What are We Praiseworthy For?

ZOË JOHNSON KING
University of Southern California

1. Why think about praiseworthiness?

Philosophers and legal scholars act like we don't need to know what praiseworthiness is. We write about moral responsibility with a near-exclusive focus on blame and punishment, rarely mentioning praise except in parentheses after an "or". Thus we write as though we think that, once we get the correct theory of blameworthiness figured out, we'll easily extract the correct theory of praiseworthiness just by crossing out all of the negatively-valenced words and writing in their positively-valenced opposites.

But a little reflection reveals plenty of reasons to think that praise is not just the positive flip-side of blame. For example, consider strict liability. We sometimes hold people accountable — at least legally, if not morally — for having brought about terrible outcomes regardless of whether they intended to do so, whether they were aware that they were doing so, and whether they could or should have been aware that they were doing so. Nothing like this seems true of praise. The mere fact that movements of someone's body were a significant event in the causal past of a good outcome is never sufficient for her to deserve praise for this good outcome, no matter how good it is. So, it looks as though at least one important element of our blaming practice has no positive analogue.

Next, consider negligence. Roughly: if you weren't aware that there was a risk of your action's having some bad feature or outcome, but you *should* have been aware of this risk, then you were negligent. Negligence is often a source of blameworthiness, and indeed much of contemporary scholarship on tort law consists of analyses and discussions of negligence. But it is hard to see what a parallel phenomenon would be for praiseworthiness. We might try simply changing the word "bad" to "good", and thus saying that if you weren't aware that there was a risk of your action's having some good feature or outcome, but you *should* have been aware of this risk, then you were the-opposite-of-negligent. But that would be an odd concept. For starters, the fact that one's action might have a good feature or outcome is not a "risk". And what could we mean when we say that you "*should* have been aware" of the chance of your action's having a good feature or outcome? In the negative case, the assumption is that you should have been aware of the risks because, if you had been, then you could (and hopefully would) have taken the appropriate precautions. But it is hard to find a parallel for this in the positive case, as we do not need to take precautions against the potential good features of our actions. Again, then, an important part of our thinking about blameworthiness has no clear analogue when it comes to praiseworthiness.

Now consider recklessness. Roughly again: if you were aware that there was a risk of your action's having a certain bad feature or outcome, but you went ahead and did it anyway without taking the appropriate precautions, then you were reckless. Again, recklessness is often a source of blameworthiness. But, again, it is hard to see what a parallel phenomenon would be for praiseworthiness. Changing the word "bad" to

"good" gives us that if you were aware that there was a risk of your action's having a certain good feature or outcome, but you went ahead and did it anyway without taking the appropriate precautions, then you were the-opposite-of-reckless. And this is another odd concept. Again, the fact that one's action might have a good feature or outcome is not a "risk", and we do not need to take "precautions" against potential good features of our actions. Nor does it make much sense to think of someone who is aware of a chance that her action will have a good feature or outcome and who performs it "anyway". In the negative case, the word "anyway" suggests that we should regret taking the risk and regret whatever bad situation obtains if it eventuates. But it is hard to find a parallel phenomenon in the positive case, since we should welcome — not regret — the good features of our actions.

This is not to say that there are no ways to patch up these concepts so that they describe intelligible and interesting phenomena. On the contrary, there are many possible ways of doing so. For instance, perhaps a true "opposite" of negligence would be a case in which someone is unaware of a good feature or outcome of her action and *shouldn't* have been aware of it, such that her lack of awareness is itself praiseworthy just as the negligent person's ignorance is itself blameworthy. (But why would being unaware of the potential good features of one's action be praiseworthy?) And perhaps a true "opposite" of recklessness would be a case in which someone is aware of the potential good features of her action but *indifferent* to them, just as the reckless person is indifferent to the risks posed by her action. (But why would it be praiseworthy to be indifferent to one's action's good features?) My point is just that it is far from obvious how this patch-up is going to go. If these varieties of blameworthiness do have recognizable analogues when it comes to praiseworthiness, the precise nature of these analogues is yet to be determined. And if they have no analogues, that's a pretty big difference between praiseworthiness and blameworthiness. Either way, then, it looks as though we aren't going to be able to simply read off the correct theory of praiseworthiness from the correct theory of blameworthiness.

Similar points apply to some restrictions on the standing to blame. For example, it is widely thought to be inappropriate to blame someone for wrongful conduct that you yourself have engaged in, as your blame would then be hypocritical. And it is widely thought to be inappropriate to blame someone for wrongful conduct that you yourself facilitated, as you are then complicit in their wrongdoing. But nothing like this applies to praise. If you regularly engage in a certain type of good conduct, it remains appropriate for you to praise others who engage in it too. And if you help someone to do something good, it remains appropriate for you to praise them for doing the good thing. Indeed, in cases of complicity it would be appropriate to "own up" to your involvement in the wrongdoing, but in the positive case it is this that would be inappropriate — emphasizing your own involvement in someone else's achievement would be hogging the limelight. Moreover, some accounts of what is wrong with complicit and/or hypocritical blame clearly fail to generate positive analogues. For instance, some philosophers think that hypocritical blame is inappropriate because it violates an equal interest that we all have in avoiding the negative reactive sentiments.[1] But we do not have an interest in avoiding the positive reactive sentiments. Other philosophers think that hypocritical blame is inappropriate because it violates a duty to engage in critical self-scrutiny.[2] But there is no duty to "scrutinize" our moral achievements as we do our moral failures, and indeed doing so may seem like a distasteful form of immodesty.[3] Again, then, praise and blame just don't seem to work in the same way.

[1] See e.g. Wallace (2010).
[2] See e.g. King (2020).
[3] On which see Bommarito (2013).

If we want to understand praiseworthiness, then, it looks as though we'll have to think about it in its own right. And we should want to understand praiseworthiness. For the facts about praiseworthiness are half of the facts about moral responsibility. I am genuinely puzzled as to why, although philosophers and legal scholars have discussed moral responsibility for millennia, we have focused so extensively on the negative. But I want to understand the positive side of the story. And I intend to take some steps toward our doing so. That is the point of this paper.

