Mark Budolfson[1]

Why Morality and Other Forms of Normativity are Sometimes Dramatically Directly Collectively Self-Defeating[2]

In a prisoner's dilemma, if everyone follows the strategy of self-interest, then everyone is certain to be worse off from the perspective of self-interest than if everyone had not followed self-interest instead. This shows that self-interest is sometimes directly collectively self-defeating, because it shows that sometimes everyone has all the relevant information, correctly follows self-interest, but thereby ends up worse off from the perspective of self-interest than they would have been if they had all followed some other antecedently identifiable strategy instead. In Reasons and Persons and On What Matters, Derek Parfit argues that it is a constraint on any plausible moral theory that morality must never be directly collectively self-defeating, and he claims that the most plausible versions of consequentialism, contractualism, and Kantian ethics all imply that morality is never directly collectively self-defeating. Some theorists not only agree with Parfit that morality can never be directly collectively self-defeating, but also believe that rationality and other forms of normativity can never have that property either. I argue against these theorists, with examples that show that morality and all other interesting forms of normativity are sometimes directly collectively self-defeating.

[1] Department of Philosophy, Department of Geography and the Environment, Population Wellbeing Initiative, University of Texas at Austin, mark.budolfson@austin.utexas.edu.

[2] Funding from Riksbankens Jubileumsfond in support of grant number: P22-0662 is gratefully acknowledged.

In a prisoner's dilemma, if everyone follows the strategy of self-interest, then everyone is certain to be worse off from the perspective of self-interest than if everyone had not followed self-interest instead. This shows that self-interest is sometimes *directly collectively self-defeating*, because it shows that sometimes everyone has all the relevant information, correctly follows self-interest, but thereby ends up worse off from the perspective of self-interest than they would have been if they had all followed some other antecedently identifiable strategy instead.

In *Reasons and Persons* and *On What Matters*, Derek Parfit argues that it is a constraint on any plausible moral theory that morality must *never* be directly collectively self-defeating, and he claims that the most plausible versions of consequentialism, contractualism, and Kantian ethics all imply that morality is never directly collectively self-defeating:

> …moral principles or theories are intended to answer questions about what *all* of us ought to do. So such principles or theories clearly fail, and condemn themselves, when they are directly self-defeating at the collective level. (2011, 306, italics in the original)

> [The assumption that morality is never directly collectively self-defeating] is either made or implied by most of the many different theories [of morality]. (1984, 113)

Some theorists not only agree with Parfit that *morality* can never be directly collectively self-defeating, but also believe that *rationality* and other forms of normativity can never have that property either. For example, the Kantian idea that our acts or principles must be willable as universal law might be taken as a way of suggesting that both morality and rationality can never be directly collectively self-defeating. And even theorists who grant that there is a narrow form of rationality that is sometimes directly collectively self-defeating often insist that there is a broader and more important form of rationality, sometimes called 'enlightened self-interest', that never has that property.

These theorists are all mistaken, because morality and all other interesting forms of normativity are sometimes directly collectively self-defeating. To see why, consider cases like the following:

> **Stampede Case**
> We find ourselves in an enormous stampede. Unless everyone immediately stops stampeding, it is clear that some of us will be moderately harmed. However, it is also clear that everyone will not immediately stop stampeding, and so anyone who does stop stampeding will be severely harmed in a way that does no good for anyone else.

In this case, from every interesting normative perspective – self-interest, enlightened self-interest, morality, benevolence, and so on – each person is *required* to continue stampeding, despite the fact that it is clear that the outcome would be *better* in every normatively interesting sense if everyone did not continue stampeding instead. This shows that morality and all other forms of normativity are sometimes directly collectively self-defeating, because it shows that there are cases in which everyone can be sure that if each person does what is required, the result will be worse from the perspective of each than if each had not done what is required instead.

Here is another example:

> **Units of Good Case**
> 1,000 people are put into isolation booths. It is common knowledge that each must choose between Options A and B, with the following outcomes: If everyone chooses A, then each receives 99 additional units of good; if everyone chooses B, then each receives 100 additional units of good; otherwise, each person who chooses A receives 10 additional units of good and each person who chooses B loses a catastrophic 100,000 units of good.

