# Parfit On What's Wrong

*By Thomas W. Pogge*

THIS PAPER COMMENTS ON DEREK PARFIT'S SECOND AND THIRD TANNER Lectures,[1] in which he discusses a dazzling array of moral formulas. Parfit treats these as competing formulas. But before we can appreciate his claims about winners and losers, we must first understand what this competition is about: What role are all these formulas meant to play? By reference to which task are we to judge their success or failure?

All formulas canvassed by Parfit substantially involve the noun or verb "act." In the second Lecture, most of the formulas also involve the adjective "wrong." Here, most formulas are criteria for judging which acts are wrong or not wrong, or about how it is wrong or not wrong to act. In the third Lecture, most of the formulas also involve the verb "ought." Here most formulas are criteria for judging which acts one ought or ought not to perform, about how one ought or ought not to act. Because Parfit does not say otherwise, we should assume that he takes the noun and verb phrasings involving "act" to be equivalent, and that he also takes "ought not" and "wrong"—and (one might add) "impermissible"—as coextensive binary predicates. An act is wrong just in case it is impermissible and just in case one ought not to perform it. And one ought to perform an act just in case it is wrong or impermissible not to perform it. We see this presupposed coextensiveness at work when Parfit tells us (338-9) that the Formula of Universally Willed Moral Beliefs (Formula 12)—"An act is wrong unless everyone could rationally will it

*Since receiving his Ph.D. in philosophy from Harvard, Thomas W. Pogge has been teaching moral and political philosophy at Columbia University. His recent publications include* World Poverty and Human Rights *(Polity Press 2002),* Global Justice *(edited, Blackwell 2001),* "What We Can Reasonably Reject" *(NOÛS 2002),* "Can the Capability Approach be Justified?" *(Philosophical Topics 2002),* "On the Site of Distributive Justice" *(Philosophy and Public Affairs 2000) and, with Sanjay Reddy,* "How Not to Count the Poor" *(www.socialanalysis.org). Pogge is editor for social and political philosophy for the Stanford Encyclopedia of Philosophy and a member of the Norwegian Academy of Science. His work was supported, most recently, by the John D. and Catherine T. MacArthur Foundation, the Princeton Institute for Advanced Study, All Souls College (Oxford), and the National Institutes of Health (Bethesda). He is currently a research fellow at the Center for Applied Philosophy and Public Ethics (CAPPE) at ANU, Canberra.*

to be true that everyone believes such acts to be permissible" (338)—can be restated as Kant's Contractualist Formula—"We ought to act on the principles whose universal acceptance everyone could rationally will" (339). In making this assertion of equivalence, Parfit is surely assuming that the formulations "an act is wrong unless..." and "we ought to act ..." are both leading up to identifying a property of acts whose absence makes it the case that the act is wrong, ought not to be performed.

The formulas Parfit canvasses clearly tell us something about when an act is wrong, is impermissible, or ought not to be performed. It is less clear whether they also tell us when an act is not-wrong (that is, right). The fact that an act lacks a property whose presence would make it wrong is compatible with this act being wrong in some other way. Even where the formulas Parfit presents are ambiguous on this point, his discussion makes clear, I believe, that he takes the formulas to give sufficient and necessary conditions for the wrongness of acts. So I read all the formulas canvassed as complete (in this sense) criteria for the wrongness of acts.

Acts here are by Parfit understood as act tokens, such as particular movements a person intentionally performs, or intentionally fails to perform, with her body at some particular time and place. It is notorious that, before acts can be judged by any of the formulas, they must be individuated. If we do not know how to do this, then we do not know how to apply any of the competing candidate criteria.[2] But, since none of the canvassed candidate criteria provides any hint as to how to solve this problem and since Parfit says nothing about it either, I will skip it here and pretend that acts are clearly and uncontroversially individuated for us.

All the candidate criteria Parfit canvasses judge act tokens on the basis of some type they belong to. This poses another notorious problem: Under what description(s) is a given act to be judged? Just as one given act type may be instantiated in indefinitely many act tokens, so one given act token may instantiate indefinitely many act types. In order to judge a token by its type we thus need to know which type. We must be able to identify correctly the type or types on the basis of which the given token is to be judged. The problem is clear when one looks at the 13 candidate criteria Parfit distinguishes in his diagram (336). All these formulas involve references to people acting in this way, or to what people believe about the permissibility of such acts. All these criteria are therefore quite meaningless unless we have additional instructions about how to identify the types that are to inform our judgment about the act tokens under examination.

