Supererogation and Conditional Obligation

Daniel Muñoz (UNC Chapel Hill) and Theron Pummer (St Andrews)

Forthcoming in *Philosophical Studies*

Abstract: There are plenty of classic paradoxes about conditional obligations, like the duty to be gentle if one is to murder, and about "supererogatory" deeds beyond the call of duty. But little has been said about the intersection of these topics. We develop the first general account of conditional supererogation, with the power to solve familiar puzzles as well as several that we introduce. Our account, moreover, flows from two familiar ideas: that conditionals restrict quantification and that supererogation emerges from a clash between justifying and requiring reasons.

1. Introduction

The point of morality—if it has one—is to guide our actions. Moral guidance usually comes in the form of obligations, which steer us away from wrongs like theft and murder. But wrongdoers need guidance, too. For their sake, morality issues *conditional obligations.* For example: if you are going to murder, you must do it gently. What makes this obligation "conditional" is that it applies given a certain condition—in this case, your murdering. What makes it "contrary-to-duty" is that the condition is a wrong action (Chisholm 1963). There are also what we call "consistent-with-duty" conditional obligations, like the obligation to wear a seatbelt if driving. These provide guidance to those who have ruled out a non-obligatory option, like taking the train.

The study of conditional obligations like these has been animated by a stock of paradoxes. The "gentle murder" paradox, for example, involves a tempting inference known as "factual detachment." *If* you are going to murder, *then* you must do it gently; *and*, in fact, you are going to murder. Does that mean you must commit a gentle murder? Clearly, this cannot follow (Forrester 1984; cf. Jackson 1985: 191-92; McNamara 2019). The classic task is to explain why.

Contemporary work on conditional obligations has been in fruitful dialogue with the flourishing study of *conditionals* (see e.g., Smith 1993; Bonevac 1998; McNamara 2010).[1] Also flourishing is the literature on *obligation*—both as a topic in itself, and especially in relation to "supererogatory" acts that lie beyond its call. But this work has not been put in contact with the inquiry into conditional moral judgments. Our task in this paper is to fill this lacuna, seeking a general account of conditional judgments fit for the supererogatory, and laying out some hard cases with which any such account must contend.[2]

We argue that every existing principle struggles with at least some of the hard cases: either consistent-with-duty conditional obligation (§2), consistent-with-duty conditional supererogation (§3), or contrary-to-duty conditional supererogation (§4). We then develop a principle that can handle them all. Our principle combines a familiar conception of conditionals (as restrictors on quantification) with a key resource from the theory of supererogation—namely, the idea that permissibility depends on the balance of "requiring reasons" and "justifying reasons" (§5).

2. From Obligation to Conditional Obligation

Let's start by thinking about conditional obligations in simple choices involving supererogation.

[1] We focus on conditionals whose antecedents, intuitively, restrict the options that an agent will consider in practical deliberation (e.g., "if you are not going to save both strangers, you must save one"), and whose consequents ascribe a deontic status to an option in their option set. We set aside "obligations," stated with anankastic conditionals, that merely express how to achieve a certain end (e.g., "if you want to get the job, you have to interview"); see Greenspan 1975; von Fintel & Iatridou 2005; Condoravdi & Lauer 2016. There is also a thriving literature on indicative conditionals whose antecedents provide information about the situation (e.g., "if the miners are trapped in Shaft A..."), and whose consequents involve obligations that are sensitive to that information; see Kolodny & MacFarlane 2010; von Fintel 2012: 22–30; Cariani et al. 2013; Charlow 2013. Another kind of conditional obligation, less central in moral theory, arises in what Schwager (2006: 242) calls a "relevance conditional," whose antecedent "filters out one of the conditions (typically relevance) under which the speech act arising from an utterance of the consequent in the given context would be appropriate." For example, "If I may be honest, you shouldn't quit your day job."

[2] Our work owes a debt to Paul McNamara, who has done more than anyone to develop a deontic logic for supererogation; see Mares & McNamara 1997; McNamara 1996, 2011a, 2011b. But his semantics in these papers, with its transitive ranking on the worlds permissible to bring about, is too restrictive for the puzzles in §§2–4. (For his latest thinking, see McNamara forthcoming.)

A supererogatory act, like a kindly favor or saintly sacrifice, is permissible and yet better than a permissible alternative—it goes "beyond the call of duty." The supererogatory contrasts with the "moral minimum," i.e., the minimally decent permissible option (McNamara 1996).[3]

We begin with a three-way choice: supererogating, doing the moral minimum, and doing wrong. Suppose that two people, both strangers to you, are trapped in a collapsing building, and while you can easily and costlessly save one of them, saving both would involve serious harm to you—say, losing your legs. You thus have three options: *Save Two* (supererogatory), *Save One* (less good but still permissible), and *Save Zero* (wrong). Clearly, you aren't obligated to *Save One*, since you may instead *Save Two*. But *if* you won't *Save Two*, *then* you must *Save One*. It would be wrong to gratuitously let the other stranger die.

We can capture this with a bridge principle between non-conditional and conditional obligations. (We write '**OB**($B/\neg A$)' to mean that B is obligatory conditional on not doing A.)

NOTHING ELSE LEFT: If A and B are your only permissible options, **OB**($B/\neg A$).[4]

The more general idea is that you are conditionally obligated to do something if, given the conditions, it is the only (non-conditionally) permissible option left.[5]

But as nice as it sounds, NOTHING ELSE LEFT appears to be open to counterexamples that

[3] We focus on "moral supererogation," which is morally better than the moral minimum, as opposed to "rational supererogation," which is rationally better than the rational minimum (see Benn & Bales 2019).

