In Database: The Philosopher's Index 1940-2003/12

TITLE: **Agent Neutrality Is the Exclusive Feature of Consequentialism****AUTHOR:** **Zong,-Desheng****SOURCE:** Southern-Journal-of-Philosophy. Wint 00; 38(4): 677-693**JOURNAL TITLE:** Southern-Journal-of-Philosophy**INTERNATIONAL STANDARD SERIAL NUMBER:** 0038-4283**ABSTRACT:** It is widely believed that consequentialism is a theory characterized basically by its agent neutrality. However, it has become fashionable in recent years among some moral philosophers to deny that agent neutrality is an exclusive feature of consequentialism. In particular, there have been attempts to show that there can be agent-neutral nonconsequentialist theories as well as agent-relative consequentialist theories. I argue in this paper that this last claim is flawed because it is based on a failure to distinguish between consequentialism as a standard of right and wrong and consequentialism as a principle of moral deliberation.**DESCRIPTORS:** AGENT-; CONSEQUENTIALISM-; ETHICS-; NEUTRALITY-; QUANTIFICATION-; UNIVERSAL-**LANGUAGE:** English**DOCUMENT TYPE:** Journal-Article**PUBLICATION YEAR:** 2000**UPDATE CODE:** 353**ACCESSION NUMBER:** 3013942

AGENT NEUTRALITY IS THE EXCLUSIVE FEATURE OF CONSEQUENTIALISM *

Desheng Zong

An idea that has attracted a lot of attention lately is the thought that consequentialism is a theory characterized basically by its agent neutrality.[i] The idea, however, has also met with skepticism. In particular, it has been argued that agent neutrality cannot be what separates consequentialism from other types of theories of reasons for action, *since* there can be agent-neutral non-consequentialist theories as well as agent-relative consequentialist theories. I will argue in this paper that this last claim is false.

The paper is divided into four sections. Section one specifies two senses in which consequentialism is agent-neutral. Section two and three examine and reject, respectively, the claim that there are agent-relative consequentialist views as well as agent-neutral non-consequentialist views. I end the paper with some remarks on the plausibility, or better, the implausibility of characterizing consequentialism in terms other than agent neutrality.

I. CONSEQUENTIALISM AND AGENT-NEUTRALITY

One fairly common and incontrovertible characterization of consequentialism is that it is a theory that identifies moral rightness with ensuring (seeing to it, making sure) that the best state of affairs be brought about. The characterization applies to all versions of consequentialism. It implies, for example, that the central claim of act-consequentialism is that moral rightness consists in ensuring (making sure, seeing to it) that the best state of affairs in one's situation be brought about, and it implies that the central claim of rule-consequentialism is that moral rightness consists in ensuring (making sure, seeing to it) that the rules that are most good-conducive be followed.[ii]

This characterization of the consequentialist notion of moral rightness allows us to see two important features of this type of moral theory.[iii] To begin with: the use of the word 'ensuring' rather than 'doing' implies that, to the consequentialist, what matters, as far as morality is concerned, is that the best states of affairs be brought about, not by whom or through whom they are brought about. A slightly differently way of saying this is that the issue of agency is relevant to morality only to the extent that it affects the results. Consequentialism, then, is characterized by its doer-neutrality.

It is important, at this point, to emphasize a distinction that is as often drawn as it is forgotten, that is, the distinction between consequentialism as a standard of right and wrong and consequentialism as a principle of moral deliberation. To say that consequentialism is doer-neutral is to say that consequentialism as a standard of right and wrong is doer-neutral; it is not to say that consequentialism does not recognize the value of agency in moral deliberation. For it almost always does. For example, if doing an act myself is the only way to ensure the coming about of the best state of affairs, then making sure that I am the person who does it is totally in agreement with the theory's emphasis on doer neutrality. But considerations of agency only occur at the level of moral deliberation, and the form of agent relativity they may take should not be confused with the doer neutrality of consequentialism as a criterion of right and wrong.

Second, the characterization also allows us to see that, on this theory, what matters, as far as morality is concerned, is that the best states of affairs be brought about; and it matters very little whose good it happens to be in that which is promoted. The 'location of good' along the dimension of people, to use John Broome's catchy term,[iv] is relevant to morality inasmuch as it affects the ultimate results. Consequentialism, then, is agent neutral with respect to values.

Again, to say that, according to consequentialism, the location of good *per se* has no moral significance is not to say that the theory does not allow the moral agent to engage in agent-relative thinking when considering what to do. If, for example, taking care of your own children is the only way to ensure the coming about of the best state of affairs in a given situation, then taking care of your own children is the right thing to do, and doing it is totally in agreement with the consequentialist criterion of right and wrong. Once again, the agent-relative character of the consideration is no indication that consequentialism as criterion of right and wrong can be agent-relative.

