation of Mario's,

[9] This is not to say that the motivation to act rightly is Maryam's only intrinsic motivation. She may have any number of other intrinsic motivations, so long as they are not operative in her decisions about what to do in the cases that make up this broad pattern. For instance, it could be that Maryam is intrinsically motivated to take care of various friends and family members, but these motivations play no part in a rationalizing explanation of her choice of chairing policy, since Maryam knows that nothing she does at the conference will affect those friends and family members. In this case, though the motivation to act rightly is not Maryam's only intrinsic motivation, it is the only one that is *operative*.

Maryam has a realizer motivation with the same object. And for every realizer motivation of Maryam's, Mario has either the same motivation or an intrinsic motivation with the same object. Most of their motivational sets are identical. The only difference between these agents lies in the structure of the very top of their motivational sets: Maryam has an extra intrinsic motivation, *to act rightly*, from which her other motivations derive, whereas Mario's motivations derive from his intrinsic motivations directed toward the right-making features (which are, for Maryam, the objects of realizer motivations). But this difference in the structure of their motivational set makes no difference to their behavior. In all actual circumstances, like CHAIRING 1 and 2, Maryam and Mario act identically – and, by stipulation, morally perfectly.

These cases compare two *successful* agents, who do what they are trying to do. Maryam tries to act rightly, and does a great job. She acts impeccably. Mario tries to promote the various things that he cares about, and does an equally great job. He promotes these things to the degree to which he cares about each of them. Moreover, since Mario's motivations align with the content of the true moral theory, he also acts impeccably. So this pair of cases is well-constructed; it compares someone motivated by rightness *de dicto* with someone motivated by rightness *de re*, holding all else fixed.

What, then, should we say about the praiseworthiness of Maryam and Mario's motivations?

The false view says that Maryam's motivations are *not at all praiseworthy*. This is not obvious, so let me spell it out. Arpaly and Schroeder repeatedly emphasize that their view is about intrinsic desires (2013, pp.6-9). On this view, then, instrumental and realizer motivations are not the sort of thing that can be praiseworthy. But this view also says that not just any old intrinsic motivation is praiseworthy; as we saw in §2, the view says that only intrinsic motivations whose objects are right-making features are praiseworthy motivations. This entails that Maryam's motivations are not at all praiseworthy. For, although Maryam is motivated by every right-making feature, none of those motivations are *intrinsic*. They are realizer motivations, deriving from her intrinsic motivation to act rightly and her true moral beliefs that these features are what moral rightness consists in. Maryam's only intrinsic motivation is directed toward rightness itself. And rightness itself is not a right-*making* feature; that would be circular. (To put this another way: the "makes it the case" relation is irreflexive.) So Maryam has no motivation that is *both* intrinsic *and* directed toward a right-making feature. Thus, according to the false view, she has no praiseworthy motivations.

This is not at all plausible. Maryam is a moral saint.[10] Her life consists in the performance of one morally right act after another. She also has all and only true moral beliefs. And neither her consistently right actions nor her perfectly accurate moral beliefs are a fluke; Maryam is this way because she is motivated by rightness *de dicto*, which has led to a great deal of careful thought, sophisticated reasoning, and moral effort on her part. She is this way because her life is guided by an unfailing, and successful, commitment to acting rightly. There may be some bad things about moral saints – one may not want to have a moral saint as one's best friend, for instance – but it is simply incredible to say that the motivations of someone as morally outstanding as Maryam are not praiseworthy to *any* degree whatsoever.

This verdict on Maryam is even less plausible when we compare it to the false view's verdict on Mario. On this view, although Maryam's motivations are *not at all* praiseworthy, Mario's are *fully* praiseworthy. This is because (like Maryam) he has a motivation for every right-making feature, and (unlike Maryam) these motivations are intrinsic. But these wildly divergent verdicts are clearly the wrong result. Maryam and Mario's motivational sets are extremely similar, differing only in the nature of their motivations directed toward right-making features – his are intrinsic, hers realizer motivations – and in the fact that Maryam

[10] See Wolf (1982). The claim that this kind of agent is a moral saint is also made by Carbonell (2013).

has an additional intrinsic motivation to act morally rightly. This is the *only* difference between them. Their other realizer motivations are all identical. They both act perfectly. And both have all and only true moral beliefs. If Maryam and Mario were to observe each other's behavior, or discuss any moral issue, they may be unable to identify any difference between them. Once these cases have been constructed so as to remove all other grounds for differences in praiseworthiness, then, the difference in the structure of the very top of Maryam and Mario's motivational sets seems far too flimsy a distinction to ground the difference between full praiseworthiness and none at all.

We can contrast motivation by rightness *de dicto* and *de re* without imagining agents as morally amazing as Maryam and Mario. Imagine Shmaryam and Shmario, who excel in conference-chairing but make lots of other moral mistakes. There are countless possible Shmaryams who try to act rightly but do not succeed as well as Maryam, as their moral beliefs get only part-way toward the truth, so some, but not all, of their realizer motivations are directed toward genuine right-making features. And for each Shmaryam, there is a corresponding Shmario who has *intrinsic* motivations directed toward the features that are the objects of Shmaryam's realizer motivations.[11] We can stipulate that the agents in each pair have identical beliefs about the right-making-feature-making-features, and have developed the appropriate realizer motivations. So these agents again have identical moral beliefs, and, again, they act identically. Using such pairs of cases, we can compare motivation by rightness *de dicto* with motivation by rightness *de re*, holding all else fixed. And for each such pair, the false view entails that Shmario's motivations are *somewhat* praiseworthy, but Shmaryam's are *not at all* praiseworthy. These pairs of verdicts are all implausible. So this is a whole family of pairs of examples, each of which provides support for my thesis: if motivation by rightness *de re* is praiseworthy, then so is motivation by rightness *de dicto*.

There is a way to modify the false view to avoid these awkward verdicts without conceding that motivation by rightness *de dicto* is praiseworthy. We might say that there are two types of praiseworthy motivation: intrinsic motivations whose objects are right-making features, *and realizer motivations whose objects are right-making features*. This would entail that the motivational sets of the agents in each (Sh)Maryam-(Sh)Mario pair are equally praiseworthy, since, for every intrinsic motivation directed toward a right-making feature that Mario has, Maryam has a realizer motivation directed toward the same feature. We would thereby avoid the unwelcome verdicts.

