Supererogation and the Limits of Reasons[1]

Nathaniel Baron-Schmitt and Daniel Muñoz

For David Heyd (ed.), Springer handbook on supererogation

1. Introduction

Supererogatory acts are good deeds beyond the call of duty, ranging from friendly favors to saintly sacrifices to risky rescues. As any reader of this handbook will have noticed, philosophers disagree deeply about what supererogation is, and whether it is even possible.

To some extent, this is a verbal dispute. "Supererogation" is not ordinary language, like "good" or "wrong." It is a "quasi-technical term" (Heyd 1982), whose meaning is somewhat up for grabs.[2] If you say that supererogation must spring from a noble motive, whereas we say that only needs to be a good thing to do, there's no point in shouting at each other about the true essence of supererogation. We are better off admitting that we just prefer different definitions.

In our view, however, the big questions about supererogation are substantive, not verbal, and they can be asked in plain language. The fundamental question here is the:

Classic Paradox of Supererogation

If A is a better thing to do than B, how could it be permissible to do B?

For example, if donating a kidney is better than keeping it, how could donating be optional rather than obligatory? How could morality let us do the worse thing? These are perfectly good questions, and we can't dodge them by fiddling with our definitions. We confront, as Dancy puts it, "a

[1] The authors were equal contributors to this project. Daniel would like to thank Kerah Gordon-Solmon and Theron Pummer (as always); from Nathaniel, thanks to Benjamin Kiesewetter and other colleagues at the Human Abilities Center in Berlin.

[2] By contrast, when Horgan and Timmons define "supererogation," they do so with an eye towards the "common-sense usage of the term" (2010: 31).

philosophical boggle" (1993: 131, emphasis added). Nor can we define our way out of more recent boggles, like Horton's (2017) All or Nothing Problem.

How do philosophers try to solve the paradoxes of supererogation? In recent decades, most attempts have drawn on one and the same source: the theory of reasons for action.[3] And so we find a flurry of distinctions between kinds of reasons: agent-relative vs. agent-neutral (Dancy 1993), perfect vs. imperfect (Portmore 2011), requiring vs. justifying vs. favoring (Archer 2016; Horgan and Timmons 2010; Little and Macnamara 2017, 2020; Tucker forthcoming), and on the list goes. By mixing and matching varieties, and by linking them to other concepts like permissibility and value, ethicists have tried to make sense of supererogation.

This chapter provides a tour of some of the main paradoxes of supererogation, as well as the main solutions provided by reasons-ologists. We end with a twist: *the paradoxes of supererogation cannot be solved with reasons alone*. Supererogation, we think, is a counterexample to the "Reasons First" program, which tries to reduce ethics to the study of reasons.

None of our arguments, by the way, will turn on the definition of "supererogation." That said, a definition will come in handy. Since we are here to talk about the classic paradox and its descendents, we will define "supererogation" as the thing that the paradox is meant to rule out: supererogatory acts are optional and better than a permissible alternative. Of course, we are talking about morality here, so we mean that supererogation is *morally* better than a *morally* permissible alternative. As for "reason," we can use it in the familiar way: to be a reason is to count in favor of

[3] Some philosophers, especially old-school consequentialists, do see supererogation as intolerably paradoxical (see, e.g., Kagan 1984, 1989). But this view has its costs. "In commonsense moral reasoning, we take it for granted that there are supererogatory acts, and it would be incredible if the very idea of supererogation turned out to be incoherent" (Dreier 2004: 145).

some way of acting, making it more choiceworthy, or in other words, making it a better thing to do.[4] Reasons, so understood, may not be enough to solve the paradoxes of supererogation.

We begin with the classic paradox.

2. The Classic Paradox

Why does supererogation feel paradoxical? Or to put it another way, why would anyone worry that heroic sacrifices and kindly favors might turn out to be obligatory?

Let's start with a textbook case of an obligatory act. Consider:

Scarce Drug

You own a scarce drug, which you can use either to save your acquaintance Alex, who needs it all to survive, or five others, who need only a fifth each. (Foot 1967)

Most people think you have to save the five, other things being equal.[5] Saving one life is good; saving five is better still; and so, people infer that you have to save the five.

Behind this easy inference looms a principle: that we have to do the best thing we can. The "best" thing, in the relevant sense, is the one that is most choiceworthy, the most favored by the balance of reasons.[6] So we get:

---

[4] By 'reason', we mean "normative reasons," as opposed to "motivating reasons," which are the grounds on which an agent acts. The two concepts can come apart. One might not be motivated by the normative reasons in favor of helping those far away (for example), and one might instead be motivated by whims or prejudices that aren't supported by normative reasons at all. For a helpful overview of these issues, see Dancy 2000.

[5] For exceptions, see Taurek (1977) and Anscombe (1967).

[6] The best thing to do needn't have the best outcome. For example, killing Alex to save five other people's lives is a bad thing to do, even if its outcome is optimal, because Alex has the right not to be killed. This fact in itself is a reason not to kill her.

The Most Reason Principle

If you have most reason to do A, then A is obligatory.

