REJECTING SUPEREROGATIONISM

BY

CHRISTIAN TARSNEY

Abstract: Even if I think it very likely that some morally good act is supererogatory rather than obligatory, I may nonetheless be rationally required to perform that act. This claim follows from an apparently straightforward dominance argument, which parallels Jacob Ross's argument for 'rejecting' moral nihilism. These arguments face analogous pairs of objections that illustrate general challenges for dominance reasoning under normative uncertainty, but (I argue) these objections can be largely overcome. This has practical consequences for the ethics of philanthropy – in particular, it means that donors are often rationally required to maximize the positive impact of their donations.

1. Introduction

Suppose I am faced with a choice whether to spend $500 on a new TV or donate the money to GiveDirectly. I am sure that donating the money is either morally obligatory or supererogatory, but unsure which. Conversely, I am sure that buying the new TV is either morally prohibited or merely permissible, but unsure which.

The central thesis of this article is that, on most plausible ways of spelling out the details of this case, I am rationally required to donate my $500 to GiveDirectly. The argument for this claim is straightforward: because donating to charity is certainly at least as good as, and possibly better than, buying the TV, the former option statewise dominates the latter. Or, put more simply, buying the TV carries risk, namely the risk of violating a moral obligation, while donating to GiveDirectly carries no such risk – and indeed, as I will argue, no risk of acting contrary to one's all-things-considered reasons.[1]

Pacific Philosophical Quarterly •• (2018) ••–•• DOI: 10.1111/papq.12239

This argument borrows its pattern from an argument given by Jacob Ross (2006b). Ross argues that it is always rational to 'reject' (meaning, to ignore for purposes of practical deliberation) 'absolutely deflationary' theories like nihilism according to which it is never the case that one practical option is more choiceworthy than another, so long as one has positive credence in an ordinary moral theory like Kantianism or utilitarianism, since under moral uncertainty, dominance reasoning will always recommend the course of action preferred by the non-deflationary portion of one's credal distribution. So in §2 we will begin with an examination of Ross's argument, focusing on two objections that, as we will see, potentially generalize to other applications of dominance reasoning under moral uncertainty. §3 then turns to supererogation. The dominance argument for rejecting supererogationism succeeds only when a supererogatory option is at least as choiceworthy, all things considered, as its merely permissible alternatives, and in this section I argue both that many ordinary cases of supererogation seem to satisfy this condition and that the most plausible theoretical understandings of supererogation allow it to be satisfied. §4 then examines analogues of the two objections to Ross's anti-nihilistic argument that arise in the case of supererogation, and argues that these objections are on balance less threatening in the latter context. §5 draws out some practical implications of rejecting supererogationism for the ethics of philanthropy, arguing in particular that, in most cases where I judge that I am certainly at least permitted to donate a particular sum of money, I am rationally required to maximize the expected value of my donation. §6 is the conclusion.

2. *Moral nihilism*

Ross argues that we may almost always disregard what he calls 'absolutely deflationary' ethical theories according to which it's never the case that one course of action is more choiceworthy than another, whatever our credence in such theories.[2] He distinguishes, within this class, between 'uniform' theories according to which any pair of options is equally choiceworthy and 'nihilistic' theories according to which 'the notions of good and bad and of right and wrong are illusions and ... objectively speaking, no option or state of affairs is better than any other, nor are any two options or states of affairs equally good' (Ross, 2006b, p. 748).

Nihilism presents the more difficult and interesting case for Ross's dominance argument, so it is here that I will focus. Ross imagines a case in which he is nearly certain that moral nihilism is true and finds himself faced with a trolley dilemma. In such circumstances, he claims, he may simply disregard his credence in nihilism for purposes of practical deliberation. He supports this claim by the following argument.

[S]uppose...that I have a degree of credence of .01 in T_L [a non-nihilistic moral theory that prescribes turning the trolley to the left], but...I have a degree of credence of .99 in a nihilistic theory T_N. And again suppose that I must decide between sending the trolley to the right and sending it to the left. In this case we could reason as follows. According to T_L, it would be better for me to send the trolley to the left than to send it to the right. And so my credence in T_L gives me *pro tanto* subjective reason to send the trolley to the left. The only way this could fail to be the most rational option would be if my credence in T_N gave me a sufficiently strong reason to send the trolley to the right. But T_N implies that there would be nothing valuable or disvaluable about either alternative. And so my credence in T_N gives me no subjective reason to favor either alternative. Hence the *pro tanto* subjective reason to send the trolley to the left is unopposed, and so this is the rational option. (Ross, 2006b, p. 748)

This line of reasoning seems straightforward, but it faces at least two significant objections.

2.1. THE CYCLICITY PROBLEM

MacAskill (2013) offers the most careful analysis to date of dominance reasoning under moral uncertainty. In formalizing Ross's argument, he initially attributes to Ross the following principle.

Dominance over Theories (DoT) – If some theories in which you have credence give you subjective reason to choose x over y, and no theories in which you have credence give you subjective reason to choose y over x, then, rationally, you should choose x over y. (MacAskill, 2013, p. 511)

DoT seems plausible on face, but MacAskill shows that it is subject to an extremely powerful objection: for an agent who has positive credence in one or more non-nihilistic theories that posit *incomparability*, i.e. that generate an incomplete ranking of options, DoT can generate preference cycles. The demonstration of this result depends on a pair of complicated thought experiments. Stated schematically, and leaving aside the details MacAskill offers to illustrate the schema, the simpler of the two thought experiments goes like this: Agent A has three options O, P, and Q, and divides her moral beliefs among three theories, T_1, T_2, and T_3. T_1 implies that O is better than P and that Q is incomparable to both. T_2 implies that P is better than Q and that O is incomparable to both. And T_3 implies that Q is better than O and that P is incomparable to both. Hence, by DoT, $O \succ P \succ Q \succ O$ (MacAskill, 2013, pp. 513–4).

To avoid cyclicity, then, we must limit ourselves to dominance principles whose conditions of application exclude incomparability. In particular, MacAskill replaces DoT with what he calls

Genuine Dominance over Theories (GDoT) – If some theories in which you have credence give you subjective reason to choose x over y, and all other theories in which you have credence give

you equal subjective reason to choose x as to choose y, then, rationally, you should choose x over y. (MacAskill, 2013, p. 518)

But this weaker principle is bad news for Ross's argument, since nihilistic theories, as Ross has characterized them, do not claim that we have equal subjective reason to choose either option from any pair, but deny that notions like value, choiceworthiness, and subjective reasons are meaningful and hence that any comparisons in terms of these notions, even comparisons of equality, can meaningfully be made. Thus, MacAskill concludes that dominance reasoning cannot justify rejecting nihilism.

It seems to me, however, that there is a relatively straightforward escape from the cyclicity problem, and that Ross has simply made a very inconvenient mistake by treating equality as a *positive* value relation, such that nihilistic theories deny that any two things are equally good. Consider by contrast John Broome's proposed definition of equality with respect to a property F: '"x is equally as F as y" means that [i] x is not Fer than y, and [ii] y is not Fer than x, and [iii] anything that is Fer than y is also Fer than x, and [iv] y is Fer than anything x is Fer than' (Broome, 1997, p. 72). If nihilism is true, then all four clauses in Broome's definition are trivially satisfied for any x and y and any evaluative property F (e.g. 'good,' 'right,' 'choiceworthy,' 'supported by objective/subjective reasons'): if nothing is better than anything else, then x is not better than y, y is not better than x, and since neither x nor y is better than anything, it is vacuously true that for anything either x or y is better than, the other is better as well. Furthermore, by virtue of these last two clauses, Broome's definition distinguishes (as Broome intends it to) between equality and other relations like parity and incomparability in the context of non-nihilistic theories.

Thus, we can strengthen MacAskill's GDoT by simply replacing the positive relation of equality, which nihilistic theories decline to attribute to any pair of options, with Broome's negative relation. (Let's call the latter 'equality*,' for ease of reference.)

Strengthened Genuine Dominance over Theories (GDoT*) – If some theories in which you have credence give you subjective reason to choose x over y, and all other theories in which you have credence give you equal* subjective reason to choose x as to choose y, then, rationally, you should choose x over y.

