Working on Yourself

Zoë Johnson King
Harvard University

DRAFT – please feel free to circulate but check with me before citing.

Abstract. This paper is about how to respond to our own moral achievements and moral failures, as well as to the moral achievements and failures of others. I begin by considering the view that we should be harsher on ourselves than on others, discussing philosophical defenses of versions of this view and an array of reasons to doubt it. I then introduce my synthesis of this thesis and antitheses: the idea of working on yourself. I give an account of what it is to work on yourself, propose that this work is (at least) an imperfect duty, and observe some metaphysical and moral asymmetries between the work that it is possible and permissible to do on oneself and on others. This account accommodates all of the bewildering array of phenomena with which we began. I close by clarifying some details of my view, applying it to tricky cases, and introducing a final puzzle to which it gives rise but that it does not answer.

1. Introduction

Nobody's perfect. We are all deeply flawed creatures whose lives are peppered with moral missteps, although we nonetheless manage to have quite a few redeeming features and to also do some good things. One of the good things that we can do, either to redeem ourselves when we have acted poorly or simply to try to mitigate our moral flaws and develop our morally good qualities, is to *work on ourselves*. That's what this paper is about.

It is a thesis-antitheses-synthesis kind of paper. I am primarily interested in how we should respond to our own *moral achievements* and *moral failures* – two notions that I'll define in a moment – as well as to the moral achievements and moral failures of other people. I begin by exploring the view that we should be *harsher* on ourselves than we are on others – in a variety of ways that I'll also spell out in a moment — and arguing that, although this view might seem initially compelling, it faces a host of problems. Section 2 explains our topic, explains this view, and then explains four ways of problematizing the view. That's the thesis and its antitheses.

Section 3 introduces the synthesis: the idea of *working on yourself.* I offer an account of what it is to work on yourself and of the moral importance of this sort of work. I then argue that my account shows not only that the view that we should be harsher on ourselves than on others contains some grains of truth, but also that

the phenomena that seem to problematize this view can be accommodated alongside those that make it seem compelling. To synthesize the thesis with its antitheses, we need an account that (a) emphasizes the moral importance of both working on yourself and supporting others' work on themselves, (b) recognizes the limits on the latter imposed by the nature of agency and the values of autonomy and privacy, and (c) also recognizes the enormous range of ways in which people might fail either to work on themselves or to support others' work and might accordingly need a pointer in the right direction. So that will be the kind of account that I give.

I'll close (in Section 4) by going into some of the details of my view, applying it to some interesting cases, and then introducing a final puzzle to which it gives rise but that it does not answer. The puzzle concerns whether we should prioritize working on ourselves over encouraging and facilitating others' work, and, to the extent that we should, what might explain this division of moral labor.

2. Thesis and Antitheses

A *moral achievement* is what occurs when someone deliberately does something good, either appreciating (at least some of) the reasons why it is good to do and doing it on that basis or appreciating the fact that it is good to do and doing it on that basis.[1] So construed, moral achievements come in all shapes and sizes. It can be a major moral achievement to devote years of one's life to promoting a worthy cause, for instance. But, happily, more minor moral achievements cross our paths every day: life is full of promises kept, hardships alleviated, people given things that they are owed, bad moods cheered up and small joys precipitated. These, too, can count as moral achievements.

The category of *moral failures* is even more disjunctive. First, there is what occurs when someone does something bad with awareness of the fact it is bad to do. Second, there are cases in which someone does something bad without awareness of the fact that it is bad to do, but they *should* have been aware, and so their lack of awareness is no excuse. Again, so construed, moral failures come in all shapes and sizes. We can see some major moral failures in the annals of history (including recent history). But, sadly, more minor moral failures also cross our paths every day: life is also full of gratuitous derogatory remarks, tiny acts of selfishness, reasonable requests ignored or denied, people put into uncomfortable situations, bottles that could have been recycled thrown into the trash. These, too, are often moral failures.

When we learn that we or someone else has acted in a manner that constitutes a moral achievement or a moral failure,[2] we can respond to this information. And our responses are themselves morally assessable. So we might be interested in assessing them – that is, we might wonder how one *should* respond to one's own moral achievements and moral failures, and to the moral achievements and moral failures of others. And one might further wonder whether there are any self/other asymmetries here.

---

[1] Philosophers disagree about what it takes for an action to be well-motivated. On one view, it is enough that the agent performs it for reasons that, in fact, make it morally right or good, appreciating that these considerations obtain but not necessarily appreciating their moral significance. On another view, it is important that the agent appreciates that the considerations make the action morally right or good. See Arpaly (2003) and Markovits (2010) for the first view, Sliwa (2015) and Johnson King (2020) for the second, and Singh (2020) and Isserow (2020) for hybrid views. I remain neutral on this debate — hence the ambiguous phrasing in the main text.

[2] This phrasing is also deliberately ambiguous. There are two ways to learn that someone acted in a way that constitutes a moral achievement: you may learn that they did something you know to constitute a moral achievement, or you may learn that something you know they did constitutes a moral achievement. Likewise for moral failures.

Here is a view about this matter that I have always found somewhat compelling albeit somewhat vague:

ASYMMETRY VIEW: We should be harsher on ourselves than we are on others.

I was raised with the impression that it is generally more appropriate to be harsh on oneself than on others, or, equivalently, more appropriate to be lenient on others than on oneself. (Anecdotally, most of those with whom I have discussed this paper report that they were raised to think something similar.[3]) And instances of this dynamic play out frequently in ordinary conversation about moral achievements and failures: we see "That was so nice of you!" countered with "Oh, it was really nothing", we see "How could I have been so stupid?" countered with "It's okay, give yourself a break", and so forth.

One may wonder, though, what *exactly* being harsher on oneself than on others amounts to. And one may wonder just what is supposed to be so good about it.

Fortunately, some precisifications of the view that in responding to moral achievements and moral failures we should be harsher on ourselves than on others have already been defended at length within Philosophy. Notably, there is a sizable literature on *modesty*, in which the central idea is that being a good person involves downplaying your achievements and (other) good qualities — including especially, although not limited to, moral achievements and morally good qualities.[4] Authors in this literature disagree as to what exactly this downplaying should consist in. Some say that you just have to avoid bragging about your achievements, while others say that you should also avoid dwelling on them in the privacy of your own mind, and still others say that you should continually remind yourself of ways in which your achievements are limited – either by limits on your degree of causal influence over the relevant events or limits on their importance in the grand scheme of things.[5] Nonetheless, everyone in the modesty literature could agree that the respect in which you should downplay your moral achievements (whatever they each think it is) has no analogue in the way you should react to others' moral achievements. It is totally fine to celebrate others' moral achievements. And it is no part of virtue to routinely point out the ways in which others' achievements are limited; on the contrary, this would be rude, callous, and in some cases ungrateful, especially when it is moral achievements that you downplay and nitpick.

---

[3] Later in the paper I will speak of ideas that are prevalent within my cultural *milieu*. I have in mind the contemporary Western world. Some of those with whom I have discussed this paper have told me that similar ideas are also present elsewhere, but I do not take myself to know enough about other times and places to talk about them.