2. Introducing the "both" view

Philosophical thinking on blameworthiness can be sorted, sweepingly and with some oversimplification, into two broad historical traditions. In one tradition are the *voluntarists.* These theorists hold that moral responsibility is centrally tied to deliberate action; we're fundamentally blameworthy for what we voluntarily *choose to do*. When it comes to negligence and recklessness, philosophers in this camp typically either deny that we're blameworthy for them or assert that our blameworthiness must be "traced" back to a prior deliberate action or omission.[4] In the other tradition are the *quality-of-will theorists.* These theorists hold that moral responsibility is centrally tied to motivations, intrinsic desires, or other conative states; we're fundamentally blameworthy for what we *care about*. Philosophers in this camp think that not only negligence and recklessness, but also deliberate action itself, is blameworthy only insofar as it manifests a poor quality of will.

These camps are often thought to be at odds with one another. They are seen, and see themselves, as vying for the "top spot" — that is, as arguing about whether deliberate actions or conative states are the more plausible candidate for being the fundamental locus of moral responsibility. Thus philosophers in both camps sometimes try to explain the other candidate's intuitive appeal in terms of its relation to their own preferred candidate. Voluntarists emphasize that we often deliberately cultivate our characters and may also recklessly or negligently fail to prevent the development of ill will.[5] Meanwhile, quality-of-will theorists observe that deliberate action reveals us to have certain motivations and that negligent or reckless behavior reveals us to be insufficiently concerned with that which we thereby endanger.[6] If someone's act or omission does not reveal ill will, then quality-of-will theorists will insist that it is not blameworthy after all. Likewise voluntarists will deny that ill will is blameworthy if it cannot be traced to a blameworthy action or omission. This is an ongoing dispute.

My view is about praiseworthiness rather than blameworthiness. And I just argued that praiseworthiness and blameworthiness come apart in all sorts of curious ways. Nonetheless, my view can be perspicuously introduced by comparing it to these two prominent traditions in our philosophical thinking about blame. Here it is easy to see what the positive analogues are: voluntarism about praiseworthiness is the view that we are fundamentally praiseworthy for what we deliberately do, with everything else being praiseworthy only insofar as it can be "traced" to the praiseworthiness of prior deliberate actions or omissions, while the quality-of-will theory of praiseworthiness is the view that we are fundamentally praiseworthy for what we care about, with everything else being praiseworthy only insofar as it manifests a good quality of will. In this context, my view can be understood as a kind of *dualism* about praiseworthiness. I think that we are fundamentally praiseworthy *both* for what we do *and* for what we care about. Thus, my view explores the possibility that the two great traditions in philosophical thinking about moral responsibility were both right

[4] On tracing see e.g. Fischer and Tognazzini (2009).
[5] See, e.g., Rosen's remarks on negligently allowing indifference to "fester" in his (2008), pp.607-609.
[6] See, e.g., Smith's discussion of the example of forgetting a birthday in her (2005).

all along – at least when it comes to praiseworthiness – except for the part where they each took the other to be wrong.

I am unswayed by purported reductions of one locus of responsibility to the other. I do think that the very concept of deliberate action entails that deliberate action requires prior motivation – a conative state that is itself eligible for praise (or blame). This might seem like fodder for the quality-of-will theorist, who can say that in praising someone's deliberate action we are "really" praising the underlying conative state. But that would be too quick. The fact that something has a precondition does not mean that, in praising it, we are "really" praising the precondition. Deliberate action has so many preconditions – the presence of oxygen in the environment, for instance – that it is easy to think of counterexamples to that idea. Likewise, I think that the quality of our wills is not fixed over time, and that someone whose good will is the result of careful work on herself over time is more praiseworthy than someone whose conative states are god-given or the result of a bump on the head. (I will say a little more about this at the end of the paper.) But this does not imply that, in praising someone's good will, we are "really" praising the effort that went into its development. In general, something can have a praiseworthy precondition or a praiseworthy causal antecedent without its praiseworthiness collapsing into that of the precondition or antecedent.

The right lesson to draw from these purported reductions, I think, is just that the two things for which we can be praiseworthy are intricately related. We are praiseworthy for the deliberate cultivation of good will. And we are praiseworthy for the good will manifested in deliberate action. So, if we were to try to calculate someone's total overall praiseworthiness, then we would discover a host of praiseworthy motivations in the course of assessing her praiseworthy actions (since praiseworthy motivations underlie praiseworthy actions), and we would probably discover some praiseworthy actions if we traced the causal history of her praiseworthy motivations (since most praiseworthy motivations are at least in part the result of our efforts to cultivate them). But none of this means that one type of praiseworthiness should be understood entirely in the other's terms. Rather, I think, it suggests that the two most historically prominent traditions in philosophical thinking about moral responsibility are each on to something.

The "both" view has some intuitive plausibility, independent of the history of philosophical thought. For it is natural to think that, in morally assessing people, we care about their accomplishments and also about the good in their hearts. Being well-motivated is great, but it is even better for someone's good motivations to lead her to do good things. Similarly, being ill-motivated is bad, but we also care about the poor choices that people make and the caution that they fail to show in virtue of their conative lacunae. And people whose outwardly identical behavior is underlain by motivations of wildly different varieties and strengths are not exactly equally praiseworthy overall; the total set of someone's motivations is relevant to how good a person she is, not just those motivations that find an outlet in her actions and omissions. Moreover, people can be praiseworthy for *trying* to accomplish something good even if they fail, as their motivations remain praiseworthy. Likewise we can be blameworthy for ill will even if fortuitous circumstances prevent it from ever affecting what we do. Hence the intuitive appeal of the view that we are fundamentally praise- and blameworthy both for what we do and for what we care about: both for our accomplishments (or failures) and for the good (or bad) in our hearts.

In the remainder of this paper, I will articulate two positive views. One is about how to evaluate the good in people's hearts — the praiseworthiness of their motivations. The other is about how to evaluate people's accomplishments — the praiseworthiness of their actions. From this point on, the paper is programmatic: rather than giving novel arguments for novel views, I aim to summarize and coherently unify the positive views that I have defended elsewhere.

