4-2020

Agent-relative consequentialism and collective self-defeat

Matthew HAMMERTON
Singapore Management University, mhammerton@smu.edu.sg

Citation

HAMMERTON, Matthew.(2020). Agent-relative consequentialism and collective self-defeat. *Utilitas,* .
Available at: https://ink.library.smu.edu.sg/soss_research/3200

Agent-Relative Consequentialism and Collective Self-Defeat[1]

Matthew Hammerton, Singapore Management University

mhammerton@smu.edu.sg

Abstract: Andrew Forcehimes and Luke Semrau argue that agent-relative consequentialism is implausible because in some circumstances it classes an act as impermissible yet holds that the outcome of all agents performing that impermissible act is preferable. I argue that their problem is closely related to Derek Parfit's problem of 'direct collective self-defeat' and show how Parfit's plausible solution to his problem can be adapted to solve their problem.

Agent-Relative consequentialism has emerged in recent years as a serious option in normative ethics. Andrew Forcehimes and Luke Semrau argue that it should not be seen as such because it violates an overwhelmingly plausible principle they call 'Non-Compliance Is Never Preferable'.[2] In this article I challenge their thesis by arguing for two claims. First, I show how, given certain assumptions, the problem they highlight is identical to Derek Parfit's problem of direct collective self-defeat.[3] Therefore, rather than being a new problem for agent-relative consequentialism, it may instead be an old problem faced by all agent-relative theories. Second, I show how Parfit's suggested solution to his problem can be updated to apply to agent-relative consequentialism, and how this solution is plausible even when the problem is stated using Forcehimes and Semrau's terms. The upshot is that 'Non-Compliance is Never Preferable' is not a serious threat to agent-relative consequentialism.

[1] Please cite the published version, which is forthcoming in *Utilitas* (10.1017/s0953820820000096)

[2] Andrew T. Forcehimes and Luke Semrau, 'Non-Compliance Shouldn't Be Better', *Australasian Journal of Philosophy* 97 (2019), pp. 46-56.

[3] See Derek Parfit, 'Prudence, Morality and the Prisoner's Dilemma', *Proceedings of the British Academy* 65 (1979), pp. 539–64, and Derek Parfit, *Reasons and Persons* (Oxford, 1984), pp. 95-110.

1. The 'Non-Compliance Is Never Preferable' Problem

Agent-relative consequentialism combines a consequentialist deontic principle (each agent must always perform the act that, of the acts available to her, results in the best consequences) with an agent-relative axiology (the correct evaluative rankings of states of affairs varies from agent to agent). Its agent-relative axiology allows it to accommodate deontic constraints. For example, consider a deontic constraint that prohibits killing an innocent person even if doing so is the only way to prevent more innocent people being killed by others. Agent-relative consequentialism can accommodate this constraint by holding that, for each agent, her killing an innocent person is *worse-relative-to-her* than other agents killing several innocent people. Let's call agent-relative consequentialist theories that employ these kinds of axiological claims to accommodate the deontic constraints found in commonsense morality 'standard agent-relative consequentialism'. Advocates of standard agent-relative consequentialism often claim that it is an especially attractive moral theory because it combines the theoretical elegance of consequentialism with the intuitive deontic verdicts of deontology.

Forcehimes and Semrau claim that these benefits come at a great cost. They suggest that the following principle is overwhelmingly plausible:

> *Non-Compliance Is Never Preferable*. A moral theory must not allow there to be any possible circumstance in which, were every agent to act impermissibly, each would have more reason (according to the theory) to prefer the world thereby actualized over the world that would have been actualized if every agent had instead acted permissibly.[4]

They then argue that standard agent-relative consequentialism violates this principle. Their argument appeals to the following case:

> *Thirst For Blood*. Charlie and Debbie have a powerful desire to kill. Without intervention, each will freely kill three young children. Fortunately, Debbie possesses a drug. If she treats herself, then her killings will be reduced by one. If instead she treats Charlie, then his killings will be reduced by two. Charlie also possesses a drug. If he treats himself, then his killings will be reduced by

[4] The wording closely follows Forcehimes and Semrau 'Non-Compliance Shouldn't Be Better', p.51.