But this modification is more trouble than it's worth. Arpaly and Schroeder restrict their focus to intrinsic motivations for a reason: it is possible for agents to develop realizer motivations directed toward right-making features by sheer fluke.

Here is an example:

> CHAIRING 3: Aarulina is chairing a session at a prestigious Philosophy conference, which is notorious for getting nasty during Q&A. Aarulina has just one intrinsic motivation: she desperately wants to act in a way that is approved of by aardvarks. (This serves no further end – just as some people care about acting in a way that is approved of by their family or friends, or by God, for Aarulina it's all about aardvarks.) Fortunately, Aarulina thinks she has figured out what aardvarks approve of: Aarulina thinks aardvarks approve of chairs at prestigious Philosophy conferences conducting Q&A so as to prioritize junior scholars

[11] For example, suppose that there are seven right-making features, and that Shmaryam has identified two of them and developed the appropriate realizer motivations. Then Shmario is intrinsically motivated by these two right-making features but not the other five.

> over senior scholars, prioritize those who have asked fewer questions over those who have asked lots already, discourage the audience from asking versions of the same question over and over again, and discourage them from battering the speaker with multiple lengthy follow-ups. Aarulina also has beliefs about the relative importance of each of these four values to aardvarks. So she devises a set of principles that she thinks embody aardvarks' concerns, and she acts accordingly. Moreover – oddly enough, and entirely unbeknownst to Aarulina – *the values that Aarulina ascribes to aardvarks correspond perfectly to the content of the true moral theory*. So, like Maryam and Mario, Aarulina acts perfectly.

Again, we can imagine that this is part of a broader pattern. For every right-making feature that is the object of Mario's intrinsic motivation and Maryam's realizer motivation, Aarulina has a realizer motivation directed toward this feature, derived from her intrinsic motivation to act in a way approved of by aardvarks and her belief that aardvarks approve of acts with this feature.

This spells trouble for the modified false view. The modified false view entails that the motivations of oddballs like Aarulina are *fully praiseworthy*. Again, this is just not plausible. Aarulina is a weirdo whose realizer motivations happen to align in content with the true moral theory. Since she is indifferent to morality *per se*, Aarulina would not even be pleased to learn that the objects of her realizer motivations were all and only the right-making features. This would strike her as nothing more than an interesting coincidence, like learning that a ball game one invented as a child corresponds perfectly to a sport played in a distant country. This coincidental orientation toward morality does not seem praiseworthy.[12]

The modified false view must hold that Aarulina's motivations are equally as praiseworthy as Maryam's. This is also implausible; Maryam is a moral saint who tries to act rightly and succeeds fantastically, while Aarulina is a benign weirdo who acts rightly by coincidence. This makes Maryam's motivations far more praiseworthy than Aarulina's. So we cannot say that just any old realizer motivation whose object is a right-making feature is praiseworthy. The provenance of these motivations matters. The modification fails.

Here we have a recipe for constructing counterexamples to the false view. This is the recipe: Pick one or more of your favorite right-making features. Imagine an agent who is intrinsically motivated by these features. Then imagine another agent who is intrinsically motivated to act rightly, has figured out that these features are right-making, and has developed the appropriate realizer motivations. Compare the two agents. Next, imagine a third agent with a wacky but morally innocuous intrinsic motivation, the belief that performing acts with your preferred right-making features constitutes achieving the object of this motivation, and the corresponding realizer motivations to perform acts with these features. Compare all three agents. Et voilà! You have a dilemma for the false view. Unmodified, it yields implausible verdicts about the first two agents. Modified, it yields implausible verdicts about the second and third agent.

So I submit that the false view should be rejected.

What caused the trouble for this view was a pair of claims: that only intrinsic motivations are praiseworthy, and that intrinsic motivation by rightness *de dicto* is not praiseworthy. Abandoning the first of these claims alone does not help – it lands us on the Aaurulina horn of the dilemma. So we should abandon the second claim as well. The appropriate response to cases of people trying to act rightly and succeeding fantastically,

[12] It is tempting to say that Aarulina acts morally rightly *by accident*. I agree. In fact, I think the same is true of Mario. I argue for this in my "Accidentally Doing the Right Thing", a draft of which is available on my website. The reader may be worried that the provenance of Maryam's motivations is, at present, equally mysterious; I discuss this in §3.3.

like Maryam, is to accept that their motivations are indeed praiseworthy. Or, at least, they are praiseworthy *if* Mario's motivations are praiseworthy. This is exactly what my conditional thesis says.

### 3.2. *Bad cases*

One popular argument against the praiseworthiness of motivation by rightness *de dicto* notes that people can be led by this motivation to act wrongly if they have false beliefs about what moral rightness consists in. In these cases, the argument goes, the agents often don't look very praiseworthy.

I accept that motivation by rightness *de dicto* can lead someone to act wrongly, if she has false moral beliefs. But this is equally true of motivation by rightness *de re*. Rightness is not the only moral property. Many "thick" moral properties are plausible candidates for being right-making features.[13] So beliefs about these properties' nature and extension – for example, beliefs about what fairness, well-being, or justice consists in – are all moral beliefs. And if someone has a false belief about what fairness consists in, then, by *trying* to act fairly, she can *in fact* act unfairly. Parallel remarks apply to promoting well-being or justice. But it is wrong to act unfairly, to undermine well-being, or to inhibit justice. So motivations whose objects are right-making features can lead someone to act wrongly, if she has false moral beliefs.

Here are three cases to illustrate this point:

> FAIRNESS: A father is coming up with a toy-sharing policy for his two daughters. He wants his toy-sharing policy to be fair. So he thinks awhile and comes up with a rudimentary theory of fairness. But he gets it wrong; he thinks that his daughters' age-difference is irrelevant to considerations of fairness, when in fact it is relevant. So he ends up instituting a policy that is in fact unfair to his younger daughter.
>
> WELL-BEING: A mother wants to promote her son's well-being. She thinks it will promote his well-being for him to learn a musical instrument. So she signs him up for piano lessons and forces him to go. But this underestimates the importance of autonomy as a component of well-being; the son doesn't want to learn piano, so her forcing him to do it in fact undermines, rather than promoting, his well-being.
>
> JUSTICE: Some parents are trying to think of a just punishment for their child, who has drawn on the walls of their house. They falsely believe that smacking is, sometimes, a just punishment. And they believe that this is one of those times. So they smack their child. But they are wrong; smacking is never a just punishment.