[4] Philosophical accounts of modesty apply to non-moral achievements and good qualities as well as moral ones. Nonetheless, our reactions to moral immodesty are distinctive. When we see folks bragging about long races or difficult climbs or elaborate pastries, the usual reaction is a roll of the eyes and a mild irritation. But if someone were to *brag* about having used no plastic for a year or driven a loved one to the airport then most people's reaction would be more visceral: it would be the sort of revulsion that we experience in response to "grandstanding" behavior (Tosi and Warmke 2016). This reaction is exacerbated for major moral achievements; it would be repellent for someone to brag about saving lives, for instance. (Note that this does not necessarily imply that it would be repellent for someone even to acknowledge that she has done good in saving lives; what exactly modesty forbids, beyond bragging, is a matter of substantial controversy – on which cf. the articles mentioned in the next footnote.)

[5] Wilson (2016) focuses on behavior, arguing that modesty consists in the disposition to "present your accomplishments/positive attributes" in a kind and sensitive way. Bommarito (2013) focuses on dwelling, arguing that modesty consists in patterns of attention. Flanagan (1989) focuses on the realistic assessment of one's accomplishments that acknowledges their unimportance "sub specie aeternitatis" and the elements of luck involved. Curiously, theorists of modesty disagree as to whether being harsher on oneself than on others is a constitutive part of this virtue; Maes (2004) and Brennan (2007) argue that it is, Statman (1992) and Ben Ze'ev (1993) that it is not.

Or so one might think. I will push back shortly. For now, I just want to observe how familiar it is to think that there is this asymmetry between the ways good people react to their own moral achievements and to those of others.

Turn now to moral failures. Here, again, one can see precisifications of the view that we should be harsher on ourselves than on others within Philosophy. For starters, a popular view in both Philosophy and Law is that we should give others *the benefit of the doubt* as far as possible, where this involves being reluctant to conclude that others' behavior amounts to a moral failure (rather than something that can be tempered with justification or excuse) and, if the evidence that it does constitute moral failure is unequivocal, then being further reluctant to ascribe this failure to an underlying character trait rather than seeing it as a one-off.[6] Trying to extend the benefit of the doubt uniformly to everyone will usually result in forming more negative beliefs about ourselves than about others, since we usually have much more information about ourselves than most others, leaving fewer candidate doubts on the table from which to benefit. For example, if a stranger pushes past you on the subway, then there are usually lots of potential justifications or excuses for this behavior that are consistent with your evidence. Maybe they are late to interview for a job that will lift their family out of poverty. Maybe their neighbors had a noisy party last night and they are sleep-deprived and grumpy. And so forth. But many of these justifications and excuses will be ruled out by your evidence when you yourself are the subway-pusher; you usually know where you are going, how you slept, whether you are ill, and so forth. You also usually know whether a particular moral failure reflects an underlying character trait much better for yourself than for someone else. As a result, it is usually harder to give ourselves the benefit of the doubt than to give it to others who act similarly, since many of the doubts from which we might benefit are ruled out by our evidence about ourselves.

I am sure that this epistemic explanation accounts for *some* of the leniency that we should show to others in response to their apparent moral failures. But the explanation goes only so far. We are not completely transparent to ourselves, and on many occasions a third party is better able to tell what is really motivating someone than is the agent herself. So, while our evidence about ourselves and others is usually lopsided, it is not all *that* lopsided. Moreover, it is implausible that we should even *try* to give ourselves the benefit of the doubt in the way we should to others. We should not respond to evidence that our own behavior constitutes a moral failure by strenuously resisting this conclusion and casting about for any possible mitigating factor(s) that we cannot rule out. We should not think, "Okay, so I know I wasn't late or grieving when I pushed that guy on the subway. But, hey, maybe I was affected by a mitigating factor whose influence is introspectively opaque! Like, maybe I was hangry. Yeah, that's consistent with my evidence! Let's say I was probably just hangry." This is an attempt to shirk responsibility, which is not a laudable reaction to one's apparent moral failure and indeed might itself amount to a further moral failure. Likewise, it is inappropriate to strenuously resist the conclusion that one's acknowledged moral failure reflects an underlying character trait and to try to dismiss it as a one-off. This resistance looks kind and patient when displayed toward others but lazy and complacent when displayed toward oneself. So, one way to precisify the asymmetry view is to hold that the default stance that we should take toward evidence of our own and others' moral failures is different in kind rather than in degree: we should treat evidence of others' moral failures with skeptical caution while being much less reticent to accept evidence of our own moral failures.

[6] The literature on how to understand the idea of the "benefit of the doubt" is large, but see Laudan (2006) for an overview, and cf. Faulkner (2018). There is also a related literature on *doxastic partiality* – roughly, the practice of giving more benefit of the doubt to one's nearest and dearest than to strangers. Cf. Keller (2004), Stroud (2006), Kawall (2013), Piller (2016), Crawford (2019).

This asymmetry in default stance might be spelled out in terms familiar from other parts of Philosophy. We can say that the right response to evidence of others' moral failures is to be *forgiving*, to show *grace*, to shrug off the moral failure and remain hopeful about the person's capacity for reform. Forgiveness has received increased philosophical attention recently and the view that there is something morally good about forgiving those who wrong us is now also quite popular.[7] Some argue further that we should be as optimistic about others' potential for moral reform as our evidence allows, responding to poor behavior and bad character by remaining hopeful that the person is able to improve and emotionally invested in the possibility of their doing so.[8] By contrast, being forgiving toward oneself can seem like another attempt to shirk responsibility for moral failures. Simply *hoping* that we'll do better next time does not seem good enough, no matter how fervent the hope or how emotionally invested we are in the possibility of our own reform. In short: strongly desiring that we reform, remaining optimistic about our capacity for reform, and showing grace in allowing ourselves room to reform all seem inadequate unless we actually start *doing* it. So, even when our evidence about our own and others' moral failures is equally strong, the thought that there is a kind of leniency that we should show to others but not to ourselves is quite compelling.

Or is it?

I feel the intuitive pull of each of these ways of spelling out the view that we should be harsher on ourselves than on others. But I also feel the pull of four ways of problematizing that view. So let's turn to those.

For moral achievements, the purported asymmetry boils down to the view that we should downplay our own moral achievements but "up-play", so to speak, those of others. And both conjuncts in that conjunction can be resisted. It is currently quite trendy – not within Philosophy in particular, but within our broader cultural *milieu* – to hold that it is a good thing to "know your worth", as the kids say, and to "celebrate your successes". This encourages us to extend precisely the attitude that the asymmetry view would have us take toward others' moral achievements also toward our own. And, *pace* the modesty literature, the idea has some theoretical merit: if acts of a certain type are praiseworthy then presumably they are praiseworthy regardless of who performs them, and the claim that we should praise the praiseworthy is close to an analytic truth.[9] However, there are also other social contexts in which it is deemed appropriate to downplay others' achievements — even moral achievements — in precisely the way the asymmetry view encourages us to downplay our own. It is sometimes seen as important to react to people's achievements by "taking them down a peg or two", to prevent them from getting "too big for their boots" and stop it from "going

[7] For contemporary work on forgiveness, see the papers in Warmke, Nelkin and McKenna (eds., 2021), Satne and Krisanna (eds., 2022), and Enright and Pettigrove (eds., forthcoming). For a classic articulation of the view that forgiveness requires recognizing the forgiven agent's blameworthiness – as opposed to giving them the benefit of the doubt – see Hieronymi (2001).