A corollary is that, if you have more reason to do A than B, B is wrong—obligatory not to do.[7]

This principle can seem undoubtable, and so it should be no surprise that we find people asking how one could possibly doubt it:

> How can one be permitted to refrain from action which is required by reason? (Raz 1975, 165)

> …how can an action that is morally best to perform fail to be what one is morally required to do? (Horgan and Timmons 2010, 29)

> …how can supererogatory acts be so valuable and important, and yet not obligatory? (Heyd 1982, 4; he calls this the "good–ought tie-up")

You get the idea.

All of this leads us to the classic paradox. The problem is that supererogatory acts, if they exist, must be better than some permissible alternative. But the alternative *can't* be permissible if supererogating would be better. Doing best is obligatory, given the Most Reason Principle, which is just a basic principle linking duties to reasons. So there can't be such a thing as a supererogatory act.

---

[7] In this paper, when we refer to options like A and B, we have in mind options that are fully specific, rather than options that can be carried out in relevantly different ways. Think: "saving Alex," not "either saving Alex or saving the five." Also, we'll stick to choices from finitely many options. (Our "corollary" wouldn't follow in infinite cases.)

And yet, such acts seem not only conceivable but actual! A friend of ours (another philosopher) once donated a kidney to a stranger, who was probably spared years of painful dialysis. This was a wonderful thing to do, far better than the alternative of not donating. But it seems extreme to say that the stranger was entitled to the kidney, or that our friend was merely doing her duty. Don't people have the freedom to decide what happens to their body, even when their decision isn't quite optimal?

More generally, people often seem to enjoy a permission to favor their own interests over the greater good of others. Contrast our earlier example with:

Alex's Scarce Drug

Alex owns a scarce drug, which she can use either to save herself or five others.

Now, saving the five does *not* seem like an obligation. It seems optional, like our friend's donation. (Though Alex *would* be obligated to sacrifice if the stakes were different. Presumably, she wouldn't be allowed to cure her own headache rather than save a million lives.)

The classic paradox, in effect, gives us two problems at once. First, we need to explain how the very idea of supererogation is coherent. This means, if we are reasons theorists, that we need to replace the Most Reason Principle with a more complex account of how reasons give rise to requirements. Second, our account needs to fit with core instances of supererogation and obligation, such as the drug cases above. Our theory might be coherent but problematic when applied to the relevant examples.

In our view, these problems are central to normative ethics, and it is a good thing that ethicists have tried to solve them using reasons.[8] The idea of a reason nicely brings out the tension in

[8] Reasons took a while to catch on. (See Pybus 1982, for an example of a rejection of supererogation that doesn't use the concept of a reason.)

the idea of supererogation. Reasons traditionally play three roles: they tend to favor, justify, and require. Supererogation seems to upset this presumption. Perhaps the reasons to supererogate merely favor (without requiring), or the reasons against supererogation merely justify (without favoring or requiring)—unless we add *something* like that to our theory, the very idea of supererogation will be incoherent.

3. Equal Weight

Is that right? Why can't we just say that supererogation involves a clash of reasons, understood in the traditional way?

Again, reasons are traditionally thought to play three roles at once:

1. Reasons favor actions, in the sense of tending to make them choiceworthy.

2. Reasons justify actions, in the sense of tending to make them permissible.

3. Reasons require actions, in the sense of tending to make them obligatory.

One simple way to think of this is with a trio of numbers. We assign to each option X a triple $<X_F, X_P, X_R>$, where these are numbers representing how much the relevant reasons favor, justify, and require that option, respectively.

A word of caution about what these numbers mean. If we give X a score of <2, 2, 2>, we aren't saying that X is *doubly* justified, or *doubly* required. These numbers get their meaning from how they figure in comparisons, which then determine permissions and requirements.

Obligation

An option is obligatory if and only if its "requirement" score beats any alternative's "justification" score.

Permission

An option is permissible if and only if its "justification" score is not beaten by any alternative's "requirement" score.

Choiceworthiness

An option is more choiceworthy than an alternative if and only if it has a higher "favoring" score than the alternative has.[9]

This three-way distinction gives us a deeper way to understand the classic paradox. Notice that the three principles above do *not* by themselves conflict *in the least* with the possibility of supererogation. For they don't entail anything like the Most Reason Principle—not until we add a

[9] We can write this out a bit more formally, if you like, using "O(X)" to mean "X is obligatory," and "P(Y)" to mean "Y is permissible."

Obligation

O(X) *iff*, for any alternative Y, $X_R > Y_J$.

Permission

P(X) *iff*, for no alternative Y, $Y_R > X_J$.

Favoring

X is more choiceworthy than Y *iff* $X_F > Y_F$.

We assume that the three dimensions can be measured using a common unit, as we can use meters to measure a thing's length, width, and height. We might also want to assume that the ratios of numbers matter, and that the scales have a meaningful zero point.

further assumption. That assumption—the real crux of the paradox—is the traditional equivalence between the three kinds of weight. We can call it:

Equal Weight

$X_F = X_J = X_R$.

In other words, reasons favor, justify, and require in equal measure. So there cannot be reasons that purely favor, purely justify, or purely require.[10]

With this, we get the classic paradox in a fancier form, which distinguishes the kinds of reasons at play. Supererogation is optional and better than a permissible alternative. So, there must be more favoring reason to supererogate than to do the alternative. It follows, given Equal Weight, that there must be more *requiring* reason to do the supererogatory act than there is *justifying* reason to do the alternative. But then the alternative cannot be permissible after all, given our principle of Permission. Contradiction!