Faced with cases like these, it is tempting to say that the parents are at least praiseworthy for *trying* to create a fair toy policy, to promote the son's well-being, and to come up with a just punishment, even if they are also blameworthy for *in fact* acting unfairly, undermining well-being, and inhibiting justice. But if we are going to say that about these cases, then we can say the same thing about trying and failing to act rightly. We can imagine analogues of FAIRNESS, WELL-BEING, and JUSTICE in which the agents want to act *rightly* and falsely believe that it is *right* to institute the unfair toy policy, force the son to take piano lessons, or smack their child. But once we construct our cases in this way – as genuine minimal pairs – it is no longer plausible that trying and failing to act rightly is *ipso facto* less praiseworthy than trying and failing to act,

[13] "Thick" properties are partly descriptive and partly normative. See Roberts (2013) for an introduction, and Väyrynen (2013) for detailed discussion. For a related idea in moral metaphysics see Leary (2017).

say, fairly. In all cases, we have people who are well-meaning but morally mistaken. It doesn't seem to make a difference whether they are mistaken about rightness or another moral property. J.S. Mill famously remarked that "there is no difficulty in proving any moral standard whatsoever to work ill, if we suppose universal idiocy to be conjoined with it" (Mill 1871, p.35); the same holds of moral motivations.

This supports my conditional thesis. If motivation by rightness *de re* is praiseworthy even when led astray by false moral beliefs, then so is motivation by rightness *de dicto*. And if motivation by rightness *de dicto* is no longer praiseworthy when led astray by false moral beliefs, then the same should go for motivation by rightness *de re*.

Here my opponents might object. Several authors have argued that false moral beliefs cannot excuse an agent from blame for wrongdoing, and that agents who are led to act wrongly by their false moral beliefs are still blameworthy (see e.g. Harman 2011). My opponents might worry that the position I am defending contradicts this view, by suggesting that such agents might, in fact, be praiseworthy.

This worry is misplaced. The question of whether moral ignorance excuses is a question about when agents are blameworthy *for acting wrongly*. To answer it, we may need to know when agents are blameworthy *for moral ignorance*.[14] But my thesis is about praiseworthy *motivations*. And motivations, beliefs, and acts are all different things. So our verdicts about them can come apart: someone may be blameworthy for one or two of them while being praiseworthy for the rest. This means that we can say that, in cases like FAIRNESS, WELL-BEING and JUSTICE, the agents are praiseworthy for *trying* to act fairly, promote well-being, or bring about justice, even if they are blameworthy for *in fact* acting unfairly, undermining well-being, or inhibiting justice. We can even say that someone is still praiseworthy for her good motivation if she is blameworthy not only for her wrong act, but also for her false moral belief. For example, it might be that the parents in JUSTICE are blameworthy both for thinking that smacking is permissible (thus displaying insufficient concern for their child's welfare) and for smacking their child, but nevertheless are praiseworthy *for wanting to find a just punishment when their child has drawn on the walls*. Even if the act and belief are blameworthy, the good motivation can still be praiseworthy.

I think that this is the right thing to say about these cases. It is natural to say "Her intentions were good", taking oneself to be mentioning a redeeming feature of an agent who has acted poorly. I think such claims are often literally true. The agent's intentions *were* good – that is to say, her motivations were praiseworthy. It is a commonplace that we can be criticizable in some respects while also having some redeeming features; people are not either wholly perfect or wholly awful. What I am suggesting is that this is true of the well-meaning but morally mistaken. Good motivations can still be praiseworthy even in agents who act poorly or hold false moral beliefs.

And if this holds for motivation by rightness *de re*, then it should also hold for motivation by rightness *de dicto*. The fact that someone was at least *trying* to act rightly can be a redeeming feature just as well as the fact that she was at least *trying* to act fairly, in cases where the agent ends up acting wrongly due to false moral beliefs. So this kind of case provides further support for my thesis: if motivation by rightness *de re* is praiseworthy, then so is motivation by rightness *de dicto*.

[14] I have a view on this. I accept, with Harman, that people are blameworthy for false moral beliefs that embody a failure to care adequately about that which is, in fact, morally valuable. But I deny that all false moral beliefs embody such failures to care. I argue for this in my "Don't Know, Don't Care?" (draft available on request). For discussions of related ideas see Calhoun (1989); Moody-Adams (1994); Smith (2005).

My opponents may now raise a different concern. Arpaly and Schroeder (2013, pp.186-7) note that false moral beliefs may erode someone's praiseworthy motivations, eventually eliminating them. They imagine someone who is initially intrinsically motivated by a right-making feature, but who becomes convinced of a false moral theory, and is also motivated by rightness *de dicto*. They imagine that she is then so concerned to act well by the lights of this false theory that the intrinsic motivation directed toward that which is truly right-making loses its grip on her. Such agents are literally *corrupted by theory*.

I agree that people can lose praiseworthy motivations when they are corrupted by theory. But this concern does not raise doubts about the value of motivation by rightness *de dicto*, nor about the thesis of this paper. That is because the risk of people's being corrupted by theory is not confined to motivation by rightness *de dicto*. It arises with equal force for motivation by rightness *de re*. For example, the parents in JUSTICE may be led by their false theory of justice to slowly lose their natural inclinations against hitting their child. Or the father in FAIRNESS may find that his natural inclination to be more lenient with his younger daughter slowly dissipates as he becomes increasingly convinced of his false theory of fairness. Again, it is natural to say that the parents are still praiseworthy for *trying* to bring about justice or to institute a fair toy policy, even though this blinds them to their initial concern for that which justice and fairness actually amount to. And, again, it is hard to see why we should not then say the same thing about trying to act rightly. Again, then, the many kinds of moral motivation are all on a par.

Here is a third, related, worry. My opponents may suggest that some agents' moral beliefs are so wildly askew that they deserve *no praise whatsoever* for trying to act rightly. If someone's conception of what is morally right is way off-track, and this leads them to commit horrific acts, then perhaps it is implausible that their motivation to act rightly is still praiseworthy.

Here, for instance, is Markovits (*op. cit.*, p.224):

> [T]he fact that Göbbels was driven by his conscience to persecute the Jews does not exonerate him, much less endow his acts with moral worth.