[8] Ryan Preston-Roedder (2013) offers a compelling defense of this idea under the moniker of *faith in humanity.* He says that "when someone who has [faith in humanity] makes judgments about people's past or current attitudes or actions, she tends to be acutely sensitive to evidence of people's decency, including evidence that others are likely to overlook" (p.667); that "when someone who has such faith forms expectations about people's future attitudes or actions, she tends to be relatively slow to judge them harshly" (p.666); and that "as reasons to doubt people's decency mount, her faith may simply dispose her to hold on to the belief that right action is attainable for these people, or in other words, a live possibility for them" (*ibid.*).

[9] What is analytic is that it is *fitting* to praise the praiseworthy. This is not the same as the claim that we should praise the praiseworthy all-things-considered, as there may be fittingness-independent reasons not to engage in fitting praise. See [redacted] for discussion.

to their head".[10] And, again, this idea has some theoretical merit: if there *is* a risk of someone's achievement leading to excessive pride then that is surely a reason not to exacerbate the process by lavishing them with praise, whether this person is oneself or someone else. Moreover, to the extent that downplaying others' moral achievements consists in accurately pointing out these achievements' limits, it is simply a way of acknowledging the detailed contours of the normative facts. It is hard to take issue with accurate moral assessment.

Likewise for moral failures. I have suggested that it is appropriate to be lenient with others but not with oneself, but, again, both of the conjuncts in that conjunction can be resisted. The view that it is appropriate to be lenient with oneself — to "talk to yourself the way you would talk to a friend", as it is sometimes put, or "see yourself as a work in progress" — is also currently very trendy. And the claim that we should talk to ourselves the way we talk to friends literally just *is* the claim that we should extend the leniency that the asymmetry view would have us show to others also to ourselves. More tellingly, some theoretical defenses of the importance of lenience toward others readily extend to lenience toward oneself. Being emotionally invested in others' potential for reform is a valuable way of standing in solidarity with them, rather than writing them off or giving up on them.[11] And writing oneself off or giving up on oneself seem roughly as bad as doing the same to someone else.[12] That said, there is also a countervailing trendy view to the effect that we should be just as harsh in response to others' moral failures as the asymmetry view would have us be in response to our own. The view that it is important to "call people out" and to "hold them accountable" for their moral failures has skyrocketed to cultural prominence in recent years. On this view, shrugging off others' moral failures in blithe optimism about their underlying good character and/or capacity for reform amounts to a kind of complicity, which again is no part of virtue and might constitute further moral failure on the complicit agent's part.

So, for each putative asymmetry in the ways we should react to our own and others' moral achievements and failures, there are two ways to break the asymmetry: leveling down or leveling up. And each of them can be made to seem roughly as attractive as the asymmetries themselves. If we want to figure out the truth about how we should react to our own and others' moral achievements and failures, then, we have our work cut out for us. One task is to figure out the extent to which the asymmetries are genuine, and, if they are at all genuine, what explains them. We must figure out whether and why being a good person involves relating so differently to one's own moral achievements and failures than to those of others. Another task is to reconcile the asymmetry view with the counterphenomena that cast it into doubt. We must figure out whether to level up, down, or not at all, and if we should sometimes level up but sometimes down then we must figure out what makes the difference and how to distinguish these times. It would also be desirable to explain the co-occurrence of these various counterphenomena within our current cultural *milieu*; if some of these views are incorrect then their popularity requires an error theory, and if they are somehow all

[10] Notice that this downplaying is appropriate only if it is well-motivated. An intervention that is genuinely aimed at preventing someone's achievement from going to their head can be appropriate, but similar-looking outward behavior can also be an inappropriate instance of what Carbonell (2022) calls "malicious moral envy".

[11] On this see especially Preston-Roedder (2013, pp.683-685).

[12] Compare Paul and Morton (2018a, 2018b) on the importance of believing in one's own and one's significant others' ability to succeed. Paul and Morton's account places a *higher* premium on not giving up on oneself than on not giving up on others: they argue that, given our unique relationship to our own agency (in particular, that *we* are the ones who decide whether to persevere with our projects in the face of ambiguous evidence of our prospects for success), we each have practical reason to estimate our own abilities and opportunities as favorably as our evidence will allow, whereas for significant others we may choose *either* to encourage their optimism by being optimistic ourselves *or* to take a more cynical attitude and privately devise back-up plans for what to do if they fail. Paul and Morton are interested in goals of all sorts, rather than specifically moral goals, but their argument applies to moral goals as a special case.

correct then we must account for the fact that they appear to be opposites. A *lot* of views in this arena enjoy some popularity and have something to be said for them. And I proceed on the methodological assumption that widespread intuitions usually point toward something that is worth taking seriously, with the upshot that they should be rejected as completely baseless only as a last resort. I intend, instead, to tie these various disparate views together.

3. Synthesis

a. *The view*

The concept with which I aim to tie the phenomena together is the concept of *working on yourself*.

So, what is working on yourself? Here's what it is:

WORKING ON YOURSELF:

i. Identifying a moral shortcoming.
ii. Trying to figure out what it would take to reduce, mitigate, or (ideally) eliminate the shortcoming.
iii. Taking some steps toward that goal.

I will refer to (i) as the *diagnostic* component of working on yourself, to (ii) as the *investigative* component, and to (iii) as the *active* component.

Let me unpack this a little.

First, working on yourself involves identifying a *moral shortcoming*. There are all sorts of things this might be. Perhaps you tend to discount the significance of a certain person's interests, or a certain cause or value, in your deliberations about what to do. Or perhaps you are not very good at even noticing when a certain something is at stake in your decision. Alternatively, perhaps a certain morally significant consideration grips you monomaniacally whenever you see that it is at stake, playing an outsize role in your deliberations – one that far outstrips its actual moral importance. Or perhaps you find yourself caring about things that are morally bad (like arbitrary hierarchies). You might also recognize something to be morally significant but know very little about it, such that you are rarely in a position to determine what, if anything, you should do in light of its being at stake. Or you may have moral shortcomings that show up outside the context of deliberation and choice; for instance, you may tend to avoid raising important moral questions because you find them uncomfortable, or to snap at your kids when you are stressed.

In sum, moral shortcomings are features that lead you to fail to *discern and respond to your actual moral reasons for action*.[13] They are things that go amiss with the process by which you decide what to do and control your behavior in accordance with your decisions. Moral shortcomings are therefore features that leave you liable to moral failure. But they count as such regardless of whether they actually lead to moral failure; if some feature of yours leads you to fail to discern and respond to your actual moral reasons, then it is a moral shortcoming, even if (fortunately) this failure in thought never produces a corresponding failure in action.

[13] I have said 'reasons' because it is still the most fashionable candidate for being a normative primitive. If you prefer a different normative primitive then you may substitute it in as appropriate.

Suppose you have identified a moral shortcoming. Working on yourself then has two further components: you try to figure out what it would take to reduce, mitigate, or (ideally) eliminate the shortcoming, and you take some steps toward that goal. Again, there are all sorts of things that these components can consist in. You could talk to other people about your shortcoming – perhaps those who know you well, or those who have had a similar shortcoming and have successfully addressed theirs to some extent. You could read articles and books, or watch movies and TV and plays, or listen to podcasts with related content. You could see a therapist. You could identify short-, medium-, and long-term goals, and develop routines, rituals, and reward systems to keep yourself on track. Or you could simply ask your partner to point out when you're doing the thing again and resolve not to get irritated with them when they do.