[4] Horton 2017: 96. He states his principle in terms of contrastive obligation (to do one act rather than another), but elsewhere (94) implies that it is equivalent to our formulation. (Note that he does not present NOTHING ELSE LEFT as a general account of conditional obligations.)

[5] When we say "permissible" (rather than "conditionally permissible") outside of the consequent of a conditional, we mean plain old *non-conditional* permissibility. *Mutatis mutandis* for "supererogatory," "obligatory," and "wrong." We prefer "non-conditional" to "unconditional" because "unconditionally permissible" sounds like it means "permissible no matter what," which is not the intended meaning.

feature certain consistent-with-duty conditionals (the principle is silent about contrary-to-duty conditionals). Consider a case known for giving rise to the "All or Nothing Problem" (Horton 2017).[6] We again have a collapsing building with two strangers trapped inside, but now you'll lose your legs whether you save one stranger or save both—saving zero remains costless. You have three options: *Save Zero* (no cost), *Save One* (costs you your legs), or *Save Two* (costs you your legs). In *this* collapsing building case you are plausibly permitted to *Save Zero*, as this is the only way to keep your legs. Moreover, it seems clear that *Save One* is wrong, because *Save Two* is far better and no costlier. Again, it's wrong to gratuitously let the other stranger die.[7]

Those are your obligations. What about your conditional obligations? Here is where the problem starts. If you won't *Save Two*, it seems you have only one permissible option left—*Save Zero*. But does it follow that, if you won't save everyone, you *must save none* (rather than saving one)? That seems perverse. This is the All or Nothing Problem, and its core is NOTHING ELSE LEFT. It is true that you have only two permissible options: saving "all" or saving "nothing," but it shouldn't follow that you must choose nothing if you won't choose all.[8]

3. Is "Next Best" Good Enough?

To solve the All or Nothing Problem, we will need a new kind of principle. We now turn to alternatives inspired by work on dyadic conditional obligations (Hansson 1969; Lewis 1973;

[6] Horton doesn't see this case as a counterexample to NOTHING ELSE LEFT; he just uses NOTHING ELSE LEFT to generate the All or Nothing Problem.

[7] For early discussions of cases like this one, see: Fried & Parfit 1979; Parfit 1982; Kagan 1989: 16. For more recent discussions, see: Wessels 2015; Pummer 2016; Pummer forthcoming; McMahan 2018; Sinclair 2018; Frowe 2019; Portmore 2019: §6.4; Bader 2019; Pummer 2019, Muñoz 2020; Rulli 2020.

[8] Some writers argue that saving zero is morally better than saving only one, which suggests that if you won't save two, you must indeed save zero. This is not our preferred solution to the All or Nothing Problem (for that see Pummer 2019; Muñoz 2020), but it is an important contender, and it should not be ruled out a priori. Later we will show how to express this kind of view—defended by Lazar and Barry ms and Tucker ms—within our framework (fn. 30).

Comesaña 2015). Here is the simplest. (We write '**PE**(B/¬A)' to mean that B is permissible conditional on not doing A; 'iff' means 'if and only if'.)[9]

NEXT BEST: **PE**(B/¬A) *iff* B is the best option compatible with ¬A.

In other words: if you won't do A, you may do the best option left on the menu besides A. (And you are conditionally obligated to choose from your conditionally permissible options.)[10]

A perk of NEXT BEST is that it can give advice to agents who will act contrary to duty. If you will murder, you must murder gently, because even though any murder is worse than no murder, gentle murder is the next best option (assuming your only options are murdering brutally, murdering gently, and refraining from murdering altogether). But NEXT BEST, like NOTHING ELSE LEFT, struggles with the consistent-with-duty permissions of the All or Nothing Problem. NEXT BEST implies that, if you will not *Save Two*, you must *Save One*. This seems too demanding. If there were just one person in the building, you would not have to sacrifice your legs to save them. This sacrifice would be supererogatory. So why should that same tradeoff be obligatory, conditional on not saving both strangers in the All or Nothing Problem? It ought to be conditionally optional.

[9] We use the dyadic conditional operator **PE**(_/_) without defining it in terms of a conditional and the monadic operator **PE**(_). We do not claim that such a definition is impossible; we just wish to abstract away from irrelevant controversies about, e.g., the proper syntax of 'if' (on which, see Kratzer 2012: Chapter 4). For simplicity, we talk as though the arguments of **PE**(_/_) as options, i.e., things that an agent could do, like *murder gently*. Officially, the arguments are propositions, like <Jack will murder gently>.

[10] We assume that obligation and permission are interdefinable: what you are obligated to do is to pick some or other permissible option, and what you are permitted to do is to pick an option consistent with your obligations. **PE**(A) *iff* ¬**OB**(¬A). This rules out dilemmas, but dilemmas are not our topic here (for discussion of contrary-to-duty scenarios and deontic dilemmas, see Kiesewetter 2018). Note also that on a dyadic interpretation of conditional obligations, the inference from '**OB**(A or B)' to '**OB**(A/¬B)' is invalid (Pummer 2019: 286). This inference, as well as various forms of deontic detachment, is valid on wide-scope interpretations—'**OB**(if not B, then A)' is equivalent to '**OB**(A or B)', where 'if' is read as the material conditional.

This leads Pummer (2019: 286) to offer the following revised principle, which like NOTHING ELSE LEFT involves a bridge from non-conditional to conditional deontic statuses:

> <u>OKAY OR NEXT BEST</u>: **PE**(*B*/¬*A*) *iff* (i) **PE**(*B*), or (ii) *B* is the best option compatible with ¬*A*.[11]

Think of it like this. Options that are non-conditionally permissible will remain conditionally permissible, but if an option is non-conditionally wrong only because it is worse than *A*, then it will be permissible conditional on not doing *A*.