In this case, from every interesting normative perspective, each person is *required* to choose A, despite the fact that it is clear that the outcome would be *better* from each person's perspective if everyone did not choose A instead. Once again, this shows that morality and all other forms of normativity are sometimes directly collectively self-defeating, because everyone can be sure that if each does what is required, the result will be worse than if each had not done what is required instead. Importantly, these conclusions follow even in cases such as these that do not involve any uncertainty or normative failure, in which it is common knowledge that: everyone will satisfy their normative requirements, everyone knows the relevant facts, and everyone knows which course of action would lead to the best outcome.[3] This ensures that these cases are genuine counterexamples to the thesis that morality is never directly collectively self-defeating in the sense intended by Parfit and others.

Why are morality and other forms of normativity directly collectively self-defeating in these cases? The answer is that a particular form of risk aversion is sometimes required: in particular, sometimes even when it is common knowledge that everyone will satisfy their requirements and that everyone is fully informed, it is also clear that

[3] On the natural and intended understanding of these cases, there is a sense in which everyone has the same options, one of which is such that each individual knows enough about what the others will do to know of that option that the outcome will be objectively best if s/he choose that option, and thus there is a particular option such that each knows that s/he will successfully follow morality only if s/he chooses that option. In this way, these cases are not *unsettled coordination problems* in which it is unknown which course of action would lead to the best outcome. (Compare the unsettled coordination problems discussed by Parfit, 1984, 53–4.)

the option that would lead to the best outcome if universally chosen is associated in a way that is salient to everyone with great risks without compensating rewards, and in some such cases each person can, by this very reasoning, know that others will coordinate on a 'risk-averse' option instead, thereby ensuring that each person is required to choose that 'risk-averse' option themselves, even if it is clear that everyone choosing that risk-averse option guarantees a worse outcome from the perspective of each than if everyone did not choose that option instead.[4] In the words of David Lewis in another context, in these cases individuals can be seen as reaching "a coordination equilibrium that is somehow salient: one that stands out from the rest by its uniqueness is some conspicuous respect. It does not have to be uniquely *good*; indeed, it could be uniquely bad. It merely has to be unique in some way the subjects will notice, expect each other to notice, and so on" (1969, 35, italics in the original).

In addition to showing that all forms of normativity are sometimes directly collectively self-defeating, the preceding considerations also show that an important research program on morality and game theory is misguided, because the essential and guiding assumption of that research program is that morality always guarantees optimal cooperation when it is common knowledge that: everyone has full information about the symmetrical choices facing everyone, will act freely, will satisfy their normative requirements including moral requirements, and knows of a unique option that the outcome would be best if that option were chosen by everyone.[5]

Having argued that morality and all other forms of normativity are sometimes directly collectively self-defeating (DCSD), it is useful to consider further implications for moral theory.

First, consider the Kantian idea that an act is permissible only if the maxim behind that act is willable as universal law. What does this mean? Suppose one does not know exactly what this means. Nonetheless, one can know on the basis of the arguments above that if this implied that morality is never DCSD, then it would be false. More generally, consider versions of 'Kantian ethics', 'rule utilitarianism', 'utilitarian generalization', 'cooperative utilitarianism', or any other view on which a notion of 'universalizability' seems to play a guiding role. Because we can show that morality is sometimes DCSD, we can show that such views would be false if they implied that morality is never DCSD. As a result, we should not interpret such views as having that implication – contrary to Parfit's claims – if we want to develop the most plausible versions of these views.

Recognizing that morality is sometimes directly collectively self-defeating might

[4] Note that this 'risk aversion' in this sense does not imply departure from standard decision theory.