As these examples suggest, the investigative and active components come in degrees; we can try to figure out how to address a moral shortcoming with more or less diligent effort and can take more or fewer of the steps that inquiry suggests are good ways to address it. Someone counts as working on themselves even if they do only a little bit of this stuff, though their doing the bare minimum will usually be unimpressive and we might think that they should *put in more work* – i.e., that they should push themselves to do more of the investigative or the active part. As I am thinking of it, someone also counts as working on themselves if their work is inept, interrupted, or otherwise incomplete, with the result that they fail to address their shortcoming satisfactorily. Working on yourself is a goal-driven activity, and, like other goal-driven activity, you still count as engaged in it even if you do not attain your goal – although it is certainly great for people to work on themselves and succeed, thereby bringing about deliberate self-improvements.

Working on yourself can be a simple and easy process or a long and difficult one, or something in between, depending on the nature and scale of the shortcoming in question. It is easy to buy a recycling bin and develop the habit of throwing empty bottles into it. It is much harder for someone who sees interpersonal relationships as things from which they can extract benefits to learn to do things for others' sakes without feeling resentful or keeping a mental score. That work will be a longer process with more twists and turns. Indeed, when you first start trying to address a moral shortcoming, you might be desperately unclear about what (if anything) it will take. This means that the investigative component does not always precede the active component. You may not be able to tell whether a course of action will reduce, mitigate, or eliminate your shortcoming without trying it. And you may need to change tack partway through if what you are trying isn't working. If your moral shortcoming is deeply entrenched, or if it is something that you have very little information about, then you might see yourself as engaged in a process of trial and error all along. But that's fine. The trial-and-error process *is* the process of working on yourself.

Now, here is a thesis:

> MAIN THESIS: Working on yourself is required in response to an egregious moral failure and is an imperfect duty the rest of the time.

I will not here take a stand on what the egregiousness of some of our moral failures amounts to; I leave the reader to form her own opinion on this. The important point is that we are sometimes required to work on specific moral shortcomings in light of their negative impact on our behavior, but, even when we have discharged all such requirements, *our work is not yet done*. Those of us who will never attain moral perfection – including all real people – need not wait to see whether our shortcomings result in egregious moral failure before addressing them. We can be proactive: we can work on our shortcomings even before they lead to disaster. (Indeed, this seems preferable to the disaster option.) Here a minor, less-than-egregious moral failure might serve as evidence, alerting us to moral shortcomings. And, ideally, we might not even need

that. We can become aware of our shortcomings in conversation or through introspection, or by noticing shortcomings in others and wondering whether we may be the same way, and so forth. The diagnostic component of working on yourself can take all manner of evidence as its input. And, since none of us will ever attain moral perfection, we will all always have plenty of material to work with. Even without an egregious moral failure to require it or a minor moral failure to inspire it, deliberate self-improvement is a worthy aim. Indeed, since self-improvement is good, it follows from the definition of 'moral achievement' (given right after the Introduction) that deliberate self-improvement is *itself* a moral achievement, over and above any subsequent moral achievements that the agent thereby brings herself into a position to attain.

As edifying as continual self-improvement may sound, though, it is just one morally important goal among many on which we might spend our time. Addressing our moral shortcomings is not more important than fixing the world beyond ourselves: it is not more important than promoting well-being and justice, for instance (though sometimes it *is* a way of promoting these things, since sometimes our own shortcomings are part of what stands in the way of them). And, barring egregious moral failure, addressing our moral shortcomings is not something that we are always required to prioritize over every personal project and everything we do for fun. This is especially so for improvements of moral shortcomings that are relatively minor and peripheral, since those are unlikely to lead to major moral failure anytime soon and so working on them will rarely be the best use of limited time. Continual self-improvement must therefore be balanced against all of the other valuable goals that we have in our busy lives. Hence the idea that it is an imperfect duty: this idea is that part of what it is to be a good person is to strive to become steadily better — not on any specific occasion, but over time, gradually, and alongside your sundry other goals.[14]

To further see the plausibility of this idea, consider someone who makes no attempt whatsoever to work on himself throughout his entire life, despite having ample time and other resources with which to do so. Assume that none of his actions amount to egregious moral failures (so he is not violating any strict moral requirement in failing to work on himself), but that plenty amount to minor moral failures and that he also acquires a good deal of other evidence of his moral shortcomings. For good measure, we may suppose that he even forms beliefs to the effect that such-and-such are his moral shortcomings. But he just doesn't *do* anything about them. Ever. For his whole life. What would you think of such a person?

My intuitions about such a person are strongly negative. For one thing, he seems to exhibit a kind of moral recklessness: he is aware that he has features that leave him liable to moral failure and yet he is apparently indifferent to this moral risk, since he fails to take readily available steps to lower it. He also seems to exhibit a worrying indifference to whichever morally significant things he thereby endangers. Moreover, he seems insufficiently concerned with his own moral character – and, more strongly, I am inclined to say that he seems insufficiently *self-respecting*. For he is aware of his moral shortcomings yet does nothing about them. He allows his moral character to be shaped entirely by whatever causal forces surround him, like a plastic bag buffeted by the wind. This seems at best complacent, at worst defeatist, and worryingly heteronomous. It seems like a flight from responsibility for that for which he should instead take responsibility. Of course, this taking responsibility need not be a matter of *constantly* working on oneself, as we have already seen. But surely he should at least do it *sometimes*, given the opportunity. Surely he should at least *care* about it. And that, I think, is because it's an imperfect duty.

[14] Kant is famously associated with the idea that we have an imperfect duty to develop ourselves. I see this proposal as being along roughly similar lines, except that I do not mind whether you learn chess or play the banjo – I just care about whether you are actively making yourself into a better person. Johnson (2011) also defends the thesis that we each have an imperfect duty to work on ourselves, which he regards as a quintessentially Kantian thesis, whereas Callard (2018) explores the nature of a similar process that she terms 'aspiration' and traces to Plato.

That's my account of the nature of working on yourself and the moral significance of this sort of work.

What about other people?

When others around you are working on themselves, there are lots of ways you can encourage or facilitate their work. You can serve as an "accountabilibuddy": someone with whom they discuss their shortcoming and brainstorm ways to address it, to whom they report their progress, and/or who congratulates them on improvements made and gives them a talking-to if they are not trying hard enough. And, as well as playing one of these supporting roles in others' work on themselves, we can directly *work on others*: we can identify their moral shortcomings, try to figure out ways to reduce, mitigate, or (ideally) eliminate them, and take steps toward that goal. Indeed, it is possible to address others' shortcomings without their explicit agreement, their consent, or even their knowledge. For instance, you can point out their moral failures and ask what they think led to these failures without their having asked you to do so; you can use reward and punishment to try to shape them in a positive direction (à la the early behaviorists); you can deploy subtle nudges to surreptitiously induce desired behaviors in the hopes of building good habits.

But the ways in which we can "directly" work on others are *less* direct than the ways we can directly work on ourselves. This is due to the basic metaphysical fact that each of us is the only person whose actions we directly control. This basic metaphysical fact means that there are components of the process of working on yourself that it is literally impossible to do to others: in general, we can choose to act or direct our attention in certain ways but cannot *choose* for someone else to act or direct their attention, since we control neither others' limbs nor their minds. We can also form intentions and plans for our own future behavior, or even resolutions about our future behavior – never to do something again, or to do better next time, or to "do the work" – in a way that we cannot do for others, since our control over others is too attenuated for such things to be intelligible. I can suggest a plan to you, but I cannot put a plan in your head. I can incentivize your action, but I cannot perform it. I can encourage you to look in a certain direction, but I cannot do the seeing for you.[15]

In addition to these ways of working on yourself that it is *impossible* to do to others,[16] lots of ways of working on others are metaphysically possible but *morally impermissible*. This holds of many of the ways of working directly on others just mentioned; trying to surreptitiously "fix" another person, when they are a healthy adult human being who has not given their prior consent, is usually impermissible since it usually involves a conniving and manipulative reach into aspects of the other's life and character that are not yours to try to control.[17] Even if you thereby succeeded in alleviating the other person's moral shortcoming, this would

[15] Thanks to [redacted] for this last way of putting the point, on which the first two are just a riff.