OKAY OR NEXT BEST, like NOTHING ELSE LEFT, gets the right answer in the simple supererogation case. If you won't *Save Two* (cost: your legs), then you must *Save One* (cost: free!). That is the next best option, and the only one that's non-conditionally permissible.

Moreover, with OKAY OR NEXT BEST, we can solve the All or Nothing Problem. Conditional on not saving two, you may save *either* one or zero. Because *Save Zero* is non-conditionally permissible, it is also conditionally permissible. *Save One*, meanwhile, is conditionally permissible for a different reason: it is the next best option, given that you won't *Save Two*. OKAY OR NEXT BEST thus implies that saving the one stranger is an instance of conditional permission—indeed, of consistent-with-duty *conditional supererogation*.[12] In this sort of case, OKAY OR NEXT BEST performs better than NOTHING ELSE LEFT.

[11] '**PE**(*B*)', of course, means that *B* is non-conditionally permissible. If desired, we could define this as permission conditional on a tautology: **PE**(*B*) $=_{\text{df}}$ **PE**(*B*/*A* or ¬*A*).

[12] Note that this conditional permission (to save one stranger) is "contrary to duty" in one sense, but not in the sense that we have been using so far. The *condition* is not a wrong action; it is the permissible omission of a supererogatory action, namely, saving two. That said, the act of saving the one is *itself* wrong, non-conditionally. So the act is itself contrary-to-duty, even though the moral permission to do it is not conditioned on a transgression of duty.

You might wonder, "Who cares about conditional permissions? What kind of guidance do we get from hearing that an option is conditionally *permissible*?" Of course, we don't get the same direct guidance—"do this!"—that we would get from non-conditional obligations. But again, this guidance isn't for everyone. Wrongdoers, like the would-be gentle murderer, need the guidance of conditional obligations as a supplement. Non-supererogators, meanwhile, need conditional permissions as a *replacement* for inapt guidance that they might otherwise receive. At the outset of deliberation, morality tells you to avoid *Save One*—it's decisively worse than *Save Two*. But that advice is outdated as soon as you exclude *Save Two* from deliberation; morality should not be guiding you away from *Save One* towards *Save Zero*. We don't want a conditional requirement to save nobody if not everybody. If that's right, conditional permissions are crucial, as they preempt the disastrous imperative to move from second-best to very worst.[13]

4. The Hardest Case: Contrary to Duty, Beyond the Call

We have seen that there are counterexamples to NOTHING ELSE LEFT in cases of wrong but conditionally permissible actions. Are there any counterexamples to OKAY OR NEXT BEST? We believe further cases of conditional supererogation do indeed yield such counterexamples.

First consider Kamm's (1985) influential example from "Supererogation and Obligation." You have three options: you can *Keep Your Promise* to meet a friend for lunch (permissible), break your promise to *Do Nothing* at home (wrong), or break your promise to *Save One* stranger's life at great cost to yourself (supererogatory). *Save One* is the best option compatible with excluding *Keep Your Promise*. So, according to OKAY OR NEXT BEST, if you won't *Keep Your Promise*, you

[13] See Pummer 2019: 285. (Of course, this point will not be persuasive to those who think that *Save Zero* is better than *Save One*; see the discussion of Lazar and Barry in fn. 8, above.)

must *Save One*, which costs you your legs.[14] But this is not plausible, as Kamm herself observes.[15] After all, even though you can't justify *Do Nothing* over *Keep Your Promise*, since there is a negligible difference in costs to you, it is much costlier to *Save One*, and so it seems you may invoke the costs of heroism to justify your choice to *Do Nothing* instead.[16]

In Kamm's case (which has not been widely discussed in the context of conditional obligations), there is a conflict between supererogation and what would ordinarily be an obligation—viz., *Keep Your Promise*.[17] It is supererogatory to *Save One*, and indeed it should remain supererogatory even if you are not going to keep the promise. Life-saving is thus conditionally *and* non-conditionally supererogatory.[18]

So why does OKAY OR NEXT BEST struggle to get the right result in Kamm's case? Perhaps the problem is that the next best option if you won't *Keep Your Promise* (namely, *Save One*) is really a supererogatory option, which shouldn't be a duty even conditionally. This suggests a tweak:

[14] The same implication follows from NOTHING ELSE LEFT. If you do not keep the promise, there is only one permissible option left—saving the life.

[15] Kamm writes: "The view is also problematic if it implies (and perhaps it does not) that if I fail to do my particular duty in order to do a supererogatory act, I am then obliged to do that supererogatory act. Although the pressure on me may increase to do the supererogatory act, I do not believe that failing my lunch date will necessarily leave me with a duty to give up my kidney to save someone" (1996: 318). (In her example, saving the life involves a kidney transplant rather than a rescue mission.)

[16] Some will say that that, since you must avoid getting your hands dirty by doing nothing, you are obligated to *Save One* if you do not *Keep Your Promise*. We disagree, but again (see fns. 8 and 30), our framework is flexible enough to accommodate this intuition.

[17] Kamm's case is famous because it seems to involve a nontransitivity: you may do nothing rather than save the life (in a pairwise choice), you may save the life rather than keep your promise, but you may not do nothing rather than keep your promise. For discussion, see Archer 2016; Portmore 2003 (314-6), 2017; Muñoz 2020.