[5] For an brief description of this research program, see Parfit, 1986, pg. 867; for more detail, see Regan, 1980, especially pp. ix –xi, and pp. 4–5, and Gibbard, 1971, especially pp. 6–9.

also lead us to reexamine beliefs about what individuals are required to do in real-world collective action problems. For example, consider the following:

Pollution Case
Each of us will do better by not reducing emissions than by reducing emissions; however, at the same time, each of us will do substantially worse if no one reduces emissions than we would if everyone reduced emissions.

Many would say that each of us is required to reduce emissions in this case because the alternative is directly collectively self-defeating. However, that is a bad argument, because morality and all other forms of normativity are sometimes DCSD. So, if individuals are required to reduce emissions in such a case, it must be for some other reason, such as the impermissibility of the *harm* that is done by those emissions.

In response to all of this, it might be claimed that although morality is sometimes *mildly* DCSD as in the Stampede Case and the Units of Good Case above, it can never be *dramatically* DCSD.

At first glance, this response might seem promising. However, it does not succeed, because morality and all other forms of normativity are sometimes dramatically directly collectively self-defeating. To see why, consider cases like the following:

Dramatic Stampede Case
We find ourselves in an enormous stampede. Unless everyone stops stampeding, it is clear that an increasing number of people will be seriously harmed and killed. However, it is also clear that everyone will not stop stampeding, and so anyone who does stop stampeding will be severely harmed or killed in a way that does no good for anyone else and simply adds to the ultimate aggregate harm caused by the stampede.

This case is representative of real-life stampedes. In such cases, individuals are not required to stop stampeding, even if continuing is dramatically directly collectively self-defeating.

Here is an infinitely dramatic example:

One Million Dollars Case
Everyone on the planet is isolated and instructed to choose a number, and a neon sign reading 'One Million' is lowered in front of each person. If everyone chooses the same number, then the standard of living of each person in the world will be increased by an amount equivalent to one one-millionth of that number of dollars; otherwise, if everyone fails to choose the same number, each person's standard of living will be dramatically reduced. All of this is common knowledge.

What number should each choose in this case? Each should choose one million, because it is common knowledge that one million is uniquely salient to everyone, which makes it common knowledge that one million is the only number that has any chance of being chosen by everyone, which makes it the case that each should choose that number, given the dramatic costs of a failure to coordinate. However, if everyone chooses one million, the standard of living of each person in the world will remain the same rather than rising by, say, $1 billion each, which it is clear that everyone could bring about by simply by choosing the number one quadrillion instead of one million (and so on for any amount whatsoever). As this shows, morality and all other forms of normativity are sometimes *catastrophically directly collectively suboptimal*, because they sometimes direct everyone to choose an option that is certain to lead to a catastrophically worse outcome than an antecedently identifiable option that they could have directed everyone to choose instead. This is truly catastrophic, because instead of solving all of the world's material problems, following morality and other forms of normativity in such a case would not do anyone any good at all.[6]

Here is another dramatic example:

> **End of the World Case**
> Aliens come to Earth and force each family on the planet to choose between 'cooperating' and 'defecting', which are known to have the following consequences: If all choose to cooperate, the aliens will leave and everyone's life will go on the same as before – but if even one family chooses to defect, in one year the aliens will destroy the Earth and every living thing on the Earth, and in the meantime will ensure that each family that chooses to cooperate has a miserable life of intense suffering, while each family that chooses to defect has a wonderful and flourishing final year on Earth.

A philosopher might insist that every family would be required to cooperate in this case. But upon reflection, it is clear that if a billion families were actually in this situation, then some of them, by this very reasoning and without making any moral mistake, would choose to defect, thereby ensuring the end of the world in one year, and ensuring that cooperation would mean a futile sacrifice of one's own family in a way that was impermissible. This illustrates the way in which morality can be *catastrophically* directly collectively self-defeating.

In response to all of this, a philosopher might insist that it is simply *absurd* to think that morality is sometimes dramatically directly collectively self-defeating. Such a

[6] This example is based on a case discussed by Schelling, 1957, who also introduces the relevant notion of *salience*. For ease of exposition, it is assumed here that one dollar would not make a difference to anyone on the planet.

thought is true in an important sense – it is true in the same sense that we sometimes find ourselves in situations that are absurd. But *absurdity* in that sense does not give rise to a reductio – just as finding ourselves in an *absurd* situation such as the End of the World Case would not show that we were not in that situation.