To get a taste of the difficulty, consider Parfit's examination of Formula 11: "An act is wrong unless everyone could rationally will that everyone acts in this way" (337). Parfit quickly dismisses this formula by pointing out that "Kant did not act wrongly ... in having no children" (337). But this seems too quick. Let us grant that not everyone can

rationally will that everyone act on the maxim of remaining childless irrespective of circumstances. But does it thereby forbid Kant's childlessness? This does not follow, because it is presumably also true that everyone can rationally will that everyone act on the maxim of remaining childless whenever this is his or her preference and the human population is either large or increasing. In order to tell whether Formula 11 does or does not forbid Kant's childlessness, we must first know which type instantiated by Kant's conduct is the relevant type, referred to by "in this way." Parfit proceeds as if he has an answer to this question, but he does not tell us what this answer is nor, more importantly, how he identified this right answer from among indefinitely many possibilities.

Looking through the whole text, we find some formulas that address this problem. Four distinct approaches are exemplified, though Parfit seems unaware of the distinction. Approach One invokes the descriptions under which the agent herself is intentionally acting. Thus, one of the formulas (named RLN) states that an act is wrong unless the agent could rationally will that everyone does whatever, in acting in this way, she would be intentionally doing (328). To make this formulation mean anything, more needs to be said. In performing some particular act, agents often have several aims in mind as things they are trying to achieve or trying to avoid. Are all these aims relevant, or are further instructions forthcoming about how this list of aims is to be whittled down? And once we have identified the relevant aims: For the act to escape wrongness, must the agent be able rationally to will that all her intentional aims be pursued by everyone, that at least one of her aims be pursued by everyone, that everyone pursue at least one of her aims, or what?

Parfit takes Approach One to be Kant's. But Kant had something quite different in mind when he made the notion of a maxim central to his moral philosophy. I have written elsewhere about Kant's view and should not restate my reading here.[3] But perhaps four short paragraphs are in order to bring out one main contrast between Parfit's reading of Kant and mine.[4]

Parfit seeks a criterion for the wrongness of act tokens which invokes a criterion for the assessment of act types in some subsidiary role. Parfit believes that Kant is pursuing the same project. But this is not so. When Kant formulates the Categorical Imperative, he is not interested in Parfit's problem: the moral assessment of act tokens. Rather, Kant is interested in the moral assessment of act types or, more precisely, of agents' maxims. The Categorical Imperative is a criterion for the permissibility of maxims, and Kant intends this criterion to play a subsidiary role in the assessment of character ("good will")—not in the assessment of act tokens.

In addition, Parfit mistakenly assumes that maxims in Kant's sense are intermediate moral principles. Witness what Parfit calls Kant's Contractualist Formula (339, cited above). The formulas Kant provides

do not deal in intermediate moral principles pronouncing on the wrongness or permissibility of act tokens. Instead, they deal in maxims, which Kant defines as subjective principles of volition or of action—that is, as personal conduct-guiding policies.[5]

So, when Kant says that it is wrong, or rather that we ought not, to act on a certain maxim, he means that it is wrong to have and wrong to act on (remain committed to) this (impermissible) maxim. From this it does not follow that each act performed pursuant to this maxim is wrong. Parfit is quite right to say (297-8) that a gangster is not performing a wrong act when he pays for his coffee merely because doing so is less trouble than stealing it. But this is no criticism of Kant. For when Kant holds that such a gangster acts wrongly he means not that her act (token) is wrong but that her maxim, and her acting on this maxim, is. In fact, Kant offers the shopkeeper example[6] to make just the point Parfit is making with his gangster example. The shopkeeper is acting according to duty: Her act tokens are permissible and so she is not acting wrongly in Parfit's sense. But the shopkeeper fails to act from duty: She is acting wrongly in Kant's sense (in violation of the Categorical Imperative), because it is impermissible to act on the maxim of unconstrained profit maximization. The shopkeeper and gangster cases illustrate Kant's point that conduct can be both right (token) and wrong (type)—that an agent performing permissible act tokens may be acting rightly or wrongly in Kant's sense, depending on the actual maxim of her conduct.[7]