3. Praiseworthy motivations

I subscribe to a version of the view that we are praiseworthy for caring about that which is, in fact, morally significant.

To spell this out clearly, we must make a brief foray into moral metaphysics.[7] When an act is morally right, the fact that it is right is not a brute fact. There is always some feature (or set of features) of the act that *makes* it right — for example, that it is fair. But, whenever an act possesses one or more of these right-making features — for example, when it is fair — the fact that it possesses the feature(s) is not a brute fact either. On the contrary, whenever an act possesses a right-making feature, there is always some further feature or set of features in virtue of which it possesses this feature. For example, it might be fair because it is meritocratic, or because it distributes resources based on need, or because it makes reparations for past injustice. And these features are not brute either. So on we might go, describing a metaphysical hierarchy of features of the act and the "makes it the case" relationships that they bear to one another. In what follows I assume that there is a uniquely correct first-order moral theory, which, if fully fleshed out, would tell us what the right-making features are, what the right-making-feature-making features are, and so on. I will refer to the metaphysical hierarchy thus described as "the true hierarchy".[8]

Some motivations are intrinsic: we pursue things for no further end, but "for their own sakes". Intrinsic motivations can give rise to other motivations, in combination with beliefs about causal or constitutive relationships that obtain between the objects of the intrinsic motivations and other objects. The more well-known are *instrumental motivations*: we pursue things because they are part of a causal chain that leads to the object of our intrinsic motivation, or at least raises its probability. For example, a runner may train for a race because she wants to win and knows that training improves her chances of winning. There are also *realizer motivations*, which are like instrumental motivations except that they are based on beliefs about constitutive rather than causal relationships between things.[9] For example, the runner may want to get her torso across the finish line faster than her competitors because that is what *constitutes* winning, as opposed to being a step in a causal chain on the way to winning. Motivations whose objects appear in the true hierarchy can be like this; for example, my motivation to be meritocratic might derive from my belief that meritocracy is a kind of fairness and my motivation to be fair, while my motivation to be fair might in turn derive from my motivation to act rightly and my belief that fairness is a right-making feature. We can therefore subdivide realizer motivations whose objects appear in the true hierarchy into the *well-derived* and the *poorly-derived* ones. Well-derived realizer motivations derive from intrinsic motivations whose objects are moral features in the hierarchy and correct beliefs about the metaphysical relationships that hold the hierarchy together – they are the motivations that result when someone cares about something that genuinely matters and figures out what it in fact consists in. By contrast, poorly-derived realizer motivations derive from intrinsic motivations whose objects are not in the true hierarchy and/or false beliefs about it, but coincidentally end up with features that actually are in the true hierarchy as their objects.

[7] I go into this view in a little more detail in my (2019). It is inspired by Leary (2017).
[8] N.B. I focus on the true metaphysical hierarchy where the thing at the top is *rightness*. But I am open to the possibility of other such hierarchies with other moral things — say, *goodness* — at the top. My view about them would be parallel. There may also be other "flavors" of normativity that people can be praiseworthy for according with; perhaps as well as moral praiseworthiness there are such things as *aesthetic* or *epistemic* praiseworthiness. If there are such things, my view about them would again be parallel, though I'm not sure which features they would have at the top.
[9] I take the concept of a realizer motivation from Arpaly and Schroeder (2013).

That's enough background to understand my view about praiseworthy motivations. My view is that we are praiseworthy for our intrinsic motivations whose objects are moral features in the true hierarchy, and also praiseworthy for well-derived realizer motivations whose objects appear anywhere in the hierarchy.[10] In short: it's good to care about what really matters, and it's also good to figure out what matters and care about it accordingly. But, on my view, it's okay if someone hasn't figured out precisely what the thing she cares about consists in and developed well-derived realizer motivations all the way down. Most of us are well-meaning but somewhat morally ignorant, identifying and caring about some parts of the true hierarchy – some things that really matter – but not the rest. On my view, that is okay. For example, suppose that a parent wants to treat their children fairly but falsely believes that it is fair to give the children exactly equal resources despite their age differences. This parent may be blameworthy for treating their children unfairly, and perhaps also blameworthy for even thinking that fairness could consist in strict equality of resources. But surely they are still praiseworthy for wanting to treat their children fairly – that's the part that they are getting right. My view is designed to accommodate this verdict. On my view, even if the parent has no motivations whose objects are the things that in fact fall below fairness in the true hierarchy, the fact that they do at least want to treat their children fairly remains praiseworthy. I call this "the partial credit approach" (2020a, p.424).

There is a distinction between moral and non-moral features within this view, so as to generate plausible intuitions about cases. Intuitively, it is good to be intrinsically motivated to act fairly. And, intuitively, it is good to be intrinsically motivated to act rightly, to understand that fairness is a right-making feature, and thus to develop a realizer motivation to act fairly. Intuitively, it is good to be intrinsically motivated to make reparations for past injustice. And, intuitively, it is also good to be intrinsically motivated to act fairly, to understand that making reparations is a way of acting fairly, and thus to develop a realizer motivation to make reparations. And so on. For moral features in the hierarchy, then, both intrinsic and well-derived realizer motivations seem praiseworthy. But the same does not hold of the non-moral features that appear once we get "low" enough.[11] For example, wanting to make sure that people get plenty of sleep because one understands that this is an essential component of human well-being would be praiseworthy, but wanting to make sure that people sleep a lot for its own sake would just be weird. Hence, my view is that, for non-moral features in the hierarchy, only well-derived realizer motivations are praiseworthy – not intrinsic motivations. In general, on the partial credit approach, the better one appreciates how and why something matters (i.e., its position in the true hierarchy), the more praiseworthy one's total motivational set will be.

This view suggests a plausible picture of moral learning and moral development. There are two ways we can revise our motivational set so that it more closely matches the true hierarchy: *top-down* and *bottom-up*. We engage in *top-down* revision of our motivational set when we reflect on something that we care about and form true beliefs about what it consists in, developing the corresponding realizer motivations. This can involve developing new beliefs and motivations where there previously were none before (if we had no idea what the thing consists in), or replacing false beliefs and poorly-derived realizer motivations with the correct and well-derived ones (if we were mistaken). For example, the parent who believes that fairness consists in strict equality may reconsider, perhaps thinking about new arguments or talking to people about the issue, and may thereby change her mind. By contrast, we engage in *bottom-up* revision of our

[10] I develop and argue for this view more fully in my (2020a).