What does contractualism say about all of this? It is unclear. Contractualism is, roughly, the view that an act is required if it is required by principles that we would agree upon in some special scenario, or, alternatively, if it is required by principles that we could not reasonably reject. One problem for contractualism is that everyone would want everyone to cooperate in the End of the World Case, and everyone would agree to cooperate if such agreement was possible and binding; but if it were supposed to follow from these facts that contractualism implies that individuals are required to cooperate in the End of the World Case, then the view would be false, and would be false because it ignores every interesting aspect of collective action problems.

In response, contractualists would quite reasonably insist that their view does not mistakenly imply that cooperation is required in the Dramatic Stampede Case and the End of the World Case. But if that is correct, then they should also be quick to admit that their view does not give us any easy answers about how to think about challenging collective action problems such as those discussed above in the way it would if such a view were never DCSD. Similar remarks apply to universalization theories: either those theories are false because they imply that morality is never DCSD, or else they do not provide any immediate guidance about how to think about such collective action problems.[7]

The upshot is that morality and all other forms of normativity are sometimes dramatically directly collectively self-defeating, which means that many influential normative theories are either false, or at least don't have the consequences that their adherents take them to have.[8] One important consequence is that morality and other forms of normativity cannot be relied upon to solve collective action problems even in a world of normatively flawless agents. In particular, even if a disaster will ensue if everyone acts in a particular way or on a particular principle, that does not settle the

[7] The cases above are counterexamples even to sophisticated never directly collectively self-defeating universalizability theories that are intended to apply to non-ideal situations. For example, consider Parfit's principle "Follow the rules whose being followed by everyone would make things go best, unless some other people have not followed these rules, in which case do whatever, given the acts of others, would make things go best" (Parfit, 2011, pg. 317). This principle seems to deliver the mistaken verdict that everyone is required to choose B in the Units of Good Case, because at the moment that everyone chooses in that case, no one has yet failed to follow optimific principles, and so the principle seems to imply that each must choose B. Similar remarks apply regarding most of the other cases above.

[8] It is worth noting that the arguments here do not depend on any controversial premises. In particular, the arguments here do not depend on the controversial premise that the *better than* relation is intransitive; compare Rachels, 1989, and Temkin, 2011. Even if the *better than* relation were intransitive, that would not show that morality is sometimes *dramatically* DCSD, as the arguments here reveal.

question of whether individuals are permitted to act in that way or on that principle. And because many of the most important questions about modern moral life are essentially questions about what individuals are required to do in such situations – for example, what individuals are required to do about climate change, what individuals are required to do when products are produced in morally objectionable ways – an important practical upshot is that such questions cannot be answered by asking 'But what if everyone did that?', or by more sophisticated appeals to 'universalizability'.

In response to the preceding arguments, theorists who are focused on Parfit's writings sometimes object that the notion of direct collective self-defeat is merely a technical notion introduced by Parfit, and as a result it is impossible to evaluate the arguments above without examining Parfit's precise definition.[9]

This objection is misguided, because as Parfit's own discussion makes clear, direct collective self-defeat is *not* a technical notion even on his view. For example, when Parfit first discusses that notion in *Reasons and Persons*, he assumes that we can all grasp that notion independent of any definition by reflecting on prisoner's dilemmas and other social dilemmas, where these examples illustrate the importance of that intuitive notion to normative theory. He then considers a provisional definition that might initially seem to capture that notion, and then immediately rejects that provisional definition. Why? Because Parfit argues that when we consider a hypothetical case, we can see that the provisional definition gives a different verdict than the intuitive notion that we care about; thus, the provisional definition must be rejected.[10] This shows that the notion of direct collective self-defeat is not a technical notion even for Parfit. Instead, Parfit correctly recognizes that direct collective self-defeat is an intuitive notion that we can all grasp on the basis of reflection on social dilemmas and independent of any stipulative definition, and that this non-technical notion is of central importance to normative theory – and when we consider how this non-technical notion applies to the cases presented above, we see that morality and all other forms of normativity are sometimes directly collectively self-defeating. (This and other issues related to Parfit's views are discussed in more detail below.)