Markovits is here arguing that, if Göbbels was trying to act rightly, this should not make us think better of his wrongful *act*. But we can equally imagine someone arguing that there is nothing of value in Göbbels' *motivations*, notwithstanding the fact that he wants to act rightly and believes that what he is doing is right. Perhaps if someone gets *really* bad, then their so-called "good intentions" are no longer a redeeming feature.

This is an important worry. I will go through four responses to it.

The first and most important thing to note is that this verdict is consistent with my thesis. My thesis is that *if* motivation by rightness *de re* is praiseworthy, then so is motivation by rightness *de dicto*. This does not require holding that motivation by rightness *de dicto* is always praiseworthy. Circumstances involving the agent's having wildly askew moral beliefs may be among the times when it is not. A problem for me would only arise if motivation by rightness *de re* is praiseworthy *under the same circumstances* – holding everything else fixed. And I very much doubt that this is the case. Once again, we can imagine similar cases of being misled by motivation by rightness *de re* and false beliefs about what the right-making features consist in. Suppose that Göbbels wanted to act *justly* and thought it *jus*t to persecute Jews. Or suppose that he was motivated by the thought of *people getting what they deserve*, and thought that Jews deserve persecution. (Either supposition might accurately characterize the actual historical Göbbels.) These versions of Göbbels seem no better than the version who is motivated by a drastically mistaken conception of moral rightness. Either way, his moral beliefs are wildly askew and his actions unconscionable, to the point where his caring

about something that it is usually good to care about does little, if anything, to redeem him. So in this case, again, whether the agent is (mistakenly) motivated by rightness *de dicto* or *de re* simply does not matter.

Here is a second response. An agent may successfully refer to such properties as rightness, justice, or desert, even if she has an incomplete understanding of their natures. But when her beliefs about what these properties amount to are *really* warped, she may fail to refer to them at all. This stems from a general feature of reference. For example, suppose that someone claims to want to visit "Detroit", but that her only belief about Detroit is that it is somewhere in England. There may be somewhere that this person wants to visit, which she calls "Detroit". But it is not Detroit; she fails to refer to Detroit. Likewise, someone who claims to care about a thing that she calls "rightness", but whose *only* belief about rightness is that it is a property of her left shoe, fails to refer to rightness. If this line of reasoning is correct, then agents whose moral beliefs are wildly askew may fail to be motivated by rightness *de dicto* at all. They are motivated by something, which they call "rightness". But it is not rightness.[15] Again, the same can be said for any of the right-making features. Göbbels' saying that he cares about "justice" or "desert" may not be enough to make it the case that he is motivated by justice or desert. The fact that his beliefs about the nature of these things are wildly off-track may prevent him from referring to them at all.

A third response is to point out that someone like Göbbels may not really believe that what he is doing is morally right. People often use moral language to advance their own interests; they use positively-valenced moral terms to describe horrific acts that they perform, order, or sanction, without believing the claims that they are making, in the attempt to manipulate others (and thereby to gain and maintain power) or to reduce cognitive dissonance. Use of moral language by agents perpetrating moral atrocities does not show that these agents care *de dicto* about rightness, fairness, justice, or anything else. It could instead suggest that people mask behavior that they know to be morally atrocious in positive terms in order to sleep at night.

A fourth response is to bite the bullet. We can say that, if Göbbels really was sincerely trying to act rightly, and wasn't faking, then this is praiseworthy. That would mean that we cannot say that there is *nothing* praiseworthy about Göbbels. But we can still say that the fact that he was trying to act rightly is the *only* praiseworthy thing about him. We can still say that he is blameworthy for his wrongful acts and false moral beliefs. So even if Göbbels' moral motivation is praiseworthy, he will still be an utterly despicable person overall. Given that, I do not think that this is the hardest bullet to bite. Indeed, we can motivate the claim that Göbbels' sincere moral motivation (if he had any) is a redeeming feature using another minimal pair; we can compare the clueless Göbbels who sincerely believes that what he is doing is right with a knowing Göbbels who is fully aware that what he is doing is deeply wrong and just doesn't care. This is a difficult comparison. But I am tempted to think that the former agent is at least slightly better than the latter. If so, then that is presumably because his intentions are good.

All four of these responses are available, so I have canvassed them in order to let the reader choose between them. I expect that different responses will be appropriate in different cases. Collectively, I expect that they will cover everything.

### *3.3. The "partial credit" approach*

It is possible for defenders of the false view to take a hard line on these matters: they can say that agents are praiseworthy only if their motivations align precisely with the *true* nature and extension of the right-

[15] This requires moving beyond the *de re*/*de dicto* distinction to a more sophisticated account of how someone may be motivated by rightness. I go some way toward this in my "How To Be a Moral Fetishist" (draft available on request.)

making features. On this approach, an agent's being praiseworthy requires more than that (e.g.) fairness is a right-making feature and she is motivated by fairness *de dicto*. On this approach, she must *also* have true beliefs about what fairness consists in, and must have developed the corresponding realizer motivations. Moreover, she must have still further true beliefs about that which that-which-fairness-consists-in *itself* consists in, and must have developed the corresponding realizer motivations. (For example, if fairness consists in distributing benefits and burdens on reasonable grounds, she must have a realizer motivation to distribute benefits and burdens on reasonable grounds, and she must have true beliefs about what this amounts to, and she must have realizer motivations directed toward whatever it amounts to.) And so on, all the way down the metaphysical hierarchy discussed in §2. And so, similarly, for each other right-making feature.

The hard-line approach entails that those who are led to act wrongly by their false moral beliefs, including those who are corrupted by theory, do not have praiseworthy motivations. On this view, it is simply false to say of the agents in FAIRNESS, WELL-BEING, and JUSTICE that their intentions were good. Their intentions *would* have been good if they were accompanied by true beliefs and realizer motivations about that which the objects of the intentions consist in (and about that which the features that they consist in consist in, etc.). But these agents have false beliefs, and their realizer motivations are misaligned. So, on this view, their motivations are not praiseworthy.