[18] Even if it were obligatory to keep the promise in Kamm's case, OKAY OR NEXT BEST would still wrongly imply that you must save the life if you won't keep the promise. Doing nothing isn't "okay," nor is it the next best option after promise-keeping. For a similar case to this interpretation of Kamm's case, see the "Two Buttons" example below.

> <u>Okay or Next Best*</u>: **PE**(*B*/¬*A*) *iff* (i) **PE**(*B*), (ii) *B* is the best option compatible with ¬*A*, or (iii) the only alternative to *B* (that is compatible with ¬*A*) is supererogatory.

The tweaked principle gets the right answer in Kamm's case: *Do Nothing* is permissible if you don't *Keep Your Promise*, because the only alternative—*Save One*—is supererogatory. (That is to say, it's non-conditionally supererogatory.)

But Kamm's case is not so easily dealt with. For one thing, the modified principle overgenerates conditional permissions. Suppose that you have three options: *Keep Your Promise* (permissible), *Save One* at great cost (supererogatory), or go on a *Murderous Rampage* (very wrong). Murder is impermissible, even conditional on not keeping the promise. But Okay or Next Best* implies that the rampage is conditionally permissible, since the only alternative left (*Save One*) is supererogatory. That is ridiculous.

What's more, both versions of Okay or Next Best seem to undergenerate conditional permissions. They fail to accommodate cases of *contrary-to-duty conditional supererogation*.

Suppose Alice is safe and Betty is in mortal danger.[19] There are two buttons before you: Buttons A and B. Pressing either button will seriously harm Alice without her consent, causing her to lose her legs. But if you *Press B*, that will have two more effects: *you* will also lose your legs, and Betty's life will be saved. While it's true that if you *Press B* you will save Betty's life, you are not required to do so. And pressing either button is wrong, given the harm caused to Alice. Now here is the key point. It is *not* true that, if you are going to harm Alice, then you must save Betty. You are not required to *Press B* conditional on pressing a button. Why not? Sacrificing your legs to save a stranger is paradigmatically supererogatory. It is also, effectively, what you are doing

[19] This case is much cleaner thanks to comments from Kerah Gordon-Solmon.

when you *Press B rather than Press A*. That is why we think *Press B* is *not* obligatory conditional on pressing a button. Instead, *Press B* is conditionally supererogatory. And since the condition is a wrong act (namely, button-pressing), we have an instance of contrary-to-duty conditional supererogation.[20] Even if one does not share these judgments, they are substantive judgments that are worth taking seriously; they should not be ruled out from the start by an account of conditional permissions and obligations.

This spells trouble for existing views of conditional obligation. OKAY OR NEXT BEST implies that, if you are going to press a button, you must *Press B*, since that is the best option besides not pressing anything. We get the same result, for the same reason, from OKAY OR NEXT BEST*. This seems extreme—a kind of fanatical moral offsetting. If you eschew the moral minimum, you may be obliged to make sacrifices that seem wildly disproportionate. (In Kamm's case, a broken lunch date compels you to sacrifice your legs. Better not skip dinner!)[21] Again, this view is substantive and controversial; it shouldn't trivially follow from our theory of conditional moral judgments.

There is an objection lurking.[22] Supererogatory acts are supposed to be good, but pressing button B is bad—how could a bad act be supererogatory, even conditionally? Our answer is that,

[20] In our Two Buttons example, the "contrary-to-duty" options all violate a deontological restriction—Alice's right against harm. This is not essential to the case. In another version, suppose that you can either do nothing, *Press C* to costlessly save 100 lives, or *Press D* to donate your kidney to save a stranger named Debbie. You aren't obligated to *Press D* if you won't *Press C*. (Deontological restrictions are also inessential, by the way, to Kamm's case, in which *Do Nothing* and *Save One* both break a promise. In another version, suppose that you are choosing between doing nothing, giving your legs to save a life, and costlessly saving another stranger's finger from being crushed. The only permissible options are *Save One* and *Save the Finger*, but conditional on not saving the finger, *Save One* is supererogatory, not obligatory.)

[21] There is also a worry about diachronic inconsistency. Consider the Two Buttons case. On the "offsetting" view, you have to sacrifice your legs to save Betty's life if you are going to harm Alice. But if the choice to harm and the choice to save take place at different times, the choice looks radically different. Harming Alice on Monday doesn't obligate you to heroically sacrifice for Betty on Tuesday. (This is especially clear if the harm to Alice is low and the cost of saving Betty is high.)

[22] We owe this objection to a helpful anonymous referee.

on the standard definition, supererogation is *not* always good in some absolute sense; it is just comparatively *better* than a permissible alternative (see Muñoz 2021b). Conditionally supererogatory acts, therefore, just need to be better than a *conditionally* permissible alternative, and *Press B* is indeed better than *Press A*.

In the All or Nothing Problem, it is wrong to save one stranger rather than both, even though saving one at great cost to yourself is better than saving zero. Likewise, in our Two Buttons case, pressing B (causing Alice to lose her legs, causing you to lose your legs, and saving Betty's life) is wrong even though it costs you greatly and is better than pressing A (simply causing Alice to lose her legs). Both actions are unambiguously wrong. Each is decisively ruled out by an alternative. And yet, each represents a remarkably good sacrifice in comparison to a third option. That is the phenomenon that we can capture with the concept of conditional supererogation. Judgments of conditional supererogation guide agents toward the best remaining options while still acknowledging that betterness may come at a serious cost, which might justify refraining. So far, we have not found any principle that can make sense of wrong acts that are conditionally supererogatory.

5. A Solution: Justifying and Requiring

Let's take stock.