Deontologists might object that this entire discussion depends on a sense of *betterness* that is foreign to their view, because (they might say) their view is concerned with *acts* rather than *outcomes*.[11] However, such an objection is misguided. If we all

[9] In conversation.

[10] Parfit, 1984, pp. 53–54. Unlike the coordination problems that Parfit uses in these passages to reject the provisional definition of DCSD, the examples in this paper are not merely cases where morality *fails to direct us toward* the morally best outcomes, but are cases where morality *directs us away from* the morally best outcomes, thereby constituting cases in which morality is genuinely DCSD (Parfit, 1984, pg. 54).

[11] For such an objection, see Adams, 1997, pg. 259.

continue stampeding in the Stampede Case, it is certain that we will cause harm, whereas if we all stop stampeding, it is certain that we will do no harm at all. As a result, it is perfectly sensible and correct to say that, collectively, continuing stampeding is deontologically worse than stopping stampeding, but that nonetheless each of us individually is required to continue stampeding, because if an individual were to stop, s/he would do something (namely, severely harm an innocent person – him or herself) that is deontologically worse than what s/he would do by continuing stampeding. That is why in the Stampede Case deontology is DCSD.[12]

More fundamentally, some deontologists might reject the judgments about cases appealed to above, and insist instead that, for example, one has a moral obligation to stop and be trampled to death in the stampede case even though doing so would do no good for anyone. In light of this possibility, the arguments above should officially be understood as conditional on the judgments appealed to above. If one accepts those judgments – as many theorists and almost all ordinary people do – then the conclusions above about direct collective self-defeat follow; if one rejects those judgments, then these arguments still establish an interesting conditional result, the consequent of which can be resisted only by endorsing verdicts on cases that many find highly counterintuitive.

A more subtle objection comes from agent-neutral consequentialists, some of whom believe that it is a clear and important virtue of their view that it is never directly collectively self-defeating. For example, Parfit argues:

> [Agent-neutral consequentialist theories] cannot be directly self-defeating, since [they are] *agent-neutral*: giving to all agents *common* moral aims. (1984, 54–55, italics in the original)[13]

> …Common-Sense Morality is often directly collectively self-defeating. [But] a moral theory must be collectively successful. [Those who believe in Common-Sense Morality] must therefore revise their beliefs, moving from [Common-Sense Morality to a form of agent-neutral consequentialism]. (1984, 111)[14]

This is Parfit's main argument for the sort of moral theory he favors in *Reasons and Persons*. Unfortunately, this argument is unsound, because agent-neutral consequentialism is sometimes directly collectively self-defeating, as illustrated by the Units of

[12] Another example from Parfit: "Suppose that each could either (1) carry out some of his own duties or (2) enable others to carry out more of theirs. If all rather than none give priority to their own duties, each may be able to carry out fewer. Deontologists can face [situations in which their theory is DCSD]" (Parfit, 1984, pg. 98).

[13] Parfit, 1984, pp. 54–55, italics in the original.

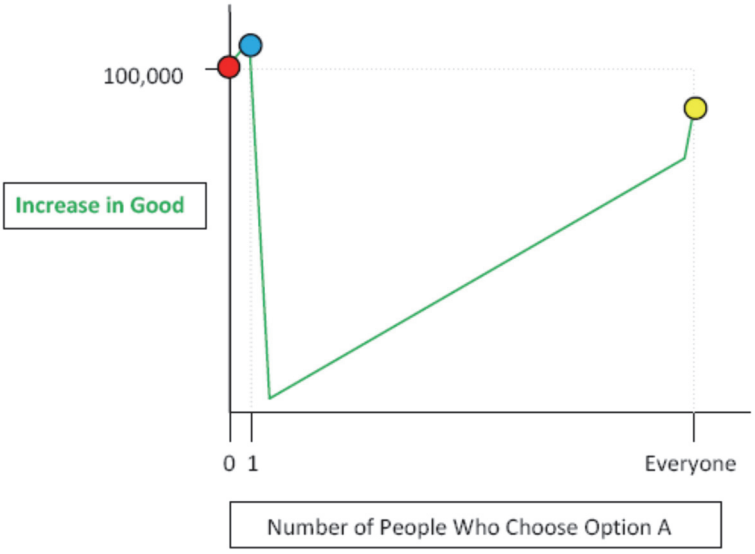
[14] See also Parfit, 2011, pg. 306.