Consequentialism is then agent-neutral in two importance senses. It is agent-neutral with respect to agency, and it is agent-neutral with respect to values. But what is even more interesting about all this is that there seems to be a purely formal way to capture the two senses of agent-neutrality identified here. Consider, for example, Thomas Nagel's well-known analysis of the logical forms of reasons for action. Nagel observes that "...every reason can be expressed by a predicate R, such that for all persons p and events A, if R is true of A, then p has prima facie reason to promote A." (*The Possibility of Altruism*, p. 90). That is:

(R1) (p, A) (RA c p has prima facie reason to promote (to ensure, to see to it that, etc.) A)

(R1) allows the definition of two distinct reason predicates, that of an agent-neutral reason and that of an agent-relative reason: a reason predicate R is an agent-neutral one iff R contains no free variable within itself; it is agent-relative otherwise.

That this formal notion of an agent-neutral reason for action captures the two senses of agent neutrality stated earlier is obvious. To begin with, the requirement that there be no free variable in R guarantees that whatever is capable of giving rise to a reason for action is able to do so independently of its relation to the agent. Second, the doer neutrality discussed earlier is guaranteed by the use of the expression '... p ... promote' in R. Note that 'promote' cannot be replaced by words that indicate performance rather than results; for example, replacing 'promote' with 'do' would change the nature of the predicate.[v] That (R1) has this feature does not suggest arbitrariness. The point is best appreciated when we compare it with an equally effective system, developed by McNaughton and Rawling, that captures the two senses of agent neutrality. Their analysis is in terms of rules rather than reasons. The definition runs as follows: "Given a rule, we begin by transforming it into the form:

(MR1) (x) (x S[...]) [reads 'for any x, x should ensure that ...']

The rule is agent-relative iff there is an occurrence of 'x' in the square brackets bound by the initial universal quantifier; agent-neutral otherwise."[vi] (MR1) achieves the same effect as 'promote' in Nagel's analysis by virtue of the restriction that there be no occurrence of 'x' in the brackets, this despite the possibility of occurrence of expressions such as 'do' within the brackets.

From the perspective of this paper, the significance of the existence of formalisms such as the two just introduced lies in that they enable us to say that what defines consequentialism *as a criterion of* *right and wrong* is an agent-neutral reason (if Nagel's analysis is used) or an agent-neutral rule (if the analysis of McNaughton and Rawling is used). For example, act-consequentialism, which says that what the moral agent ought to do is to ensure that the best state of affairs gets promoted, is captured by the following rule:

(MR2) (x) (x should ensure that [(y) (z) (y is the best state of affairs in z's situation C z does y)])

With appropriate modifications, an agent-neutral rule can also be formulated that captures the central claim of rule-consequentialism (something along the line of, say, (x) (xS[(y) (z) (y is a set of rules the compliance of which tends to promote the overall good C z follows y)]). For convenience's sake, I will call the claim that agent-neutrality is an exclusive feature of consequentialism the agent neutrality thesis.

II. THE IDEA OF AGENT-RELATIVE CONSEQUENTIALISM

Could there be theories that are consequentialist in nature and yet agent relative? What might such theories look like? Suppose we try this: a consequentialist theory is agent-relative if it is based on an agent-relative conception of the good. There are two cases to consider regarding this suggestion, since the term 'agent-relative value' allows two different readings, 'agent-relative non-moral value' and 'agent-relative moral value'. I will consider both these possibilities. Start with the non-moral value version first. If the claim that there are consequentialist theories based on agent-relative non-moral value is going to be true, two things must be the case. First, the word 'consequentialism' must mean what most people in the recent analytic tradition take it to mean, i.e., moral theories which Bentham, Mill, Sidgwick and others have championed and which Rawls, Williams, Foot and others have criticized.

Second, it must be the case that the alleged examples of agent-relative consequentialism are truly theories that are based on, and only on, agent-relative values that are non-moral in nature. Could there be any moral outlooks capable of satisfying these two conditions? The following is a suggestion made by Arthur Kuflik. Concerning Derek Parfit's claim that all consequentialist theories make the same central assumption, i.e., there is one ultimate moral aim: that outcomes be as good as possible (*Reasons and Persons*, p. 24), this author claims that some forms of egoism constitute examples of agent-relative consequentialist theories. Here is Kuflik in his own words:

> [A] Consequentialist theory can be agent-relative (if it rests on an agent-relative conception of what is good) and an agent-neutral theory can be Non-Consequentialist (if the common aim which it gives to all moral agents is not to maximize what is good)... The most obvious example of a theory that is agent-relative yet Consequentialist is universal egoism: each person ought to do whatever makes his life as good as possible *for him* (italic original)... [vii]