How can we make room for supererogation? Nowadays, the most popular move is to deny Equal Weight, opting for reasons that purely justify or purely favor.[11] But another option, more popular in the 1970s and 1980s, did not rely on any such revision to our idea of reasons. Let us take a closer look at this move, along with its problems.

---

[10] Equal Weight is, strictly speaking, overkill. It implies that we can make comparisons of strength between favoring reasons and justifying/requiring reasons, using a common unit. (Like saying "X's height in meters = Y's length in meters.") To get the Paradox, we don't need such comparisons. We just need to say that $X_F > Y_F$ entails $X_R > Y_J$. But this principle is a bit unfamiliar, and the technicalities here don't matter much, so in the text we use the stronger, simpler Equal Weight.

[11] For a different approach, see Bedke's (2011) "Millian inversion," which analyzes requirements in terms of "reasons to require," not reasons to do the thing required. See Snedegar (2016) for a powerful critique. (Snedegar, like us, pushes back against Reasons First, though with different targets and motivations. For example, he critiques Gert (2004), and he focuses on the need for an account of 'ought' rather than the need to solve the traditional paradoxes.)

4. The Shadow Paradox

As we have just seen, there is no way to get supererogation, as we have defined it, given the traditional view of reasons. But remember: what counts as "supererogation" is up for grabs, so even if supererogation is impossible *as we define it*, it might be possible given a less stringent definition. Here we have to use our judgment to see whether the less stringent claim is still interesting.

Consider, for example, the idea that self-sacrifice involves a conflict between two kinds of moral reasons: reasons to help others and reasons to respect duties to oneself.[12] Given Equal Weight, this sort of view cannot say that self-sacrifice is supererogatory (in our sense). At best it can say that, in some cases, self-sacrifice is optional. Giving one's kidney can be optional if the reason to keep it is equal to the reason in favor of giving; the duty to help the recipient balances off against the donor's duty to self.

But this move does not get us very far, for two reasons.

First, the optionality is not robust enough. If the reasons to give *exactly* balance off against the reasons to keep, then—assuming that the weights of reasons are measured with single numbers, like physical weights—any increase in one side will tip the scales. If we were to make an optional sacrifice even barely less costly, that would make it obligatory. This is an absurd result; self-sacrifice should be optional across a range of costs and benefits, not just at one precise balance point (Hurka and Shubert 2012, 8; Kagan 1989, 378–79).

To address this problem, reasons-ologists have come up with several ideas as to how a tie between reasons might be stable over small additions to either side. For example, some say that the

[12] This idea is popular in Kantian ethics. Patricia McGoldrick, for example, argues that heroic sacrifices are not obligatory because such an obligation "would come into conflict with our obvious duty to recognize our own intrinsic worth" (1984, 527), and she says that, in the ethics of self-sacrifice, the "heart of the matter is the Kantian argument that we have duties to ourselves as well as others" (ibid.: 527).

two clashing reasons are only *imprecisely* equal in weight, and that small boosts don't break imprecise ties (Parfit 2011: 137–41).[13]

But even if these moves can solve the classic paradox—and we have doubts[14]—they lead us straight into a second, deeper paradox. Let's grant that the reason against donating undercuts the duty to donate. Given Equal Weight, won't it also undercut the *betterness* of donating? On a view like Parfit's, there is nothing more choiceworthy about making the sacrifice. We save the optionality of supererogation at the cost of its superiority.

We call this the:

Shadow Paradox of Supererogation

If there is a strong reason against typical supererogatory acts, why should these acts be more choiceworthy than the moral minimum?[15]

As with the classic paradox, the shadow paradox has a conceptual side as well as a normative one. Our theory has to be coherent while also being a good fit for the relevant cases, most notably self-sacrifice.

Any theory of supererogation, then, has to steer between these two paradoxes, the classic and its shadow. We need to find something that can make supererogation optional in the relevant cases without detracting from its value.

---

[13] Portmore has a more complicated view: the clashing reasons are *imperfect*, in the sense that they "do not support performing any specific alternative, but instead support performing any of the alternatives that would each constitute an equally effective means of achieving the same worthy end" (2011: 156). We focus on Parfit's view because it is simpler and more general.

[14] See Muñoz (ms.) for a critique of the appeal to imprecise weights.

[15] Some views face a particularly grave version of the shadow paradox called the "Wrongness Problem" (Muñoz 2021a: 615, Muñoz and Baron-Schmitt ms.). If the reason against donating is weightier than the reason to donate, then, given Equal Weight, the sacrifice we wanted to call supererogatory is in fact *wrong*. Postow (2005: 246) discusses a version of the shadow paradox where the sacrifice is "irrational"—opposed by decisive non-moral reasons.

5. Favoring Reasons

To solve the paradoxes, we have our work cut out for us. We need to prevent the reasons to supererogate from generating a duty (classic), while also ensuring that supererogation is favored over the moral minimum (shadow). There are two basic ways to do this. We could say that the reasons to supererogate favor more than they require, or that the reasons against supererogating justify more than they favor. Let's start with the first option.

According to the "pure favoring view," supererogation is possible because the reasons to supererogate count in favor without tending to require anything. Such reasons are "purely favoring reasons."[16] These reasons are almost tailor-made to solve our pair of paradoxes. Because pure favorers do not even tend to ground requirements, acting on them is optional by default; they do not need to be offset by contrary reasons, which would then threaten the value of supererogating.