[16] Strictly speaking, several of the things that I have discussed are not things that we do *to* ourselves/others but things that we do *for* or *with respect to* ourselves/others. It sounds odd to describe forming an intention as something that we do *to* ourselves, for instance. I say "to" in the main text for brevity, but a disjunction of prepositions is intended here.

[17] My impression is that common sense morality takes there to be severe limits placed on the work we may permissibly do on others by considerations of autonomy and privacy. For example, consider Anthony Burgess's novel *A Clockwork Orange*. The novel is told from the point of view of its main protagonist, Alex, who is thereby revealed to be a horrific amoralist. It is evident to the reader that Alex acts about as badly as possible and is wholly unrepentant. Nonetheless, when he is arrested and subjected to forced corrective brainwashing in the attempt to "cure" him of his bad character, we react with revulsion rather than approval. Even though it is clear that this direct work on Alex greatly alleviates his moral shortcomings, we are appalled by the autonomy violation. I suspect that this is part of a broader pattern: the more heinous someone's moral shortcoming is, the more morally dubious are the steps that someone else would have to take to alleviate it without their willing cooperation.

not constitute a moral achievement and would likely constitute a moral failure – perhaps an egregious failure, if your invasion of their privacy or violation of their autonomy is severe. This is a significant disanalogy between working on yourself and working on others: it is only in bizarre fringe cases that you can invade your own privacy or violate your own autonomy by working on yourself, and you do not need to seek your own prior permission in order to work on yourself permissibly.

There may be ways of working directly on others that do not invade privacy or violate autonomy so badly as to be morally impermissible. I return to this possibility in the final section. What *is* clearly permissible (barring unusual circumstances, like threats of the world-destroying demon of philosophical imagination) is to support another person's work on themselves in precisely the way that they have asked you to do. In such a case, you may aid in their deliberate self-improvement. Indeed, this is a nice sort of case in which moral achievements abound. For their deliberate self-improvement constitutes a moral achievement, and your deliberate assistance with their deliberate self-improvement constitutes a further, additional moral achievement on your part. Everyone's a winner.

b. *How it synthesizes*

Now we can synthesize the thesis with its antitheses.

We can start by identifying some grains of truth in the asymmetry view. Recall that this view holds that we should be more lenient in response to others' moral failures than in response to our own. In my view, the correct way to respond to our own moral failures is to see them as an *impetus* to work on ourselves. This is required if the moral failure is egregious and is otherwise not required but still shown to be a good idea. However, we should *not* see others' moral failures as an impetus to work on them. The metaphysical and moral self/other asymmetries imply that much of the work we might do on ourselves is either impossible or impermissible to do to others. The correct way to respond to others' moral failures is to encourage and incentivize them to work on themselves, offering to help inasmuch as you are able to do so. But you cannot force another person to improve – in both the metaphysical and moral senses of the word 'cannot'. Hence there is indeed a significant difference in the default stance that we should take toward (evidence of) our own and others' moral failures: we should be spurred by our own moral failures into working on ourselves, but should not be spurred by others' moral failures into barging in and trying to "fix" them.

The asymmetry view also holds that we should downplay our own moral achievements but not those of others. In my view, the correct way to respond to one's own moral achievements is by recognizing them as achievements while remaining mindful of their limitations and the work one has yet to do. This is because the duty to work on yourself is imperfect, and so, regardless of what you have achieved so far, you're far from done. There is nothing inherently wrong with feeling pride. But there is something wrong with *resting on your laurels* – that is, allowing pride in your moral achievements to date to take precedence over thinking about what good to do next. However, there is no real danger of resting on someone else's laurels, since that makes no sense. Laurel-resting involves luxuriating in one's own virtue. It would be a conceptual error to take this attitude toward someone else's moral achievement (unless you had encouraged or supported them, in which case it would be this distinct achievement of your own in which you luxuriate). This would be like feeling guilty for something someone else did in the distant past or feeling grateful for a benefit conferred by one stranger to whom you bear no relation on another stranger to whom you bear no relation. But, since there is no danger of resting on someone else's laurels, we can go ahead and celebrate others' moral achievements. The work that they have yet to do is theirs, not ours, and so we need not focus on it.

Those are the grains of truth in the asymmetry view. In my view, these grains arise from some fundamental differences between what it is possible and permissible for us to do to ourselves and what it is possible and permissible to do to others. These differences imply that the right ways to react to our moral achievements and moral failures are things that we either cannot or may not do to others, although there are similar-in-spirit things that we can and may do.

What about all the counterphenomena, then?

I think that these are *also* chiefly explained by the nature and importance of both working on yourself and supporting others' work on themselves.[18] As we have seen, the co-occurrence of these disparate phenomena within our cultural *milieu* can be puzzling. It can be difficult to know what to do with the simultaneous injunctions to celebrate your successes *and* take others down a peg or two, or to talk to yourself the way you would talk to a friend *and* hold others accountable for their bad deeds and bad character. Nonetheless, I suggest that these injunctions are not just a motley crew of independently-motivated ideas that awkwardly clash with one another. There is more underlying unity to them than is immediately apparent. For they are a suite of psychological checks and balances whose simultaneous cultural prominence arises from the fact that people differ massively in what it will take to get us to work on ourselves and help others to work on themselves. We can thus explain their co-occurrence using an Aristotelian model according to which virtuous behavior lies in between vicious extremes.

The central Aristotelian idea is that there exist traits associated with each virtue that one can possess and manifest either to a deficient degree or to an excessive degree, in which case they are not virtuous. This holds of downplaying our moral achievements, up-playing others' moral achievements, and showing lenience in response to others' moral failures but not in response to our own. Downplaying your own moral achievements is all well and good until it veers into a spiral of negative self-talk – which can make it harder for you to keep working on yourself, rather than easier, since it undermines your sense of self-efficacy.[19] Likewise, it is no good to shrug off your moral failures and cast about for any mitigating factor consistent with your evidence, but it is also no good to berate yourself endlessly, beyond what is deserved and long past the point where doing so might yield new insight about how your poor behavior came to be or how you could prevent it in future. Being *too* self-critical inhibits working on yourself rather than prompting it. And something similar holds of lenience with others. Being forgiving and optimistic in response to others' moral failures can reassure them that they are capable of reform, but if it looks as though they *won't* reform when left to their own devices then we may need to push them to do the work. Likewise, celebrating others' moral achievements can be a way of giving credit where it's due, perhaps thereby reinforcing their sense of self-efficacy, but it is no good to start fawning over them obsequiously. This encourages them to rest on their laurels rather than keeping up the good work.