[11] Depending on how the debate between metaethical naturalists and non-naturalists turn out, there may be a level in the hierarchy below which *all* features are non-moral features. I will avoid taking a stand on this for present purposes. On any plausible view, there will be at least some non-moral features in the hierarchy, since acts' moral features depend at least in part on their non-moral features.

motivational set when we reflect on a set of things that we care about and come to realize that they have something important in common that ultimately explains why they all matter, which we thereby come to care about. For example, someone who is initially intrinsically motivated to be meritocratic, to make reparations for past injustice, and to distribute resources based on need, may on reflection come to see that these things are united insofar as they are all realizers of fairness. She may then decide that it is fairness that she ultimately cares about. In such cases the agent replaces an initial set of disparate-seeming intrinsic motivations with a set of realizer motivations directed toward the same objects, now reconceptualized as morally significant *qua* realizers of a more general feature (e.g. fairness) by which she is henceforth intrinsically motivated. On my view, both of these kinds of moral development increase an agent's overall praiseworthiness.

The partial credit approach also fits well with our everyday evaluations of people's motivations. Most of us are not completely certain of a particular precise first-order moral theory. Yet we often think that, once we get to know someone reasonably well, we have a reasonably clear sense of how good a person they are. My view explains this by observing that, once we know someone reasonably well, we usually take ourselves to have a reasonably clear sense of at least some of their intrinsic motivations. And that is enough information to make an initial rough assessment of their praiseworthiness. For, once we know the object of someone's intrinsic motivation, we need not be certain of exactly where this object falls in the true hierarchy in order to tell whether it is a praiseworthy motivation. If it is a moral feature, then as long as we are confident that it is in fact morally significant, we can be confident that it appears somewhere or other in the true hierarchy. On my approach, that is sufficient for our friend's motivation to be a praiseworthy motivation. If the object of our friend's intrinsic motivation is a moral feature, but one that we doubt is in fact morally significant – like chastity or revenge – then we can be confident that the motivation is not praiseworthy. And if the object of their intrinsic motivation is a non-moral feature, then we can again be confident that it is not a praiseworthy motivation. Though we may not be certain of the exact structure of the true hierarchy, we usually take ourselves to have a fairly decent grasp of what really matters and what doesn't, and thus of which features appear somewhere in the hierarchy and which ones don't. According to the partial credit approach, this is all we need to make rough initial assessments of people's praiseworthiness once we are aware of some of their intrinsic motivations.

One striking implication of my view is that a maximally praiseworthy agent would have *lots* of motivations. This is because it is possible to have both an intrinsic motivation and a realizer motivation with the same object. If the object is a moral feature that appears in the true hierarchy, then, on my view, both motivations are praiseworthy. For example, someone may be intrinsically motivated to act fairly while also being intrinsically motivated to act rightly and recognizing that fairness is a right-making feature, and thus having a well-derived realizer motivation to act fairly. I give them credit for both motivations. On my approach, then, a maximally praiseworthy motivational set – one that gets "full credit" – will include both intrinsic and well-derived realizer motivations directed toward every moral feature in the true hierarchy, as well as well-derived realizer motivations directed toward every non-moral feature. This is a demanding view of what it takes to be fully praiseworthy. But I am not troubled by that implication, since, on my approach, it is very easy to be *somewhat* praiseworthy. While we cannot realistically expect that any actual people will attain full praiseworthiness, we can realistically expect that plenty of actual people will attain this more modest goal. And my approach is designed to give plausible verdicts about the overall praiseworthiness of actual agents' motivational sets, most of which (I assume) miss out on vast quantities of the praiseworthy motivations that we could have had while still including a moderate number of praiseworthy motivations. This reflects what I take to be the evident fact that most of us are deeply flawed individuals who nonetheless manage to have quite a few redeeming features. The appropriate response to this evident fact, I think, is to give credit where it's due.

All of this is complicated by the fact that motivations have *strengths* – some are stronger than others. This means that there are two quite different ways for our motivations to "match", or fail to match, with moral reality. First, these motivations' objects could be or fail to be things that are in fact morally significant (i.e., those in the true hierarchy), as we have been discussing. Second, these motivations' strengths could correspond or fail to correspond to the facts about the *relative* moral significance of these things. The true hierarchy, as I have described it, is a metaphysical hierarchy: it tells us what each of the things that matter consists in. It thus orders things by fundamentality rather than by moral importance. But it is highly plausible that some of the features in the hierarchy are in fact more morally important than others – or, equivalently, it is implausible that everything that matters does so to the exact same degree as everything else that matters. So, suppose that someone cares about everything in the true hierarchy, but that her degrees of concern massively fail to correspond to these things' actual relative moral significance. For example, suppose that she cares about individual self-expression far more than anything else, but individual self-expression is not in fact far more important than everything else. Intuitively, this person's motivations could be more praiseworthy, even though they have all the right objects; they could still be more in sync with the facts about what actually matters *more* than what.

This point is all the more compelling if we consider people who are out of sync in their motivations whose objects are particular instances of morally significant things – tropes of the relevant properties, if you like. For example, imagine someone who is ardently committed to reparations for one historically-oppressed group, but displays only a lukewarm interest in reparations for all other such groups. This person's motivations clearly are not maximally praiseworthy. And they might even seem blameworthy; for example, if she is passionate about reparations for her own racial group but only barely interested in the fate of any other racial groups, then she seems racist. The same goes for someone who is strongly motivated to promote educational opportunities for their own children but only very weakly interested in promoting educational opportunities for any other children. In cases of *selective caring* such as these, the agent's motivations definitely do not seem maximally praiseworthy. But these agents might still have *some* degree of motivation directed toward everything that in fact appears in the true hierarchy. What seems off about their motivations is not their objects, but their (relative) strengths.