The pure favoring view, we think, is conceptually coherent. But does it fit the cases? It seems like a good fit for low-stakes kindnesses, gifts, and personal favors, which are the focus of Horgan and Timmons (2010). They argue that pure favoring fits the phenomenology of kindnesses, such as in their case of Olivia:

> Olivia…meets a recently widowed woman, Mary, a neighbor who lives a few doors down…[who] lost her husband to cancer…[and who] is an avid baseball fan…But without anyone to go with, [Mary] doesn't go [to baseball games] anymore. The next day, it occurs to Olivia that it would be a nice gesture to offer to go to a Cardinals game with Mary, although [Olivia] herself has no particular interest in the game. But she thinks: "Here is a chance to do

[16] Other terms for "favoring" include "merit-conferring" (Horgan and Timmons 2010) and "commendatory" (Little and Macnamara 2017). See also Dancy (2004a) on "enticing reasons" and Heyd (1982: 171-72) on the "commendatory" vs. "prescriptive" senses of "ought."

> something nice for someone…Why not?" She calls Mary, who is delighted by the invitation, and they end up going to a game. (2010: 47)

Horgan and Timmons suggest that "[in] contrast to cases of obligation, Olivia does not experience the reasons she has to [invite Mary to the game] as requiring her to do so, although, of course, the reasons in question are experienced as favoring [inviting her]. Such reasons, then, are experienced differently than are the reasons involved in experiences of obligation" (2010: 48) And if the experience is accurate, we can conclude that the reasons to perform such kindnesses are purely favoring.[17]

What about costly beneficence? Horgan and Timmons do not try to extend the pure favoring view to risky rescues and saintly sacrifices. Dreier (2004), however, suggests that these can be morally optional if beneficence is purely favoring. He does not think all moral reasons purely favor; reasons of justice, in his view, do ground requirements. But beneficence in particular is purely favored, and that is why we don't have to make costly sacrifices.

The pure favoring view is coherent, and it seems plausible in certain cases. But not all cases (Snedegar 2016: 165). What if beneficence is cheap? Suppose you could rescue a child at the cost of getting your suit muddy (Singer 1972). Or suppose it is 100 children, at no cost at all. Surely *costless* rescues are obligatory.[18] Beneficence, on reflection, is only optional when the costs are fairly high relative to the benefits. This suggests that costly sacrifices are optional precisely *because they are costly*, not because the reasons to help others are inherently non-requiring. It is not that beneficence is inherently optional, but rather that one can sometimes justifiably refuse to pay big costs.

---

[17] The argument shows that Olivia's reason is, in a sense, non-requiring. But must it be *intrinsically* non-requiring? What if it only fails to require because Olivia is justified in spending her time as she likes? We'll come back to this idea later when we discuss justifying reasons and prerogatives.

[18] Here we are following Archer (2016: 460) and Portmore (2008: 381), who press this objection against Dreier's (2004) version of the pure favoring view. Our objection to the primarily favoring reasons view, in the next section, is original.

We take this "costless rescue problem" to be a decisive objection to the pure favoring view. Pure favoring cannot be the solution to the paradoxes of supererogation, because many cases of supererogation do not involve pure favoring. But some writers have tried to amend the view to get around the problem. They posit *primarily* favoring reasons, which favor more than they require, while still requiring to some degree. Little and MacNamara (2017) take this approach (see also Archer 2016: 459): they say that the reasons to save lives have *some* requiring weight, but not as much as they have favoring weight: $X_F > X_R > 0$. So, rescues are optional when costly but obligatory when cheap.

The "primarily favoring view," as we'll call it, is the best view we've seen yet. It can handle the classic and shadow paradoxes and a whole range of cases, including costless rescues. But we don't think the view ultimately works. It struggles with a more complicated paradox, Joe Horton's All or Nothing Problem, to which we will now turn.

6. All or Nothing: The Need for Justifying Reasons

Horton gives a case with three options:

> *Two Kids*
>
> Suppose that two children are about to be crushed by a collapsing building. You have three options: do nothing, save one child by allowing your arms to be crushed, or save both children by allowing your arms to be crushed. (2017: 94)

Saving neither ("0") seems to be permissible, as does saving both ("2"); but saving only one ("1") is wrong. You can't justify saving only one when you could save a further child at no extra cost. Yet, despite its wrongness, 0 is surely not more choiceworthy than 1. If you save a life by sacrificing your

arms, that does not seem worse than doing nothing and saving no one. In our terms: $0_F \leq 1_F$.[19] This is the All or Nothing Problem.

In our view, solving the problem means finding a way to make these moral "seemings" fit together. But the primarily favoring view by itself cannot do this. If we suppose that justifying weight and favoring weight are equal, it follows from $0_F \leq 1_F$ that $0_J \leq 1_J$. Since 0 is not favored over 1, 0 is not more strongly justified. But since 0 is permissible, whereas 1 is not, the justifying reason for 0 must be greater than the justifying reason for 1. $2_R > 1_J$ and $2_R \leq 0_J$, so $0_J > 1_J$. Contradiction!