GDoT* allows Ross's anti-nihilistic argument to go through, while remaining immune to the preference cycles brought on by the incautious principle DoT: in MacAskill's case described above, T_1 does not imply that there is equal* subjective reason to choose O as to choose Q, since there is greater subjective reason to choose O than to choose P but not greater subjective reason to choose Q than to choose P, nor is there equal* subjective reason

to choose P as to choose Q, by precisely the same reasoning (and likewise, *mutatis mutandis*, for T_2 and T_3).

Thus, the dominance argument for rejecting nihilism can escape the threat of cyclicity. However, as we will see in §4.1, MacAskill's problem does reappear as a challenge to the dominance argument for rejecting supererogationism, somewhat constraining the scope of that argument.

2.2. NIHILISTIC UNDERMINING

There is another difficulty that strikes me as more worrisome for Ross's argument and which has yet to receive substantial attention in the literature. This is the worry that principles like DoT, GDoT*, and any other dominance principle to which Ross's argument might appeal seem to be *undermined* by the very same nihilistic hypotheses that the argument is meant to justify rejecting – or at least, by some of those hypotheses.

Consider a form of nihilism I will call *Democritean normative error theory* (DNET). According to this view, only fundamental entities exist (flatworldism[3]) and all fundamental entities are physical (physicalism). 'Reasons,' 'obligations,' 'goods,' 'values,' etc., are not fundamental physical entities, so no reasons, obligations, goods, values, etc. exist. For any normative/evaluative proposition to be true, at least one such entity must exist to serve as its truthmaker. Therefore, DNET holds, no normative/evaluative propositions are true.[4]

Is such an extreme view plausible? Though I don't assign it the greater part of my own credence, it seems to me that DNET is a more reasonable contender for truth than many popular normative views (I leave it to the reader to fill in examples), and my impression is that a significant number of philosophers accept something like it (though, by self-selection, few of these are normative ethicists). It combines two widely held views in physicalism and flatworldism, and draws a plausible implication from the conjunction of these views.[5] I will assume, then, that DNET is plausible enough, as nihilistic theories go, to merit some investigation.[6]

Now, Ross's argument suggests that no matter how high my confidence in a nihilistic theory like DNET, I should nevertheless accept a second-order dominance principle like GDoT*.[7] The problem is this: since GDoT* expresses an affirmative normative proposition (a 'should' claim), DNET $\vdash \neg$GDoT*. So if I accept DNET with subjective probability p, I must assign GDoT* a degree of belief no greater than $(1 - p)$. If $p > .5$, then C(GDoT*) $< .5$. And, assuming that my beliefs ought to be consistent, if I *believe* DNET, I ought to disbelieve GDoT*.[8] But how can it be rational, let alone rationally required, to act on a principle that I regard as more likely false than true, or that I believe to be false?

It seems to me that the solution to this undermining worry, if there is one, must involve taking the relevant principle of rational requirement – a

dominance principle like GDoT* or some stronger principle from which it can be derived – to have 'external' normative force, in the sense given by Weatherson (2014). An external norm is one to which an agent is subject regardless of her beliefs and evidence, i.e. even if she justifiably rejects that norm. Weatherson argues that unless at least some normative principle has this sort of external force, we are faced with a vicious regress: if any norm is only incumbent on me to the extent that I believe it, then whenever I am uncertain of the correctness of some norm, I must find some higher-order norm that tells me how to act in light of my uncertainty. But since, plausibly, a rational agent will be at least a little uncertain about *any* proposed norm, including higher-order norms of choice under normative uncertainty, this commits her to considering an infinite hierarchy of higher-order norms in order to arrive at any rationally guided decision (Weatherson, 2014, pp. 155–6).

The escape from this regress, plausibly, is to suggest that there is at least one norm on which it is rational – indeed, rationally required – for an agent to act, regardless of her beliefs with respect to that norm. One candidate for such a regress-stopping, external norm is the requirement that agents choose options with maximal expected choiceworthiness.[9] Since maximizing expected choiceworthiness implies GDoT* as a limiting case, this would allow the dominance argument for rejecting nihilism to avoid the threat of undermining: it would still be rational for an agent to act as GDoT* prescribes, even if she rationally disbelieves GDoT*.[10]

Whether we should be comfortable with the notion of belief-independent rational requirements, and what the content of those requirements might be, is too large a question to resolve here. But we will return to the challenge of undermining when it re-arises in the context of the dominance argument for rejecting supererogationism. There I will argue that the versions of supererogationism that threaten to undermine the relevant dominance principles are less plausible than the corresponding versions of nihilism, and hence that the issue of normative externalism is at least less urgent in this context.

3. *Supererogation*

Let's turn, then, to the case of supererogation. In the opening section of the article, I suggested the following: When an agent A must choose between options O and P, where O is certainly either obligatory or supererogatory and P is certainly either forbidden or merely-permissible, she is rationally required to choose O over P, despite having some (perhaps quite high) degree of belief that P is objectively permissible. (As we will see, this simple claim will have to be sharpened and qualified before it takes its final form, given at the end of §4.) Paralleling – or rather, parroting – Ross, an initial

argument for this claim can be put like this: Call the moral theory according to which O is objectively morally required T_R and the theory according to which it is merely supererogatory T_S. According to T_R, A is morally required to choose O and forbidden to choose P, so A's credence in T_R certainly seems to give her *pro tanto* subjective reason to choose O.[11] The only way this could fail to be the most rational option would be if A's credence in T_S gave her sufficiently strong reason to choose P instead of O. But although T_S implies that it is *permissible* to choose P, it does not imply that it would be *better* to choose P than to choose O. At most, it implies that the two options are equally choiceworthy. Hence the *pro tanto* subjective reason to choose O is left unopposed, and so this is the rational option.

3.1. 'ALL-THINGS-CONSIDERED' SUPEREROGATION

This argument crucially assumes that, according to the supererogationist theory T_S, A has equal or greater all-things-considered reason to choose the supererogatory option O as to choose the merely-permissible option P. Let's call option O_S *morally* supererogatory iff for some alternative option O_P, O_S is morally better than O_P, but O_P is morally permissible; and let's call O_S *all-things-considered* supererogatory iff for some alternative option O_P, O_S is better than O_P all-things-considered (that is, supported by stronger all-things-considered reasons) but O_P is still all-things-considered permissible (that is, permissible as a matter of objective rationality). In these terms, then, the dominance argument for rejecting supererogationism assumes that we are dealing with a case of all-things-considered supererogation. The argument will be practically interesting, therefore, only if plausible versions of supererogationism do in fact regard many, most, or all ordinary cases of supererogatory behavior as cases of all-things-considered, and not merely moral, supererogation.[12]

This claim is *prima facie* plausible when we reflect on paradigm cases of supererogation, like donating a moderately substantial portion of one's income to charity. Even if we regard such actions as merely supererogatory, we do not regard an agent who performs them as acting irrationality, contrary to the all-things-considered balance of reasons. We can of course describe cases where we might be inclined to say that, although an agent has acted well from the moral point of view, she has acted on comparatively weak moral reasons and against stronger prudential reasons, and therefore has acted irrationally overall: think of Ned and Maude Flanders sending a once-disheveled stranger off in Ned's best suit, with the promise to sleep on card tables if he ever wants the use of their master bedroom.[13] But such cases go well beyond ordinary acts of supererogatory altruism, and our attitude toward them is quite different from our attitude toward, say, someone who donates five percent of her weekly paycheck to GiveDirectly. Unlike the absurd self-abnegation of the Flanders family, common sense regards

these more reasonable forms of altruism as admirable and praiseworthy, not just from a moral point of view, but *simpliciter*.[14] Though many of us judge that even moderate altruism (e.g. at the 5% level) is supererogatory rather than obligatory, it is hard to imagine that any plausible theory of all-things-considered objective reasons (so long as it incorporates non-egoistic moral considerations at all) will judge that we have less overall reason for making such sacrifices, when they will greatly improve the lives of the very badly off, than we have for spending the money on ourselves.

