Article type : Special Issue

Corresponding author mail id: prailton@umich.edu

Toward a more adequate consequentialism

Peter Railton

Introduction

Philip Pettit has been one of the pioneering figures in the contemporary development of consequentialism toward an approach to ethical theory more adequate to our understanding of ourselves and of our lives together. He has shown how consequentialism has resources for contending with many of the most serious criticisms that have been leveled against it. Two fundamental elements of this development make an appearance in "Three Mistakes". The first element is broadening the base of intrinsic goods in terms of which consequentialism makes its fundamental evaluations of acts

and outcomes. At the outset of "Three Mistakes", Pettit urges that consequentialists should go beyond the hedonic concerns of Benthamite utilitarianism to include among fundamental intrinsic goods friendship, promise-keeping, truth-telling, and the kind of respect for others shown when we impose upon ourselves limits on how we are willing to treat them. So sensible is this recommendation, that pluralism about intrinsic value is now perhaps the majority view among consequentialists. The second element, which is the primary focus of "Three Mistakes", and which builds upon such a pluralistic theory of intrinsic value, is less often fully acknowledged: there are diverse *ways* in which our acts can contribute value to the world, including constitutive and dispositional as well as strictly causal ways of "doing good (and bad)" through our acts.

Pettit offers an elegantly structured way of thinking about the alternative ways of "doing good (and bad)". Most immediately, there is intrinsic value that is realized *in* the act, and not as a separable "effect" of it. For example, when one returns a loaned book loaned one has promised to return, this act does not *cause* one to keep a promise, but partly constitutes the promise-keeping. Secondly, the act of returning a loaned book can be done from a variety of *motives* or *dispositions*, which might include keeping one's reputation intact, but also might include a sense of obligation to keep a promise, or a commitment to a friend, or a feeling of gratitude for the favor the friend has done in loaning it. *Which* motives are involved in the act, Pettit argues, can partially determine the value the act contributes to the world—it matters not just *that* one is keeping a promise, but *why*. In an example of Pettit's, when one does a favor for a friend from a feeling of friendship, as opposed to, say, acting from mere, generic helpfulness or from a desire to induce reciprocation, the act can "bestow friendship as well as favor" (p. 19). Pettit calls this kind of benefit "demanding", presumably because it asks more of the agent performing it—not simply doing the favor, but doing it in a certain spirit. And third, there is the enhancement or enriching of such a "demanding" good when the

doing of the favor is modally robust—"other things being equal, the more robustly you produce the contingently demanding benefit, the more good you do" (p. 28).

While these non-causal ways of contributing value through our actions are often overlooked, Pettit argues, there is nothing in consequentialism's basic motivation that requires this. All of these ways of contributing value through acting count, for Pettit, as "consequences" of the action, in "a perfectly natural sense of consequence" that does not require a causal reading (p. 13). Common sense might bridle at this a bit, and many criticisms of consequentialism arise from ignoring this "perfectly natural sense", but Pettit points out that we can speak without oddness of someone's immunity to a virus as a "consequence" of her possession of certain antibodies, even though the immunity and the possession of antibodies are not separable as cause and effect (p. 13), so let us assume the broadened usage of 'consequence' for the purposes of the present discussion. Moreover, Pettit suggests, by recognizing both the plurality of fundamental intrinsic values and this variety of ways in which our actions can *in themselves* contribute such value, some of the intuitive evidence that has led people to abandon consequentialism, or to introduce into it "agent-centered value", can be accommodated within a "neutralist consequentialism" after all (p. 10).

Moral vs. non-moral fundamental goods

Non-consequentialists often are skeptical of such attempts to enhance the intuitive "fit" of consequentialism by complicating it beyond simple hedonic utilitarianism. They tend

to see these ways of "complicating" consequentialism as in fact diluting consequentialism's claim to constitute a distinctive approach in moral theory.[1]

The challenge for consequentialists is thus to show that their ways of "complicating" their view do not alter its distinctive character. Views differ on what is core to consequentialism, so I cannot proceed on an uncontroversial understanding. My best sense, however, is that the distinctive character of consequentialism as an approach to ethical theory lies in its *explanatory* structure: it seeks to account for the various forms of moral appraisal, such as moral rightness and moral goodness or virtue, in terms of tendencies to realize fundamental, intrinsic *non-moral* good, whether directly or indirectly assessed. Simply moving from a monistic to a pluralistic theory of intrinsic value need do nothing to alter this explanatory priority with respect to moral assessment—so long as the pluralism does not treat as basic any form of *moral* value.