We have raised problems for two principles of conditional obligation. The first, NOTHING ELSE LEFT, holds that we are conditionally obligated to pick from our remaining permissible options—if there are any. This principle fails in a case of consistent-with-duty conditional supererogation: the All or Nothing Problem, where saving one stranger is permissible, indeed supererogatory, conditional on not saving two. This problem can be solved with a second principle, OKAY OR NEXT BEST, which conditionally permits saving one because it is the best option left, if you won't save two. But this principle, even if modified, cannot handle contrary-to-duty conditional supererogation. Even if you wrongly do harm, you are not obligated to harm in the best possible way if the costs to you are disproportionate.

Why don't these principles work? NOTHING ELSE LEFT ignores *betterness*. For example, in the All or Nothing Problem, if you will not save both, NOTHING ELSE LEFT forbids *Save One* even though this is better than *Do Nothing*.[23] OKAY OR NEXT BEST, meanwhile, ignores *costs*. In the Two Buttons case, if you are going to press a button, OKAY OR NEXT BEST obligates you to *Press B*, even though the benefits of doing so (saving Betty) are not enough to outweigh the costs to you (losing your legs). If we are to make sense of these cases, we need a principle that is more flexible and powerful—something that directly factors in not only the justification we have to choose better options, but also the justification afforded by costs to the agent, whether those costs are suffered beyond the call (as in normal supererogation) or beneath it (as in conditional supererogation).

Thankfully, the key resources needed are already present in the literature on supererogation and normative reasons for action. In particular, we appeal to the distinction between requiring reasons and justifying reasons. (Sometimes this distinction it put in terms of a reason's "requiring strength" versus its "justifying strength.") A *requiring reason* tends to make actions obligatory. A *justifying reason* merely tends to make actions permissible.[24] To illustrate, suppose I can save

[23] There is also a more formal diagnosis of the trouble that NOTHING ELSE LEFT has in the All or Nothing Problem. The case, as described, violates a principle that Sen (2017: Chapter 1.6*) calls BETA. According to BETA, if *A* and *B* are both permissible, then adding more options cannot make *only one* of them wrong. But this is what seems to happen in the All or Nothing Problem: adding the permissible *Save Two* makes *Save One* wrong while leaving *Do Nothing* permissible. In such a case, NOTHING ELSE LEFT won't work. The only permissible options are *Save Two* and *Do Nothing*, and yet you needn't *Do Nothing* conditional on ¬*Save Two.* NOTHING ELSE LEFT has trouble with Kamm's case for the same reason: adding the permissible *Keep Your Promise* makes *Do Nothing* wrong while leaving *Save One* permissible.

[24] By "tends to make," we mean "contributes towards making," not "usually makes." Our distinction is basically drawn from Gert 2004, 2007 (for a related distinction, see Greenspan 2005; for a moral version,

Chico's life at the cost of my legs. I have a requiring reason to help Chico. If helping were costless, I would *have to* do it. But I don't actually have to help, since that would cost me my legs, and I have a powerful reason not to harm myself. This reason isn't itself a requiring reason. (More accurate: I have *more* justifying reason not to self-harm than I do requiring reason.) But my reason can still counterbalance the reason to help Chico, blocking a requirement to give aid.

We are now in a position to observe that an act is permissible *iff* the justifying reason in favor can outweigh the requiring reason to do otherwise.[25] More officially:

> J&R: **PE**(*B*) *iff* for any alternative *A*, the justifying reason to do *B* (rather than *A*) can outweigh the requiring reason to do *A* (rather than *B*).

"J&R" is short for (you guessed it) "justifying and requiring." We will assume that justifying and requiring reasons are somewhat, though not entirely, independent of each other. Any requiring reason doubles as a justifying reason, but not vice versa. There is always at least as much justifying reason to do an option as there is requiring reason. However, it is possible for the justifying reasons in favor of an option to outstrip the requiring reasons.[26]

see Portmore 2011). See also Hurka and Shubert (2012) on prima facie duties (which tend to favor and require) versus prima facie permissions (which tend to justify), as well as Muñoz (2020) on moral reasons versus prerogatives. We can take (merely) justifying reasons as a primitive (Hurka & Shubert 2012), or we could try to understand them in other terms. One idea is that they act as "disabling conditions" on requiring strength (see Dancy 2004 on disablers). For a more developed view, see Muñoz 2021a on moral defense.

[25] We are using 'reason' here as a mass noun rather than a count noun.

[26] Since this is a paper about *conditional* obligation, we do not aim to defend J&R at length over rival views of *non-conditional* obligation. But we should discuss one issue. We believe that an action's choiceworthiness—how strongly "favored" it is—depends only on how much *requiring reason* there is to do it. Now, some will object that purely justifying reasons are also favorers. But if that is so, it is hard to see how supererogation could be favored over the moral minimum. The justifying reason to keep one's kidney, if it can outweigh the requiring reason to donate, will also strongly favor selfishness (Hurka & Shubert 2012: 9). Another objection to our view is that some reasons are *purely commendatory*—they favor without justifying or requiring (Horgan & Timmons 2010; Archer 2016; Little & Macnamara 2017). But this idea seems to invite odd recombinations. Suppose we sweeten a wrong act with commendatory reasons

This is exactly what happens in cases of supererogation. We already saw this in my simple choice between keeping my legs or saving Chico. But the real payoff will be applying J&R to many-option cases. So consider our first supererogation case: you can *Save Zero* (no cost), *Save One* (no cost), or *Save Two* (at the cost of your legs). Here, it is permissible and best to save both strangers. This is what there is the most requiring reason to do. It is also permissible to save just one, it seems, because you have a weighty justifying reason to keep your legs. But it would be wrong to save no one. There is more requiring reason to *Save One* instead, and no justifying reason to compensate. *Save Zero* and *Save One* are both worse than *Save Two*, but only the latter is *justifiably* worse, in the sense that it is worse but still supported by enough justifying reason to keep it permissible. What makes an option wrong is being *unjustifiably* worse than an alternative, i.e., not being supported by enough justifying reason to make up for the deficit in requiring reason.