Kuflik's claim does not sustain examination. The example allows two different readings. On reading one, by 'Consequentialism' Kuflik means 'consequence-oriented theories'(some examples of this type of theory are Sidgwick's Rational Egoism, Parfit's Present-aim Theory, and Kuflik's 'universal egoism' of course). But this is bad reading. The following shows this: the proponent of the agent neutrality thesis first says: "Consequentialism is agent neutral." Since Utilitarianism is a version of consequentialism, the proponent of the thesis is entitled to say: "Utilitarianism is agent neutral." The current reading makes Kuflik come out saying: "Not true! Because egoism is agent-relative." But this is not very helpful; it's like when you say: "Cats chase mice," and Kuflik answers: "Not true! Because some dogs chase bones." This first reading must be rejected. Could Kuflik be using the term in its conventional sense? This is a possibility. There are three pieces of evidence for this. First, the article is basically a response to some of the claims Parfit has made about common-sense morality and consequentialism, and Kuflik's use of the term 'consequentialism', except for the passage quoted above, is consistent throughout. He used 'Consequentialism' to refer to consequentialism, not to consequence-oriented theory. No where in the article did he state that he was using 'consequentialism' in both senses. Second, in the context where the quoted passage occurred, Kuflik was concerned with one of Parfit's claims about consequentialism, viz., the latter's agent neutrality. Parfit's use of the term 'consequentialism' in *Reasons and Persons* is consistent throughout. He means by that term what most philosophers in the recent analytic tradition take it to mean. So if Kuflik's discussion here is intended to be a discussion of Parfit, he must mean by 'Consequentialism' what Parfit means by it. Third, Kuflik inherited Parfit's idiosyncrasy of capitalizing the first letter of 'consequentialism' (Parfit does this with respect to many other theories; he uses, for example, 'Common-sense morality', 'Present-aim Theory', 'Self-interest Theory', so does Kuflik), and he did this throughout the whole article. All of this shows that 'Consequentialism' in Kuflik's article was intended, at least initially, to be taken to refer to consequentialism, not to consequence-oriented theories. But Kuflik's example still makes little sense on this reading, for Kuflik is here claiming: "Some form of egoism is an example of (Classical Utilitarianism or Welfare Utilitarianism or Perfectionist Utilitarianism or...). Kuflik's example makes no sense on both readings; it must be rejected. Moreover, our diagnosis of the alleged example allows us to make a more general point, that is: for those who claim that there are agent-relative consequentialist theories based on a non-moral conception of good, either they are not talking about the type of theories to which the term 'consequentialism' is conventionally used to refer, or they have simply confused consequentialism with consequence-oriented theories such as egoism. In both cases, although it is true that there can be agent-relative 'consequentialist' theories, the truth of the claim will amount to nothing in terms of settling the issue that concerns us here, i.e., what characterizes the type of theories exemplified by, say, Classical Utilitarianism.

I turn now to the claim that there are consequentialist theories based on agent-relative moral values. I will try to show that this suggestion is no more plausible than the one we just considered. A few words about the notion of agent-relative moral value before we turn to the key issues. That there are agent-relative values is a claim held by many moral philosophers, especially non-consequentialist philosophers. The alleged examples fall largely under two groups. Those in the first group all feature some aspect of the agent, or the notion of agency *per se*, while those that fall under the second group are often things that are more or less related to the agent. An example of the first group is not to commit murder myself. It is said that, to me, not causing deaths through my own agency has special moral significance that the state of affairs consisting of others' committing murders themselves does not. An example of the second group is taking care of one's children oneself (vs. letting others take care of them). The former is said to have special moral value that the latter does not.

Whether such a notion of good is plausible or not need not concern us here. One thing is for sure: if there are going to be any agent-relative consequentialist theories of the type under consideration here, the truth of something like what I just said must be assumed. What is worth emphasizing, however, is that to someone who claims that there are agent-relative consequentialist theories, admitting the existence of agent-relative moral value marks only the beginning of a project, not its completion. This is so for the following reasons. First, the alleged agent-relative consequentialist theories would not be the only ones that embrace the idea of agent-relative moral values, many non-consequentialist theorists, for example, will also give such values a place in their theories. So the mere embracing of agent-relative values cannot be what defines consequentialism. Second, and more important, if the proponent of the idea of agent-relative consequentialism is to get what she is after, she must see to it that the teleological character be preserved in her account. By 'teleological character' I mean the idea that (1) states of affairs can be compared and ranked in terms of better and worse, and (2) moral rightness consists in bringing about the best state of affairs. To accomplish this she must ensure that her account preserve conceptual room for comparing and ranking states of affairs. Third, and no less important, her account must make room not only for the idea that states of affairs can be compared and ranked, but also that the comparing and ranking can be, and are often done in terms other than ‘good for me’ (or ‘good from my point of view’). An alleged example of consequentialist theory will not be truly consequentialist unless it satisfies conditions 2 and 3. With these preliminary remarks in mind, let us look at some alleged examples of agent-relative consequentialist views.