One might think, however, that this must be a mischaracterization of our commonsense judgments, since the notion of all-things-considered supererogation is self-evidently and intolerably paradoxical: if rationality always requires us to choose the option that is supported by the strongest all-things-considered reasons, then the definition of all-things-considered supererogatory options as more strongly supported by reasons and yet rationally optional is a simple contradiction in terms, and so cannot be what is posited (even in some cases) by plausible supererogationist moral theories.

There are a great many proposals in the literature that aim to resolve this 'paradox of supererogation' and render the notion of the supererogatory coherent with a general theory of rationality. Fortunately, most of these proposals are consistent with if not actively supportive of the commonsense judgment that ordinary supererogatory acts are better and more admirable than their merely-permissible alternatives, not just from the moral point of view, but *simpliciter*. Portmore (2008), for instance, considers five possible ways of understanding supererogation as part of a general theory of rational options, several of which allow us to make sense of the idea of all-things-considered supererogation. Specifically, we may understand all-things-considered supererogation in terms of a *satisficing conception of rationality* (that treats the merely-permissible alternative to a supererogatory act as rationally permissible because it meets some satisficing threshold, even though there is stronger overall reason to perform the supererogatory act); in terms of a distinction between the *justifying strength* and *requiring strength* of reasons (according to which the reasons that make the supererogatory option better supported by reasons overall nevertheless lack the requiring strength to produce a rational (or moral) requirement)[15]; or in terms of *imperfect reasons* (which, in supporting the supererogatory act, make it at least as rational all-things-considered as any alternative, but do not generate a rational requirement insofar as they can be satisfied by the agent performing other morally good acts in other circumstances).[16] In a similar vein, Raz (1975) proposes to understand supererogation in terms of what he calls 'exclusionary permissions,' which empower an agent to set aside or exclude the reasons that, on balance, favor a particular course of action.[17]

Finally, Urmson's classic article in defense of supererogation (Urmson, 1958) proposes that supererogatory acts be understood as differing from acts of moral obligation simply in that the difficulty of complying with the more

demanding principles of supererogatory moral behavior means that requiring such behavior of one another, treating supererogatory acts as obligatory under our shared moral scheme, is likely to undermine the normative force of moral requirements in general to an extent that outweighs any potential social gains from compliance. On this sort of broadly pragmatic understanding of supererogation there is, likewise, no difficulty in the idea that many if not all supererogatory acts are at least as strongly supported by all-things-considered reasons as their merely-permissible alternatives.

To come at the point from the opposite direction, there are just three possible understandings of supererogation with which the argument in the preceding section is incompatible. For ordinary cases of a putatively supererogatory option O_S and a merely permissible alternative O_P, in claiming that the all-things-considered reasons to perform O_S are equal to or stronger than all-things-considered reasons to perform O_P, I have thereby denied that (a) there is greater reason to perform O_P than to perform O_S, that (b) the reasons favoring the two options are incomparable, or that (c) the reasons favoring the two options are on a par (a fourth relation of evaluative comparison, supplementing the traditional trichotomy of better/worse/equal, proposed and defended by Ruth Chang, e.g. in Chang (1997). (I must deny (b) and (c) as well as (a) in order to get by on a dominance principle that avoids MacAskill's cyclicity problem, as we will see in the next section.)

The first of these possibilities, that supererogatory acts are generally opposed by the balance of all-things-considered reasons, is strongly contradicted by our commonsense evaluative judgments, and has not found advocates in the philosophical literature.[18] The second possibility, that of incomparability, is likewise highly implausible, since an enormous number of our everyday moral judgments presuppose comparability between the sort of altruistic reasons that commonly favor supererogatory acts and the sort of prudential reasons that oppose them. For instance, the judgments on the one hand that one has stronger reason to save the drowning child in Singer's pond than to preserve one's new pair of shoes (Singer, 1972), and on the other hand that Ned Flanders acts unreasonably in sacrificing his every personal interest for the sake of whatever needy stranger lands on his doorstep, both indicate that moral and nonmoral reasons are far from incomparable.

The most plausible understanding of supererogation that my argument must exclude, then, is one that draws on the notion of parity. On such an understanding, the strength of all-things-considered reasons for choosing some supererogatory option O_S is neither greater than, nor less than, nor equal to the strength of all-things-considered reasons for choosing some alternative O_P, but a sufficient strengthening or weakening of the reasons on either side (for instance, an increase in the good that might be done for others by, or in the personal costs of, performing O_S) would generate a definite inequality, making one option more choiceworthy overall.

It seems to me, however, that this parity-based account of supererogation is still inconsistent with our ordinary attitudes toward supererogatory acts. Parity implies a kind of symmetry between options: Neither option is better than the other, and either option could come to be better by the addition of reasons in its favor, or the subtraction of reasons for its alternative. But our normal attitudes toward the agent who gives more of her income to charity than morality strictly requires do not reflect such a symmetry: our admiration for her willingness to forgo creature comforts for the sake of others in greater need is not mirrored by a deprecating judgment that she is somewhat soft-headed or imprudent, nor by an equal admiration for the steely-eyed prudence of her colleague who only ever gives the morally obligatory minimum. Again, we view the agent who acts supererogatorily as having acted well and admirably, not simply from one point of view among many, but *simpliciter*.

It seems to me, then, that the most plausible understandings of supererogation will not treat (typical) supererogatory options as on a par with, incomparable with, or less overall choiceworthy than their merely permissible alternatives. Rather, they will endorse all-things-considered supererogation, and regard most ordinary supererogatory options as more overall choiceworthy than their alternatives, though not all-things-considered required or obligatory. This gives the dominance argument for rejecting supererogationism at least *prima facie* plausibility.

4. *Rejecting supererogationism: cyclicity and undermining*

Revealingly, however, the anti-supererogationist application of dominance reasoning faces a pair of objections that closely parallel the objections to Ross's anti-nihilistic argument discussed in §2.

4.1. CYCLICITY AND SUPEREROGATION

MacAskill's cyclicity worry, remember, arises insofar as nihilistic theories are construed as treating any two options as *incomparable* in terms of objective choiceworthiness. An analogous worry arises for the anti-supererogationist argument insofar as supererogation is construed as a matter of either incomparability or parity between supererogatory and merely permissible options. Suppose, for instance, I have some credence in a theory according to which the option of donating $500 to GiveDirectly is neither more, nor less, nor equally as choiceworthy as the option of buying a new TV (i.e. the two options are either incomparable or on a par) and some credence in a theory according to which I am obligated to donate to GiveDirectly. Then the principles GDoT and GDoT* will be inapplicable and will not generate a dominance argument for donating to GiveDirectly.

The principle DoT will generate such an argument, but at the cost of preference cycles, as we saw in §2.

How much does a restriction to acyclic principles like GDoT* limit the anti-supererogationist argument? One the one hand, we have just seen reasons why the problematic theories of supererogation as incomparability or parity are implausible. On the other hand, to whatever extent these versions of supererogationism do seem plausible to a particular agent, there is no easy way of saving the dominance argument against supererogation as there is with nihilism. A normative theory that regards the options of buying a new TV and donating to GiveDirectly as all-things-considered incomparable does not treat these options as equally choiceworthy even in the 'negative' sense of equality described in §2.1, the sense in which nihilistic theories *do* necessarily regard any two options as equally choiceworthy. Thus, if an agent does have positive credence that these options are overall incomparable or on a par, dominance reasoning alone cannot generate a rational requirement to perform the morally safe option of giving to charity, on pain of inviting cyclicity.

4.2. UNDERMINING AND SUPEREROGATION

But just as the greatest worry for Ross's anti-nihilistic argument is that certain forms of nihilism will undermine the very dominance principles on which it depends, so the greatest worry about the anti-supererogationist argument is that certain forms of supererogationism will undermine those same principles – albeit more subtly than radical skeptical theories like DNET.