The above examples suggest that we should evaluate people's motivational strengths in relative terms: we should be concerned with the degree to which their motivations' strengths "match" the facts about the relative moral significance of their objects. But we should also evaluate our motivational strengths in absolute terms. Some familiar ideas from the literature on distributive justice illustrate this point: if we assess the strengths of motivations solely in relative terms, then we will say counterintuitive things about *leveling down* and about *Pareto improvements.* To illustrate the former, suppose that someone's praiseworthy motivations reduce in strength across the board — she becomes apathetic, caring about absolutely everything that matters much less than she previously did. But suppose that her degrees of concern reduce by different amounts, such that their relative strengths come ultimately to better match the facts about their objects' relative significance. Intuitively, we should not say that this agent's total motivational set has gotten more praiseworthy through a strict deterioration in the degree to which she cares about everything. Similarly, suppose that someone's relative degrees of concern initially perfectly match the facts about their objects' relative significance, but she then comes to care just a little bit more about one morally significant thing (honesty, say) without reducing her degrees of concern for any of the others. Intuitively, we should not say that this Pareto improvement in the agent's degrees of concern has made her less praiseworthy.

Or should we? Most tiny Pareto increases to one's degrees of concern seem like improvements. But not all of them do – a tiny Pareto increase to one's degree of concern for one's own racial group is not obviously

an improvement and may be a deterioration, as we just saw. Similarly, if someone was initially fanatically committed to individual self-expression, then certain ways of leveling down might intuitively improve her motivations overall — if her concern for everything else that matters decreases only slightly while her concern for individual self-expression decreases to an appropriate amount, then it seems plausible that this absolute deterioration can be offset by the dramatic improvement to her relative degrees of concern. I think that, these considerations collectively suggest another "both" view about the praiseworthiness of people's motivational strengths. Evaluated individually, a stronger praiseworthy motivation is more praiseworthy than a weaker motivation with the same object: for each thing that matters, it is better to care about it more rather than less. And, evaluated as a set, a closer match between the relative strengths of someone's motivations and the facts about their objects' actual relative significance is more praiseworthy than a less-close match. So, if we try to assess the impact on total praiseworthiness of any particular revision to someone's motivational strengths, the devil will be in the details. Some absolute deteriorations can be counterbalanced by relative improvements, and some relative deteriorations can be counterbalanced by absolute improvements. On other occasions the deterioration is too dramatic for the improvement to make up for it. Figuring out exactly where in these details the devil lies is just another hard task for first-order ethical theory.

This is far from being the only hard task that we will have to complete if we are to calculate the exact overall praiseworthiness anyone's actual total motivational set. To do that, we would have to figure out how to make *trade-offs* between different objects of concern. We would have to decide, for instance, how to think about someone who dedicates her life to mitigating climate change and so has few spare emotional resources with which to tend to the needs of her family members, or someone who dedicates her life to tending to the needs of her family members and as a result can't bring herself to do much about climate change.[12] We would also have to figure out how to make trade-offs between our evaluations of the sheer number of praiseworthy motivations that someone has and of these motivations' strengths; we would have to decide whether it is better to have few strong motivations with the right objects or many weak ones, and whether it is better to have many motivations with the right objects whose relative degrees are out of whack or few such motivations whose relative degrees are on point. I don't have the answers to these questions. I mention them only in order to highlight promising avenues for future research. But I hope that it counts as philosophical progress to have articulated a framework within which we can explore questions like these.

4. Praiseworthy actions

As we've seen, my view about praiseworthy motivations is quite complicated. My view about praiseworthy actions is, on its face, simpler: I hold that we are praiseworthy for deliberately doing good things.

Why "good things"?

Because we aren't praiseworthy for doing bad or morally neutral things. To be more precise: our actions all have very many features, and we can be praiseworthy for their good features even if they also have bad ones. For example, we can be praiseworthy for doing something that helps a student even if the same action also makes us late for a lunch meeting (which might be blameworthy).[13] And, if we are well-meaning but morally mistaken, we might even be praiseworthy for doing things that *seem* good but are not — for

[12] I discuss these and related issues about trade-offs in my (2020c).

[13] Here I speak of agents performing actions with multiple features, rather than of agents performing multiple actions at once. But everything I say can be translated into the latter conceptual framework, for those who prefer it.

instance, we might be praiseworthy for doing something that seems like it will help a student even though it actually hinders the student. But we cannot be praiseworthy *for* hindering the student or *for* making ourselves late to a lunch date. Those are bad features of our actions, and thus not the sort of thing for which we are eligible for praise. Nor could we be praiseworthy for eating a carrot or for sitting down, since these are morally neutral. The only features of our actions for which we can merit praise are the good ones.

Why "deliberately"?

Because there is wisdom in the old voluntarist idea that we deserve praise and blame for what we do, but not for what just happens to us. This idea can be further developed: we don't deserve praise for what just happens *through* us. As we noted earlier, there is no strict liability for praise – not just any old positive thing to which our actions are causally related is something for which we merit praise. For example, I once had an annoying colleague who routinely tried to take credit for improvements in educational attainment among disadvantaged students that followed curricular changes he had made. This was annoying because the colleague had not made the changes in order to bring about these improvements, and indeed he had no idea that they would follow when he implemented the changes. He brought about the improvements entirely by accident. So, intuitively, he was not praiseworthy for bringing them about, though his deliberate activity was in their causal past. The lesson to draw from examples like this is that deliberately doing something (making curricular changes) that in fact brings about a good outcome (educational attainment among disadvantaged students) is insufficient for deliberately bringing about the good outcome, and insufficient for praiseworthiness. Parallel remarks apply to positive features of our actions that are *constitutively* (rather than causally) related to that which motivates us. For example, suppose that I promise to meet you at our local coffee shop at 3pm, forget about my promise, and later independently decide that I want to go to the coffee shop at 3pm. In doing so, I keep my promise to you. But I am not praiseworthy for keeping my promise, since I forgot all about it and kept it only accidentally.[14]

What it takes to do something deliberately is a complex question in the philosophy of action to which I do not have a definite answer. But we can hone in on this concept by considering examples of things that do not fall under it and explaining why they don't. The above examples suggest that, at a minimum, one does not do something deliberately if one has no idea that one is doing it at the time when one does it. Further data come from a well-worn example from the literature on intentional action: if a CEO knows that a company policy will increase profits but also harm the environment, and she approves the policy because she doesn't care about the environment and just wants profit, then most people say that she intentionally harms the environment and is blameworthy for doing so, whereas if she knows that a policy will increase profits and also benefit the environment, but she approves it just because she wants profit and is indifferent to the environmental benefits, then most people deny that the CEO intentionally helps the environment and say that she is not praiseworthy for doing so.[15] This suggests an asymmetry in what we recognize as intentional action. It suggests that we take mere awareness of a feature of one's action to suffice for intentionally performing an action with this feature if the feature is bad, but not if it is good.