What went wrong? The issue here is not that favoring is allowed to outstrip requiring; it is that justifying is not allowed to outstrip favoring. The primarily favoring view, like the pure favoring view, leaves intact the assumption that $X_F = X_J$. It is this assumption, we think, that we need to reject if we want to solve the All or Nothing Problem.[20]

Let's suppose that there can be "purely justifying reasons," so that an option's justifying weight can outstrip its favoring weight.[21] Then we can say that at least some of the self-centered reasons to do nothing justify more than they favor (though not so for the ordinary moral reasons for saving one, which, let's assume, all justify and favor to the same extent). In other words, $0_J > 0_F$. This

[19] Horton himself would not put the point this way. His preferred term is "ought rather," and the claim he denies is that one ought to save zero rather than only one. While we are on the subject, one understated part of Horton's view is his eschewal of Reasons First. He thinks we have "justifications" not to pay costs, which may or may not be "reasonable" to appeal to in defense of one's actions (2017: 98), but he never relies on the idea of reasons for action, much less does he try to reduce everything to reasons. This is just one further way in which Horton's paper is original.

[20] See Rulli 2020 and Muñoz 2021b, which go into more depth about the idea that a wrong option could be no worse than a permissible alternative.

[21] The recent literature on "justifying reasons" starts with Joshua Gert (2007), who distinguishes two distinct "strength values" a reason may have: justifying and requiring. But the idea of pure "justifying strength" is not so new. Before Gert, "prerogatives" (Kamm 1996: chap. 8; Scheffler 1994), or "agent-relative permissions" (Parfit 1978), played the same role as justifying reasons, but without the name "reason." Others have developed similar concepts; Snedegar's (2021) version of a justifying reason is one that counts "for" an option but not "against" alternatives; Greenspan (2005) distinguishes "positive" from "negative" reasons; Portmore (2011) applies the justifying/requiring distinction to moral reasons in particular; and so on. The story, in short, is that pure justifiers were developed before Reasons First, and they were rediscovered later and only then conceived as a part of the theory of reasons, rather than something beyond its limits.

allows us to say that $0_J > 1_F = 1_J > 0_F$. This is an intuitive idea, once you get past the symbols. It follows from the natural thought we touched on earlier: costly sacrifices are optional because of their costs. Although 1 is at least as choiceworthy as 0, 1 is more costly, making 0 easier to justify.

With purely justifying reasons, we can tell a perfectly coherent (and overly simple) story about the All or Nothing Problem. The weights of reasons could be as follows:

	Save 0	**Save 1**	**Save 2**
Favoring	0	5	10
Requiring	0	5	10
Justifying	15	5	10

Of course, this is much too simplistic. Most of us think there is *something* favoring the choice to do nothing; we might fix this by making the zeros into, say, twos. (See the discussion below of "primarily justifying reasons.") Some might be skeptical of the idea that *any* numbers could accurately represent the weights of reasons; in that case, we would need a more complicated formalism. (See the discussion of imprecise reasons above.)

Still, even if the details need to be finessed, the above table tells us something important. Notice that the favoring weight for all three options is equal to the requiring weight. We can give a coherent story about the All or Nothing Problem without distinguishing favoring weight from requiring weight. Solving the problem doesn't require primarily favoring reasons (or purely favoring reasons). We just need one departure from Equal Weight: a reason to do nothing that justifies more than it favors.

This is a striking result. With purely justifying reasons, we can solve Horton's problem, as well as the classic and shadow paradoxes and the costless rescue problem.[22] We do not need to distinguish requiring weight from favoring weight, or to allow that requiring (or favoring) weight can exceed justifying weight.[23] (In fact, we may want to deny that this last inequality is possible, since it allows for moral dilemmas. If $X_R > Y_J$ and $Y_R > X_J$, then one must do each of two incompatible things.)

7. Are Justifying Reasons Reasons?

Purely justifying reasons allow us to solve the paradoxes of supererogation. Supererogation is better, yet optional, because the reason against merely *justifies* refraining; it doesn't undercut the betterness of going ahead, much less make it wrong to do so. This, in our view, is the most promising approach for the "reasons first" account of supererogation. But there is something funny, we think, about the very idea of purely justifying reasons: they don't *favor*. Isn't favoring, understood as making an option more choiceworthy, supposed to be the essential mark of a reason?

Many ethicists think so:

> *A* is better than *B* iff there is more reason to choose *A* than to choose *B*, given the choice between *A* and *B*. (Snedegar 2017: 93)

[22] Just in case it's not clear how justifying reasons help with the other paradoxes: costless rescue is obligatory, whereas costly rescue is optional, because one has a purely justifying reason not to pay the costs. (The reason to rescue can have equal justifying, favoring, and requiring weight.)

[23] One further paradox does complicate things: Kamm's Intransitivity Paradox (Archer 2016; Dorsey 2013; Kamm 1996; Kamm 1985; Portmore 2017). Elsewhere, one of us argues that we can only solve the paradox if justification is *comparative*, in the following sense: how far I can justify doing X rather than Y is not intrinsic to X, but depends on the relative costs of X and Y; so, I might have a powerful justification to do nothing rather than give my arms to save a life, but no such justification to do nothing rather than costlessly save somebody's hand (Muñoz 2021b). The idea that justification is comparative is, we think, fairly intuitive. The idea that it is nontransitive is goes back to Parfit's (1982) reply to Kavka (1982). Parfit's insight has been overlooked, though his example has become famous again because it foreshadowed (wait for it) the All or Nothing Problem.