To be sure, Kant held beliefs not only about his questions: "When is a maxim morally wrong?" and "When does a person have a good will?", but also about Parfit's question: "When is an act token morally wrong?". But Kant does not provide a clear path from the first to the last question. The path cannot be this: An act token is morally wrong just in case it is performed on an impermissible maxim. The shopkeeper and gangster examples refute this. The path must be something like this: An act token is wrong (contrary to duty) just in case any maxim on which it might be performed is impermissible.[8] Let us call this Approach Two. None of the criteria Parfit considers is of this kind. But my interest here is in Parfit, not Kant. So I will not try to develop Kant's answer to Parfit's question about when act tokens are wrong.

Ending the digression, let us proceed to the next approach to judging act tokens through a criterion that invokes a subsidiary criterion for the assessment of act types. This approach effects the binary sorting of act tokens via a binary sorting not of act descriptions, nor of maxims, but of intermediate moral principles.[9] Each such moral principle defines a certain type of act and then declares such acts to be right or to be wrong. Of course, there are indefinitely many such principles, often mutually inconsistent. Intermediate moral principles can nonetheless help us achieve a binary sorting of act tokens, provided two conditions are satisfied:

We can tell of at least some of the intermediate moral principles

that they are valid.

The set of valid intermediate moral principles is consistent, so that no act token is judged wrong by one valid principle and also judged right by another valid principle.

If the binary sorting is to extend to all act tokens, then a third condition must be satisfied:

For each act token, there is at least one intermediate moral principle that is both known to be valid and applicable to that act token (entailing either that it is wrong or that it is right).

Needed for this approach to work is a subsidiary criterion for judging the validity of intermediate moral principles. The formulas Parfit canvasses in his third Lecture are meant to fulfill this role. It is worth noting that when he discusses any candidate formula for this role, he ignores the question of whether this formula satisfies conditions 2 and 3.

Yet Parfit may nonetheless be addressing this question indirectly. For many of the formulas he considers speak of "principles" in the plural. One candidate formula, for instance, declares valid "the principles whose universal acceptance everyone could rationally choose" (361). This formulation is ambiguous between a distributive and a collective use of the plural, and the present approach thus splits into two. Approach Three embraces the distributive interpretation: Each intermediate moral principle is tested individually and independently from the others to determine its rational choosability. The winning principles are then conjoined into a set about which one must ask whether it satisfies conditions 2 and 3. Approach Four embraces the collective interpretation: Whole candidate sets of intermediate moral principles are tested for rational choosability.[10] Here one might well lay down from the start that a set of principles is rationally choosable only if it satisfies condition 2 and perhaps 3 as well. We may call any set of intermediate moral principles that satisfies 2 a moral code and any set of such principles that satisfies 2 and 3 a complete moral code.

Approach Three runs into a great problem: It is very hard to show that all winning principles are mutually consistent (condition 2) and form a complete set (condition 3). Approach Four also runs into great problems: Moral codes are most unwieldy entities—quite tedious to specify in detail and also quite difficult to assess (for rational choosability or whatever determines their validity). Moreover, there is also the problem of uniqueness. It seems highly unlikely that there should be only a single rationally choosable moral code. And this may spell trouble when persons who adopt different valid codes interact in the same world. The fact that each valid moral code is internally consistent does not guarantee that valid moral codes are mutually harmonious.

But perhaps this problem with Approach Four can be turned to advantage. Consider how Parfit criticizes Kant for giving the wrong answer on tyrannicide—holding that, pace Kant, it would have been