For example, hedonistic act-utilitarian explains the *moral rightness* of an act in terms of its probable or actual contribution to the net realization of pleasure, without regard to whether the pleasure is *morally deserved* or *morally appropriate*. That we have such a notion of the non-moral value of pleasure is easily seen: even retributivists who see the incarceration of criminals as morally deserved and appropriate, agree that incarceration is deserved and appropriate because it is a *bad state* to be in, and therefore a form of

[1] Pettit suggests that *if* hedonic utilitarianism were the only theory available to consequentialists, then they would be unable to avoid the "three mistakes" he has identified. Classical utilitarians, for example, "think the only good is pleasure" and "this is inevitably a causal consequence of what you do" (p. 14, n. 12). But, on "classical" views in aesthetics, when one engages in active aesthetic appreciation of a work of beauty, this is an intentional activity that is partly constituted by the pleasure one is experiencing *in* it—it would not be an act of appreciation without this intrinsic reward, and so the pleasure is not, strictly speaking, a causal consequence of it. Hedonic utilitarians can, I think, count such constitutive contributions in favor of performing such actions.

punishment. The sense in which this morally appropriate state is nonetheless bad is the non-moral sense. The structure of the act-utilitarian's explanation of moral rightness remains the same if we add other intrinsic, non-moral values to the evaluation of actual or probable consequences, and similar remarks hold for rule- and motive-utilitarian explanations of moral rightness.

However, critics will have a point if would-be consequentialists seek better fit with moral intuitions by adding *moral* goods to the plurality of intrinsic values used for assessing consequences. Unless these moral goods are in turn explained in terms of intrinsic non-moral value, the explanatory structure of consequentialism has been compromised at a fundamental level. Thus I worry about Pettit's suggestion that, among the intrinsic goods consequentialism should recognize are "keeping promises made, satisfying another's trust, and telling the truth, as well as goods like respect, honesty and friendship" (p. 2). Are all of these genuinely non-moral goods?

How might we tell? John Stuart Mill glossed the notion of "non-moral value" as potential aims or "pleasures" toward which "all or almost all" of those with wide experience would have "a decided preference, irrespective of any feeling of moral obligation to prefer [them]" (Mill, 1962, p. 259). It seems plausible that *friendship* would be such a good—preferred as such, independently of any notion of distinctively moral obligation (for example, it is sometimes said that the Ancient Greeks lacked the notion of distinctively moral obligation, yet friendship was for them a central component of a good life in its own right). But what about promise-keeping and truth-telling? Consider a promise to keep silent, made earnestly and solemnly by a recent college graduate who has learned that her close friend is taking drugs. If she steadfastly remains silent, even as her friend spirals down into addiction and overdose, it seems unlikely that we'd say her keeping this promise was, in itself, at least one good thing about these events. It is not, I think, that keeping the promise is an intrinsic good outweighed by the harms of

doing so. Of course, keeping a promise is often a manifestation of such goods as friendship, loyalty, gratitude, or integrity, so the keeping of a promise can in a given instance contribute these goods to the world. But keeping a promise when friendship, loyalty, gratitude, and integrity all lie on the side of breaking it—is that something that still holds good in itself?

Surely, one can protest, such a promise not to reveal the drug-taking is not binding, or unconscionable, nullifying the promise. But such normative delegitimation or loss of bindingness seem to me to be ethical notions rather than ideas of intrinsic value. Recall the criminal suffering incarceration—even when the normal constraints against holding a person against his will, or imposing pain upon another, are nullified by his commission of a crime, so that it is *not* obligatory not to constrain him and *legitimate* that he suffer the pain of imprisonment, still, the intrinsic (dis)value of the incarceration remains the same. Similarly, the lack of an obligation to keep a promise or the legitimacy of breaking it should not alter the intrinsic value of promise-keeping. In a later discussion of such goods, Pettit speaks of conditions in which not keeping a promise or telling the truth is *excused* (pp. 17-18), but excuse, too, seems to be a moral or ethical notion. The holding of a criminal against his will is excused, but this is still a state it is bad to be in—otherwise it wouldn't be punishment.