But the hard-line approach is unappealing, since it is likely to entail that no actual person has praiseworthy motivations. Whether it does so depends on what moral theory turns out to be true. The hard-line approach will grant moral praiseworthiness only if the right-making features turn out to be things that everybody is moved by and that we all understand perfectly. This is, of course, exceedingly unlikely. For any plausible candidate for being a right-making feature, normal agents have only an inchoate grasp of this feature, rather than detailed beliefs about its precise nature and extension with corresponding realizer motivations. For example, many people are intrinsically motivated by justice *de dicto*. But there are surely very few people who are motivated by each person's having the highest degree of basic liberties compatible with equal liberty being granted to all and by social and economic inequalities' being (a) distributed to benefit the least well-off and (b) open to all under conditions of fair equality of opportunity, with (b) taking lexical priority over (a). Yet the most famous and influential theory of justice (Rawls 1971) says that this is what justice consists in. If anything like this theory is true, then only a few people – who have read Rawls, were persuaded, and remember his account in detail – are motivated by justice *de re*. This generalizes: the true moral theory, if fully spelled out, would provide us with accounts of the nature and extension of the right-making features that far surpass ordinary agents' understanding of them, and that are objects of motivation for nobody. So, by making this understanding and these motivations a necessary condition of our ordinary motivations' being praiseworthy, the hard-line approach effectively decrees that our ordinary motivations are not praiseworthy.

People's motivations are often praiseworthy. So we should not take the hard-line approach.

Instead, I propose that we take what I will call a "partial credit" approach. This approach says that we are praiseworthy for having motivations whose objects *approximate* the content of the true moral theory, and we are more praiseworthy the closer the approximation is.

Here is what that means. In §2 I described a metaphysical hierarchy of right-making features, right-making-feature-making-features, and so on. The true moral theory, fully spelled out, would tell us a large part of what this hierarchy is. It would exhaustively specify the right-making features, clarifying the relationships between them and any conditions on their being right-making, and it would tell us what metaphysically

constitutes these features. At least, it would tell us these things about those of the right-making features *that are themselves moral features* – like honesty, equality, desert, fairness, well-being-promotion, and so on. These being moral features, the tasks of specifying their nature and figuring out what it takes for them to obtain are part of moral theory. So, when I say that people are praiseworthy for having motivations whose objects approximate the content of the true moral theory, I mean that people are praiseworthy for having motivations whose objects are the moral properties in this metaphysical hierarchy. And when I say that people are more praiseworthy the closer the approximation is, I mean that someone is more praiseworthy, the more of the moral properties in this hierarchy are objects of motivation for her.

For example, let us continue to suppose that fairness is a right-making feature. In this case the partial credit approach says that anyone motivated by fairness *de dicto* is somewhat praiseworthy. But someone is *more* praiseworthy the more accurate her conception of fairness is, and thus the more her realizer motivations align with the true nature of fairness (i.e. are directed toward the properties that fall below fairness in the metaphysical hierarchy.) Now return to the father in FAIRNESS. He does not have realizer motivations that align with the true nature of fairness, so he is not as praiseworthy as he could be. But he is, at least, *trying* to act fairly. So he gets partial credit; he has an intrinsic motivation directed toward fairness *de dicto*, which is a praiseworthy motivation. Assuming that fairness is a matter of distributing benefits and burdens on reasonable grounds, the father gets a bit more credit; he has figured this much out, and has developed a realizer motivation to distribute benefits and burdens – in this case, toy playtime – on reasonable grounds. That is a further praiseworthy motivation, on the partial credit approach. This is so even though the father is mistaken about what sorts of grounds are reasonable, so that his realizer motivations from this point on diverge sharply in content from the true moral theory and are not praiseworthy.

Here is another example. Imagine someone who cares about fairness, knows that it consists in distributing benefits and burdens on reasonable grounds, and has developed the appropriate realizer motivation. But imagine that she thinks that only considerations of increased future utility can ever be reasonable grounds for anything. Let's stipulate that she is wrong about this: in fact, considerations of increased future utility are *among* the reasonable grounds on which to distribute benefits and burdens, but are not the whole story, since there are also considerations of merit and of reparations for past injustice. This agent then gets partial credit. She is motivated to act fairly, to distribute benefits and burdens on reasonable grounds, and to take considerations of increased future utility into account. All of this is praiseworthy, on my view (given those simple stipulations about what fairness in fact consists in). But our agent would be more praiseworthy if she were *also* motivated (a) to take merit into account and (b) to make reparations. That is how the partial credit view works.

A qualification is needed here. Not just *any* motivation whose object is one of the moral features in the true metaphysical hierarchy is a praiseworthy motivation. The provenance of these motivations matters; that was one of the lessons of §3.1. A motivation with one of these features as its object is praiseworthy if it is *either* an intrinsic motivation *or* a well-derived realizer motivation. By "well-derived" I mean that the realizer motivation derives from an intrinsic motivation directed toward a moral feature further up in the hierarchy, plus true beliefs about the metaphysical relationships that hold the hierarchy together. For example, take someone who is motivated to distribute benefits and burdens on reasonable, non-arbitrary grounds. She is praiseworthy for this if she cares about it intrinsically, or if she cares about it because she cares about fairness and knows that this is what fairness consists in, or if she cares about it because she cares about acting rightly, knows that fairness is a right-making feature, and knows further that this is what fairness consists in. But she is not praiseworthy for caring about this feature if she does so because she is intrinsically motivated to please aardvarks or make it rain frogs on Jupiter, or the like, and she believes that distributing benefits and burdens on reasonable grounds constitutes attaining one of these unusual goals.

Realizer motivations are praiseworthy when they are based on accurate – though perhaps incomplete – appreciation of the moral significance of their objects.

A stronger version of this qualification applies to motivation by non-moral features that may appear lower down in the metaphysical hierarchy.[16] For these non-moral features, I suggest that only well-derived realizer motivations are praiseworthy. It would be odd, and not especially praiseworthy, for these features to be the objects of *intrinsic* motivation, rather than well-derived realizer motivation. For example, it is praiseworthy to have a realizer motivation to ensure that we all have enough oxygen to breathe, having recognized that this is a vital human need and being intrinsically motivated to contribute to the satisfaction of people's needs. But it is odd, and not particularly praiseworthy, to be *intrinsically* motivated to make sure people have plenty of oxygen to breathe. Divorced from any beliefs about the moral significance of oxygen, wanting to ensure that people breathe plenty of oxygen *for its own sake* would just be weird. And this generalizes. So we should say that it is praiseworthy to have well-derived realizer motivations directed toward non-moral features that realize the moral features in the hierarchy, but we should deny that it is praiseworthy to have intrinsic motivations directed toward the non-moral features. In general, an agent is praiseworthy for caring about something that matters morally iff she has figured out at least part of the story about why it matters morally, and she cares about it on this basis.