permissible to assassinate Hitler during the Second World War (321). To be sure, had all Germans believed this to be permissible, Hitler would have been on his guard—no assassination attempt would have succeeded and the Nazis would have been an even greater menace. But Parfit declares this fact irrelevant. He is thereby assuming, in effect, that it is bad if all Germans take tyrannicide to be impermissible, that it is even worse if they all take tyrannicide to be permissible, and that it is best if tyrannicide is taken to be impermissible by a great majority and taken to be permissible by a small clever minority. But how can a morality deliver this result? How can one morality tell its adherents different things about what they may and must not do in identical circumstances? Parfit gives no formula that even attempts to solve this problem which he deems fatal to Kant's view. The trick might be accomplished by a move Parfit does not consider. This move builds on Approach Four in that relevant types of acts are defined by intermediate moral principles which are assessed collectively, as moral codes. The innovation is to construct formulas whose instruction to the agent about which moral code she should follow involves essential reference to the moral codes of other agents. This innovation replicates the conditionalization move I made earlier to defend Kant's childlessness against condemnation by Formula 11. Just as Kant might have acted from a maxim that makes his preferred childlessness conditional upon the actual maxims and conduct of others, so a plausible criterion of wrongness might permit a German to follow a moral code permitting Hitler's assassination just in case the vast majority of Germans follow a moral code forbidding Hitler's assassination.

I lack the space to present or defend a formula that exemplifies this variant of Approach Four. But it deserves study, I believe. It is important that persons choose different professions. So the question, "Which profession is it best for everyone to choose?" starts us off in the wrong direction. If it is desirable that agents follow diverse moral codes, then the question, "Which moral code should everyone follow?" is similarly misguiding.

Let us take stock. I have identified Parfit's project as that of classifying act tokens as either right (permissible) or wrong (impermissible). After pointing out that Parfit fails to address the individuation of act tokens, I have outlined four distinct approaches to his project. Approach One classifies an act token on the basis of a subsidiary criterion that applies to the descriptions under which the agent is intentionally acting. Approach Two, Kant's, classifies an act token on the basis of a subsidiary criterion that applies to the maxims on which agents might perform this act. Approach Three classifies an act token on the basis of a subsidiary criterion that applies to intermediate moral principles permitting or forbidding this act. Approach Four classifies an act token on the basis of a subsidiary criterion that applies to moral codes permitting or forbidding this act. These four

approaches to the classification of act tokens as right or wrong are quite different from one another. To be successful, Parfit's discussion needs to bring out these differences, or so I believe.

I conclude with a final reflection on the question of the range of the sought criterion for sorting act tokens into those that are wrong and those that are not wrong. Is this criterion meant to apply (a) to all acts by all agents at all times in all possible worlds, or (b) to the acts merely of human beings, or (c) only to the acts of human beings living under a just legal order, or (d) solely to the acts of humans living under a just legal order in a world whose agents all comply with the same intermediate moral principles... or what?

As far as I can tell, Parfit has not attended to this question and has different answers in mind at different times. (It is interesting to observe that—starting around page 328—his wording of the formulas he considers switches from "it is wrong" type formulations to "our act is wrong" or "we ought to" formulations. The use of the first person plural suggests that Parfit is here beginning to think not in terms of what code any one agent should follow, given the actual conduct of the others, but in terms of what code all agents should follow.) This unclarity is unfortunate, because the question is of great importance. If the range of a formula is (d), or even (c), then, even if correct, it is of no use in the world we inhabit. In this world, we absolutely need a morality that guides us plausibly to adjust our conduct to existing imperfect social institutions and to the conduct of other agents—those who share our morality, those who follow different moralities, and those who are amoral or immoral. φ

## *Notes*

This paper was first presented at a Rutgers University conference (April 2003) which, honoring Derek Parfit on the occasion of his 60th birthday, was entirely devoted to his Tanner Lectures. Larry Temkin organized this memorable and philosophically very productive event. I have reworked my paper so as to accommodate changes that Parfit has made before the publication of his lectures. In doing so, I have greatly benefited from discussions with Rüdiger Bittner and especially Sam Kerstein.

[1] All page references in simple parentheses are to these lectures, entitled "What We Could Rationally Will," as printed in The Tanner Lectures on Human Values XXIV, ed. Grethe Peterson (Salt Lake City: University of Utah Press, 2004), 285-369.

[2] For a brief discussion, see my "What We Can Reasonably Reject" in NOÛS Philosophical Issues 11 (2001): 118-147, Section II, the "first problem."

[3] Compare my "The Categorical Imperative" in Grundlegung zur Metaphysik der Sitten, ed. Otfried Höffe. Ein kooperativer Kommentar (Frankfurt: Vittorio Klostermann 1989), 172-193; reprinted with revisions in Kant's Groundwork of the Metaphysics of Morals, ed. Paul Guyer (Totowa: Rowman and Littlefield, 1998), 189-213.