1. In his book *Weighing Goods,* John Broome complains that the recent debate surrounding the notion of agent neutrality has distracted the attention from the real issue about consequentialism. According to Broome, what characterizes consequentialism is not its agent neutrality, but its teleology (about this term, see below). To support his claim, Broome offers an example of what he calls ‘agent-relative teleological view’. The kind of ethical view Broome has in mind is effectively captured by the following schema:

(R2) (p) (p ought to see to it that [p does what counts more for p rather than what counts more for others])

Broome uses promise-keeping to illustrate (R2): Suppose that you are faced with two options: to keep your promise, or to break it to prevent five promises from being broken by others. According to Broome, if you decide ‘that I ought to keep my promise because my promise breaking counts more for me than other people’s’ (p. 8), then the chances are that you are offering an agent-relative consequentialist view. The reason:

> This is an agent-relative opinion. It is also teleological...It treats the wrongness of promise breaking as a bad feature of an act. It weighs the badness of my promise breaking against the badness of other people’s, to determine the overall badness of the alternative acts. For me, my promise breaking weighs more than other people’s. In the end, for me, breaking my promise

> comes out worse overall. Therefore, I ought not to do it. This is a teleological argument -- an agent-relative one. (p. 8)

Is what Broome dealing here an example of agent-relative consequentialist theory? The answer is No. Broome claims that teleological theories (this is Broome's own term for 'consequentialist theories') treat the rightness (wrongness) of something as determined by its good feature (bad feature). On this point we need not to argue with him. He claims that promise-breaking is a bad feature of an act. We can also agree with this. However, one needs to be careful here, in view of the following: every intentional act performed by an agent can be viewed from two different perspectives. On one of these, an act is an mere event, something that has, or will take place. On the other, an act is more than an occurrence; it is something done by an agent. It matters which of these two perspectives one takes when one is trying to determine the moral significance of something such as promise-breaking. The 'event perspective' makes it possible for one to compare, from an objective standpoint, the badness of two situations in which promise-breaking is a factor. Viewing from this perspective, one may reasonably, and intelligibly, conclude that a world that contains fewer instances of promise-breaking (as event) is better than one in which there are more instances of promise-breaking. Broome calls promise-keeping a 'general good', which indicates that he does not deny that promise-keeping can be an impersonal good. In contrast to the event perspective, the 'doing perspective' focuses on the agency aspect of a situation. Since an act's being performed by one agent automatically rules out the possibility of its being performed by another, the perspective allows one to say that, when viewed not just as things that take place, but as actions performed by oneself, an act, such as the breaking of a promise, acquires extra moral badness which might not be reflected in the value an impersonal ranking might assign to it. Again, Broome has no problem accepting this point; as he puts it himself, “For me, my promise breaking weighs more than other people’s. In the end, for me, breaking my promise comes out worst overall.” (p. 8)

The key issue here is this: When Broome claims that ‘my promise breaking comes out worst overall’, what are we to take him to mean by the ‘worse overall’? There are three possibilities to consider (in what follows I will say nothing about the self-interest interpretation of ‘worst overall’; as I have made clear earlier, I am here concerned only with the notion of moral good). First, ‘worst overall’ means ‘containing the greatest number of instances of promise-breaking as occurrence’. This interpretation allows one to say that the world that contains one instance of promise-breaking (due to my failure to keep some promise that I made) is better than the one that contains five instances of promise-breaking (all due to the failure of five other people to keep their promises). Second, ‘worst overall’ means ‘worst overall after the special moral weight of the issue of agency has been appropriately counted, against the background of the impersonal ranking’. This type of evaluation is usually reached in two steps: one first weighs the objective badness of the two situations (keeping my promise and letting five promises be broken by others vs. breaking my own promise to keep five promises from being broken by others). The result of the evaluation is then adjusted by taking into account the extra moral badness of *my* breaking my own promise. On this second reading, the world that contains one instance of promise-breaking (due to my failure to keep my own promise) need not be better than the one that contains five instances of promise-breaking due to others; it could actually be worse. Third, ‘worst overall’ means ‘what is worst overall when the situation is viewed from, and only from the doing perspective’. In this third case, that actions performed by me can be viewed and ranked in terms of their objective value or disvalue is out of the window; there is no such thing as the ‘event perspective’ when it comes to the right and wrong of actions. Situations are judged good (meaning ‘right’) and bad (meaning ‘wrong’) only by a ‘Did I, or did I not do such and such’ criterion. The worst is that I commit wrong (e.g., break my promises), and the best is I do the right thing (e.g., keep my promises). Another way to describe the third view is to say that agency is so special a moral factor that, by its very presence in a situation, it renders the other factors morally irrelevant.