---

[18] To be clear: I don't think that the importance of working on yourself *exhaustively* explains the presence of these phenomena within our current cultural environment. Some also serve other purposes. For example, calling people out and holding them accountable facilitates reconciliation between wrongdoers and the wronged. The value of this moral repair is independent of that of any deliberate self-improvement undertaken as part of the process; reconciliation and repair are not merely incidental side-effects of our valuable work on ourselves. Similarly, taking people down a peg or two is sometimes part of the kind of jokey banter that bonds friends together. The value of banter among friends has nothing to do with working on yourself. This is just an independent end that the same behaviors also serve.

[19] The concept of self-efficacy was popularized by psychologist Albert Bandura (1977, 1997). It refers to the degree to which an individual sees herself as able to take action to achieve whatever goals she may have, dealing with obstacles and impediments as necessary.

Each of these "excesses" (to continue with the Aristotelian model) corresponds to one of our counterphenomena. Spiralling negative self-talk about the limits of your moral achievements might be prevented by a gentle reminder to know your worth and celebrate your successes – at least once in a while, and especially when you're starting to doubt yourself. Likewise, a gentle reminder to talk to yourself the way you would talk to a friend can help you to stop catastrophizing over your moral failures and focus on the work you have to do. Similarly, when there is a risk of your ignoring others' failures for the sake of personal comfort, but in fact you could easily and safely prod them toward addressing their shortcomings, a reminder of the importance of calling people out can be the prod that *you* need to get the conversation going. And a reminder that it can be appropriate to take people down a peg or two can prompt you to take others' achievements within their proper context and encourage them to do the same. When there is a risk of our harshness toward ourselves or lenience toward others' becoming excessive, these adages can steer us back, so to speak, to the right path – namely, the path of careful, gradual, sustained work on ourselves and well-placed support of those around us doing the same.

Per the Aristotelian model, these "excesses" have opposing "deficiencies": negative self-talk is opposed by self-congratulatory laurel-resting; berating yourself endlessly by shirking responsibility; ignoring others' moral failures by hounding them mercilessly; obsequious fawning by jerkily pointing out the limitations of others' moral achievements at every available opportunity. However, these deficiencies are precisely the behaviors that the asymmetry view guards against – they are ways of being too lenient on oneself or too harsh on others. In all four respects, then, the phenomena that seem to problematize this view can instead be understood as ways of preventing us from taking it too far. I suspect that the asymmetry view is more culturally prominent than the counterphenomena just because the vices against which it guards are more prevalent: more people are too lenient on themselves and too harsh on others than are too lenient on others and too harsh on themselves.[20] Still, people vary. Plenty are too lenient on others and harsh on themselves. And a single person can stray from the mean in different directions at different times. So we are well-served by a smorgasbord of opposing adages about how to respond to our own and others' moral achievements and failures. Each of them is sometimes just what we need to hear in order to point us in the right direction: forwards.

4. Further bits and bobs

That was the synthesis in a nutshell. Here I will clarify some details, discuss some interesting cases, and introduce a final puzzle.

Working on yourself involves *activity*: both mental activity – wondering why you acted as you did, noting another person's shortcoming and asking yourself whether you may be similar, estimating the efficacy of different mitigation strategies, mulling over your therapist's comments, and so forth – and also physical activity – journaling, asking your friends for an honest opinion, recruiting an accountabilibuddy, going to see a film about an important moral issue that you know next-to-nothing about, and so forth. This is all a

[20] A bit of empirical data supporting this conjecture is what is sometimes called the "fundamental attribution error" (see Ross 1977 for the phrase and Jones and Harris 1967 for the original study). This is the purported tendency of normal adult humans to reach much more readily for explanations of others' behavior in terms of stable character traits and much less readily for explanations that point to non-recurring features of the circumstances. One meta-analysis (Malle 2006) found that people tend to explain others' behavior more in terms of personal traits than situational factors and their own behavior more in terms of situational factors than personal traits *when the behavior is poor*, but this asymmetry was reversed for good behavior, thus indicating a "self-serving bias".

matter of engaging in processes rather than being in static states. So it contrasts with the *immediate reactive attitudes* that one might take in response to moral achievement or moral failure, such as guilt, shame, anger, indignation, pride, self-respect, esteem, and commendation. One might therefore wonder whether my account has anything to say about the immediate reactive attitudes. In particular, one might wonder about another way of precisifying the asymmetry view: when someone has equally strong evidence of her own and another person's moral achievement or failure, and when those two achievements/failures are equally good/bad, should she adopt the immediate reactive attitudes toward herself to a different *degree* to that to which she adopts them toward others? Should she blame herself more strongly and others less so, or praise others more and herself less?[21]

This might be what some people have in mind when they think of the asymmetry view. But I see no cogent rationale for it. Once we stipulate that the agent's evidence concerning two actions is *exactly* equally strong and that the actions are *exactly* equally good or bad, I see no reason why she should not praise/blame them to an exactly equal degree, notwithstanding the fact that one action is her own and the other someone else's. And I see a compelling reason why she should indeed praise/blame them to an equal degree: symmetrical attitudes are, by stipulation, fitting in light of her evidence.

Nonetheless, my view does suggest some other asymmetries in the immediate reactions we should have to our own and others' moral achievements and failures. For I have argued that the appropriate response to your own moral failure is to see it as an impetus to work on yourself, and that, since the duty to work on yourself is imperfect, no moral achievement does away with this ongoing duty. This suggests that there are fitting immediate reactions to one's own moral achievements and failures *besides* the reactive attitudes. The fitting response to one's own moral failure includes not only guilt but also a redoubled commitment to working on oneself and a sense of the importance of this goal – especially if the failure is an egregious one. And the fitting response to one's own moral achievement includes not only pride but also being mindful of the place of this achievement within one's continual process of work on oneself, recognizing not only what it says about one's work so far that one was able to act well on this occasion (of which one may rightly be proud) but also what the next steps might be. These further attitudes have no analogues in the fitting reactions to others' moral achievements or failures, since their work is not ours to do. So, my view suggests that there are some differences in the immediate reactions we should have to our own and others' moral achievements and failures, but that these differences do not lie in the degrees of the traditional reactive attitudes.[22]

The Aristotelian model suggests some further differences, not in the attitudes that it is fitting to adopt, but in which among these attitudes we should focus on at different times. If you find yourself veering into deficiency or excess, either in your reactions to yourself or to someone else, then it is appropriate to focus on the aspects of the case whose recognition elicits a reaction that brings you closer to the mean. If you find yourself silently fuming over a minor moral failure then it is appropriate to direct your attention toward its mitigating factors. If you find yourself blowing a moral achievement way out of proportion then it is appropriate to direct your attention toward its limits. And so forth. All of this holds regardless of whether it is your own or another person's moral achievement or failure to which you react. So it does not generate significant self/other asymmetries. But it does generate differences in the appropriate responses to moral

[21] Here I focus on *degree* of attitude because the reactive attitudes that constitute self-blame and self-praise are always different in kind from those that constitute blame and praise of other people: self-blame consists in guilt and self-praise consists in pride, which are phenomenologically qualitatively different from blame and praise of others.

[22] I am grateful to [redacted] for pressing me to clarify this point so insistently and repeatedly that I eventually came to see the need for it.

achievement and moral failure across cases, if we construe 'response' broadly enough to include not only the attitudes that it is appropriate to take but also the ways in which it is appropriate for the agent to direct her attention. Here what makes different responses appropriate are not differences between the actions or agents, but differences between the reactions that one is currently having to them.