One way this might happen is if the same arguments that support the existence of first-order supererogation – in particular, arguments from demandingness – can be simply re-run as objections to dominance principles like GDoT*: that is, GDoT* should be rejected, just like maximizing utilitarianism, because the demands it places on moral agents are excessive. But this line of argument is unconvincing for at least two reasons. First, even if one thinks it plausible that principles of morality must conform to a requirement of undemandingness, it is much harder to see why principles of *rationality* like GDoT* should conform to such a requirement. Morality, perhaps, is a human practice that must be designed to accommodate human weakness and imperfection. But the canons of rationality derive from more abstract standards of consistency or coherence that are not obviously sensitive to such pragmatic concerns. Second, appeals to arguments for first-order supererogationism are question-begging, or something very much like it, when our goal is to formulate principles for choice under conditions of first-order normative uncertainty, where one is *inter alia* uncertain about the soundness of those very arguments. That is to say, the conditional claim that a 'second-order' theory of rational choice under moral uncertainty

should be rejected if it is highly demanding is much less plausible than the conditional claim that a 'first-order' moral theory should be rejected if it is highly demanding, because the former conditional would generate illicit 'upward dependence' of principles for choice under moral uncertainty on the very first-order moral questions we are uncertain about.

But there is another, more difficult way in which the undermining worry can arise. To see it, we must introduce a further distinction among conceptions of supererogation, at a somewhat higher level of generality than the distinctions discussed in the last section. Suppose we accept the existence of all-things-considered supererogation, i.e. of option-pairs O_S and O_P such that O_S is more all-things-considered choiceworthy than O_P, but O_P is nevertheless all-things-considered permissible. Presumably we would like to hold that an agent who is *certain* that O_S is supererogatory in this sense is rationally permitted to choose O_P (i.e. permitted as a matter of subjective as well as objective rationality). There are, it seems to me, two broad ways to make sense of this.

First, suppose one accepts the following two intuitively appealing principles: (i) an agent is always rationally required to do what she has most subjective reason to do and (ii) an agent's subjective reasons derive from her beliefs about her objective reasons – more specifically, that belief in an objective reason of a given strength gives rise to a subjective reason of proportionate strength. Then the most natural way of understanding all-things-considered supererogation is to suppose that, although there is more objective reason *available* to support the supererogatory option O_S, some of the reasons that favor O_S are of a sort that the agent has the power to set aside or eliminate: 'justifying' or 'enticing' reasons, imperfect reasons, or reasons that come along with exclusionary permissions. That is, the agent is rationally permitted to perform O_P, according to the supererogationist theory, because she has the power to alter the initial balance of objective reasons by setting aside certain optional reasons, and thereby to alter the balance of subjective reasons as well in favor of O_P.[19] Since this approach is exemplified by the idea of exclusionary permissions, let's call it the 'exclusionary conception of supererogation' (ECS).

The alternative, it seems to me, is to deny (i) and hold that supererogation arises because, in a certain class of cases, agents are rationally permitted to act against the balance of subjective reasons. That is, one is rationally permitted to choose O_P even though one has more subjective reason to choose O_S. Since this approach is exemplified by understandings of supererogation in terms of a satisficing conception of subjective rationality, let's call it the 'satisficing conception of supererogation' (SCS).

Just as DNET posed an undermining threat to the dominance argument for rejecting nihilism, so SCS poses an undermining threat to the dominance argument for rejecting supererogationism.[20] To make this threat concrete,

suppose that I have .6 credence in a version of SCS that implies the following principle.

Egoistic Permissions Principle (EPP) – An agent is always rationally permitted to choose the option that is prudentially best in expectation, out of those options in a given choice situation that violate no moral constraint (e.g. against killing the innocent, breaking a promise, etc.), even when there is greater all-things-considered subjective reason for her to choose some other option.

Just as in §2.2 we saw that DNET $\vdash \neg$GDoT*, so here EPP $\vdash \neg$GDoT*. For consider a situation in which I can decide to spend some money harmlessly on myself or donate that money to charity, and in which I divide my beliefs between a moral theory on which donating the money is obligatory and one on which donating the money is supererogatory. In this situation, GDoT* implies that I am rationally required to donate the money, while EPP implies that I am not. Just as with DNET, then, it would seem that if I accept EPP with subjective probability p, I must assign GDoT* a degree of belief less than or equal to $(1 - p)$; that if $p > .5$, then C(GDoT*) $< .5$, and that if I *believe* EPP, I must disbelieve GDoT*. So again we may ask: how can I be subject to the rational requirements of a principle, GDoT*, that I regard (justifiably, we may stipulate) as more likely false than true?

This undermining worry can be mitigated in several ways. First, of course, there remains the option of taking a principle like GDoT*, or some stronger principle from which it derives, to have belief-independent rational requiring force. Thus, we might say that, even if I justifiably believe that I am not rationally required to perform the potentially-obligatory option O because I have greater than .5 credence in a principle like EPP, I can nevertheless be so required.

But moreover, there is reason to think that the versions of SCS that undermine dominance principles like GDoT* are uniquely implausible, indeed more implausible than undermining forms of nihilism like DNET. It strikes me as a basic conceptual truth that I am rationally required to do what I have most subjective reason to do. To think otherwise is to multiply basic normative concepts beyond necessity, allowing the concept of rational requirement to float free of the concept of subjective reasons. And as I suggested above, the sort of demandingness argument that might support a satisficing conception of rationality in the context of potentially-supererogatory acts seems much more plausible in the domain of first-order normative ethics than in the domain of rationality.[21] It is much more plausible, then, to posit the existence of certain objective reasons that I have the power to set aside or exclude from consideration, such that they do not generate subjective reasons in the first place, than to hold that I can simply ignore the balance of subjective reasons I do have.

Nevertheless, SCS does pose an undermining threat to the dominance argument for rejecting supererogationism, for an agent who assigns the

majority of her credence to this version of supererogationism, and if we deny that a principle like GDoT* has belief-independent normative force. But what about ECS, the rival conception of supererogation according to which an agent can *alter* the initial balance of reasons so that she no longer has all-things-considered subjective reason to choose the supererogatory option? This seems to me the most plausible way of understanding supererogation, but if so, it raises an interesting complication for the dominance argument for rejecting supererogationism. It seems plausible to suppose, on this conception, that a fully rational agent exercises her power to alter the balance of reasons in favor of the merely permissible option iff she actually *chooses* the merely permissible option: that is to say, it seems at least irrational if not simply inconceivable for an agent who has the power to alter the balance of reasons in favor of a merely-permissible option to choose that option but fail to alter the balance of reasons in its favor, or to so alter the balance of reasons (setting certain moral reasons aside) but then choose the supererogatory option anyway. Indeed, by what means could an agent alter the balance of reasons to favor the merely-permissible option, other than by *making a rational choice of that option*?[22]

If this is so, then ECS treats cases of supererogation as exhibiting a kind of extreme act-state dependence: an agent A who is certain of ECS, and certain as she chooses between options O and P that O is supererogatory and P is merely-permissible, will be certain that whichever option she chooses will be the one favored by the balance of reasons. If she is exactly .6 confident that she will choose P, then she is exactly .6 confident that the balance of reasons will, in the end, turn out to favor P.[23] So, if A is not certain that she will choose O, then she is not certain that the balance of reasons at the time she makes her choice will not favor P, and so it looks as if dominance reasoning cannot serve to eliminate P from consideration.

But there is something intuitively wrong with this line of reasoning: for although A is not certain that her reasons to choose O will turn out to be equal to or greater than her reasons to choose P, she is certain that *if she chooses* O, then there will be equal or greater reason to choose O – indeed, she is specifically certain that there will be greater reason. If she is certain that her supererogationist theory is right, then she has the same certainty about P, i.e. that conditional on her choosing P, P is the option favored by the balance of reasons. But if she has some degree of belief in a more rigorous moral theory according to which O is obligatory, i.e. on which she lacks the power to alter the balance of reasons in favor of P, then she can no longer have this certainty that if she chooses P, P will be the option most supported by her reasons. This creates an asymmetry in favor of O that seems to make O the most rational option.

And indeed, a slight reframing of the dominance principle GDoT* that accounts for the problem of act-state dependence allows us to accommodate

this intuition. The fully precise way of expressing the idea of dominance reasoning over moral theories, it seems to me, is as follows.