It is clear that the first and the third readings are incapable of providing support for Broome's claim that the example he gives here is an example of agent-relative consequentialism. Will the second interpretation do the trick? The answer is still No. It is clear that, while agency is given some moral weight on the second view, it is regarded nevertheless as but one value, to be weighed against other values in the world (which might well include values on some impersonal ranking of states of affairs). What we still do not have is a plausible example of agent-relative consequentialist view.

2. Some philosophers have suggested that there is a logical space for a version of (agent-relative) consequentialism, according to which what matters is that *I* bring about the best state of affairs. This is a more general claim than the one we just considered. Instead of claiming that some aspect of the issue of

agency is capable of grounding a consequentialist view, it claims that agency *per se* provides ground for a agent-relative consequentialist view.

I take the point of the advocate of this view to be this: agency occupies an unusually high place in morality such that, given an impersonal (agent-neutral) ranking of states of affairs from the best to the worst, it is always better if I am the one who actually brings about the best state affairs than it would be if someone else were to do it. Note that this line of reasoning is possible exactly because the idea of there being a ranking of states of affairs from the best to the worst in agent-neutral terms is already in place (the expression 'the best state of affairs' tells this). As in Broome's case, the advocate of this view is faced with a dilemma. That is, if she says that by 'the best...' she means 'best from the objective point of view', then she introduces agent neutrality right back into her account; if she chooses an agent-relative interpretation of the expression, then she abandons her teleological position by denying that the goodness and badness of states of affairs can be compared and ranked in terms other than 'good (bad) from my point of view'. I do not see how she can get herself out of this dilemma.

So far, I have only considered two examples of the so-called ‘agent-relative consequentialism’. I have not said anything about cases that are, or might be built on agent-relative values of the second class as pointed out at the beginning of this section. It might be said that the issue cannot be sufficiently dealt with in this manner, given the range of possible cases to be considered. But this is not so. No matter what aspect of the notion of agency (or aspect of the life of the agent) is taken to be capable of grounding an agent-relative consequentialism, in order to stay consequentialist, the proponent of the view under discussion here must admit that the results of evaluations from perspectives other than the agent-relative one constitute part of the background on which her moral thinking must be based. This is a point our discussion of Broome’s example has made clear. It follows that whichever aspect of the agent is taken to be worth accommodating into the consequentialist framework, it is always into an agent-neutral context that it is accommodated.

III. THE IDEA OF AN AGENT-NEUTRAL NON-CONSEQUENTIALISM

I now turn to the claim that there could be agent-neutral non-consequentialist views. I begin with some remarks on semantics. First, a few words about the term ‘agent neutral’. As I pointed out earlier in the paper, two important meanings of agent neutrality are (1) neutrality with respect to agency (doer neutrality), and (2) neutrality with respect to value. I also showed that consequentialism embodies both these features. Could we have missed some other possible meanings of the notion of agent neutrality, and therefore have failed to see other senses in which consequentialism might be agent neutral? The possibility need not be denied, but then what we might have missed is perhaps inconsequential. In any case, I am claiming that if a claim of agent neutrality is to be worthy of the name, then it must squarely address the two senses of agent neutrality that I have identified, and this is also how I am going to take the claim that there are agent-neutral deontological theories. Now a few words about the term ‘non-consequentialism’. A common practice nowadays seems to take this to be more or less the same as 'deontological theories' (or 'theories that contain deontological rules'). If this common practice is followed (I do not see why we should not do so, especially in the context of this paper), then the question of whether there are any agent-neutral non-consequentialist theories becomes the question of whether there are any agent-neutral deontological theories.

In what sense can a deontological theory be agent-neutral? It is clear that a full discussion of the issue will consist of two parts, one part addresses the question of whether there could be deontological theories that embody what I have called 'doer neutrality', and a second part addresses the question of whether there could be deontological theories that embody what has been called 'impersonal good'. But my discussion below will only focus on the first task. This is because even if the reason to do good a deontological theory will typically recognize can be said to be based on impersonal good (and this is debatable), if we can show that no deontological theories will embody doer neutrality, the claim that there could be agent-neutral deontological theories will still be unsustainable. It is also for this reason that most alleged examples of agent-neutral non-consequentialism found in literature are accounts designed to show that non-consequentialist theories can be doer neutral.