This leads me to a final possible class of objections. The reader might think that my above remarks about motivation by the non-moral features in the hierarchy apply equally to motivation by rightness *de dicto*. I have suggested that it is not particularly praiseworthy to care about one of the non-moral features with no appreciation of the moral features above it in the metaphysical hierarchy, as these are the features that lend it moral significance. The reader might think that, similarly, it is not particularly praiseworthy to care about moral rightness with no appreciation of the moral features *below* it in the metaphysical hierarchy – the right-making features, right-making-making features, etc. – as these are the features that lend moral rightness its significance. She might say that what makes moral rightness significant are the features that it consists in, and that one fails to see why rightness matters if one does not appreciate these features. So, she might say, intrinsic motivation by rightness *de dicto* is not enough for praiseworthiness; it must be accompanied by appreciation of the right-making features. Relatedly, she might worry that motivation by rightness *de dicto* with no appreciation of the right-making features would be empty of content. Or she might wonder how someone could come to have such a motivation.

The first thing to note about these objections is that – as usual – they generalize. These are general worries about the praiseworthiness of *de dicto* intrinsic motivation: they will arise for intrinsic motivation by any of the right-making features *de dicto*, just as for motivation by rightness *de dicto*. For example, take kindness. I confess to being unable to articulate exactly what kindness consists in. But I do care about kindness, and I want to act kindly. I assume that I am not unusual in these respects. The reader might now allege that it is not praiseworthy to care about kindness with no appreciation of the features below it in the metaphysical hierarchy – of kind-making features, kind-making-making features, etc. She might say that it is the things

[16] Whether there are non-moral features in the hierarchy, and how low-down they are, depends on the truth of various moral and metaethical theories. For example, if the sort of robust realism defended by Enoch (2011) is true, then the moral is fundamental, and there is no level in the hierarchy such that the levels below it contain only non-moral features. The same holds if the right-making features include thick properties – as I have assumed – and if the "anti-disentanglement" argument is correct; on this see Roberts (2013), pp.680-681, and cf. McDowell (1998); Putnam (2002). By contrast, if the true moral theory is a simple maximizing consequentialism with only non-moral things in its theory of the good, then the right-making-feature-making-feature is a non-moral feature: acts are made right by their being value-maximizing, and this is so in virtue of their maximizing the various non-moral things.

that kindness consists in that make kindness morally significant, and that we fail to see why kindness matters if we do not appreciate these features. So she might say that intrinsic motivation by kindness *de dicto* is not praiseworthy; it must be accompanied by appreciation of the kind-making features. Relatedly, she might worry that motivation by kindness *de dicto* with no appreciation of the kind-making features would be empty of content. Or she might wonder how a person could ever come to have this motivation.

I think that no part of either class of objections is correct. I will now explain why.

It is not, in fact, all that easy to have one of the moral properties in the true metaphysical hierarchy be the object of an intrinsic motivation. First, for one of the things in the hierarchy to be the object of any attitude, the agent must be able to refer to it. This places some constraints on what can count as intrinsic motivation by a feature in the hierarchy *de dicto*. If someone says that she cares about acting X-ly, but can say nothing whatsoever about X, then there is nothing to make it the case that her term "X" refers to rightness, kindness, or any other particular moral feature, rather than referring to anything else. So the agent must grasp at least part of the nature of the thing that is the object of her motivation, to be able to refer to it. But she need not attain this grasp by knowing what falls below the relevant feature in the true metaphysical hierarchy. There is a difference between the *nature* of moral rightness and the things that rightness consists in. This difference explains why people who have very different views about what the right-making features are can still enter into substantive disagreement with one another, rather than talking past each other. They disagree about what the property of moral rightness consists in, but they share an understanding of its nature. The same holds, *mutatis mutandis*, for disagreement over what constitutes one of the right-making features.

There are ways of elucidating the nature of moral rightness that are neutral as to what the right-making features consist in, relying on conceptual connections between rightness and other normative concepts, which can be what these people have in mind. For example, one might think that the morally right is that which could secure the agreement of all reasonable persons, or that the morally right is that which a suitably idealized observer would recommend, or that the morally right is what which we would be subject to fitting blame for failing to perform. Or one might characterize moral rightness as the property of being required by the true moral theory, or the property of being supported by the balance of moral reasons, or the property of responding adequately to all the morally significant features of one's situation – provided one understands enough about the nature of morality to distinguish *moral* theory, reasons, and significance from other kinds of theory, reasons and significance (of prudence, say, or of nutrition). For present purposes I will remain neutral on whether any of these glosses on the concept of moral rightness is correct,[17] and on the question of exactly how many and which of them one must have in mind in order to grasp the concept of moral rightness. I think it is plausible that the concept MORAL RIGHTNESS is a cluster concept. But one must grasp *something* along these lines for moral rightness to be the object of one's motivation.

This helps us to see how someone could become intrinsically motivated by rightness *de dicto* without simply caring about some right-making feature/s and seeing moral rightness as a property constituted by it/them. Someone could initially have various intrinsic motivations whose objects are right-making features, with reference to which she comes to acquire the concept of moral rightness. Once she has acquired this concept, she can re-conceptualize the objects of her motivations, coming to understand them as realizers of moral rightness. She can then kick away the ladder; she can begin to think that it is acting morally rightly that she cares about, whether or not it turns out to consist in the things that she initially used to grasp the concept. For example, someone could acquire the concept of moral rightness with reference to instances of fairness

[17] For defenses of various versions of these glosses, see Railton (1989), (1993); Gibbard (1990); Smith (1994); Darwall (2010); Scanlon (1998); Stratton-Lake (2002); Cuneo and Shafer-Landau (2014).

and kindness – sharing things among classmates at school or among siblings at home, say – and could later wonder whether perhaps those things were never morally right and moral rightness consists in maximizing utility alone. This person cares about moral rightness, but does not merely see it as a property constituted by one or more concrete features that she cares about, since she understands that she might be mistaken about whether these features are right-making (though she still thinks that they are right-making, and cares about them on this basis). Thus, she cares about rightness *de dicto*. Again, parallel remarks apply to right-making features. For example, someone who initially cares intrinsically about meritocracy and reparations for past injustice can re-conceptualize these things as realizers of fairness, and can subsequently decide that it is fairness that she really cares about, whether or not it consists in meritocracy and reparations (though she still believes that it does, and still cares about these things on that basis). This person has come to be motivated by fairness *de dicto*.