> I will take the idea of a reason as primitive. Any attempt to explain what it is to be a reason for something seems to me to lead back to the same idea: a consideration that counts in favor of it. (Scanlon 1998: 17)

> …reasons…count in favor of possible *properties* of agents—things that agents can do, in a very broad sense of 'do'. (Schroeder 2021: 34)

> [I]t's commonly accepted that a normative reason for action is *a consideration that counts in favor of the action* (Markovits 2014: 2)

Now, to be sure, some think there are deeper ways to spell out "counting in favor" (perhaps in terms of "ought," as in Hurka 2014: 32, though see Dancy 2004, Chapter 2); others find a subtle difference between favoring and being a reason, which applies in certain many-option cases (see Schroeder 2021; Snedegar 2013: 42).

Still, we think most readers will agree that the business of reasons is to count in favor of actions. But then how could there be purely justifying reasons? The whole point of such reasons is that they do *not* favor (or require). They only justify.[24]

There is a possible way around this problem. Perhaps there are no *purely* justifying reasons. But there might be *primarily justifying reasons*—reasons that "justify more than they require" (and favor), in the words of Chris Tucker.[25] The idea here is that the reasons against supererogation count

[24] It won't help to say these reasons "count in favor" of an action's being permissible, or merely that these reasons "compete" with other reasons to determine what is permissible (see Schroeder 2021b: section 3). Favoring means counting in favor of *doing* an action, making it more choiceworthy.
[25] Tucker (forthcoming) defends the existence of such reasons, which he calls "justifying heavy requiring reasons."

as genuine reasons, because they do genuinely favor, but only a bit, so it is still better overall to act against them. And so we might think that the cost of being the hero, in Two Kids, is a primarily justifying reason not to save anyone. The costs favor somewhat ($0_F > 0$), and they justify even more than they favor ($0_J > 0_F$).

But we do not think this move can save Reasons First.[26]

For one thing, even if costs favor, not all justifiers work like costs. Consider body rights. You have the right to keep your spare kidney, if you choose, because it is *yours*. But this in no way counts against donating. As Jeremy Waldron puts it:

> ...the fact that P has a right to do A does not of itself give rise to any reason in favor of A which is capable of competing with and being balanced against the reason for not doing A. (Waldron 1981: 28)

The fact that your organs are rightfully *yours*, then, would appear to be a pure justification for not donating.[27] (See Muñoz (2021a) for more on this idea.) In our view, the same can be said about body rights in Two Kids; the fact that it's *your arms* at stake is a pure justification for not sacrificing them.

---

[26] There are ways to appeal to justifying reasons. Dancy thinks supererogation is opposed by an agent-relative reason we can "discount" at will. This is like our view, except we prefer waivable rights to discountable reasons, because such reasons would not be "stable" over the choice of whether to act on them (see Muñoz 2020: 700; 2021a: 616). Another view is Raz's (1975: 167). He thinks supererogation is made possible by "exclusionary permissions," which come from reasons that "entitle disregarding" ordinary reasons to act. The balance of (first-level) reasons favors supererogating, but one may disregard the balance (thanks to second-level exclusionary reasons). Raz's view struggles with the shadow paradox: if one has a (decisive) reason to disregard the reasons to give one's kidney, then giving the kidney isn't choiceworthy, and may be irrational. (Raz might also have to reject the principle, mentioned above, that reasons must be "stable" over our decisions.)

[27] Waldron himself would not say that body rights "justify." For Waldron, to justify an action is "to show that the standard to which in the circumstances it conformed or the worthiness of the goal that it was intended to advance" (1981: 28). Clearly, "it's mine" does not show that keeping a spare organ is a *worthy* goal. But it certainly does help show that keeping the organ is *justifiable*, and that is all we mean by saying that it "justifies." (Perhaps "it's mine" reduces the need for justification, rather than providing a justification.)

Second, even if costs are "primarily justifying reasons," such reasons are an awkward fit for Reasons First. For they seem to be *more* than just reasons. Costs, on this view, are supposed to justify actions more than they favor. So how can we reduce their justifying to their favoring, that is, to their being reasons? Their excess justifying weight is a further independent feature that they have, in addition to their being reasons.

We do agree with Tucker about some things. In Two Kids, there is some favoring weight behind saving no one, and still more justifying weight. But from this, we don't want to infer that *one and the same thing* is doing the favoring and the justifying, or that the justifying *reduces to* the favoring. The fact doing the favoring could be the cost of heroism; the source of the excess justification could be a pure justifier, like the right not to harm oneself; and at any rate, the justifying seems like an independent factor, not reducible to the favoring.

What are we to make of the appeal to "justifying reasons," if they aren't just reasons? We could give up on anything like them, on the grounds that Reasons are First. But then we would leave some of the paradoxes unsolved. Solving the paradoxes, in our view, is more important than keeping Reasons First. And we're close to a solution! After all, justifying reasons are almost what we need. They help with the paradoxes because *they justify more than they favor*; it is only this feature of them, not their status as reasons, that we need for our theory.