I start with an account suggested by F. M. Kamm in her book *Morality, Mortality*. Kamm claims that one of John Taurek's arguments in "Should the Numbers Count" is subject to an interpretation which renders Taurek as arguing along the following line: there is no such thing as 'impersonal good' ('good, period'); everything that is good is so only relative to some agent to whom it matters. Furthermore, if something is good relative to some agent, then everybody is permitted to aid that person with respect to the good. According to Kamm, this view is agent-relative, because the value at issue is an agent-relative good; it is also agent-neutral, because "[T]he permission to act is not relative to one agent rather than another; it is an agent neutral permission" (*Morality, Mortality*, p. 78).

Kamm's suggestion has not escaped some perceptive philosophers' notice. Temkin, for example, has made reference to it, with apparent approval, although no pagination was given.[viii] But is what we have here an example of agent-neutral non-consequentialism? The answer is No. The account reveals its true nature when a little formalism is applied to it. Using the formal tool we used earlier, we can reformulate Kamm's proposal in the following way:

(K) (p) (p should ensure that [(q) ((∃x) (x is a good to q) → p is permitted to help q with respect to x)])

That is, for any person p, p should ensure that, for any person q, if there is an x such that x is a good to q, then p is permitted to help q with respect to x. The rule is of the form '(p) (p should ensure that [... p ...])'; it is *not* an agent-neutral rule.

That the example should turn out to be agent-relative rather than agent-neutral as claimed should come as no surprise. After all, Kamm's suggestion is based on an interpretation of Taurek. But in the latter's original argument, the view that anyone may aid someone with respect to something that matters to her is, for Taurek, derived from a more basic view, i.e., the claim that anyone may aid herself when some good of herself is at stake. Taurek stresses the importance of the distinction between 'loss to a person' and 'loss of a person'; and it is this emphatic approach – which views things from the standpoint of the victim – that leads him to the adoption of the view under discussion here . (K) is agent-relative exactly because it is the result of viewing things from the point of view of the person whose good is at stake.

Let us consider next an example offered by Arthur Kuflik in the article mentioned earlier. Kuflik's account goes like this:

> An example of agent-neutral Non-Consequentialist thinking would be a theory which tells each and every agent to see to it that 'justice' is done -- but insists that injustice can only be brought about by means that are themselves just. The principles of justice would be treated as basic *constraints* on conduct which *anyone* ought to try to get *everyone* to observe, but which *no one*

> is at liberty to violate. A theory of this sort would be agent-neutral, for it would assign the same goal to everyone: that is, “universal justice through just means only”; but it would be Non-Consequentialist, for the goal would *not* be to *maximize* “the net sum of good minus bad.” (p. 798, italic original)

This example is worth careful analysis, not because it is a proof of the existence of agent-neutral non-consequentialist theories, but because it highlights many of the traps that trick supporters of ‘agent-neutral non-consequentialist theories’ into believing in something that does not exist. Let me explain. Suppose we take theories of reason for action (theories of morality being a subset of this) as theories of behavioral rules. Given this, it seems clear that in order to get a theory that is agent-neutral yet non-consequentialist, you need to take care of two things. First, you need to make sure that the theory you will eventually get includes, among others, agent-neutral rules. To ensure that there will be at least one agent-neutral rule in your theory, you could start with something like the following:[ix]

(R3) (p) (p should see to it that ...)

(R3) gives you what I have called in this paper ‘doer-neutrality’. The second step is to give (R3) a non-consequentialist content. This could be done in a number of ways, since a deontological theory may well include more than one deontological item. Suppose you are interested in getting an agent-neutral yet deontological view. You could, following Kuflik, choose justice (anything else, such as promise-keeping, will do). This will give you (R4):

(R4) (p) (p should see to it that justice be done)

However, (R4) is not a deontological view; in fact, as it stands, it might just as well be a consequentialist rule that tells you to promote justice, even if doing so means violating justice yourself. So to ensure that (R3) will turn out to be what you want, i.e., a deontological rule, you will have to add restrictions on the ways justice is to be promoted; say, you choose the condition 'through just means only'. Let us call this 'the constraint clause'. (R4) plus the added constraint clause now gives you (R5):

(R5) (p) (p should see to it that justice be done through just means only)

This completes the steps of constructing an agent-neutral yet non-consequentialist view. Or so you would think if Kuflik were right. But he is not! For what you got, from (R4) plus the added condition, is actually a different rule, (R6), which is nothing but a badly formulated version of (R5):

(R6) (p) (p should see to it that p does not seek justice through unjust means)

(R6) is known to be an agent-relative view. It says that for any p, p does not violate justice herself. The sleight of hand is completed, by the added constraint clause at the time (R5) was derived from (R4). That is, when the condition of 'through just means only' was added, it caused a crucial change in (R3), rendering it a rule of entirely different logical form:

(R7) (p) (p should see to it that ... p ...)