> one. If instead he treats Debbie, then her killings will be reduced by two. Taken together, the drugs completely eliminate the desire to kill.[5]

What ought Debbie and Charlie do in such circumstances? Should they treat themselves or treat each other? Forcehimes and Semrau point out that if we assume standard agent-relative consequentialism then this case has a Prisoner Dilemma structure. According to standard agent-relative consequentialism, each agent is required to treat herself. Debbie is required to treat herself because, whatever Charlie does with his drug, treating herself will minimize the killings that she performs. Charlie is required to treat himself because, whatever Debbie does with her drug, treating himself will minimize the killings that he performs.[6] However, if Debbie and Charlie obey this requirement and treat themselves then each will end up killing two innocent people. By contrast, if each violates this requirement and treats the other then each will end up killing only one innocent person. Yet, according to standard agent-relative consequentialism, that Debbie and Charlie each kill one innocent person is a better outcome relative to each of them than the outcome of each killing two innocent people. Thus, in *Thirst For Blood* standard agent-relative consequentialism requires agents to act a certain way even though it also tells them that a better outcome would result were none of them to act this way. In other words, it violates *Non-Compliance Is Never Preferable*. Let's call this the 'non-compliance' problem.

2. The 'Direct Collective Self-Defeat' Problem

Derek Parfit introduces a problem for agent-relative moral theories that he calls 'direct collective self-defeat'.[7] He understands all moral theories as giving agents certain substantive aims, which he calls 'theory-given aims'. According to Parfit, a theory is *directly collectively self-defeating* if and only if when *all* of us successfully follow that theory, we thereby cause our theory-given aims to be worse achieved than they would have been if *none* of us had successfully followed that theory.[8]

[5] Forcehimes and Semrau 'Non-Compliance Shouldn't Be Better', p.53.

[6] See the table in Forcehimes and Semrau 'Non-Compliance Shouldn't Be Better', p.53.

[7] See Parfit 'Prudence, Morality and the Prisoner's Dilemma', and *Reasons and Persons*, pp. 95-110.

[8] Parfit *Reasons and Persons*, pp. 53-54.

Parfit presents several cases where commonsense morality appears to be directly collectively self-defeating. For example, commonsense morality appears to give each parent the substantive aim that her children are not harmed. Yet in a case called the *Parents Dilemma*, two parents each have the options of either saving their own child from a lesser harm or saving the other's child from a greater harm.[9] The agent-relative requirement that each parent prevents her child being harmed results in this case having a Prisoner Dilemma structure. Each parent is required to protect her child from the lesser harm because, whatever the other parent does, this will result in less harm to her child. However, if both parents obey this requirement then their children are worse off than they would have been had each parent instead protected the other's child from the greater harm. Hence, their theory-given aims are better achieved if both disobey the agent-relative requirement.

3. Are They the Same Problem?

One interesting feature of *direct collective self-defeat* and the *non-compliance problem* is that they appear to apply to exactly the same cases. Forcehimes and Semrau demonstrate the non-compliance problem with *Thirst for Blood*. However, *Thirst for Blood* is also an example of commonsense morality being *directly collectively self-defeating*.[10] The constraint on killing gives each agent a theory-given aim that she does not kill. In *Thirst for Blood* this constraint requires Debbie and Charlie to treat themselves. Yet Debbie and Charlie would better satisfy their theory-given aims if they instead treated each other. Similarly, Parfit's *Parents Dilemma* is a case in which the non-compliance problem arises. Each parent is required to save her own child from the lesser harm. Yet, their children would fare better if they instead acted impermissibly and saved the other's child from the greater harm. Thus, acting impermissibly is morally preferable. In this example, it is notable that a moral theory does not need to be a version of agent-relative consequentialism to face the non-compliance problem. It only needs to be an agent-relative theory holding that there are things agents are morally required to prefer. Therefore, just like direct collective self-defeat, the non-compliance problem appears to apply to agent-relative theories more generally.

[9] Parfit *Reasons and Persons*, pp. 95-98.

[10] They appear to acknowledge this in Andrew T. Forcehimes and Luke Semrau *Thinking Through Utilitarianism* (Hackett, 2019), p. 71.

These similarities suggests that *direct collective self-defeat* and *non-compliance is never preferable* may actually be the same problem, stated using slightly different language. In fact, it is possible to produce statements of each of them that mirror one another:

> **The Non-Compliance Problem**: The moral theory entails that in certain circumstances, were every agent to act impermissibly, each would have more reason (according to the theory) to prefer the world thereby actualized over the world that would have been actualized if every agent had instead acted permissibly.
>
> **Direct Collective Self-Defeat**: The moral theory entails that in certain circumstances, were every agent to act impermissibly, each agent's theory-given aims would be better achieved then they would if every agent had instead acted permissibly.

Each of these statements concerns a fundamental clash between what a theory requires individually of each agent and what best realizes the moral aims or moral preferences of all agents. Whether they are in fact equivalent depends on whether the following biconditional is true:

> (1) An agent has more (moral) reason to prefer the actualization of W1 to W2 if, and only if, the agent's theory-given aims would be better achieved if W1 rather than W2 is actualized.

Many will find this biconditional plausible. Intuitively, if a moral theory says that an agent morally ought to prefer something, then it should also make that thing the agent's substantive moral aim. Intuitively, if a moral theory gives an agent a particular substantive moral aim then it ought to require that agent to prefer the realization of that aim over its non-realization.