8. If Not Reasons, Then What?

We have argued that the best view of supererogation is one where justification can outstrip the favoring power of reasons. If giving one's spare kidney is supererogatory, for example, then one has a sufficient justification to keep it, even though one has more reason to give than to keep. On such a view, Reasons are not really First. Alongside reasons, we need pure justifiers—which are traditionally

known as prerogatives (Kamm 1996; Muñoz 2021a; Scheffler 1994), or "agent-relative" permissions (Parfit 1978: 287; Slote 1984).[28]

But what is a prerogative, if not a kind of reason? Hurka and Shubert (2012) argue that we should take the concept of a prerogative, like that of a reason, to be primitive. Prerogatives cannot be derived from any other element of ethics, though they can be described with analogies and metaphors. Hurka and Shubert give a helpful physics analogy: whereas reasons are like independent forces acting on an object, prerogatives are like friction (2012: 7-8). Friction cannot move an object on its own; it can only mitigate other forces. Prerogatives, analogously, cannot favor or require an action on their own; they can only prevent reasons from grounding an obligation.

Hurka and Shubert argue that prerogatives are worth having even if they have to be left primitive. We agree. But in our view, there is still hope for deriving prerogatives from something deeper—not reasons, but *rights*. In our view, prerogatives can be derived from waivable rights against oneself (see Muñoz 2021a for the details).[29] The key is that whether such rights are in play depends on what you decide; your kidney is off limits if, and only if, you decide not to donate.[30]

To solve the paradoxes of supererogation, we need not only reasons, which count in favor of actions, but also prerogatives, which purely justify. The classic and shadow paradoxes can be solved using purely or primarily favoring reasons, but these lead to other problems: the costless rescue problem and Horton's All or Nothing Problem. Justifying "reasons" solve the paradoxes, but they

---

[28] Hurka and Shubert (2012) use "prima facie permissions," meaning "things that tend to make permissible," as an homage to W. D. Ross's (1930) "prima facie duties," which are things that tend to make an option a duty; the modern term for this is "moral reason." (Although we do not follow Hurka and Shubert in taking prerogatives to be primitive, we have learned much from their paper.)

[29] See also Kagan 1989: 206-216; Muñoz and Baron-Schmitt ms. For more on waiving duties to/rights against oneself, see Kanygina 2022, Muñoz and Baron-Schmitt forthcoming, Muñoz 2020, Schaab 2021, and Schofield 2021.

[30] We think it might also be possible to derive prerogatives from the happiness of merely possible people (supposing their happiness only matters *if you create them*); here we take inspiration from Spencer (2021) on attractive permissions.

aren't really reasons. Instead, we need something other than reasons to justify: either primitive prerogatives, rights against oneself, or some other justifiers waiting to be discovered.

Bibliography

Anscombe, G.E.M. 1967. "Who Is Wronged? Philippa Foot on Double Effect: One Point." *The Oxford Review* 5: 16–17.

Archer, Alfred. 2016. "Moral Obligation, Self-Interest and The Transitivity Problem." *Utilitas* 28(4): 441–64.

Bedke, Matt. 2011. "Passing the Deontic Buck." In *Oxford Studies in Metaethics, Volume 6*, ed. R. Shafer-Landau. Oxford: Oxford University Press, 128–52.

Dancy, Jonathan. 1993. *Moral Reasons*. Oxford: Blackwell.

———. 2000. *Practical Reality*. Oxford: Oxford University Press.

———. 2004a. "Enticing Reasons." In *Reason and Value: Themes From the Moral Philosophy of Joseph Raz* eds. R. Jay Wallace, Philip Pettit, Samuel Scheffler, and Michael Smith. Oxford: Clarendon Press, 91–118.

———. 2004b. *Ethics Without Principles*. Oxford: Oxford University Press.

Dorsey, Dale. 2013. "The Supererogatory, and How to Accommodate It." *Utilitas* 25(3): 355–82.

Dreier, James. 2004. "Why Ethical Satisficing Makes Sense and Rational Satisficing Doesn't." In *Satisficing and Maximizing*, ed. Michael Byron. Cambridge University Press, 131–54.

Foot, Philippa. 1967. "The Problem of Abortion and the Doctrine of the Double Effect."*Oxford Review* 5: 5–15.

Gert, Joshua. 2007. "Normative Strength and the Balance of Reasons." *Philosophical Review* 116(4): 533–62.

Greenspan, Patricia. 2005. "Asymmetrical Practical Reasons." In *Experience and Analysis: Proceedings of the 27th International Wittgenstein Symposium*, eds. J. C. Marek and M. E. Reicher. ÖBV and HPT, 387–94.

Heyd, David. 1982. *Supererogation: Its Status in Ethical Theory*. Cambridge: Cambridge University Press.

Horgan, Terry, and Mark Timmons. 2010. "Untying a Knot from the inside out: Reflections on the 'Paradox' of Supererogation." *Social Philosophy and Policy* 27(2): 29–63.

Horton, Joe. 2017. "The All or Nothing Problem." *Journal of Philosophy*114(2): 94–104.

Hurka, Thomas. 2014. *British Ethical Theorists from Sidgwick to Ewing*. Oxford: Oxford University Press.

Hurka, Thomas, and Esther Shubert. 2012. "Permissions to Do Less Than Best: A Moving Band." In *Oxford Studies in Normative Ethics, Volume 2*, Oxford: Oxford University Press, 1–27.

Kagan, Shelly. 1984. "Does Consequentialism Demand Too Much? Recent Work on the Limits of Obligation." *Philosophy and Public Affairs* 13(3): 239–54.