(R7) is not (R3)! How does the constraint clause do this? The reason is very simple, and yet very instructive: The aim a rule gives us can be a common aim only if what it enjoins us to aim at is some result, such as the coming to pass of some state of affairs (the coming about of an ideally perfectly just world in Kuflik's case). On the other hand, one's doings, and how one acts, can never be a common aim for everybody. For example, when I am given the instruction that I keep my head straight when saluting, my keeping my head straight is not something anybody else can aim at doing; all they can do is to keep *their* heads straight, not mine. It is for this reason that rules such as 'Do not lie even if...', 'Do not coerce even if...', 'Do not kill even if...', 'Do not betray trust even if...', etc., which are basically restrictions on acts themselves, can only give agent-relative aims. However, when the clause 'universal justice be done' is added to (R3), the result, i.e., (R4), remains an agent-neutral rule because justice is here taken to be an unrealized state to be brought about by the effort of everybody. Yet, when, to get (R5), one adds the condition 'you can only get everyone to observe justice through just means', one changes the nature of the rule by making justice a matter of everybody conducting *their own* behavior in such and such a way. The aim of the rule, (R5), now becomes an agent-relative one exactly because what we are now required to aim at is no longer 'seeing to it that universal justice come about', but rather 'conducting yourself in ways that are not violations of justice'. The focus of (R5) is one's doings, not the result everyone is to aim at in their doings: *Your* not committing murder, lying, betrayal, etc. in your attempt to get everyone (including yourself) to observe justice can only be your aim; *my* not committing murder, lying, betrayal, etc. in my attempt to get everyone (including myself) to observe justice can only be my aim.

Kuflik, then, like Kamm, is mistaken in believing that he has produced an example of an agent-neutral non-consequentialist view embodying the doer neutrality feature. Far from idiosyncratic, the account is the result of following a path along which many false examples of 'agent-neutral non-consequentialist views' can be generated. Despite the fact that the deceptive steps deployed in producing the false example takes some formalism to expose, the real root of the problem lies in a failure to recognize a simple distinction, i.e., the distinction between what can be aimed at as a common goal (i.e., results), and what can only be aimed at by each agent herself (the performing of some acts by herself). No formalism is needed in order to see the difference.

I now turn to our last example, a claim made by John Broome concerning the so-called 'side-constraints'. Most people believe that if there are any agent-relative rules at all, side-constraints must be among them. Broome disagrees. According to him, "... what makes side-constraint theory nonteleological is not agent relativity. It is the way it takes ethical considerations to work: side constraints determine *what ought to or ought not to be done* directly, and not by determining goodness." (p. 10, italic added). Not only that, for if it is not agent-relativity that makes side-constraints what they are, then it is reasonable to believe that there might be side-constraints that are agent-neutral! And indeed, Broome has offered the following as an example of an 'agent-neutral side-constraint':

> Consider the view that whenever a miner is in mortal danger trapped in a mine, all available resources should be devoted to rescuing him. This will reduce the resources devoted to safety measures in mines, and so lead to the deaths of more miners in the future. Nevertheless, it is what ought to be done. This is a side-constraint view. But it is agent-neutral. (*Weighing Goods*, p. 9-10)

I find Broome's discussion here problematic. Let us start with his analysis of side-constraints, i.e., that 'what makes side-constraint theory nonteleological is not agent relativity. It is the way it takes ethical considerations to work: side constraints determine what ought to or ought not to be done directly, and not by determining goodness'. This statement of Broome's contains two crucial phrases, 'what ought to or ought not to be done directly' and 'directly'. How is one to take the first phrase? Broome's miner rescue example provides the clue for understanding what he really means by that. It turns out that what he really means is 'side constraints determine what ought to or ought not to be done directly *regarding some desired end*', namely, the rescue of the trapped miners, not 'side constraints determine what ought to or ought not to be done directly'. In Broome's miner example, i.e., 'trapped miners must be rescued no matter what it takes', which Broome takes to be an example of agent-neutral side constraint, there is an end that the rule tells us to bring about, namely, the rescue of the trapped miners; furthermore, the rule enjoins everyone to do whatever it takes (including using all available sources) to achieve that goal. This is a bona fide example of consequentialism; it bears little resemblance to what are commonly called 'side-constraints'. The latter are not rules that determine what ought to or ought not to be done *regarding* a desired end, but restrictions on what one can or cannot do *regardless of* the goals one may be aiming at.

Broome's own example is also helpful in understanding what he means by that mysterious 'directly' in his definition of side constraints. What Broome really means by that term, it turns out, is something like 'without first comparing the utility of the means one is resorting to in hopes of achieving the desired end and the utility of the end itself'. It appears that, to Broome, every case of (self-consciously produced) cost-benefit *in*efficiency is an example of a side-constraint at work. However, there is a better way to describe Broome's idea of side-constraints, that is, he confuses acting like a bad utility maximizer with acting on a side constraint. The defender of side-constraints may be, as many critics of deontology have tried to remind us, short-tongued when it comes to its justification; she may be unable to come up with convincing reasons to support her adoption of this type of rules. But whatever little reason she can come up with, she is not so naive as to put 'promoting the good' on her defense list, as what Broome would have her do here.