Nonetheless, it is possible to find moral theories that reject (1). For example, some deontologists hold that it is morally wrong to kill one to save five, yet agree that killing one would bring about a morally better outcome. Such deontologists are often interpreted as saying that it is wrong to kill the one even though the outcome of this act is morally preferable.[11] Yet, according to Parfit's account of how moral theories provide substantive aims, this deontological view gives agents the substantive aim that they do not kill. Therefore, we have a case where a moral theory gives an agent the substantive aim that she does not kill yet holds that she morally ought to prefer the outcome in which she does kill. Such a theory entails that there are cases in which a moral theory is directly collectively self-defeating, yet does not face the non-compliance problem.[12] Therefore, whether we should regard these problems as equivalent or distinct will depend on the theoretical commitments we endorse. However, even on views in which these two problems are technically distinct, there seems to be enough similarities between them to suggest that a solution to one will generally be applicable to the other. Parfit presents a solution to the problem of direct collective self-defeat.[13] Yet Forechimes and Semrau do not anticipate any solutions to the non-compliance problem. Therefore, an obvious next step is to see whether Parfit's solution also applies to the non-compliance problem.

4. Parfit's Solution

Parfit's solution makes two revisions to commonsense morality so that it no longer requires agents to act in ways that are self-defeating. Revision R1 concerns what agents *ideally* ought to do, whereas R2 concerns what agents are required to do when not everyone will do what they ideally ought to do:

> (R1) When obeying a rule in this theory is self-defeating, we should all ideally do what will cause the theory-given aims of each to be better achieved.
>
> (R2) When obeying a rule in this theory is self-defeating, each agent should do what we all ideally ought to do if at least k agents will do this.

Parfit clarifies that R2 has two special features: (i) if k or more agents do what we ideally ought to do then each will better achieve their theory-given aims than they would have if no one did this. (ii) If less than k do what they ideally ought to do then

[11] For example, see Casper Hare *The Limits of Kindness* (Oxford 2013), p.91.

[12] Theories that entail the converse are also possible, although they seem less plausible. For example, a theory that requires you to kill the one, yet says that you have most reason to prefer that you do not kill them.

[13] Parfit *Reasons and Persons*, pp. 100-110.

each will do worse at achieving their theory-given aims than they would have if no one had done this.[14]

An agent-relative theory that incorporates R1 and R2 escapes the problem of *direct collective self-defeat*. For example, in the *Parents Dilemma* a theory containing these rules says that the parents ideally ought to save each other's child from the greater harm rather than saving their own child from the lesser harm. It also says that each parent ought to do what he ideally ought to do if the other parent will do what she ideally ought to do. Therefore, if I know that you will save my child from the greater harm then I ought to save your child from the greater harm.

If non-compliance is the same problem as direct collective self-defeat, then Parfit's solution equally applies to it. On the other hand, if they are distinct problems, we only need to slightly modify R1 and R2 for them to directly address the non-compliance problem:

> (R1′) When obeying a rule in this theory violates *Non-Compliance is Never Preferable*, we should all ideally do what will cause each to better achieve what the theory holds to be preferable.
>
> (R2′) When obeying a rule in this theory violates *Non-Compliance is Never Preferable*, each agent should do what we all ideally ought to do if at least k agents will do this.

A theory containing these principles requires agents to cooperate in the *Parents Dilemma* (and other similar cases) whenever it is the case that the other parent will also cooperate. The result is that all agents acting in a way that will bring about outcomes that ought to be preferred by all is no longer impermissible.

Parfit claims that his principles, R1 and R2, can be added to any agent-relative theory. However, he does not consider agent-relative consequentialism, which was not properly developed as a theory at the time that he was writing.[15] Yet R1 and R2 cannot

[14] See Parfit *Reasons and Persons*, pp. 100-102. Parfit states these conditions in terms of the *Parent's Dilemma*. I have rewritten them to apply to all cases of *direct collective self-defeat*. I have also left out a third revision (R3) that further elaborates on R2.