If this line of thought is right, then it *would* dilute the consequentialist character of a view to admit promise-keeping or truth-telling as fundamental intrinsic values. Moreover, consequentialism seems to be well-poised as an ethical theory to *explain* why promising is a ubiquitous feature of human societies, even though promise-keeping as such is not a basic human good like happiness, friendship, trust, or affiliation. For example, consequentialists can explain both how effective an institution of promise-keeping can be in fostering these fundamental human goods, and also why promise-keeping as an institution is typically hedged about with so many limitations and

restrictions, to exclude practices of promise-keeping that would be antithetical to promoting fundamental human goods. Moreover, consequentialism can explain why promise-keeping practices take a variety of forms and degrees of importance in a diverse array of societies, as a reflection the needs, possibilities, and modes of interaction in such societies—no one form is intrinsically best. Perhaps similar points can be made concerning truth-telling and respect (understood as the acceptance of certain limitations on how we affect others).

So, if we are to prevent our ways of "complicating" consequentialism from leading to a consequentialism that isn't, we must restrict the dimensions of fundamental evaluation to those intrinsic goods that are plausibly non-moral. Non-consequentialists are right to worry that consequentialists who seek to achieve a better "fit" with intuition by exercising too free a hand in admitting types of fundamental intrinsic good may undermine, rather than shore up, consequentialism as a distinctive approach to moral theory.

Ways of adding value

Let us turn to Pettit's account of different ways of realizing value, and of "demanding goods", which forms the main subject of "Three Mistakes". This account is reminiscent of Aristotle's discussion of virtuous action. On a standard interpretation, Aristotle recognizes as virtuous action that is done (1) non-accidentally, (2) "at the right time, in the right way, with the right feeling, and toward the right end", *and* (3) from a stable disposition of character (*Nicomachean Ethics*, Books 3 and 4). Venerable as it is,

however, this account of virtuous action has been challenged by arguing that, while (1) and (2) are important for showing that the agent is indeed aptly *reasons-responsive* in acting, (3) seems to have a different role—it is relevant to evaluating the virtue of the *agent*, but not of the *action* as such. One might raise a similar question about Pettit's view that the counterfactual robustness of the disposition from which one acts is partly constitutive of the value realized *in* that very action.

Consider the following kind of case. Suppose that I have asked a friend to look in on my pet schnauzer while I am away for a day. To simplify things, consider only those cases where looking in on my schnauzer involves nothing morally objectionable in itself and does not conflict with any moral duties my friend independently has. When my friend follows through, and does look in on Frida, a favor is done for me. If, moreover, he does so out of the motives of friendship and concern for Frida, a good is *realized* that is greater than just a favor. This latter good is, I believe, what Pettit would call a "demanding good", since it depends not only upon the benefit done, but also whether my friend performs the act "out of a suitable disposition". Acting from suitable dispositions is a form of *reasons-responsiveness* on my friend's part:

> Absent excusing obstacles [or indications that it is "trumped by reasons of a manifestly weightier kind", my friend acts] … out of responsiveness to the reasons that argue for giving respect, for being honest or for acting as a friend. [p. 18]

Moreover, the resulting action will realize a "robustly demanding good" if this reasons-responsiveness exhibits a certain kind of *counterfactual resilience*:

> … it is not only the case that [my friend] would provide the restraint or truth-telling or favor in actual circumstances; [he] would provide it also under a range of suitably varying, counterfactual scenarios. [p. 18]

Now let's consider a particular friend, Alec. Alec is a close friend and a lovely person, but reliability is not his strength. Not because he is devious, deceitful, selfish, or lazy, but because he is *un peu dans la lune*—spacey, preoccupied, and prone to lose track of things. I wouldn't normally ask Alec to look in on my dog, but in a particular case I happen to have no good alternative. So, when I leave, I carefully set out enough food and water that Frida will not be hungry or thirsty even if Alec loses track of the date and forgets to look in. As it happens, the precaution wasn't necessary—Alec does keep the date straight, and, out of friendship and concern for Frida's well-being, he looks in on her, and, tender-hearted soul that he is, Alec hears the somewhat forlorn tone in Frida's whimper, and gives her a much-appreciated backrub to cheer her up.

Suppose that we compare the value realized in this act of looking in on Frida, including noticing that she is lonely, rubbing her back, etc., with the value realized when my much less spacey friend Beryl looked in on Frida several weeks ago. Beryl acts from the same dispositions of friendship and concern, present in equal intensity, and equally notices Frida's loneliness, rubbing her back generously. Is the value realized *in* Beryl's act of looking in any greater than the value realized in Alec's? True, Beryl would "provide the favor" to me in a much wider range of counterfactual circumstances than Alec, since Beryl is good remembering things, keeping track of dates, giving himself reliable reminders, and so on. And it's also true that Beryl's reliability makes his friendship useful—and therefore valuable at least instrumentally—in a number of ways that Alec's is not. But does this translate into any greater value when Beryl looks in on Frida than when Alec does—is this a value to be found in the individual act itself? For example, to shift the way of putting it, my forgetful friend Alec's particular act of looking in seems to be just as *reasons-responsive* as my date-conscious friend Beryl—each performs the right act "at the right time, in the right way, with the right feeling, and toward the right end", without any admixture of "suitable" motives. Moreover, even though Alec is less

reliable than Beryl, when he does look in on Frida, he does not do so *unintentionally* or *accidentally*.