Final Dominance over Theories (FDoT) – If (i) every theory in which an agent *A* has positive credence implies that, conditional on her choosing option *O*, she has equal* or greater subjective reason to choose *O* as to choose *P*, (ii) one or more theories in which she has positive credence imply that, conditional on her choosing *O*, she has greater subjective reason to choose *O* than to choose *P*, and (iii) one or more theories in which she has positive credence imply that, conditional on her choosing *P*, she has greater subjective reason to choose *O* than to choose *P*, then *A* is rationally prohibited from choosing *P*.

If some principle like FDoT is correct, that allows dominance reasoning to accommodate the sort of act-state dependence generated by exclusionary permissions, then the dominance strategy will succeed against ECS versions of supererogationism. To avoid the undermining problem, therefore, it is only required that the agent have less than .5 credence in the less plausible satisficing conception of supererogation, or that the relevant dominance principles have belief-independent normative force.[24]

Having defended the notion of all-things-considered supererogation and considered the cyclicity and undermining challenges in this new context, we can now state the final conclusion of the dominance argument for rejecting supererogationism as follows: For any agent *A*, if either (i) a dominance principle like FDoT has external normative force (and is thus incumbent on *A* regardless of her normative beliefs) or (ii) *A* has less than .5 credence in an SCS version of supererogationism that would undermine such a dominance principle, then whenever, given options *O* and *P*, *A* is certain that *O* is either obligatory or all-things-considered supererogatory (and, correspondingly, that *P* is either prohibited or merely permissible), but uncertain which, *A* is rationally required to choose *O* and prohibited from choosing *P*. The practical scope of this conclusion will depend mainly on how often an agent is justified is being certain that some practical option is either all-things-considered supererogatory or obligatory. I turn to this issue in the next section, and consider some important practical cases to which the anti-supererogationist argument is plausibly applicable.

5. *Practical upshots*

To the extent that the anti-supererogationist argument is effective, it sheds new light on some contested practical questions concerning the ethics of philanthropy. Suppose you are a typical member of the middle class in the developed world, and that on careful examination of the available evidence you judge that donating to well-chosen charities can save a life in expectation at a cost on the order of a few thousand dollars.[25] The upshot of a

dominance principle like FDoT seems to be that you are rationally required to donate a substantial portion of your income to such charities, even if you think that such donations are probably supererogatory. More specifically, as per FDoT, you are rationally required to give at least to the point at which you judge that, by giving more, you *might* no longer be acting for the best, all things considered. In contrast to Peter Singer's famous injunction to 'give to the point of marginal utility' (Singer, 1972), we might say that you are required (at least) to give 'to the point of all-things-considered uncertainty.' If you are not yet at this point, then you do not risk acting contrary to the overall balance of reasons by giving another dollar to charity. But if you keep the dollar, you do run such a risk, since have some credence that your all-things-considered objective reasons require you to donate. You may believe, with great confidence, that your donation is objectively supererogatory, but there is no (overall) harm done by performing a supererogatory action on the grounds that it might be obligatory, while there is (overall) harm done by refraining from an objectively obligatory action on the grounds that it might be supererogatory. The latter is a risk to be avoided, the former is not.

Giving to the point of all-things-considered uncertainty will generally be quite a bit less demanding than giving to the point of marginal utility. But just where is the point of all-things-considered uncertainty, for a typical well-off agent? Here intuitions will no doubt diverge. We can easily imagine cases where any reasonable agent would be at least a little uncertain whether she is acting for the best, all things considered, especially where there are competing moral considerations in play: moving your family into a hovel in order to squeeze more deworming money out of your Wall Street paycheck, for instance, might well be an overall mistake. Of course, this risk of wrongdoing against your family might be outweighed by the risk of wronging the least well off by not giving to the point of marginal utility, but it might not be – dominance reasoning alone cannot speak to this question. In any case, it should be readily conceded that there is a point *somewhere* before the point of marginal utility at which the possibility of stronger countervailing obligations or non-moral reasons comes into play, and dominance reasoning can go no further.

Still, common sense suggests that few First Worlders are presently in any danger of going past this point. Is there a significant risk that a typical American would violate a moral obligation, or otherwise act against the overall balance of reasons, by deciding that she and her family should live on $50,000 rather than $60,000 per year, if by this sacrifice they could save the equivalent of three lives every year in expectation? Even here, perhaps, intuitions may diverge. But I imagine no one will think that an agent in this position who donated, say, $1000 a year to charity would be acting for the worse all things considered. And so by the above reasoning, this much at least is rationally required.

More clearly, perhaps, the dominance argument holds implications for *where* one's charitable efforts should be directed. Suppose that a billionaire philanthropist every year makes a $1,000,000 donation to the local art museum. One year, an admirable friend suggests to her that this money would be better spent supporting some life-saving health initiative in the developing world. Here, I don't think reasonable intuitions can diverge. It seems simply obvious that, between the options of donating a million dollars to the art museum and donating a million dollars to the health initiative, there is more all-things-considered reason to choose the latter, and the former option can be permissible only if the latter is supererogatory. Put another way, it is extremely implausible that our imagined agent would violate a moral obligation or otherwise act for the worse by choosing to support the health initiative instead of the museum. In so doing, she might deprive the local museum-going public of the chance to feign enjoyment of one more Jackson Pollock canvas, but it is quite hard to believe that her obligations in this respect outweigh her obligations to the global poor.

Here, then, is a straightforward practical conclusion: when an agent is confident that it is all-things-considered permissible for her to devote certain resources (time, money, etc.) to philanthropic endeavors, and has no reason to think herself under any other obligation to support some specific philanthropic endeavor (e.g. by having *promised* the money to the art museum), the possible obligation to support *efficient* endeavors – those that do the most good per dollar spent – is enough to generate a rational requirement that she choose these efficient endeavors over their less efficient alternatives.[26]

Rejecting supererogationism may have practical consequences in other domains as well. For instance, an agent might be convinced that non-human animals have enough moral standing that vegetarianism is at least as choiceworthy, all things considered, as meat consumption, and yet be unsure whether vegetarianism is morally required or supererogatory. Likewise, a citizen of a democratic society might be certain that she has at least as much overall reason to vote as to stay home on election day, and yet be unsure whether voting is morally required or supererogatory. For such agents, the dominance argument implies that they are rationally required to choose the morally safe option (vegetarianism or voting, respectively).

A common theme among the consequences of rejecting supererogationism – favoring altruism and philanthropy in general, efficient philanthropy in particular, and increased regard for potential moral subjects like non-human animals – is their broadly utilitarian character. The effect of rejecting supererogationism is largely pro-utilitarian, I suspect, because utilitarian theories are the only plausible moral theories that are *totalizing*, in the sense of rendering every or nearly every choice a matter of moral obligation. Thus, wherever a potentially obligatory action fails to maximize utility, it is potentially forbidden, i.e. there is a possible contrary objective obligation to counterbalance it (even if this possible obligation is outweighed in the last analysis)

– so for any agent who has positive credence in classical utilitarianism, dominance reasoning cannot imply a rational requirement to choose any non-utility-maximizing course of action. By contrast, we often face choices where the only potential moral obligation one way or the other is that implied by a maximizing utilitarian theory, and in these cases dominance reasoning suggests that we are often rationally required to meet the utilitarian demand.

6. *Conclusion*

I have argued that, when an agent is uncertain whether some practical option is obligatory or all-things-considered supererogatory, even if she thinks it is most likely supererogatory, she may nevertheless be rationally required to choose that option. And I have argued that, according to the most plausible forms of supererogationism, ordinary cases of moral supererogation are also cases of all-things-considered supererogation. The major qualification that must be added to these conclusions, as we have seen, is that the agent must not have credence greater than .5 in satisficing versions of supererogationism that endorse strongly permissive principles like EPP – that is, unless it turns out that the normative force of dominance principles like GDoT* or FDoT is simply independent of the agent's beliefs. But because principles like EPP are implausible – and less plausible, I think, than the versions of nihilism that threaten to undermine Ross's argument – it seems to me that the dominance argument for rejecting supererogationism is on a surer footing than the dominance argument for rejecting nihilism.