IV. CONCLUDING REMARKS

The claim that there are agent-relative consequentialist theories as well as agent-neutral non-consequentialist ones is essentially a denial that agent-neutrality is what uniquely characterizes consequentialism (as the term is used in this paper). What I believe I have shown in this paper is that you cannot take away agent-neutrality from an essentially consequentialist theory without turning it into a theory that is not consequentialist in nature, and that you cannot make agent-neutrality a fundamental feature of a theory without turning it into a consequentialist theory. However, arguing the way I did in this paper is not the only strategy available to the defender of the agent neutrality thesis. An equally forceful way of defending it will be to ask opponents of the thesis the following question: "If agent-neutrality is not what characterizes consequentialism, what else is?" Due to the space limit, I cannot pursue the issue here. However, judging from the difficulties involved in the few proposed alternatives,[x] it does not seem to be an exaggeration to say that the prospect of finding something other than agent-neutrality that uniquely characterizes consequentialist is not very encouraging.

NOTES

* I would like to take this opportunity to thank Professor Bruce Brower for commenting on an earlier draft of this paper, and for the help and encouragement he has given me over the years.

[i] The rise of interest in the notion of agent neutrality in the discussion of consequentialism is largely due to the writings of Derek Parfit, Thomas Nagel, Samuel Scheffler, and Amartya Sen. See especially,

Nagel, *The Possibility of Altruism*, and *The View From Nowhere*; Parfit, *Reasons and Persons*, Oxford, 1984; Sen, “Rights and Agency” (*Philosophy and Public Affairs* 11 (Winter 82)); and Scheffler, *The Rejection of Consequentialism.* Also see David McNaughton and Piers Rawling, “Agent-Relativity and the Doing-Happening Distinction”, *Philosophical Studies 63*: 167-85, 1991; Jonathan Dancy, *Moral Reasons*, ch. 11, Blackwell, 1993.

ii. Although the definition I am giving here is in terms of maximizing, it is meant to include the so-called ‘satisficing consequentialism’. This is not to deny that the two versions of consequentialism are genuinely different; rather, the point is that since the two differ only in their attitudes toward the issue of how the good is to be promoted, not the issue of what kinds of good are to be promoted or how they are to be ranked, little is lost in the current context if we ignore the difference. This will become clear when the two senses in which consequentialism is said to be agent-neutral are introduced below. For an insightful discussion of the difference between maximizing and satisficing, see Michael Slote, “Satisficing Consequentialism”, *Proceedings of the Aristotelean Society* 58 (1984), pp. 139-63. I thank the anonymous readers for this journal for pointing out the importance of making this point clear.

iii. The term ‘consequentialism’ is to be taken to mean, throughout this paper, act-consequentialism.

iv. See *Weighing Goods*, Oxford: Basil Blackwell, 1991. All quotations of this author in this paper are from this book.

v. Replacing ‘promote’ or its cognates will cause (R1) to lose its doer-neutrality feature. The reason is simple: the word ‘do’ says nothing about the consequences of what one does; ‘promote’, on the other hand, is result-oriented. That ‘promote’ cannot be replaced by ‘do’ has been sufficiently shown by John Broome in his criticism of J. Skorupski. See John Broome, "Skorupski on Agent-Neutrality", *Utilitas* Vol. 7, No. 2, November 1995.

vi. David McNaughton and Pier Rawling, "Agent-Relativity and the Doing-Happening Distinction", *Philosophical Studies* 63 (1991), p. 175.

vii. Arthur Kuflik, "A Defense of Common-Sense Morality", *Ethics* 96 (July 1986), p. 798. All quotations of this author in this paper are from this article.

viii. "Weighing Goods: Some Questions and Comments", *Philosophy and Public Affairs* 23, p. 352

9. The formalism used in this section is due to McNaughton and Rawling. The discussion could employ Nagel's analysis introduced in section one. The two are basically equivalent, but McNaughton and Rawling's analysis in terms of rules is more straightforward.

x. Two accounts that I am aware of are Rawls' much criticized suggestion that consequentialism is characterized by the idea that the good is to be specified independently of the right, and John Broome's recent proposal that consequentialism can be defined in purely structural terms. For criticism of Rawls, see Will Kymlicka, "Rawls on Teleology and Deontology," *Philosophy and Public Affairs* 17 (Summer 1988). For criticism of Broome, see the article by Temkin cited in note 8.