Note that the control theory of intentional action, which Pettit very plausibly defends as part of his account of correcting for errors about doing good, enables us to put this point a bit more precisely. For both Alec and Beryl, the act of looking in Frida is performed "under the control" of the "guiding dispositions" of friendship and concern—they model the situation in terms of an *evaluative representation* in which friendship and concern make the act salient and eligible, guide its shape, and supply the motivation through which it is performed.[2] As they act they implicitly attend to feedback as to whether their behavior is actually realizing these values in the circumstances—which partially explains why both notice Frida's loneliness and adjust their behavior to provide some affectionate contact. Thus a "feedforward-feedback" control structure is in place, centered around the value each attributes to friendship and to Frida's welfare. Whether all this is operative in their behavior does not depend, however, on whether, in some alternate circumstance, the same kind of act would be performed. To be sure, control structures typically have counterfactual resilience, so that such resilience often serves as *evidence* of the presence of control. But one lesson that might be drawn from the literature on the "Principle of Alternate Possibilities" is that a behavior can be controlled actively by an agent, and the agent can therefore responsible for it, even if a small change in circumstances would have led to control being pre-empted or lost. Or, to draw upon the Pettit's example of temperature regulation, a thermostat can be controlling the flow of current into a heating unit via active feedback from whether the temperature is at a given set-point, and thereby controlling the current temperature in the room, even if the thermostat is fragile and would fail in response to relatively minor changes in ambient temperature or line voltage. Of course, such a fallible device would

[2] For discussion of action under the control of an evaluative representation, see Railton (2017).

not be very good for the purposes we usually have in mind in installing a thermostat, and so we would not say that such a fragile device is a good thermostat. But fragile thermostats on the verge of failure can still control temperature, and do so "in response to" the relevant "reasons". In the high-precision environment of a scientific experiment, devices or processes can be used to control an important magnitude, even though their operation depends upon maintaining within very strict limits a complex conjunction of experimental conditions.

To return to the question of control in intentional action, let's look at a different kind of case. Suppose that Nancy is naturally timid, and easily put upon. She has a hard time defending her own interests, and tends to flinch when a powerful person tries to intimidate her. When she is subject to sexual harassment at the workplace, she is afraid to discuss this with anyone. Fortunately for her, her boss is promoted and moves to a branch office in another town. Indeed, her former boss rises rapidly, and becomes a well-known personality in the business world and public life. She has spoken of the harassment only once, to a workmate at the time who has since moved to the Sunbelt and with whom she has lost touch. Nancy has been silent as she observed her former boss's climb to renown. But one day, out of the blue, a reporter calls. Nancy's name had been given to the reporter by her former workmate, who had later herself experienced harassment by the same boss after he "took her with him" to his new position after his first promotion. The reporter has also contacted three or four other women who have had similar experiences with this boss, and who are thinking about allowing the reporter to publish their stories if enough women will come forward simultaneously. "Would you be willing to join them?" the reporter asks. Nancy is terrified by the thought—what would her husband, children, or friends think? "No," she replies, "I really can't." The reporter says, "Take your time, I know this wouldn't be easy for you. Most of the women I've contacted have been reluctant at first, but they also felt

they couldn't allow this man to continue to get away with this kind of harassment and intimidation. So they've agreed to step forward if enough others will join them. I'll be back in touch in a little while and you can tell me what you think. No pressure—only do this if you feel it's right for you. Here's my number if you want to reach me."

Nancy puts down the phone, trembling. She doesn't know what to think. She had hoped this was all behind her, forgotten and buried in time. But slowly, she finds a feeling of resolve growing—if the others can do it, so can she. When the reporter calls back, she finds herself saying, with great difficulty, "Yes, I'll join the others."

Now it seems to me that Nancy acts "at the right time, in the right way, with the right feeling, and toward the right end"—the evaluative model of the situation that is guiding and motivating Nancy's behavior assigns to the act of joining the other women the values of justice and compassion, so that her eventual agreement is under the control of "suitable dispositions". However, her act would not have occurred without an unprecedented conjunction of events. Should we say that Nancy's act therefore is lower in such "demanding goods" as justice, courage, and solidarity than the action of other women less timid than her?