In the end, our best philosophy of action might not vindicate this asymmetry in folk intuitions about what it takes to do something intentionally.[16] But, regardless of what the philosophers of action decide, ethicists

[14] I discuss these and related examples in more detail in my (2020b).

[15] This example is originally from Knobe (2003).

[16] Knobe's research challenges the orthodox view that whether someone can be said to have brought about a side-effect intentionally depends only on their attitudes toward the effect (e.g. whether they were trying to bring it about, whether they were aware of it). But, as he points out, "ordinary language does not here constitute a final court of appeal" (p.190).

should acknowledge the underlying folk intuitions about what it takes to deserve credit or blame for features of one's action. These underlying intuitions are that being aware that one's action has a bad feature is sufficient for being blameworthy for performing an action with that feature, whereas mere awareness is much too weak of a relationship to a good feature of one's action to render one praiseworthy for performing an action with this feature. And that point is just a version of my earlier observation that there is no opposite of recklessness: someone who is aware of the good features of her action but unmoved by them, and who performs the action "anyway", is not praiseworthy for performing an action with these good features. So, I think that one lesson to draw from the CEO example is that, in order to do something deliberately in the sense relevant for praise (rather than blame), the fact that your action has a certain feature must be part of *why you're doing it* — that is, part of what motivates you to perform it.

I say "part of" because I think you can be praiseworthy for performing an action with a good feature even if the fact that your action has this feature is not your *sole* reason for performing it. We must say this to secure plausible verdicts about cases in which someone is motivated by two or more good things about an action. For example, if our CEO recognizes that a policy change will help the environment, help the global poor, and help non-human animals, and she is motivated to institute the change by all three considerations, then it would be absurdly harsh to say that she cannot be praiseworthy for bringing about any of these good outcomes just because none was her sole reason for instituting the policy change. It is possible to "kill two birds with one stone" — that is, to recognize multiple appealing things about an action and be motivated by all of them. In such a case, if the appealing things are all good things that the agent does deliberately, then she can be praiseworthy many times over.

However, once we allow that a good feature need not be an agent's sole reason for acting in order for her to merit praise for performing an action with this feature, a striking implication follows. The implication is that, by developing praiseworthy motivations of the sort discussed in section 3, we can ratchet up the total praiseworthiness of our subsequent actions. For example, suppose that the CEO has engaged in some "top down" and "bottom up" expansions of her motivational set, such that she now cares about meritocracy both intrinsically and *qua* realizer of fairness, which she cares about both intrinsically and *qua* realizer of rightness. Now suppose that she institutes a new policy on the grounds that it is meritocratic, *which* is fair, *which* is right, such that each of these considerations is part of what motivates her. On my view, she can then be praiseworthy thrice over: for deliberately doing something meritocratic, deliberately doing something fair, and deliberately doing the right thing. In short, on my view, the deeper our appreciation of how and why an action matters morally, the more praiseworthy (in total) we will be for performing it.

Not all agents are not doing such a great job of developing praiseworthy motivations and manifesting them in praiseworthy actions. Indeed, even someone who is motivated to perform actions with a good feature and succeeds in performing an action with this feature may nonetheless fall short in either of two ways. First, she may be mistaken about the features in the true hierarchy that fall *below* the one she is interested in. For example, she may be motivated to promote justice but mistaken about what justice consists in, say by being committed to an excessively retributivist view. Second, the agent may be mistaken about the features in the true hierarchy that lie *above* the one she is interested in. This will be the case if she has a poorly-derived realizer motivation: a motivation derived from intrinsic motivations whose objects are not in the hierarchy and/or false beliefs about the metaphysical relationships that hold the hierarchy together. For example, she may think that avoiding eating certain foods matters because it is commanded by God, when in fact it matters because it is eco-friendly and cruelty-free.

These ways of being mistaken are a motley crew. But I have grouped them as I have because I think it is helpful to distinguish *lower-down* from *higher-up* mistakes. Lower-down mistakes pertain to features that

are metaphysically "below" the one the agent is interested in. They are mistakes about what this feature consists in. Higher-up mistakes pertain to features that are metaphysically "above" the one the agent is interested in. They are mistakes about the feature's normative significance — about why it matters.

Lower-down mistakes can limit the degree to which someone is praiseworthy for performing an act with a good feature. This is just because they can limit the degree to which she does something with this feature *deliberately*. And that remains the case even if the action's having this feature was part of what motivated her to perform it, since trying to F and F-ing are not jointly sufficient for deliberately F-ing. For example, if I am trying to send an SOS in Morse code, but I am both wrong about what Morse code is and wrong about what I am typing, and yet — by sheer coincidence — what I am in fact typing does in fact spell out SOS in the real Morse code, then it seems wrong to say that I deliberately sent a Morse SOS. I was too confused about why what I was doing constituted sending a Morse SOS to count as having done so deliberately. Such cases have a Gettier-like feel. But we should be careful not to over-generalize from them, as someone need not have a perfect understanding of why an action has a feature in order to deliberately perform an action with the feature. For example, I deliberately turn on a light by flipping a switch even if I have only the vaguest understanding of how electricity works, incorporating various false beliefs that my school science teachers imparted to me as close-enough approximations of the truth. There is vagueness here; deliberate action can withstand a few minor false beliefs, but the more egregiously mistaken someone is about why her action has a certain feature, the less true it seems to say that she performs an action with that feature deliberately. On my view, this correspondingly lessens the degree to which she is praiseworthy for so acting.