With J&R in place, we can offer our proposed principle of *conditional* permissibility.

> CONDITIONAL J&R: **PE**(B/$\neg A$) *iff* for any alternative C that is compatible with $\neg A$, the justifying reason to do B (rather than C) can outweigh the requiring reason to do C (rather than B).

To see if B is permissible conditional on $\neg A$, we need to know how B compares to the options that are still being considered. If B is the best remaining option, or tied for best, it is permissible. If B is worse than some C, we have to ask: is it unjustifiably worse? If so, then B is conditionally wrong. If B is justifiably worse, then it may still be conditionally permissible.

The core idea here is that we have a two-step process for determining permissibility

until it is the best option—even better than the permissible alternatives. Then the optional option will be wrong. This seems absurd. A virtue of our view is that it rules out such possibilities.

conditional on ¬*A*. First, remove *A* from the set of options. Second, take each option that remains, and ask whether it can be justified over each of the remaining alternatives; *iff* the answer is "yes," the option is conditionally permissible. The conditionally permissible options are the ones that are permissible to choose from the restricted set of options. What does it mean to say that we "remove" the option of doing *A*, when we condition on ¬*A?* One possibility is that we imagine a counterfactual scenario in which the restricted menu *is* the agent's entire option set, that is, we consider the case in which *A* is not available as an option at all. This is not how we are conceiving of things. Rather than considering a counterfactual scenario in which *A* is off the menu, we are holding fixed *A*'s presence on the menu, but ignoring it in that we are not counting options as impermissible simply because they are unjustifiably worse than *A*.[27] In this sense we *exclude* the option of doing *A* from consideration, although we still suppose that *A* is *available*: we are still talking about a scenario in which the agent has the ability to do *A*.[28]

With CONDITIONAL J&R laid out, we can next ask whether it gets the right answer in our hard cases. We believe it does.

First, unlike NOTHING ELSE LEFT, our principle solves the All or Nothing Problem. Recall your three options: costlessly save no one, save just one stranger at the cost of your legs, or save this same stranger and another at no greater cost to yourself. CONDITIONAL J&R has the plausible implication that saving just one is permissible conditional on not saving both, since saving one is

[27] For more on conditionals as domain restrictors, see Jackson (1985: 191–92) and Kratzer (1986).

[28] Here is an example to illustrate why it might be important to hold fixed the availability of an excluded option. Suppose you have three options: dress up and attend a costume party (*Costume*), go to the party without dressing up (*Casual*), or skip the party and stay at home (*Skip*). Since *Costume* is an option, it would be wrong to pick *Casual*; it expresses disrespect to the host (let's suppose). Intuitively, it is true here that you must not attend if you are not going to wear a *Costume*, i.e., if *Costume* is excluded. But if *Costume* were *unavailable*, there would be nothing disrespectful about *Casual*. The fact that one *could* dress up colors the choice between dressing casually and skipping out. That is why it is important not to imagine that *Costume* is unavailable when excluding it from consideration. (Our thanks to Joe Horton for this example and for many other helpful ideas.)

the best option compatible with not saving both. Moreover, saving zero is *also* permissible conditional on not saving both, since it is *justifiably* worse than saving one (which is the only other option left). This gives CONDITIONAL J&R an advantage over NOTHING ELSE LEFT.

Second, CONDITIONAL J&R does better than OKAY OR NEXT BEST in Kamm's case. Here your options are: *Keep Your Promise* (permissible), break your promise to *Do Nothing* (wrong), and break your promise to heroically *Save One* (supererogatory). According to OKAY OR NEXT BEST, if you do not *Keep Your Promise*, you are obligated to *Save One. This* seems awfully demanding. CONDITIONAL J&R issues no such demand. For staying home is *justifiably* worse than life-saving (despite being unjustifiably worse than promise-keeping).[29]

Finally, CONDITIONAL J&R is the first principle that can handle our Two Buttons case, where the options are: *Do Nothing* (permissible); *Press A,* harming Alice (wrong); and *Press B,* harming Alice and saving Betty at a big cost to you (also wrong). Conditional on pressing a button, it is intuitively optional, not obligatory, to *Press B*. Pressing B would be contrary-to-duty conditionally supererogatory. CONDITIONAL J&R has a neat explanation. Although it is worse to *Press A* than to *Press B*, pressing A is *justifiably* worse, given the cost to you of pressing B. *Press B* is thus conditionally supererogatory. It is conditionally permissible and better than a conditionally permissible alternative. Since the condition is a wrong act (namely, pressing a

[29] Because *Do Nothing* and *Keep Your Promise* are (roughly) equally costly, but much less costly than giving your legs to *Save One*, there is an interesting result. You have a powerful justifying reason to *Do Nothing* rather than *Save One*, but not to *Do Nothing* rather than *Keep Your Promise*. This is what Muñoz (2020) calls a "comparative prerogative," or what we might call a "contrastive justifying reason" (see Snedegar 2017 on contrastive reasons, but note that Snedegar's (forthcoming) treatment of supererogation does *not* commit to these reasons, or to any treatment of three-option cases like Kamm's and Horton's). Admittedly, "comparative prerogative" sounds like some heavy-duty jargon. But the idea behind it should be uncontroversial, and the term "comparative" is inessential. The point is just that *relative* costs are what determine justifying reasons (at least, the cost-based ones). Pointing to the costs of option A, even if they are steep, will not justify doing B in the slightest, if B's costs are just as big or bigger. (We would like to thank an anonymous referee, and Justin Snedegar, Brendan de Kenessey, Chris Tucker, and Benjamin Kiesewetter for helpful comments here.)

button), we have an instance of contrary-to-duty conditional supererogation.