[4] There are two other major ways in which my reading is at variance with Parfit's. First, I believe that when Kant stresses the equivalency of his formulas (Kant, Immanuel, Grundlegung zur Metaphysik der Sitten (Preußische Akademieausgabe IV), 436), he

is not making an assertion, which can be easily set aside as implausible, but issuing a prescription: The various formulas make distinctive contributions to the clarification and specification of the Categorical Imperative—they gradually enrich its meaning, until at last its full import can be understood. Once fully understood, the Categorical Imperative can then be read back into each of these formulas so as to make them equivalent as Kant demands. Second, I think Parfit departs from Kant by plugging into Kant's formulas his (Parfit's) own account of what one can rationally will or want. (Unlike Kant, Parfit does not distinguish these expression from each other or indeed from what one "could rationally share" (292, 306), "could rationally consent to" (292-5, 298-301, 312-14, 337-8, 352, 359), "could rationally choose" (293-5, 338, 348ff), "to whose acceptance it would be rational to agree" (339, 348).) This is distorting insofar as Kant—especially in the discussion of his second formula—provides his own elaborate account of what a rational being must will and cannot will. Still, in this brief comment, I want to focus on the merits of Parfit's discussion of the many formulas he considers, not on how Kant's view is different from all of them.

[5] Kant, Immanuel, Grundlegung zur Metaphysik der Sitten, op.cit. note 4, 400n, 420n.

[6] Kant, Immanuel, Grundlegung zur Metaphysik der Sitten, op.cit. note 4, 397. Sam Kerstein has forcefully argued that Kant thinks of the shopkeeper as acting on a permissible maxim. If he were right, I wound need to find other evidence to support against Parfit my claim that Kant understood that a person acting on an impermissible maxim may yet produce permissible act tokens.

[7] Likewise, in remaining childless, Kant himself acted rightly (token) and either rightly or wrongly (type)—for instance, rightly on the maxim "to remain childless whenever this is my preference and the human population is either large or increasing," or wrongly on the maxim "to remain childless irrespective of circumstances."

[8] We see here how very hard it would be to show what Kant, at times, seems to have believed—that all act tokens that involve lying are wrong. To show this, one would have to show the impermissibility of each and every maxim pursuant to which certain lies are to be performed under certain conditions. Many of these indefinitely many possible maxims would not even mention lying.

[9] Parfit associates Rawls with this approach. Rawls did indeed make two brief remarks about "rightness as fairness" in his A Theory of Justice (Cambridge MA: Harvard University Press, 1999 [1971]), 15 and 95f. Such a view was worked out by David A. J. Richards, A Theory of Reasons for Action (Oxford: Oxford University Press, 1971). But Rawls repudiated the idea later, for example in his Justice as Fairness: A Restatement (Cambridge MA: Harvard University Press, 2001), 186-8. I should add that Parfit also makes deeply mistaken assumptions about how such a view would work when he writes, "Rawls... tells us to suppose that, when we were choosing moral principles, everyone's main aim would be to promote their own well-being" (342-3). I believe this mistake is due to an isolated reading of section 27 of A Theory of Justice, where Rawls is sketching not his own view, but a contractualist justification of average utilitarianism. Rawls's own view is different in that the parties in the original position are given to know that those they represent have three higher-order interests—roughly, to develop and exercise their capacities for a sense of justice and a conception of the good and to be successful in the pursuit of the particular conception of the good they have chosen (whose content is not known in the original position). See John Rawls, Political Liberalism (New York: Columbia University Press, 1996 [1993]) 74, cf. 19.

[10] While Parfit is—intentionally or inadvertently—ambiguous, Scanlon embraces both possibilities. His book provides exactly two full formulations of his "general criterion of wrongness" (T. M. Scanlon, What We Owe to Each Other (Cambridge MA, Harvard University Press, 1998), 11). The first holds that "an act is wrong if and only if any principle that permitted it would be one that could reasonably be rejected" (ibid., 4). Later he states his criterion as "an act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behavior that no one could reasonably reject" (ibid., 153).