Higher-up mistakes don't limit the deliberateness of someone's performing an action with a feature. But, intuitively, they can still limit the agent's praiseworthiness. For example, take the agent who believes that avoiding eating certain foods matters because it is commanded by God, when in fact it matters because it is eco-friendly and cruelty-free. This person is clearly not as praiseworthy as she could be: it would be better if she recognized that avoiding the foods matters because it is eco-friendly and cruelty-free and then did it for this reason. More strongly, her deliberately avoiding the foods might not be praiseworthy at all. Her realizer motivation is poorly-derived, but just happens to alight upon something that in fact appears in the true hierarchy. Thus, this case also has a Gettier-like feel. The fact that what the agent does deliberately – avoiding eating the foods – is a *good* thing to do seems like a coincidence, given her ignorance about why this feature matters. Her ignorance therefore seems to undermine her praiseworthiness for avoiding the foods, despite the fact that she does so deliberately. This observation suggests a modification to the simple statement of my view given above: we are praiseworthy for deliberately doing good things, but it is not enough that we deliberately perform an action with feature F and F is in fact a good feature. Rather, our performing an action with the feature and it's being a good feature must be related in the right way – which excludes poorly-derived realizer motivations. And this means that our motivation to perform an act with a certain good feature must be either an intrinsic or a well-derived realizer motivation. Hence, as I said earlier, praiseworthy motivations and praiseworthy actions are intricately related: praiseworthy actions must have praiseworthy motivations driving them.

What about a case in which someone is not *mistaken* about the features that fall below or lie above the one she is interested in, but merely *ignorant* of these features? For example, a moral novice can learn that an act is fair via testimony, without understanding why it is fair, and can then choose to perform it on the basis of its fairness. This could be a case of lower-down ignorance without mistake, as the agent could have no beliefs whatsoever about what falls below the relevant feature in the true hierarchy. And there can also be cases of higher-up ignorance without mistake: if someone is intrinsically motivated to act in a certain way, then she may have no beliefs at all about what lies above the relevant feature in the true hierarchy.

My view is that ignorance, without mistake, does not undermine praiseworthiness. I think this just because it seems like the right thing to say about these examples. If someone isn't sure which action has a feature that she cares about, and she lacks the time or moral acuity to figure it out, but she can ask someone better-positioned and act accordingly, then doing so is an appropriate way of acting on her motivation. We are not moral islands, and deferring to others about the extension of important moral properties is sometimes the right way to go. But we can still count as deliberately performing an action with a certain feature when we defer to an expert about the presence of the feature – indeed, this is commonplace, as much of our deliberate activity relies on information from other people, signs, apps, and so on, about what will happen if we do things. So, if the feature is a good feature, then our deliberately performing an action with this good feature can be praiseworthy. In such cases, the fact that the agent is ignorant of what the feature consists in just doesn't seem to be a problem.

Cases of higher-up ignorance without mistake are subtler. Imagine someone who is intrinsically motivated to act fairly, who has figured out precisely what fairness consists in, and who thus succeeds in acting fairly. But imagine that she has no grasp whatsoever of why fairness matters — no inkling that fairness is a right-making feature and no understanding of the relationships it bears to the other right-making features. There are probably no real people like this, since real people usually care about fairness because they take it to be a right-making feature. So this case is hard to imagine. But someone who accepts an anti-fairness moral theory and yet instinctively feels the pull of considerations of fairness, similar to some popular descriptions of Huckleberry Finn in the contemporary literature,[17] could be like this. It seems clear that such a person deliberately acts fairly, since her understanding of why her action is fair is impeccable. And the case doesn't quite have the Gettier-like feel of cases involving poorly-derived realizer motivations, since the agent's motivation is intrinsic. Nonetheless, there is something suboptimal about this agent. On my view, this suboptimality is reflected in the fact that, although she may be praiseworthy *for acting fairly*, she is not praiseworthy *for acting rightly*. This is because the fact that her act is morally right is, from her point of view, an accidental offshoot of what she truly cares about, much as the original CEO's helping the environment is from her point of view an accidental offshoot of her pursuit of profit. Both agents are eligible for praise for what they do deliberately, but not for the other good features of their actions — including their moral rightness.

Like the partial credit approach, this view also raises a lot of unanswered questions. One concerns whether there are factors that modify the amount of praise we deserve for deliberately doing good things, besides those that modify the deliberateness of our doing them. For example, the *skill* with which we deliberately do good things may make us more praiseworthy. Or, the *effort* we put in to deliberately doing good things may make us more praiseworthy. Or both. Or something else. Another question concerns the *guises* under which one may deliberately do good things. There are pairs of terms that express the same concept, such as 'fair' and 'juste'. And, clearly, French-speakers can be praiseworthy for acting fairly. But, even within a single language, there may be various words or phrases that can express the same moral concepts and thus be the descriptions "under" which we deliberately do good things. For example, someone who doesn't know the term "eco-friendly" can presumably express the same concept using a phrase like "good for the planet". Perhaps someone who doesn't know the term "cruelty-free" could use "vegan" to express the same concept, even though the extensions of these terms come apart. And, on some metasemantic theories, there

[17] This character would be a little different from all contemporary construals of the Huck Finn case that I am aware of, insofar as she has figured out precisely what fairness consists in. Whatever is supposed to be motivating Huck – be it considerations pertaining to Jim's personhood, their friendship, or what have you – it seems safe to say that he has not figured out precisely what these considerations consist in and developed the corresponding realizer motivations.

can be pairs of concepts that refer to the same moral property – just as, for example, the concept TIPPING BIRDS and the concept SINGLE-LEG DEADLIFT refer to the same exercise. If one of these metasemantic theories is true, there may be quite a range of descriptions under which people deliberately do good things. Again, I raise these unanswered questions just to point out promising directions for future research. A fuller discussion will have to wait for now.[18]

5. Coda: Working on Yourself

I have said a lot about one of the ways in which praiseworthy motivations and praiseworthy actions are related: praiseworthy motivations underlie praiseworthy actions. But I have said comparatively little about the other way: praiseworthy motivations can be developed through praiseworthy actions. So I will close by saying a little more about that.