On the partial credit approach, we can have praiseworthy motivations throughout this re-conceptualization process. Since intrinsic and well-derived realizer motivations directed toward features in the hierarchy are both praiseworthy, the agent's motivation remains praiseworthy when it alters from being an intrinsic to a well-derived realizer motivation. And the intrinsic motivation directed toward a higher-up moral feature in the metaphysical hierarchy that she develops as a result of this process is also praiseworthy.

I will make one last point about praiseworthiness on the partial credit approach before closing. What I think is praiseworthy is *motivation* by a moral feature in the metaphysical hierarchy. This is not easy to come by. For example, motivation by rightness *de dicto* requires more than just sitting around saying "I love rightness". Motivation is a complex state that gives rise to the four dispositions discussed in §2. So, an agent is motivated by rightness *de dicto* to the extent that she is disposed to spend time thinking about what moral rightness consists in, to notice the moral quality of her acts, and to choose to perform some acts and refrain from others on the grounds that this is what is morally right. These dispositions come in degrees, because motivation comes in degrees. And I am happy to say that a motivation's praiseworthiness also comes in degrees, corresponding to its strength. So it takes more to be praiseworthy, on my approach, than a cynical reader may imagine. Once again, parallel remarks apply to motivation by each of the right-making features.

There is a great deal of work still to be done in spelling out the partial credit approach. All real people clearly fall far short of full credit: our motivations are directed toward some but not all of the features in the metaphysical hierarchy, and also come in degrees. We need a way to compare the amounts of credit that different people get when their motivations align with different parts of the true moral theory, and are present to different degrees. And some motivations might count more than others, if their objects are more important. In this case, the total praiseworthiness of someone's motivations will be a weighted sum of the strength of each of her motivations whose object is a moral feature in the true metaphysical hierarchy, weighted by the importance of the feature. But even this aggregative model might still be too simple, if there are combinatorial effects analogous to those we see in the literature on normative reasons.[18] Working this all out is far beyond the scope of the present paper. But, happily, this is not a task that I face alone. The question of how total praiseworthiness is to be calculated arises whether or not the partial credit approach is correct, and whether or not motivation by rightness *de dicto* is praiseworthy, so long as there are at least two praiseworthy motivations. Even someone who accepts the false view and the hard-line approach faces the task of calculating and comparing overall praiseworthiness as long as she thinks that there is more than

[18] See, for instance, Horty (2007); Nair (2016); Wassel (*ms.*). I argue that there is a hitherto under-appreciated class of combinatorial effects arising from metaphysical relationships between normative reasons, of the sort that I described in the hierarchy discussed in §2, in my "We Can Have Our Buck And Pass It Too" (available on my website).

one right-making feature. This task is a bigger and more complicated task on my view than on some others. But it is a task that everyone faces. I look forward to facing it in future work.

## 4. Conclusion

Motivation by rightness *de dicto* looks bad if we compare an agent who is trying to act rightly and failing with one who is trying to perform acts with a certain right-making feature and succeeding. But these cases are not minimal pairs. They vary in whether the agent is succeeding or failing at what she is trying to do, and, crucially, in whether she is acting rightly or wrongly. They do not isolate the key issue of motivation by rightness *de dicto* vs. *de re*.

I have argued that, when we compare correctly constructed cases, motivation by rightness *de dicto* looks every bit as praiseworthy as motivation by rightness *de re*. To deny this yields unduly harsh verdicts about agents who try to act rightly and even partly succeed, especially as compared with those who manage to act rightly without trying. The false view entails that the motivations of moral saints like Maryam are not at all praiseworthy, while those of people like Mario are fully praiseworthy. These extreme differences in praiseworthiness seem arbitrary and unmotivated. We could avoid this by saying that realizer motivations whose objects are right-making features are praiseworthy, but this yields unduly positive verdicts about oddballs like Aarulina. We should instead hold that motivation by rightness *de dicto* is praiseworthy.

Turning to cases involving agents who try to act rightly but fail, I have argued that all reasons to question the praiseworthiness of their motivations apply equally well to agents who are motivated by right-making features but fail to perform acts with these features. Any of these motivations can lead a person with false moral beliefs to act wrongly, and any can "corrupt" a person by slowly eroding her instinctual concern for that which really does matter morally. We should respond to this by distinguishing the praiseworthiness of motivations, acts, and beliefs, acknowledging that someone's good intentions can be a redeeming feature even if they believe and act badly. We might want to make an exception to this approach for agents whose beliefs and acts are completely beyond the pale, though we could instead argue that they do not really have the moral motivations that they claim to have, either because they fail to refer to moral properties or because they are insincere.

Lastly, I have argued that when evaluating agents who are motivated by some but not all of the features in the true metaphysical hierarchy, we should take a "partial credit" approach, not a hard-line approach. The partial credit approach's evaluations of real people are more lenient, in comparison with the hard-line approach, the more people there are who fail to grasp the true nature and extension of all the right-making features, while still caring about those features. The partial credit approach will give these people credit for their *de dicto* motivations, while the hard-line approach will not. But most people have only an inchoate grasp of the right-making features, which leads them to make moral mistakes. So the hard-line approach implausibly entails that almost no-one has any praiseworthy motivations, while the partial credit approach enables us to recognize the extent of each agent's moral success.

This paper is for the souls whose intentions are good (*de dicto*). I hope they will no longer be misunderstood.

REFERENCES

Aboodi, Ron (2016). "The Wrong Time to Aim at What's Right: When is *De Dicto* Moral Motivation Less Virtuous?" *Proceedings of the Aristotelian Society* 115(3), 307-314.

Adams, Frederick (1995). "Trying: You've Got to Believe". *Journal of Philosophical Research* 20, 549-561.

Arpaly, Nomy and Timothy Schroeder (2013). *In Praise of Desire*. Oxford: Oxford University Press.

Berridge, Kent (2003). "Pleasures of the Brain". *Brain and Cognition* 52, 106-28.

Calhoun, Cheshire (1989). "Responsibility and Reproach". *Ethics* 99(2), 389-406.

Carbonell, Vanessa (2013). "De Dicto Desires and Morality as Fetish". *Philosophical Studies* 163, 459-477.