[15] Amartya Sen 'Rights and Agency', *Philosophy and Public Affairs* 11 (1982), pp. 3–39 is widely regarded as the first presentation of agent-relative consequentialism, however, he sketches the view

be added to standard agent-relative consequentialism because consequentialism, by definition, contains only one moral requirement—to maximize the good—and cannot be supplemented with additional rules. What we can do is incorporate the idea behind R1 and R2 into standard agent-relative consequentialism by building this idea into the axiology. This, of course, is the standard move in the consequentializing programme. Ask consequentialists to accommodate a new rule into their theory and they will do so by adjusting their theory of the good so that, when agents maximize the good, they always end up complying with the rule. Here are two axiological principles related to R1 and R2 that do this:

> (A1) In circumstances like *Thirst for Blood* and *Parents Dilemma* where everyone complying with standard agent-relative consequentialism is worse relative to each agent then everyone not complying with it, the outcome in which all agents cooperate to bring about what is best relative to all is ranked higher on each agent's relativized rankings than all outcomes where they do not cooperate.
>
> (A2) In circumstances where some agents do not cooperate to bring about what is best relative to all, the outcome in which you are one of at least k agents who cooperates is *better-relative-to-you* than all outcomes in which you are not one of at least k cooperating agents.[16]

Any version of standard agent-relative consequentialism that includes A1 and A2 in its axiology will escape the direct collective self-defeat and non-compliance problems. Thus, Forcehimes and Semrau are wrong to claim that the problem they highlight undermines agent-relative consequentialism. They fail to recognize that a plausible agent-relative consequentialism will contain A1 and A2 and thereby avoid this problem.

very broadly. Parfit *Reasons and Persons* does not engage with Sen's work and appears to assume that consequentialism is necessarily agent-neutral.

[16] Parfit's two provisos for R2, suitably adjusted, obviously apply here.

3. Defending Parfit's Solution

I will finish by defending Parfit's solution against one possible line of criticism.[17] A critic might reject the use of principles like R1 and R2 or A1 and A2 on the grounds that they are ad hoc additions to a moral theory, carefully designed to account for just those cases in which the problem arises.

I think that there are three promising lines of response to this complaint. First, it may be pointed out that making small revisions to an otherwise plausible moral theory in order to address a serious objection against that theory is a legitimate move to make and cannot be so quickly dismissed as ad hoc. Many agent-relative moral theories already generally favour adopting a diverse and complex range of deontic or axiological principles in order to account for commonsense morality. Therefore, adopting further such principles in this instance does not seem especially problematic.

Second, Parfit suggests that there is a deeper justification for adopting his revisionary principles.[18] He argues that a common mistake in our moral theorizing is to focus only on individual acts, neglecting the effects of sets of acts that we perform together (he calls this the 'second mistake in moral mathematics'). He then argues that moral theories are directly collectively self-defeating in part because they make this mistake. Therefore, revisions like R1 and R2, or A1 and A2 are justified because they bring the consideration of action at the collective level to moral theories that have wrongfully ignored it. Correcting such a mistake in a moral theory is not an ad hoc move to make.[19]

Finally, several commentators have argued that Parfit's solution to the problem of direct collective self-defeat is unnecessary because commonsense morality already contains rules or assumptions that prevent cases of self-defeat from arising.[20] If these

[17] I am not aware of anyone who has raised this criticism in print. However, it seems to be a serious enough concern to deserve a response.

[18] Parfit *Reasons and Persons,* p. 108.

[19] This response will only appeal to those who are willing to accept Parfit's 'second mistake in moral mathematics'. It is not an entirely uncontroversial thesis and has been disputed by Frank Jackson 'Which Effects?' *Reading Parfit*, ed. Johnathan Dancy (Oxford, 1997), pp. 42-53.

[20] See: Bart Gruzalski 'Parfit's Unified Theory of Morality', *Philosophical Studies* 50 (1986), pp. 143-152; Arthur Kuflik 'A Defense of Common-Sense Morality', *Ethics* 96 (1986), pp. 784-803; Lanning Sowden 'Parfit on Self-Interest, Common-sense Morality and Consequentialism', *The Philosophical Quarterly* 36 (1986), pp. 515-535; and Kieran Setiya 'Parfit on Direct Self-Defeat', *The Philosophical*

critics are correct then no revisionary principles are required and thus the concern that these principles are ad hoc does not even arise. This applies not only to R1 and R2 but also to A1 and A2. Standard agent-relative consequentialism builds the rules and assumptions of commonsense morality into its axiology. Therefore, if commonsense morality already avoids direct collective self-defeat then standard agent-relative consequentialism will avoid it as well and thus does not need to include special additional axiological principles like A1 and A2.

In summary, a strong case has been made that the non-compliance problem is not a serious threat to agent-relative consequentialism. The problem appears to apply to agent-relative theories more generally, and may well be identical to the problem of direct collective self-defeat. Furthermore, Parfit's solution to the latter problem can be adapted to apply to agent-relative consequentialism, and is no less convincing when it is adapted in this way.[21]

Quarterly 49 (1999), pp. 239-242. See also Derek Parfit 'Comments', *Ethics* 96 (1986), p. 851 for a significant concession to these critics.

[21] I am grateful to Ryan Cox, Sovan Patra, and two anonymous referees from this journal for helpful comments on an earlier draft of this article.