If this is right, then we might understand the "more-robustness principles" Pettit develops in a different way. The value of a friend's favor, or of an act like Nancy's, might be greater *not* when more counterfactually robust, but when more fully attuned to the reasons present in the situation of acting—just as an instance of *appreciating* aesthetic or moral value might have greater value in itself the more attuned it is to the relevant value-making features, even if experiencing this kind of appreciation is not a stable disposition of the individual's character. Kant writes in the *Critique of Practical Reason*:

> … before a humble common man in whom I perceive uprightness of character in a higher degree than I am aware of in myself *my spirit bows*, whether I want it or

> whether I do not and hold my head ever so high that he may not overlook my superior position. [Kant, 1996, p. 202; 5:77]

The gentleman who holds his head "ever so high" in relation to those who are humble and common does not have a stable disposition to show this kind of respect for them, yet in this instance of mental "bowing" Kant finds a paradigm case of mental action that is responsive to reasons of moral worth, since it is not tinged with any "self-conceit" or self-interest—on the contrary. Kant calls such respect "attunement", akin, he writes, to the kind of attunement found in aesthetic appreciation.

This idea of attunement makes for a symmetry of sorts between action and appreciation—the core of the value realized lies in the extent to which the value-constituting features are fully and purely in play in the shaping of the act or appreciative experience themselves, however much we might care for other reasons about whether there exist stable dispositions to bring about such actions or appreciative states.[3]

Pettit's way of talking about the special contribution made by counterfactual robustness at one point suggests that its contribution might lie elsewhere than in deepening the value contributed by reasons-responsiveness. He writes:

> There is no plausible way of measuring the graded or range good that you bring about as you increase the robustness with which you generate such a benefit; in that sense, there is no discrete consequence associated with your action. But it is plausible that, other things being equal, the more robustly you produce the contingently demanding benefit, the more good you do. By producing the benefit more robustly, you provide me with more robust access to enjoying it.

[3] Christopher Bennett has suggested to me that there is an analogy between these considerations and internalist vs. externalist views of justification in epistemology—is justification an external matter of reliability across counterfactual situations or an internal matter of attunement to rational considerations?

> And you provide me with such enhanced access in virtue of the disposition out of which you act … . [p. 28]

Having more robust *access* to a good is certainly something worth caring about. But it does not sound to me like the enlarging, deepening, or "enriching" the good itself—by contrast, greater attunement to relevant reasons, less admixed with "unsuitable dispositions" does seem to enhance or enrich the value that is realized *in* an act.

Pettit is certainly right, I think, that the dispositions from which an act is performed can contribute to its value, and that understanding how this works brings us to the notion of action being under the control of a guiding disposition, and not merely a causal upshot of it. My modest amendment would treat counterfactual robustness as evidence of the qualities of dispositions or agents, and locate the enriching of the value of action as such in the extent of reasons-responsiveness or attunement.

This might be an amendment to the account Pettit welcomes, and it might even already be present in his account in those places where he develops it more fully. Certainly, it is no objection at all to his core point that consequentialists should not only be pluralists about the kinds of value, but also about the ways of realizing value. This, it seems to me, is right and important, yet frustratingly often ignored. What matters to consequentialists, it seems to me, should be how well or ill the sentient world fares, not whether these goods or harms are caused or constituted by the actions or dispositions in virtue of which they come into being. Consequentialists are free to use the term 'consequence' in this broad way, so long as they attend to the overall explanatory structure of their view, and to the challenge of identifying correctly the nexus where value is constituted. The result might leave their views somewhat less close to intuitive

deontologies than they might have hoped, but perhaps afford them the right amount of distance from such views to be able to explain them, rather than presupposing them.[4]

References:

All unattributed page numbers refer to Philip Pettit, "Three Mistakes about Doing Good (and Bad)", this volume.

Kant, Immanuel (1996), *Critique of Practical Reason*, trans. and ed. by M.J. Gregor, Cambridge: Cambridge University Press.

Mill, John Stuart (1962), *Utilitarianism and Other Essays*, ed. by M. Warnock, New York: Meridian.

Railton, Peter (2017), "At the core of our capacity to act for a reason: The affective system and evaluative model-based learning and control", *Emotion Review*, *9*, 335-342. DOI: 10.1177/1754073916670021.

[4] I am grateful to Christopher Bennett for very helpful comments on an earlier draft of this paper.