———. 1989. *The Limits of Morality* Oxford: Oxford University Press.

Kamm, Frances M. 1985. "Supererogation and Obligation." *Journal of Philosophy*82(3): 118–38.

———. 1996. *Morality, Mortality: Volume II: Rights, Duties, and Status*. New York: Oxford University Press.

Kanygina, Yuliya. 2022. "Duties to Oneself and Their Alleged Incoherence." *Australasian Journal of Philosophy* 100(3): 565–79.

Kavka, Gregory S. 1982. "The Paradox of Future Individuals." *Philosophy and Public Affairs* 11(2): 93–112.

Little, Margaret Olivia, and Coleen Macnamara. 2017. "For Better or Worse: Commendatory Reasons and Latitude." In *Oxford Studies in Normative Ethics, Volume 17*, ed. Mark C. Timmons. Oxford: Oxford University Press, 138–60.

———. 2020. "Non-requiring reasons." In Chang and Sylvan, eds., *The Routledge Handbook of Practical Reasons*. New York: Routledge.

Markovits, Julia. 2014. *Moral Reason*. Oxford: Oxford University Press.

McGoldrick, Patricia M. 1984. "Saints and Heroes: A Plea for the Supererogatory: Discussion." *Philosophy* 59(230): 523–28.

Muñoz, Daniel. 2020. "The Paradox of Duties to Oneself." *Australasian Journal of Philosophy*98(4): 691–702.

———. 2021a. "From Rights to Prerogatives." *Philosophy and Phenomenological Research* 102(3): 608–23.

———. 2021b. "Three Paradoxes of Supererogation." *Noûs* 55(3): 699–716.

———. Manuscript. "Values as Vectors."

Muñoz, Daniel, and Nathaniel Baron-Schmitt. Forthcoming. "Wronging Oneself." *Journal of Philosophy*.

———. Manuscript. "Why Isn't Supererogation Wrong?"

Parfit, Derek. 1978. "Innumerate Ethics." *Philosophy and Public Affairs* 7(4): 285–301.

———. 1982. "Future Generations: Further Problems." *Philosophy and Public Affairs* 11(2): 113–72.

———. 2011. *On What Matters: Volume One*. Oxford: Oxford University Press.

Portmore, Douglas W. 2008. "Are Moral Reasons Morally Overriding?" *Ethical Theory and Moral Practice* 11(4): 369–88.

———. 2011. *Commonsense Consequentialism: Wherein Morality Meets Rationality*. New York: Oxford University Press.

———. 2017. "Transitivity, Moral Latitude, and Supererogation." *Utilitas* 29(3): 286–98.

Postow, B. C. 2005. "Supererogation Again." *Journal of Value Inquiry*39(2): 245–53.

Pybus, Elizabeth M. 1982. "Saints and Heroes." *Philosophy* 57(220): 193–99.

Raz, Joseph. 1975. "Permissions and Supererogation." *American Philosophical Quarterly* 12(2): 161–68.

Ross, W. D. 1930. *The Right and the Good*. Oxford: Oxford University Press.

Rulli, Tina. 2020. "Conditional Obligations." *Social Theory and Practice* 46(2): 365–90.

Scanlon, Thomas. 1998. *What We Owe to Each Other*. Cambridge: Harvard University Press.

Schaab, Janis. 2021. "On the Supposed Incoherence of Obligations to Oneself." *Australasian Journal of Philosophy*99(1): 175–89.

Scheffler, Samuel. 1994. *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*. Oxford: Oxford University Press.

Schroeder, Mark. 2021a. *Reasons First*. Oxford: Oxford University Press.

———. 2021b. "The Fundamental Reason for Reasons Fundamentalism." *Philosophical Studies* 178(10): 3107–27.

Schofield, Paul. 2021. *Duty to Self: Moral, Political, and Legal Self-Relation*. Oxford: Oxford University Press.

Singer, Peter. 1972. "Famine, Affluence, and Morality." *Philosophy and Public Affairs* 1(3): 229–43.

Slote, Michael. 1984. "Morality and Self-Other Asymmetry." *Journal of Philosophy*81(4): 179–92.

Snedegar, Justin. 2013. "Reason Claims and Contrastivism about Reasons." *Philosophical Studies* 166(2): 231–42.

———. 2016. "Reasons, Oughts, and Requirements." In *Oxford Studies in Metaethics, Volume 11*, ed. R. Shafer-Landau. Oxford: Oxford University Press, 155–81.

———. 2017. *Contrastive Reasons*. Oxford: Oxford University Press.

———. 2021. "Reasons, Competition, and Latitude." In *Oxford Studies in Metaethics Volume 16*, ed. Russ Shafer-Landau. Oxford University Press, 134–56. https://doi.org/10.1093/oso/9780192897466.003.0006 (September 19, 2022).

Spencer, Jack. 2021. "The Procreative Asymmetry and the Impossibility of Elusive Permission." *Philosophical Studies* 178(11): 3819–42.

Taurek, John. 1977. "Should the Numbers Count?" *Philosophy and Public Affairs* 6(4): 293–316.

Tucker, Chris. Forthcoming. "Parity, Moral Options, and the Weights of Reasons." *Noûs*.

Waldron, Jeremy. 1981. "A Right to Do Wrong." *Ethics* 92(1): 21–39.