That completes our argument for CONDITIONAL J&R. The view gives powerful explanations and plausible verdicts on cases beyond and beneath the call of duty. The view is also principled. It is not an ad hoc concoction, but the natural product of a view of conditionals and a view of supererogation: conditionals restrict quantification, and supererogation emerges from the clash between justifying and requiring reasons. We do not claim that CONDITIONAL J&R is the only principle to avoid embarrassment in the cases we have considered above (one could mimic the results of CONDITIONAL J&R without appealing to the distinction between justifying and requiring reasons, and without even appealing to reasons at all).[30] But, in order to keep things tidy, CONDITIONAL J&R is the only such principle we consider here. There may be even lovelier principles left to discover. But given the simplicity and popularity of J&R, we think CONDITIONAL J&R is a natural place to start looking for a theory of conditional supererogation.

6. Conclusion

In this paper, we presented a series of puzzles for any theory of moral conditionals that ventures beyond obligation into the realm of the supererogatory. We began by presenting cases featuring consistent-with-duty conditional obligation (§2), consistent-with-duty conditional supererogation (§3), and contrary-to-duty conditional supererogation (§4), arguing that no existing principle can capture plausible intuitions about at least three of our cases. We then presented CONDITIONAL J&R and showed how it can capture plausible intuitions about all the cases (§5). Moreover,

[30] If we assume there is a (decisive) requiring reason to avoid doing what's non-conditionally wrong (Darwall 2010), we can get CONDITIONAL J&R to mirror the implications of NOTHING ELSE LEFT. It would then imply, for instance, that if you are not going *Save Two*, you are obligated to *Save Zero* (in the case animating the All or Nothing Problem). As noted earlier, we do not find this to be a plausible implication, but some people do, and our account has the flexibility to accommodate it.

CONDITIONAL J&R has a principled rationale; it combines a familiar conception of conditionals (as restrictors on quantification) with the supererogationist's insight that permissibility depends on the balance of requiring reasons and justifying reasons. It would seem that our examples—our three "hard cases," and the more familiar cases of conditional obligation like the gentle murder paradox—for all their differences, can be given a surprisingly systematic treatment. Still, we do not claim that CONDITIONAL J&R is the only hope for understanding moral conditionals in the realm of supererogation. Our conclusion is more modest. We have found *one* way to solve the puzzles, though there may be more solutions, and indeed more puzzles, yet to be discovered. There may even be puzzles for our own view, and we may need to make revisions—but with any luck, they will be gentle.[31]

References

Archer, Alfred (2016). "Moral Obligation, Self-Interest, and the Transitivity Problem," *Utilitas*, 25, 355–82. doi: 10.1017/S095382081600009

Bader, Ralf (2019). "Agent-Relative Prerogatives and Suboptimal Beneficence," *Oxford Studies in Normative Ethics* 9: 223–250.

Benn, Claire and Bales, Adam (2019). "Rationally Supererogatory," *Mind*. Early Online. doi: 10.1093/mind/fzz055

Bonevac, Daniel (1998). "Against Conditional Obligation," *Noûs*, 32: 37–53.

Cariani, F., Kaufmann, M. & Kaufmann, S. (2013). "Deliberative Modality Under Epistemic

[31] We would like to thank two anonymous referees for their thoughtful and constructive comments. We also thank Ralf Bader, Joe Bowen, Roger Crisp, Brendan de Kenessey, Kerah Gordon-Solmon, Peter Graham, Brian Hedden, Joe Horton, Benjamin Kiesewetter, Paul McNamara, Per Milam, Tina Rulli, Justin Snedegar, Jack Spencer, Chris Tucker, and our audience at the All or Nothing Workshop Miniseries. (Special thanks to those who realized that, even if they weren't going to every talk in the series, they should still go to ours!)

Uncertainty," *Linguistics and Philosophy*, 36: 225–59. doi: 10.1007/s10988-013-9134-4

Charlow, Nate (2011). "What We Know and What to Do," *Synthese*, 190: 2291–2323. doi: 10.1007/s11229-011-9974-9

Chisholm, Roderick M. (1963). "Contrary-to-Duty Imperatives and Deontic Logic," *Analysis*, 24: 33–36.

Comesaña, Juan (2015). "Normative Requirements and Contrary-to-Duty Obligations," *Journal of Philosophy*, 112, 11: 600–26.

Condoravdi, Cleo and Lauer, Sven (2016). "Anankastic Conditionals Are Just Conditionals," *Semantics & Pragmatics*, 9: 1–69. doi: 10.3765/sp.9.8

Dancy, Jonathan (2004). *Ethics Without Principles*. Oxford: Oxford University Press.

Darwall, Stephen (2010). "But It Would Be Wrong," *Social Philosophy and Policy* 27, 135–57.

Forrester, James William (1984). "Gentle Murder, or the Adverbial Samaritan," *Journal of Philosophy*, 81: 193–96.

Fried, Charles and Parfit, Derek (1979). "Correspondence," *Philosophy and Public Affairs* 8, 393–397.

Frowe, Helen (2019). "If You'll Be My Bodyguard: Agreements to Save and the Duty to Minimize Harm," *Ethics* 129: 204–229.