Good Case above.[15] For a further example that may be useful in anticipating various avenues of reply, consider the following complicated variant of the Units of Good Case, and the graph that follows, which represents the possible outcomes in this more complicated case:

Complicated Units of Good Case
1,000 people are put into isolation booths, and each must choose between Options A and B, with the following outcomes: If everyone chooses A, then each receives 99 additional units of good; if everyone chooses B, then each receives 100 additional units of good; otherwise, every person who chooses A receives 10 additional units of good and every person who chooses B loses a catastrophic 100,000 units of good, unless one and only one person chooses A, in which case every person receives 101 additional units of good. All of this is common knowledge, as is the fact that everyone will successfully follow agent-neutral consequentialism.

The graph in Figure 1 is a simplified representation of the possible outcomes in this case:

Figure 1. Complicated Units of Good Case.



[15] Rabinowicz, 1989 argues that *some* versions of agent-neutral conquentialism can be directly collectively self-defeating. My argument aims to show that *all* plausible versions of agent-neutral consequentialism are sometimes directly collectively self-defeating.

In this case, it is common knowledge that everyone will satisfy their requirements, that everyone is fully informed, and that everyone can see that the option that would lead to the best outcome if universally chosen (B) is associated in a way that is salient to everyone with great risks without compensating rewards; as a result, each can be certain that the others will tend to coordinate on the risk-averse option A, which ensures that each person is *required* to choose that risk-averse option A themselves (because they know that their choice would otherwise make the outcome worse on agent-neutral consequentialist grounds), even though it is clear that everyone choosing that risk-averse option A guarantees a worse outcome from the perspective of each (bringing about the yellow dot outcome on the graph) than if everyone did not choose that option instead (bringing about the red dot outcome on the graph). As a result, agent-neutral consequentialism is directly collectively self-defeating in this case because everyone can be certain that: *if we all successfully follow agent-neutral consequentialism by each doing what is actually required, we will thereby cause our agent-neutral consequentialist aims to be worse achieved than they would have been if none of us had done what is actually required*.[16] In more detail: because of what each knows about the situation and thus what each knows about how the others will choose, each can be certain that choosing A will make the outcome objectively better than choosing B, and thus agent-neutral consequentialism requires each to choose A; at the same time, each can be certain that if each failed to do what is required and therefore chose B, the outcome would be better even though no one would then satisfy agent-neutral consequentialism (because for each it would be true that there is something else s/he could have done (namely, choose A) that would have led to more good (by bringing about the blue dot outcome). Thus, agent-neutral consequentialism is sometimes directly collectively self-defeating.

Is this a bad result for agent-neutral consequentialism? No. It would be a bad result for agent-neutral consequentialism if it were never DCSD, because we've seen that all plausible normative theories are sometimes DCSD.

Why then does Parfit think that agent-neutral consequentialism is never DCSD? Parfit offers the following sufficient conditions for direct collective self-defeat:

> A theory T is directly collectively self-defeating when:
> (i) it is *certain* that, if we all successfully follow T, we will thereby cause our T-given aims to be worse achieved than they would have been if none of us had successfully followed T, or

[16] Compare (i) on page 54 of Parfit, 1984.

(ii) our acts will cause our T-given aims to be best achieved only if we do not successfully follow T.