Our motivational sets aren't permanently fixed. On the contrary, they change a lot over the courses of our lives. And these changes are not just responses to external influences. *We* can change our motivational sets. For one thing, we can engage in "top down" or "bottom up" expansions of our motivational sets. These revisions include epistemic improvements: figuring out the nature and extension of a feature that we care about, or figuring out that a bunch of things we care about are united insofar as they are all realizers of a more general thing. And it is rare for epistemic improvements of either sort to come to the agent as a sudden epiphany. It is much more common for these revisions to our motivational set to be preceded by moral reflection that is conscious, deliberate, and effortful. Moreover, these periods of "reflection" need not be a matter of the agent sitting alone staring into space; they can involve talking to others, reading, writing, and actively thinking through the issues. In short: expansions of our motivational sets often involve a great deal of *activity* on our part, which we engage in because we want to figure out the moral truth and act accordingly. And the fact that an action will help us to figure out the moral truth and act accordingly is clearly a good feature of it. So, the bouts of activity that typically lead to expansions of our motivational sets consist of our deliberately doing good things. Thus, on my view, they consist of praiseworthy actions.

Top-down and bottom-up expansions involve changing our motivations by changing our beliefs. But sometimes our beliefs are doing just fine, and the problem is that our motivations aren't following in their stead. Someone can believe that something matters and yet find herself with little or no motivation to act accordingly. Or she can find herself with motivations whose objects she takes to not really matter at all – motivations to do morally unimportant things like watching funny videos and shopping for sneakers, or motivations to do morally bad things like making snide remarks about people behind their backs. When we find ourselves motivated in these morally suboptimal ways, we have the opportunity to *work on ourselves*. We cannot change our motivations and their strengths at the drop of a hat, but they are amenable to gradual influence through concerted effort over a period of time. Again, these efforts are typically active. We can think about how our less desirable motivations might have developed, what might sustain them, what might trigger them, and what might help us to ignore or override them, with a view to taking steps to limit their grip on us. And we can figure out tiny ways to begin developing good habits, perhaps enlisting others to support us, remind us, or hold us accountable, with a view to inculcating the motivations that we wish we had. Again, the fact that an act helps us to develop praiseworthy motivations (or to mitigate the influence of blameworthy ones) is clearly a good feature of it. So, the activity that constitutes working on ourselves also consists of our deliberately doing good things, and thus of performing praiseworthy actions. That's a nice thing about working on yourself: there are moral achievements not only in the end-product,

[18] I give a somewhat fuller discussion of the guises question in my (ms).

but at every step along the way. On my view, then, there is some truth in the virtue-theoretical refrain that we become good by (deliberately) doing good things.

Of course, this rosy picture is not the whole story of how our motivations develop. It is possible to develop praiseworthy motivations in un-praiseworthy ways. Some motivations seem to travel by osmosis, getting surreptitiously absorbed from the people around us. And others are deliberately inculcated, but not in a praiseworthy way; we can be motivated to cultivate what is in fact a praiseworthy motivation by a nefarious or morally neutral aim, such as winning the adulation of some people we want to impress. Still other motivations are inculcated in us by our families, friends, colleagues, and other associates, without anything in the way of deliberate cultivation on our part. I think the right thing to say here is simply that these praiseworthy motivations do not have praiseworthy actions as part of their backstory. That does not make the motivations any less praiseworthy, so long as they are intrinsic motivations directed toward moral features in the true hierarchy or well-derived realizer motivations directed toward features that appear anywhere in the hierarchy. But it does mean that the agent is less praiseworthy overall than someone whose identical motivational set is the result of careful cultivation on her part. And, of course, most people's praiseworthy motivations are partly the result of praiseworthy cultivation and partly the result of other factors. That is another way in which most of us deserve partial credit.[19]

[19] I have been working on these ideas, on and off, for the better part of five years, and have received a tremendous amount of help along the way. I am grateful to everyone who has read or listened to drafts of my previous papers and provided feedback. For comments on an earlier draft of this paper, I am especially grateful to Kimberley Brownlee, Nico Cornell, Tom Dougherty, Daniel Fogal, Chris Howard, Alex King, Steph Leary, Gina Schouten, Keshav Singh, Daniel Wodak, and Alex Worsnip.

REFERENCES

Arpaly, Nomy (2003). "Moral Worth". *Journal of Philosophy* 99 (5): 223-245.

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

Bommarito, Nicolas (2013). "Modesty as a Virtue of Attention". *Philosophical Review* 122 (1): 93-117.

Fischer, John Martin and Neal A. Tognazzini (2009), "The Truth About Tracing". *Noûs* 43 (3): 531-556.

Johnson King, Zoë (2019). "We Can Have Our Buck and Pass It, Too". In *Oxford Studies in Metaethics*, *vol. 14*, ed. Russ Shafer-Landau. Oxford University Press.

Johnson King, Zoë (2020a). "Praiseworthy Motivations". *Noûs* 54 (2): 408-430.

Johnson King, Zoë (2020b). "Accidentally Doing the Right Thing". *Philosophy and Phenomenological Research* 100(1): 186-206.

Johnson King, Zoë (2020c). "Don't Know, Don't Care?". *Philosophical Studies* 177, 413-31.

Johnson King, Zoë (*ms*). "Deliberation and Moral Motivation". Provisionally forthcoming in *Oxford Studies in Metaethics*.

King, Matt (2020). "Attending to Blame". *Philosophical Studies* 177 (5): 1423-1439.

Knobe, Joshua (2003). "Intentional Action and Side Effects in Ordinary Language". *Analysis* 63(3), 190-94.

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

Rosen, Gideon (2008). "Kleinbart the Oblivious and Other Tales of Ignorance and Responsibility". *The Journal of Philosophy* 105 (10): 591-610.

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

R. Jay Wallace (2010). "Hypocrisy, Moral Address, and the Equal Standing of Persons". *Philosophy and Public Affairs* 38 (4): 307-341.