The practical scope of the anti-supererogationist argument depends mainly on how often we can be *certain* that some potentially-supererogatory option is at least as objectively choiceworthy as its less morally attractive alternative, e.g., that donating to charity is at least as choiceworthy as spending the same money on luxury goods for myself. If one is disposed to think that we should assign positive credence, however slight, to every or nearly every proposition, then of course dominance principles will have no scope at all for application. But this limitation of dominance reasoning is no more operative in the context of moral uncertainty than in the context of empirical uncertainty – if anything, less so, since it seems more plausible that we should always have positive credence in all empirical propositions than that we should always have positive credence in all moral propositions. Dominance principles in their native, game-theoretic context depend for their relevance on an idealized assumption of partial certainty that rarely if ever describes our real choices.[27]

In the moral as in the empirical domain, then, statewise dominance principles might be best understood as idealized limiting cases of some stronger principle, like first-order stochastic dominance or expected value maximization. That is, if I have some substantial positive credence that O is more

choiceworthy than *P*, then as my credence that *P* is more choiceworthy than *O* approaches zero, it becomes increasingly certain that *O* has higher expected value and is the more rational choice. Statewise dominance reasoning, one might think, is just a useful simplification of this sort of situation.[28] If this way of thinking about dominance reasoning is correct, however, then the practical relevance of dominance arguments in the context of moral uncertainty will depend on the success of some more ambitious and general (e.g. expectational) theory of rational choice under moral uncertainty.[29]

This is just one of several relevant open problems that I have left unaddressed, or only partially addressed, in this article. What is the strongest, fully precisified principle of dominance reasoning under normative uncertainty (e.g., how should we individuate 'normative theories' for purposes of dominance reasoning?), and what (if any) more general theory of decision-making under normative uncertainty does this principle derive from? Does either the dominance principle or the more general theory have belief-independent normative force, or is it incumbent on agents only to the extent that they believe it? And how will the true account of decision-making under normative uncertainty interact with each of the various, structurally diverse first-order normative theories that have proposed to accommodate supererogation (and other related forms of normative permissiveness, e.g. the 'small improvements' cases that motivate the literature on parity)? But although there is a great deal more to be said, I hope to have at least drawn attention to a previously neglected context in which considerations of normative uncertainty may prove practically significant. Even if I incline toward a normative worldview that gives me broad permission to pursue my own interests or projects at the expense of the greater good, I should be conscious of the possibility that I am subject to more stringent moral obligations, and of the risk that I violate my obligations by acting in ways that strike me as probably permissible. If we care about doing the right thing, such possibilities demand our attention.[30]

Faculty of Philosophy
University of Groningen

NOTES

[1] These are of course potentially distinct lines of argument, though closely related since to say that option *O* dominates an alternative *P* is just to say that there is some risk of acting for the worse by choosing *P* but no such risk in choosing *O*. My focus in this article, however, will be on the first formulation, the argument from dominance. Whether there is, say, a principle to the effect that *ceteris paribus* one ought not risk violating a moral obligation, that might stand even if the relevant dominance principles fall, is a question I will not take up.

[2] Throughout the article, I will use 'choiceworthy' just to mean 'supported by objective reasons,' such that to call one option more choiceworthy than another is to say that it is supported by stronger objective reasons.

[3] I borrow this term from Karen Bennett (2011).

[4] The name I have given this view is not meant to imply an attribution to Democritus. But the spirit of the view is neatly captured by his famous adage: 'By convention sweet and by convention bitter, by convention hot, by convention cold, by convention color; but in reality atoms and void' (Taylor, 1999). The DNETist just adds to this list: 'by convention right, by convention wrong, by convention rational, by convention irrational...but still, just atoms and void.'

[5] Of course, even if one accepts physicalism and flatworldism, one might still try to make normative propositions true by letting atoms and void alone serve as their truthmakers. This could be done by paraphrasing away existential claims concerning reasons and the like ('atoms arranged reason-wise,' perhaps), by eschewing such talk entirely, or by defending it as ineliminable but ontologically non-committing. Since my goal is not to defend DNET, I won't attempt to assess these strategies individually. I do, however, find them at least mildly implausible. Suffice it to say that normativity in the most interesting sense (features of the world that *count in favor of* actions and make our actions *matter*) must be something that carves the world at a fairly fundamental joint, and analyses of normative concepts that make them *logically* or *conceptually* (rather than metaphysically) supervenient on atoms and void appear to carve no such joint.

[6] In a footnote, Ross attempts to avoid the undermining worry by fiat, restricting the scope of his argument to 'a kind of nihilism that denies that there are any objective reasons for action but that concedes that if one held that there were such reasons, then it would be subjectively rational to act in accordance with them' (Ross, 2006b, p. 749n). In fact, Ross's stipulation must go a bit further than this, since his dominance argument is meant to apply not only to agents who *hold* (i.e. fully believe) that there are objective reasons for action but to any agent who *has positive degree of belief* that there are such reasons. Thus, the version of nihilism to which Ross restricts his attention must deny that there are any objective reasons for action but also *affirm that there are subjective reasons for action*, and that any agent with non-zero credence in the existence of objective reasons has subjective reasons. There is a significant *prima facie* tension, however, between denying the existence of all objective reasons and asserting the existence of subjective reasons, and Ross gives no argument that plausible nihilistic theories should be understood in this way. So it seems that this self-imposed restriction would limit the scope of Ross's argument to a relatively implausible subset of nihilistic theories.

[7] A variety of other dominance-like principles could be offered to facilitate Ross's argument. MacAskill considers a second principle, suggested by Ross, called 'Modified Dominance over Theories' (MDoT) that achieves the same result (MacAskill, 2013, p. 515). Weatherson considers a dominance-like principle he calls 'ProbWrong,' then suggests a schematic version called 'General Principle' that could be filled in by various evaluative predicates to facilitate the anti-nihilistic argument (Weatherson, 2014, p. 146). But the undermining objection described in this section applies equally and straightforwardly to all these principles.

[8] If GDoT* is read as a material conditional, this argument becomes slightly more complicated – GDoT* then is not itself an affirmative normative proposition, but implies such a proposition when conjoined with empirical propositions about my credence over theories that I presumably accept (and that I must accept, for GDoT* to apply to me and for Ross's argument to gain traction). But this does not affect the substance of the undermining problem: GDoT* is still inconsistent with the conjunction of empirical facts about my credal state (which make its antecedent true) and DNET (which makes its consequent false).

[9] This principle is proposed by Wedgwood (2013) as the 'fundamental principle of rationality' and endorsed by MacAskill (2014) in the context of decision-making under normative uncertainty. In ch. 7 of Tarsney (2017), I elaborate Weatherson's regress argument for externalism and argue that the best way of stopping the regress is to attribute external normative force either to the principle of maximizing expected choiceworthiness or to a weaker stochastic dominance principle. Weatherson himself prefers a view on which most or all first-order *moral* norms have external force, so that an agent's moral beliefs and uncertainties (at least about 'thin' moral

properties like *rightness*) are generally irrelevant to what she ought to do. But the regress problem on its own does not force us to adopt this extreme form of externalism.

[10] Ross himself has defended an externalist view of rational requirements (Ross, 2006a, pp. 273–7), as have Broome (2013) and Bykvist (2013), *inter alia*. If this externalist view is correct, and if the relevant dominance principles are among the belief-independent requirements of rationality, then Ross has the resources to avoid the undermining problem.

[11] I assume here, and throughout the article, that if A is morally obligated to choose O then she has most all-things-considered objective reason to choose O. This assumption is widely accepted, but not entirely uncontroversial – it is denied, for instance, by Foot (1972). For a recent defense, see Portmore (2011, pp. 38–51).

[12] Here and elsewhere I ignore the possibility of exact equality in the strength of reasons. If a supererogationist theory regards a supererogatory option O_S and its merely permissible alternative O_P as equally choiceworthy, all things considered, this would of course suffice for purposes of the dominance argument. But equality of reasons is, in general, extremely fragile, being disturbed by any increase or decrease in the strength of reasons favoring either option. Since it is fragile, equality is also extremely rare. Thus, while the dominance argument is compatible with the possibility that there is equal all-things-considered reason to perform O_S as to perform O_P, it will far more often be the case that there is at least slightly greater reason to perform O_S.