Based on these conditions, Parfit offers the following argument that agent-neutral consequentialism is never DCSD:

> [Agent-neutral consequentialism] cannot be directly self-defeating, since it is *agent-neutral*: giving to all agents *common* moral aims. If we cause these common aims to be best achieved, we must be successfully following this theory. Since this is so, it cannot be true that we will cause these aims to be best achieved only if we do not follow this theory. (1984, 54—55)

In the last sentence of the preceding quote, Parfit concludes that it is necessarily false that: our acts will cause our agent-neutral consequentialist aims to be best achieved only if we do not successfully follow agent-neutral consequentialism, which is an instance of (ii), where 'agent-neutral consequentialism' replaces 'T'. From this, it is supposed to follow that agent-neutral consequentialism is never DCSD.

At this point, someone might object to Parfit's argument as follows: "On Parfit's analysis, a theory can be DCSD in either way (i) or way (ii), and Parfit has shown only that agent-neutral consequentialism cannot be DCSD in way (ii); so, it doesn't follow from Parfit's premises that agent-neutral consequentialism cannot be DCSD in way (i), and so it doesn't follow that agent-neutral consequentialism is never DCSD."

In reply to this objection, Parfit would presumably insist that (i) is to be understood in such a way that (i) implies (ii). If that's right, then Parfit's demonstration that agent-neutral consequentialism can never be DCSD in sense (ii) also shows that it can never be DCSD in sense (i).

However, even granting such a reply, Parfit's argument still faces a decisive objection. To see the problem, note that even if (i) implies (ii), Parfit's argument is still invalid as stated:

> If C is ever (i) or (ii), then C is sometimes DCSD.
>
> C is never (ii).
>
> Therefore, C is never (i), since (i) implies (ii).
>
> Therefore, C is never DCSD.

If the problem is not immediately apparent, it might help to combine the two middle claims:

If C is ever (i) or (ii), then C is sometimes DCSD.

C is never (i) or (ii).

Therefore, it is not the case that C is sometimes DCSD.

This argument is invalid because it denies the antecedent. To get a valid argument, we would have to understand the first premise as a biconditional, and thus we would have to interpret (i) and (ii) as together yielding a full analysis of direct collective self-defeat. However, Parfit explicitly claims that (i) and (ii) provide only sufficient conditions for direct collective self-defeat, and not a full analysis.[17] As a result, Parfit's argument is invalid, because it has the invalid form above.

Of course, this raises the question of whether (i) and (ii) can in fact yield a full analysis of direct collective self-defeat – in other words, it raises the question of whether the following is true:

A theory T is DCSD when *and only when* either (i) is true or (ii) is true, where (i) and (ii) are understood in the way that Parfit intends.

This *Implicit Analysis* is false, because it does not capture the essence of direct collective self-defeat, including the essential idea that a theory is DCSD when it *directs us toward outcomes that are certain to be worse* (1984, 54). In particular, the Implicit Analysis fails to deliver the correct verdict on the cases discussed above in which:

(iii) it is common knowledge that: everyone knows the relevant facts, will act freely, will satisfy their normative requirements, and everyone can also be certain that: if each does what T actually requires, the T-given aims of each will be worse achieved than they would have been if none had done what T actually requires.

At the very least, such cases show that (iii) is an additional sufficient condition for direct collective self-defeat, which means that Parfit's argument that agent-neutral consequentialism is never DCSD cannot be salvaged, because (iii) together with the units of good cases show that agent-neutral consequentialism is sometimes DCSD.

In response, a defender of the Implicit Analysis might say "But consider the possibility that in the Complicated Units of Good Case one and only one player chooses Option A; then, each person successfully follows agent-neutral consequentialism and brings about the best outcome; this shows that even in the Complicated Units of

[17] Parfit makes this explicit in the following passage, where he explains how he intends the phrase "[A theory T is] directly collectively self-defeating when..." to be understood: "By 'when' I do not mean 'only when'" (1984, 54).

Good Case agent-neutral consequentialism does not direct us toward outcomes that are certain to be worse."