This point also creates the potential for cases in which the attitudes one should *express* are a proper subset of the attitudes one should *adopt*. If you are someone's accountabilibuddy and they are erring on one side of the mean, but you yourself are not erring similarly, then it can be appropriate for you to remind them of the things on which they consequently ought to focus and to express *only* the fitting attitudes toward those things, although it remains appropriate for your private mental states to include a richer body of attitudes (since you are not erring in the same way as the agent). The aim of the game here is not to express attitudes that one does not hold. It is rather to judiciously select attitudes to express from among those one holds, so as to facilitate the other person's work on themselves in precisely the way they need in a given moment. This makes it especially useful to have the whole panoply of counterphenomena in one's toolkit: you urge the other person to talk to themselves as they would talk to a friend if you find them catastrophizing but you call them out if you find them shirking responsibility, and you encourage them to celebrate their successes if they're underdoing it but take them down a peg or two if they're overdoing it. None of this is phony or insincere. It is a selective-but-still-accurate response to the other as you find them.

These adoption/expression discrepancies arise only when it is *possible* for you to facilitate another person's work on themselves. And that is the kind of case that I have had in mind in the paper up until this point. But it is interesting to think about how we should react to the moral achievements and failures of those over whom we can have no causal influence whatsoever, such as people who lived in the distant past and are now long-dead. In such cases, there is nothing anyone can do to support or encourage the agent's work on themselves – it's far too late for that. So my account does not apply to such cases. I suspect that the thing to do is simply to have whichever reactive attitudes are the fitting responses to the beliefs about the past person that one's total evidence supports. That might turn out to be incorrect, but if it is incorrect then the nature and importance of working on yourself is not what explains its incorrectness.

That was one kind of exceptional case. Another arises from the fact that we can bear relationships to others that make us partially responsible for their moral development. If you are someone else's parent, teacher, therapist, or life coach, for example, then the work that they have yet to do *is* partially yours – that is to say, your role includes not only supporting them but also pushing them to do the work, with the result that there are fewer moral constraints (but not no constraints!) on what you may do to get them to do it. In such a case the usual self/other asymmetries are absent; it is entirely appropriate for you to remind yourself of their moral achievements' limitations and the work that they have yet to do, and it is appropriate for you to see their moral failures as an impetus to facilitate their work in whatever way is possible and permissible. In short, to the extent that someone else's work *is* your responsibility – and thus that your relationship to their moral shortcomings resembles your relationship to your own – the appropriate reactions for you to take toward their moral achievements and failures pattern on the 'self' side of the self/other asymmetries. I think that this is good evidence for my view: it is because the nature and importance of working on yourself chiefly explains those asymmetries.

There is also a third, much more intricate, kind of exceptional case. It arises from the fact that there can be *group agents*: a family, a friendship group, an academic department, a government, a company, a band, and all sorts of other groups can constitute agents, such that we meaningfully speak of their having attitudes and performing actions. Some of these group actions constitute moral achievements or moral failures. And versions of all of the phenomena discussed in the previous section obtain for group agents, just as they do

for individual agents. When a group action constitutes an egregious moral failure – when a department fails to process the stipends for any of its graduate students, for instance, or when negligent governmental funding decisions lead to a fire in a housing project that kills 72 people – the group is, plausibly, required to work on itself. And groups need not wait for egregious moral failures to work on themselves; they can identify and attempt to address their moral shortcomings the rest of the time, too. But there are significant metaphysical limits on the work that most groups can do on most other groups, as well as significant moral limits on the permissibility of one group's attempting to "fix" another's moral shortcomings. There are even group-level analogues of the exceptional cases: just as an individual parent, teacher, or therapist can be partially responsible for another individual's moral development, so too can an inspection agency, set of consultants, or team of ombudspeople be partially responsible for the moral development of another group. The analogies are very close.

Since groups are comprised of suitably-related individuals, these group-level phenomena can give rise to individual-level phenomena that complicate the picture. Most strikingly: if another person's egregious moral failure constitutes an egregious moral failure of a group to which you belong, and if their moral shortcomings *qua* group member constitute moral shortcomings of the group itself, then you can be required to work directly on them in virtue of your role in the group. If a department member fails to meet their important service obligations then the Chair is required to chase them up, and sometimes then to use nudges or similar techniques to inveigle the wayward member into fulfilling their role. If a fight breaks out in a mosh pit then others in the pit are required to break it up, and sometimes then to physically isolate the people who got into the fight from the rest of the crowd until they calm down. And so forth. Metaphysically speaking, what I think is going on here is that the group agent is required to work on itself, but, given the metaphysics of groups, this state of affairs is *realized* through one or more of its members working directly on one or more others. This creates exceptions to the usual moral restrictions on what we are permitted to do to one another. One individual can be required, *qua* part of a group agent, to do to someone else, *qua* other part of the group agent, what it would otherwise be impermissibly invasive or paternalistic to do to them *qua* individuals.

This is an interesting kind of exceptional case. But I do not think it arises all that often. More often, what is required of group members in response to egregious group moral failures is to work on some aspect of the group dynamics. There will be metaphysical facts about the group's nature and composition that explain why some particular member's egregious moral failure constitutes an egregious failure of the group agent – e.g., the fact that they are in a position of leadership or authority or are a figurehead for the group. And there will sometimes be moral shortcomings of the group *per se*, which do not reduce to the shortcomings of any individual and which partially explain the occurrence of the moral failure in question – e.g., the fact that the group lacks adequate checks and balances in its decision-making processes, or that it has not done enough to curb the development of a pernicious ideology among its members. When a group agent is required to work on itself in light of its egregious moral failures, it will usually be this sort of internal work that the group is required to do (and that its members, *qua* group members, are required to implement).

Those were the interesting cases. Here comes the final puzzle. Consider this principle:

> THE OXYGEN MASK RULE: We should prioritize working on ourselves over encouraging and facilitating other people's work on themselves.

This is a purported moral principle. I cheekily call it "The Oxygen Mask Rule" because it proposes, in effect, that one should make sure that one's own virtue is securely fastened before attempting to assist others.

Is the Oxygen Mask Rule true?

One might think that it fits nicely with the position I have defended in this paper. For I have argued that there are things we can do to ourselves that it is impossible or impermissible to do to others and that can be part of the process of working on yourself. However, it is quite a leap from these theses to the Oxygen Mask Rule. Plenty of ways in which we can encourage and facilitate others' work on themselves are both possible and permissible. And there might be minor ways in which we can directly work on others that do not invade privacy or violate autonomy so severely as to be impermissible. For example, simply sending someone an article that happens to be related to one of their moral shortcomings seems permissible. Our relationship to an agent might render permissible further minor ways of directly working on them – putting on a related movie and watching it with them if we are their friend or partner, asking them to complete a reflective worksheet in detention if we are their classroom teacher, and so forth. But all of the time that we spend working on others in these permissible ways, as well as the time that we spend supporting their work on themselves, is time that we could instead have spent working on ourselves. So, the question of relative priority still arises, notwithstanding the asymmetries in what we can and may do to others and to ourselves. We can ask: out of all of the permissible actions in this domain, which should we prioritize?