[13] *The Simpsons*, S3 E24, 'Brother, Can You Spare Two Dimes?'

[14] Likewise, it is the Flanders-like extremes of altruism that give intuitive plausibility to Susan Wolf's skepticism about 'moral saints' (Wolf, 1982). As Wolf argues, a life of moral perfection may be genuinely better from the moral point of view than a more ordinary life, but either worse or at least incomparable all-things-considered because of the nonmoral goods that must be sacrificed in the pursuit of moral perfection. But one can grant this conclusion without denying that *some* voluntary altruism in excess of the bare moral minimum may be good and praiseworthy from an all-things-considered (and not merely a moral) point of view.

[15] Horgan and Timmons, among others, have defended this understanding of supererogation: 'What the case of Olivia suggests, then, is the idea that not all good moral reasons for an agent to perform some action, even reasons that are plausibly considered "best," are such as to require that she perform that action, even prima facie. Some moral considerations clearly do have a requiring force, but (we submit) others need not' (Horgan and Timmons, 2010, p. 50).

[16] The other two possibilities Portmore considers, drawing on the ideas of parity and incomparability/rough comparability between reasons respectively, are incompatible with the argument I have given, as we will see below. But, as I will argue, they are also incompatible with our commonsense judgments of the merits of supererogatory acts.

[17] 'A person may have an exclusionary permission to perform an act even though there are conclusive reasons for him not to perform it, provided that he is entitled not to act for those reasons, to exclude them from his considerations. In other words: to say that a person is permitted to perform an act is to say that he may perform it, i.e. that he does nothing wrong in performing it. He is permitted to perform the act because there are no conclusive reasons against doing so or because he may exclude such reasons from his considerations' (Raz, 1975, p. 163).

[18] Portmore has come nearer than others to defending this view, but disclaims it in Portmore (2008, p. 382n). In rejecting this general view of supererogation, however, we need not deny that there are *some* acts we might intuitively describe as supererogatory that we have most all-things-considered reason not to perform: Think again of Ned Flanders-like extremes of altruism, and more generally of altruistic acts that harm the agent more than they help the beneficiary. (I thank an anonymous reviewer for this point.) Whether we are inclined to call these acts supererogatory is, I think, a verbal matter on which nothing of philosophical substance should turn. My central claim is that most *ordinary* cases of behavior we intuitively regard as morally supererogatory (e.g. donating a modest portion of one's paycheck to GiveDirectly) are also intuitively cases of all-things-considered supererogation.

If one did take the view that, in all or some of these ordinary cases, we have more all-things-considered reason to perform the merely permissible option rather than its supererogatory

alternative, one might still reach the same practical conclusions I defend in this article, so long as the balance of all-things-considered reasons against the supererogatory option is slight. This conclusion derives *prima facie* support, for instance, from Ross's (2006a) argument that we have *pro tanto* reason to reject 'unselective'/'top-heavy' normative theories that regard many options as almost-maximally good (Ross, 2006a, pp. 39–42). This might be one way of spelling out the argument from moral risk suggested in footnote 1.

[19] Alternatively, one might deny the objective-subjective linkage principle (ii), and hold instead that agents have the power, not to *eliminate* certain objective reasons, but rather simply to prevent them from generating subjective reasons. As far as I can see, everything I say about the view described in the main text will apply to this variant as well.

The idea that we can sometimes alter our reasons by an act of will is rendered plausible by other sorts of cases where this seems to happen: for instance, when an agent forms or abandons a plan or life project, she may be thought of as creating or eliminating reasons for herself. Supererogatory moral reasons, on this conception, might be thought of as reasons 'out there' in the environment that an agent can, but need not, adopt as her own. For defense of the claim that agents can have some voluntary control over their objective reasons, see the discussion of 'will-based reasons' in Chang (2013). (On Chang's view, we can sometimes *create* reasons for ourselves by an act of will, though in so doing we can never reverse a definite inequality in our 'given' reasons. But once we accept the general idea that agents can have some voluntary control over their objective reasons, I see no obvious reason why we shouldn't accept that we can sometimes exclude or eliminate non-will-based reasons by an act of will, so long as the implications of the resulting theory accord with common sense.)

[20] I thank Shelly Kagan for bringing this problem to my attention.

[21] Another sort of argument for a satisficing conception of rationality invokes choice situations in which there are infinitely many options, none of which is maximally choiceworthy (Landesman, 1995). But it is hard to see how a satisficing view motivated by cases of this kind could support a theory of supererogation.

[22] Thus, on the exclusionary conception of supererogation, an agent who performs a merely-permissible option instead of its supererogatory alternative does what she has most overall reason to do, having set aside her potential reasons in favor of the supererogatory option. This is what explains why an agent who is *certain* that *O* is supererogatory and *P* is merely permissible is rationally permitted to choose *P*. This raises a question, however: why, on this account, do we not feel that such an agent merits the same sort of all-things-considered praise and admiration as she would if she had acted supererogatorily? Perhaps because, although she does what she has most reason to do, she does not do what she had most *available* reason to do. But more importantly, we praise and admire agents who do what they have most reason to do *only when this involves some kind of difficulty*. Choosing to spend $500 on a new TV rather than donating to GiveDirectly does not require any unusual fortitude, willpower, or ability. So as in other cases where doing what one has most reason to do is easy (e.g. depositing a paycheck in my bank account rather than burning it), the merely-permissible choice merits neither praise nor blame.

[23] I assume that deliberation does not 'crowd out prediction,' i.e. that an agent can have credences concerning what her own choice will be even as she deliberates about that choice. For defense of this view, see for instance Joyce (2002).

[24] In a structural sense, therefore, versions of supererogationism on which non-moral reasons are stronger than or incomparable with moral reasons pose a more substantial threat to the dominance argument, since *any* positive credence in these theories will block the application of GDoT* or FDoT. This is symptomatic of a more general worry, that the application of any statewise dominance principle requires an unrealistic assumption of certainty and hence that dominance reasoning will always or nearly always be derailed by small credences in recalcitrant theories. I return to this worry in the concluding section.

[25] Figures in this range are well-supported by recent research. See for example GiveWell's estimate of the cost-effectiveness of anti-malarial bed net distributions (GiveWell, 2013).

[26] Thus, the dominance argument for rejecting supererogationism lends support to the practical program of *effective altruism*, according to which our philanthropic and other altruistic activities should aim not simply to do *some* good but to do the *most* good possible given the resources at our disposal (MacAskill, 2015).

[27] This draws attention to another point we have passed over, though I believe justifiably. The mundane reason why I might not be certain that donating to GiveDirectly is at least as choiceworthy as buying a new TV is that my reasons for donating to GiveDirectly depend on the consequences of that donation, and I should plausibly have at least some credence that those consequences will be net negative. I have been assuming so far that, for purposes of dominance reasoning over moral theories, these empirical uncertainties can be suppressed. That is, the dominance principles we have considered say roughly that if each *moral* theory to which I assign positive credence implies that my empirical-belief-relative reasons either favor option *O* over option *P* or are neutral between the two, then I should choose *O* over *P*. (This is accomplished, in the dominance principles considered by MacAskill and my own elaborations of them, by talking about what each normative theory implies about an agent's *subjective* reasons.) It seems permissible to reason in this way, rather than attempt the fruitless task of applying dominance principles over all maximal conjunctions of moral and non-moral possibilities, because particular normative theories will have their own internal methods for handling empirical uncertainty, which are often much stronger than dominance reasoning. Thus, from the standpoint of an agent deliberating about what to do given her *moral* uncertainties, she can simply take it as input, for instance, that utilitarianism assigns greater *expected* utility to *O* than to *P*, suppressing any further consideration of the empirical uncertainties that contribute to the utilitarian assessment.