Copp, David (1997). "Belief, Reason, and Motivation: Michael Smith's *The Moral Problem*". *Ethics*, 108(1), 33-54.

Cuneo, Terence and Shafer-Landau, Russ (2014). "The Moral Fixed Points: New Directions for Moral Nonnaturalism." *Philosophical Studies* 171(3), 399-443.

Darwall, Stephen (2010). "But It Would Be Wrong." *Social Philosophy and Policy* 27(2), 135-157.

Davidson, Donald (1973). "Freedom to Act". In T. Honderich, ed., *Essays on Freedom of Action*. London: Routledge.

Dreier, James (2000). "Dispositions and Fetishes: Externalist Models of Moral Motivation". *Philosophy and Phenomenological Research*, 60(3), 619-638.

Enoch, David (2011). *Taking Morality Seriously*. Oxford: Oxford University Press.

Enoch, David (2014). "A Defense of Moral Deference". *The Journal of Philosophy* 111(5), 229-258.Fleary

Gibbard, Allan (1990). *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.

Harman, Elizabeth (2011). "Does Moral Ignorance Exculpate?" *Ratio* 24, 443-468.

Harman, Elizabeth (2015). "The Irrelevance of Moral Uncertainty". In R. Shafer-Landau, ed. *Oxford Studies in Metaethics*, vol. 10. Oxford: Oxford University Press.

Hedden, Brian (2016). "Does MITE Make Right? Decision-Making Under Normative Uncertainty." In R. Shafer-Landau, ed., *Oxford Studies in Metaethics*, vol. 11. Oxford: Oxford University Press.

Horty, John (2007). "Reasons as Defaults". *Philosophers' Imprint* 7(3), 1-28.

Korsgaard, Christine (1983). "Two Distinctions in Goodness". *The Philosophical Review* 92(2), 169-195.

Leary, Stephanie (2017). "Non-Naturalism and Supervenience". In R. Shafer-Landau, ed., *Oxford Studies in Metaethics*, vol. 12. Oxford: Oxford University Press.

Lillehammer, Hallvard (1996). "Smith on Moral Fetishism". *Analysis*, 57(3), 187-195.

Lockhart, Ted (2000). *Moral Uncertainty and its Consequences*. Oxford: Oxford University Press.

Ludwig, Kirk (1995). "Trying the Impossible: Reply to Adams". *Journal of Philosophical Research* 20, 563-570.

Markovits, Julia (2010). "Acting for the Right Reasons". *Philosophical Review* 119(2), 201-242.

McDowell, John (1998). "Non-Cognitivism and Rule-Following". In *Mind, Value, and Reality.* Cambridge, MA: Harvard University Press.

Mill, John Stuart (1871). *Utilitarianism*, 4th ed. London: Longmans, Green, Reader, and Dyer.

Moody-Adams, Michele (1994). "Culture, Responsibility, and Affected Ignorance." *Ethics* 104(2), 291-309.

Morillo, Carolyn (1990). "The Reward Event and Motivation." *The Journal of Philosophy* 87, 169-186.

Nair, Gopal Shyam (2016). "How Do Reasons Accrue?" In E. Lord and B. Maguire, eds., *Weighing Reasons*. Oxford: Oxford University Press.

Nissan-Rozen, Ittay (2015). "Against Moral Hedging". *Economics and Philosophy* 3, 1-21.

Olson, Jonas (2002). "Are Desires *De Dicto* Fetishistic?" *Inquiry* 45(1), 89-96.

Putnam, Hilary (2002). "The Entanglement of Fact and Value". In *The Collapse of the Fact/Value Dichotomy and Other Essays*. Harvard, MA: Harvard University Press.

Rabinowicz, Wlodek, and Ronnow-Rasmussen, Toni (2000). "A Distinction In Value: Intrinsic and For its Own Sake". *Proceedings of the Aristotelian Society* 100(1), 33-51.

Railton, Peter (1986). "Moral Realism". *The Philosophical Review* 95(2), 163-207.

Railton, Peter (1993) "Noncognitivism about Rationality: Benefits, Costs, and an Alternative." *Philosophical Issues* 4, 36-51.

Rawls, John (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Roberts, Debbie (2013). "Thick Concepts". *Philosophical Compass* 8, 677-688.

Scanlon, Thomas (1998). *What We Owe To Each Other.* Cambridge, MA: Harvard University Press.

Schroeder, Severin (2001). "The concept of trying". *Philosophical Investigations* 24(3), 213-227.

Sepielli, Andrew (2009). "What to do when you don't know what to do". *Oxford Studies in Metaethics* 4, 5-28.

Sepielli, Andrew (2013). "What to do when you don't know what to do when you don't know what to do…" *Noûs* 47(1), 521-544.

Sinhababu, Neil (2017). *Humean Nature*. Oxford: Oxford University Press.

Sliwa, Paulina (2016). "Moral Worth and Moral Knowledge". *Philosophy and Phenomenological Research* 93(2), 393-418.

Smith, Angela (2005). "Responsibility for Attitudes: Activity and Passivity in Mental Life". *Ethics* 115(2), 236-271.

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

Smith, Michael (1996). "The Argument for Internalism: Reply to Miller". *Analysis*, 56, 175-184.

Strandberg, Caj (2007). "Externalism and the Content of Moral Motivation". *Philosophia* 35, 249-260.

Svavarsdóttir, Sigrún (1999). "Moral Cognitivism and Motivation". *Philosophical Review* 108, 161-219.

Toppinen, Teemu (2004). "Moral Fetishism Revisited." *Proceedings of the Aristotelian Society* 104(3), 305-313.

Väyrynen, Pekka (2013). *The Lewd, The Rude and The Nasty: A Study of Thick Concepts in Ethics*. Oxford: Oxford University Press.

Wassel, Damian (*ms*.) "When Are Accruals of Reasons Stronger Than Their Elements?"

Weatherson, Brian (2014). "Running Risks Morally". *Philosophical Studies* 167(1), 141-163.

Wolf, Susan (1982). "Moral Saints". *Journal of Philosophy*, 79(8), 419-439.

Zangwill, Nick (2003). "Externalist Moral Motivation". *American Philosophical Quarterly*, 40(2), 143-154.

Yaffe, Gideon (2010). *Attempts: In the Philosophy of Action and Criminal Law.* Oxford: Oxford University Press.