Here it is important to keep in mind a difference between the Oxygen Mask Rule and the guidance about real oxygen masks that we hear on airplanes. If you do not ensure that your own mask is securely fastened before attempting to assist others, then you may suffocate and so end up assisting nobody. But there is no such thing as moral suffocation. It is possible for even the worst among us to work on themselves. And no amount of work on others, nor of helping them with their work, removes your ability to work on yourself in future. Having worked on yourself a little bit first might make it easier to then encourage and facilitate others' work, perhaps by making you more trustworthy or giving you better ideas about what interventions are worth attempting. But it also might make things harder, perhaps by making you less relatable to your "improvees" or leading you into a one-size-fits-all approach premised on the mistaken assumption that what worked for you will also work for them. So, there is no simple expediency argument for the Oxygen Mask Rule. It is not at all clear that this division of labor maximizes total moral improvement of people. Nor, for that matter, is it clear that maximizing the total moral improvement of people is a worthy aim. If the Oxygen Mask Rule is to be defended, then, its defense must appeal directly to our normative intuitions about the (in)appropriateness of devoting more time and effort to work on oneself than on others.

I sometimes find the Oxygen Mask Rule highly intuitive. Again, this principle feels consonant with what I was raised to think (and, again, many of those with whom I have discussed the material in this paper report the same reaction). There is something intuitively attractive about the idea that we are each entrusted with the cultivation and care of just one moral character – viz., our own. And, relatedly, there is something attractive about the idea that we are not required to take on the load of others' faults and failings;[23] it is a relief to think that other people's moral shortcomings are *their problem*, so to speak, and that we cannot be required (even imperfectly) to take on additional work just because we are in the vicinity of someone else who otherwise cannot or will not work on themselves. More strongly, one might think that it is positively morally objectionable to be equally as concerned with others' moral shortcomings as one is with one's own; echoes of this thought appear in the literature on hypocrisy, in which it is often said that the problem with hypocrites is that they criticize others when they should instead focus primarily on themselves. These lines of thought each suggest that we should indeed prioritize working on ourselves, exactly as the Oxygen Mask Rule contends.

[23] Thanks to [redacted] for suggesting this excellent phrase.

But I can also get into a frame of mind in which I doubt the Oxygen Mask Rule. Even within the hypocrisy literature, authors disagree as to whether we should focus *primarily* on our own moral faults and failures or should focus on everybody's faults and failures *equally* (*modulo* differences in their severity).[24] The former is consonant with the Oxygen Mask Rule but the latter explicitly disavows it. One might also think that, since moral improvement is good, it is inegalitarian to distribute this good based on the notoriously-arbitrary criterion of personal identity. Indeed, one might even think that it is *generous* to prioritize securing this good for others over securing it for oneself.[25] There is something inspiring about the idea of a society of people who are all dedicated to everyone's collectively becoming as good as possible, utterly devoid of partisanship in the distribution of our moral efforts. From this point of view the Oxygen Mask Rule looks excessively individualistic, like a pedantic fixation on the limits of each person's moral agency. And yet, within the mindset of the Oxygen Mask Rule, this collectivist position seems to ignore precisely the limits on individual moral agency that are most important: it seems to make other people our responsibility in a way that they fundamentally are not.

Thus we find ourselves with yet another thesis and antithesis, each of which can be given some theoretical backing. My account of the nature and importance of working on yourself cannot resolve these tensions since it is presupposed by them. So I do not know how to answer the question of relative priority. I leave this puzzle for the reader.

Good luck, dear reader. Let me know if I can help.

[24] For example, Fitz and Miller (2018) see hypocrites as a species of the more general genus of "inconsistent blamers" (p.131), into which one falls whenever one is disposed not to display equal blame toward equal wrongdoing, whereas King (2020) proposes that "we have reasons to *prioritize* attending to our own faults *over* those of others" (p.1425, italics added) and that "[the hypocrite's] moral priorities should be focused primarily on cleaning up his own act" (p.1436).
[25] Thanks to [redacted] for suggesting this framing.

REFERENCES

Nomy Arpaly (2003). *Unprincipled Virtue: An Inquiry Into Moral Agency.* Oxford University Press**.**

Albert Bandura (1977). "Self-efficacy: Toward a Unifying Theory of Behavioral Change." *Psychological Review 84* (2):191–215.

--- (1997). *Self-Efficacy: The Exercise of Control*. New York: Freeman.

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

Jason Brennan (2007). "Modesty Without Illusion". *Philosophy and Phenomenological Research* 75 (1):111-128.

Agnes Callard (2018). *Aspiration: The Agency of Becoming.* Oxford University Press.

Vanessa Carbonell (2022). "Malicious Moral Envy." In Sara Protasi (ed.), *The Moral Psychology of Envy*. Rowman and Littlefield. pp. 129-146.

Lindsay Crawford (2019). "Believing the best: on doxastic partiality in friendship." *Synthese* 196 (4):1575-1593.

Robert Enright & Glen Pettigrove (eds.) (forthcoming). *Routledge Handbook of the Philosophy and Psychology of Forgiveness.* London: Routledge. Currently available for pre-order, ISBN 9780367030728.

Owen Flanagan (1990). **"**Virtue and ignorance**".** *Journal of Philosophy* 87 (8):420-428.

Kyle G. Fritz & Daniel Miller (2018). "Hypocrisy and the Standing to Blame." *Pacific Philosophical Quarterly* 99 (1):118-139.

Pamela Hieronymi (2001). "Articulating an uncompromising forgiveness". *Philosophy and Phenomenological Research* 62 (3):529-555.

Jessica Isserow (2020). **"**Moral Worth: Having It Both Ways**."** *Journal of Philosophy* 117 (10):529-556.

Robert Johnson (2011). *Self-Improvement: An Essay in Kantian Ethics*. Oxford University Press.

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

Simon Keller (2004). "Friendship and Belief." *Philosophical Papers* 33 (3):329-351.

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

Larry Laudan (2006). *Truth, Error, and Criminal Law: An Essay in Legal Epistemology*. Cambridge University Press.

Bertram F. Malle (2006). "The actor-observer asymmetry in attribution: A (surprising) meta-analysis". *Psychological Bulletin* 132 (6): 895–919.

Julia Markovits (2010). **"**Acting for the right reasons**".** *Philosophical Review* 119 (2):201-242.

Sarah Paul & Jennifer Morton (2018). "Grit". *Ethics* 129 (2):175-203.

--- (2018). "Believing in Others". *Philosophical Topics* 46 (1):75-95.

Ryan Preston-Roedder (2013). "Faith in Humanity". *Philosophy and Phenomenological Research* 87 (3):664-687.

Lee Ross (1977). "The Intuitive Psychologist and His Shortcomings: Distortions in the Attribution Process". *Advances in Experimental Social Psychology* 10:173-220.

Krisanna Scheiter & Paula Satne (eds.) (2022). *Conflict and Resolution: The Ethics of Forgiveness, Revenge, and Punishment.* Switzerland: Springer Nature.

Keshav Singh (2020). **"**Moral Worth, Credit, and Non-Accidentality**."** In Mark Timmons (ed.), *Oxford Studies in Normative Ethics, Vol. 10*. Oxford University Press.

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

Daniel Statman (1992). "Modesty, pride and realistic self-assessment." *Philosophical Quarterly* 42 (169):420-438.

Sarah Stroud (2006). "Epistemic Partiality in Friendship." *Ethics* 116 (3):498-524.

Justin Tosi & Brandon Warmke (2016). **"**Moral Grandstanding**".** *Philosophy and Public Affairs* 44 (3):197-217.

Brandon Warmke, Dana Kay Nelkin & Michael McKenna (eds.) (2021). *Forgiveness and Its Moral Dimensions.* New York: Oxford University Press.

Alan T. Wilson (2016). "Modesty as Kindness." *Ratio* 29 (1):73-88.