[28] If we accept the understanding of statewise dominance as a limiting case of a more general decision theory, along with the justification for suppressing empirical uncertainty suggested in the previous footnote, then we can reformulate somewhat more strongly the conclusions reached in last two sections. The strengthened conclusion is this: If I have credence less than .5 in a satisficing version of supererogationism, or the relevant principles of rationality have belief-independent normative force, then in cases where (i) I am certain *or nearly certain* that the true normative theory implies that option *O* is either obligatory or all-things-considered supererogatory (in an *empirical-belief-relative* sense that allows, e.g., for substantial uncertainty about the consequences of doing *O*) and (ii) I have *substantial* credence that *O* is obligatory, I am *nearly always* rationally required to choose *O*.

The practical scope of this strengthened conclusion, then, depends mainly on how often an agent is justified in being *nearly* certain that some practical option is either all-things-considered supererogatory or obligatory, in the appropriate empirical-belief-relative sense – for instance, how often a typical resident of an affluent country can be *nearly* certain that, according to the true normative theory, she has at least as much empirical-belief-relative reason (before exclusions) to donate a given sum of money to a cost-effective charity like GiveDirectly rather than spending it on herself or on lower-impact philanthropic causes.

It seems clear to me that agents are often in at least this more modest sort of epistemic position. But some with whom I've discussed these arguments have expressed skepticism. It's hard to make much progress on this point without doing a lot of normative ethics, but it may be helpful, in light of the reformulated conclusion suggested above, to consider what a few of the most widely accepted normative theories imply about the cases of moderate altruism on which I have focused, like the choice between donating to GiveDirectly and buying a new TV. Utilitarianism and other standard consequentialist theories, of course, imply that I have most reason to support GiveDirectly. Kantianism holds that, in supporting GiveDirectly, I fulfill an imperfect duty of beneficence, and presumably in so doing do not act against my all-things-considered reasons. Other deontological theories that recognize duties of aid or beneficence will presumably say something similar. Virtue theories and commonsense morality are harder to pin down, but their implications can be gauged by our ordinary attitudes toward supererogatory acts and the agents who perform them, which I have argued consist of all-things-considered approbation. Commonsense views may hold (depending on whose common sense they are attempting to capture) that we are often

justified in giving extreme preference to friends and family, or to personally significant causes like the local arts scene, but this is because common sense regards the more cost-effective alternatives as all-things-considered supererogatory, not because it regards them as *worse* (or so our ordinary attitudes suggest). The only prominent normative theories according to which one's empirical-belief-relative reasons are *against* supporting GiveDirectly, as far as I can see, are egoistic theories – either ethical egoism or an egoistic theory of rationality according to which one only has reason to act altruistically insofar as it will contribute to the satisfaction of one's desires or other motivational states. I won't claim that no ordinary agent is ever justified in having substantial credence in these egoistic theories, but it seems to me that most ordinary agents do *not* have substantial credence in such theories, and are justified in so distributing their credence.

[29] Expectational theories of rational choice under moral or other forms of normative uncertainty face a number of special challenges, in particular the need to make cardinal comparisons between the value scales of rival normative theories. For discussion of this 'problem of intertheoretic value comparisons,' see Lockhart (2000); Ross (2006a); Sepielli (2009); MacAskill (2014); and Riedener (2015), *inter alia*. My own preferred approach to intertheoretic value comparison is outlined in Tarsney (2018), and described more fully in ch. 6 and Appendix B of Tarsney (2017).

[30] For helpful feedback on earlier versions of this article, I thank Samuel Kerstein, William MacAskill, Dan Moller, Christopher Morris, Caleb Pickard, Douglas Portmore, Julius Schönherr, Joshua Shepherd, several anonymous referees, and audiences at the Long Island Philosophical Society conference, the Rocky Mountain Ethics Congress, and the Oxford Applied Ethics Graduate Discussion Group.

REFERENCES

Bennett, K. (2011). 'By Our Bootstraps,' *Philosophical Perspectives* 25(1), pp. 27–41.

Broome, J. (1997). 'Is Incommensurability Vagueness?' in R. Chang (ed.) *Incommensurability, Incomparability, and Practical Reason*. Cambridge, MA: Harvard University Press, pp. 67–89.

Broome, J. (2013). *Rationality Through Reasoning*. Oxford: Wiley Blackwell.

Bykvist, K. (2013). 'Evaluative Uncertainty, Environmental Ethics, and Consequentialism,' in R. I. Hiller, Avram and L. Kahn (eds) *Consequentialism and Environmental Ethics*. Abingdon: Routledge, pp. 122–135.

Chang, R. (1997). 'Introduction,' in R. Chang (ed.) *Incommensurability, Incomparability, and Practical Reason*. Cambridge, MA: Harvard University Press, pp. 1–34.

Chang, R. (2013). 'Grounding Practical Normativity: Going Hybrid,' *Philosophical Studies* 164(1), pp. 163–187.

Foot, P. (1972). 'Morality as a System of Hypothetical Imperatives,' *The Philosophical Review* 81(3), pp. 305–316.

GiveWell (2013). 'Mass Distribution of Long-Lasting Insecticide-Treated Nets (LLINs)'. http://www.givewell.org/international/technical/programs/insecticide-treated-nets.

Horgan, T. and Timmons, M. (2010). 'Untying a Knot from the Inside Out: Reflections on the 'Paradox' of Supererogation,' *Social Philosophy and Policy* 27(2), pp. 29–63.

Joyce, J. M. (2002). 'Levi on Causal Decision Theory and the Possibility of Predicting One's Own Actions,' *Philosophical Studies* 110(1), pp. 69–102.

Landesman, C. (1995). 'When to Terminate a Charitable Trust?' *Analysis* 55(1), pp. 12–13.

Lockhart, T. (2000). *Moral Uncertainty and Its Consequences*. Oxford: Oxford University Press.

MacAskill, W. (2013). 'The Infectiousness of Nihilism,' *Ethics* 123(3), pp. 508–520.

MacAskill, W. (2014). *Normative Uncertainty*. DPhil thesis, University of Oxford.

MacAskill, W. (2015). *Doing Good Better: How Effective Altruism Can Help You Make a Difference*. New York: Gotham Books.
Portmore, D. W. (2008). 'Are Moral Reasons Morally Overriding?' *Ethical Theory and Moral Practice* 11(4), pp. 369–388.
Portmore, D. W. (2011). *Commonsense Consequentialism: Wherein Morality Meets Rationality*. New York: Oxford University Press.
Raz, J. (1975). 'Permissions and Supererogation,' *American Philosophical Quarterly* 12(2), pp. 161–168.
Riedener, S. (2015). *Maximising Expected Value Under Axiological Uncertainty: An Axiomatic Approach*. DPhil thesis, University of Oxford.
Ross, J. (2006a). *Acceptance and Practical Reason*. PhD thesis, Rutgers University-New Brunswick.
Ross, J. (2006b). 'Rejecting Ethical Deflationism,' *Ethics* 116(4), pp. 742–768.
Sepielli, A. (2009). 'What to Do When You Don't Know What to Do,' *Oxford Studies in Metaethics* 4, pp. 5–28.
Singer, P. (1972). 'Famine, Affluence, and Morality,' *Philosophy & Public Affairs* 1(3), pp. 229–243.
Tarsney, C. (2017). *Rationality and Moral Risk: A Moderate Defense of Hedging*. PhD thesis, University of Maryland, College Park.
Tarsney, C. (2018). 'Intertheoretic Value Comparison: A Modest Proposal,' *Journal of Moral Philosophy* 15(3), pp. 324–344.
Taylor, C. (1999). *The Atomists: Leucippus and Democritus. Fragments, A Text and Translation with Commentary*. Toronto: University of Toronto Press.
Urmson, J. O. (1958). 'Saints and Heroes,' in A. I. Melden (ed.) *Essays in Moral Philosophy*. Seattle, WA: University of Washington Press, pp. 198–216.
Weatherson, B. (2014). 'Running Risks Morally,' *Philosophical Studies* 167(1), pp. 141–163.
Wedgwood, R. (2013). 'Akrasia and Uncertainty,' *Organon F* 20(4), pp. 484–506.
Wolf, S. (1982). 'Moral Saints,' *The Journal of Philosophy* 79(8), pp. 419–439.