Gert, J. (2004). *Brute Rationality*. Cambridge: Cambridge University Press.

———(2007). "Normative Strength and the Balance of Reasons," *Philosophical Review*, 116 (4): 533–62. doi: 10.1215/00318108-2007-013

Greenspan, P.S. (1975). "Conditional Oughts and Hypothetical Imperatives," *Journal of Philosophy*, 72, 10, 259–76.

———. (2005). "Asymmetrical Practical Reasons," in Reicher, M.E. and Marek, J.C. (Eds.), *Experience and Analysis*. Vienna: Öbv & Hpt, pp. 387–94.

Hansson, Bengt (1969). "An Analysis of some Deontic Logics," *Noûs* 3, 373–98.

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

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

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

Jackson, Frank (1985). "On the Semantics and Logic of Obligation," *Mind* XCIV, 374: 177–195.

Kagan, Shelly (1989). *The Limits of Morality* (Oxford: Clarendon Press).

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

———(1996). *Morality, Mortality: Volume II*. New York: Oxford University Press.

Kiesewetter, Benjamin (2018). "Contrary-to-Duty Scenarios, Deontic Dilemmas, and Transmission Principles," *Ethics* 129: 98–115.

Kolodny, Niko & MacFarlane, John (2010). "Ifs and Oughts," *Journal of Philosophy*, 107, 3, 115–43.

Kratzer, Angelika (1986). "Conditionals," in Anne M. Farley, Peter Farley and Karl E. McCollough (Eds.), *Papers from the Parasession on Pragmatics and Grammatical Theory*. Chicago: Chicago Linguistics Society, pp. 115–135.

Lazar, Seth and Barry, Christian (ms). "Supererogation and Optimization."

Lewis, David (1973). *Counterfactuals*. Cambridge, Mass: Harvard University Press.

Little, Margaret and Macamara, Coleen (2017). "For better or worse: commendatory reasons and latitude," *Oxford Studies in Normative Ethics* 7: 138–160.

Mares, Edwin and McNamara, Paul (1997). "Supererogation in Deontic Logic: Metatheory for

DWE and Some Close Neighbours," *Studia Logica* 59, 397–415.

Massoud, Amy (2016). "Moral Worth and Supererogation," *Ethics* 126, 690–710.

McMahan, Jeff (2018). "Doing Good and Doing the Best," in *The Ethics of Giving*, ed. Paul Woodruff (New York: Oxford University Press): 78–102.

McNamara, Paul (1996). "Must I Do What I Ought? (or Will the Least I Can Do Do?)", in M. Brown and J. Carmo (eds.), *Deontic Logic, Agency and Normative Systems*, pp. 154–73. New York: Springer.

———(2010). "A Bit More on Chisholm's Paradox." Supplement to McNamara 2019.

———(2019). "Deontic Logic," *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2019/entries/logic-deontic/>.

———(forthcoming). "A Natural Conditionalization of the DWE Framework," in M. Brown, A. Jones, and P. McNamara (eds.), *Festschrift for Risto Hilpinen*. Springer.

Muñoz, Daniel (2020). "Three Paradoxes of Supererogation," *Noûs* https://doi.org/10.1111/nous.12326.

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

———(2021b). "Infinite Options, Intransitive Value, and Supererogation," *Philosophical Studies* 178 (6): 2063–075.

Parfit, Derek (1982). "Future Generations: Further Problems," *Philosophy and Public Affairs* 11, 113–72.

Portmore, Douglas (2003). "Position-Relative Consequentialism, Agent-Centered Options, and Supererogation," *Ethics*, 113, 303-332.

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

———(2017). “Transitivity, moral latitude, and supererogation,” *Utilitas*, 29, 286-298. Doi: 10.1017/S0953820816000364

———(2019). *Opting for the Best: Oughts and Options* (Oxford: Oxford University Press).

Pummer, Theron (2016). “Whether and Where to Give,” *Philosophy and Public Affairs* 44, 77–95.

———(2019). “All or Nothing, But If Not All, Next Best or Nothing,” *Journal of Philosophy*, 116, 278–91.

———(forthcoming). “Impermissible yet Praiseworthy,” *Ethics*.

Rulli, Tina (2020). “Conditional Obligations,” *Social Theory and Practice*, 46, 365–90.

Schwager, Magdalena (2006). “Conditionalized Imperatives,” in C. Tancredi, M. Kanazawa, I. Imani, & K. Kusumoto (Eds.), *Proceedings of SALT XVI*. Ithaca, NY: CLC Publications (pp. 241–58).

Sen, Amartya (2017). *Collective Welfare and Social Choice: Expanded Edition*. Cambridge: Harvard University Press.

Smith, Martina (1993). “Violation of Norms,” *Proceedings of the Fourth International Conference on Artificial Intelligence and Law*. Amsterdam, The Netherlands, 60–65.

Snedegar, Justin (2017). *Contrastive Reasons*. Oxford: Oxford University Press.

———(forthcoming). “Reasons, Competition, and Latitude,” in Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics, Vol. 16*. Oxford: Oxford University Press.

Sinclair, Thomas (2018). “Are We Conditionally Obligated to Be Effective Altruists?” *Philosophy and Public Affairs* 46: 36–59.

Tucker, Chris (ms). *The Weight of Reasons*. (Tentative title.)

von Fintel, Kai and Iatridou, Sabine (2005). “What to Do If You Want to Go to Harlem: Anankastic Conditionals and Related Matters.” Draft, MIT.

Wessels, Ulla (2015), “Beyond the Call of Duty: The Structure of a Moral Region,” *Royal Institute of Philosophy Supplement* 77, 87–104.