This reply gives the phrases 'direct us toward' and 'successfully follows' a meaning that is very different from their meaning in the intuitive thought that a theory is DCSD when it directs us toward outcomes that are certain to be worse, or when it is certain that the outcome would be worse if each successfully followed the theory than if each did not. More specifically, this reply involves a backward-looking conception of *directing an agent toward an outcome* and *successfully following a theory* that is irrelevant to any interesting normative concept. To see why, return to the players in the Units of Good Case and assume that all the players evaluate options and make their decisions simultaneously as well as independently. Now consider the point in time as they are about to make their decisions. At that point in time, does agent-neutral consequentialism direct the players toward a particular outcome? It does in every intuitive sense – namely, the outcome in which everyone chooses Option A: after all, even before anyone chooses Option A, each player *knows* that choosing Option A will lead to an objectively better outcome, and thus agent-neutral consequentialism directs each player to choose Option A, and thus each player successfully follows agent-neutral consequentialism only if that player chooses Option A.

The Implicit Analysis denies all of this. Instead, on that analysis agent-neutral consequentialism gives the players no direction at all before their decisions are made, on the grounds that there are multiple combinations of choices that would result in satisfaction of agent-neutral consequentialism. That is how the Implicit Analysis insists that agent-neutral consequentialism does not direct the players away from the best outcome: according to the analysis, there are no facts about what agent-neutral consequentialism directs the players to do until after everyone has made their decision, at which point the theory 'directs' everyone to have chosen in such a way that they now each satisfy agent-neutral consequentialism. However, this is a revisionary account of how agent-neutral consequentialism directs agents toward outcomes – because it entails, contrary to the claims of all actual consequentialists, that what agents know about the consequences of their choices has no relevance to what consequentialism directs them to do – and more importantly it is also an unacceptable account, because any interesting normative theory must provide direction for our decisions, and not only after they are made.

In response, a defender of the Implicit Analysis could attempt to bite the bullet and simply insist that agent-neutral consequentialism offers no direction in such cases until after decisions are made. However, the costs of such a stance prove unacceptably high when applied to other cases, especially cases that involve physical indeterminacy with no residual epistemic uncertainty. For example, consider a case that is similar to

the Units of Good Case, but where the uncertainty of the outcomes derives entirely from physical indeterminacy:

> **One-Player Units of Good Case**
> You know that you alone must choose between the following two options, and that your choices will have the following consequences for yourself and 999 other innocent people:
>
> Option A: 99% chance that everyone receives 99 additional units of good; 1% chance that everyone receives 10 additional units of good.
>
> Option B: 1% chance that everyone receives 100 additional units of good; 99% chance that everyone receives negative 100,000 units of good.

Suppose that the chances in this case are purely physical and that there is no residual epistemic uncertainty. (For example, suppose that physicists have designed a non-deterministic pleasure and pain dispensing device to have these properties; you will simply choose whether to press the 'Option A' or 'Option B' button.)

On any sensible interpretation, agent-neutral consequentialism directs you to choose Option A in this case, which means that you successfully follow agent-neutral consequentialism only if you choose Option A. Would defenders of the Implicit Analysis agree? If they do not, then they are committed to the view that agent-neutral consequentialism never provides any actual guidance to our decisions, because physical indeterminacy always underlies all of our decisions. So, to avoid this result, they would presumably agree that agent-neutral consequentialism directs you to choose Option A in this case.

But if that is right, then there is a powerful argument that agent-neutral consequentialism directs each player to choose Option A in the original Units of Good Case. For consider that, for each player in that original case, there is some distribution of credences that that player ought to have, given his or her evidence, about how the other players will choose. Given that distribution of rational credences, we can imagine a one-player game with the same outcomes and probabilistic structure, but where the probabilities arise from physical indeterminacy with no residual epistemic uncertainty as in the One-Player Units of Good Case. If, as we are assuming, agent-neutral consequentialism directs you to choose Option A in the One-Player Units of Good Case, then it also directs each player to choose Option A in the one-player game that is derived in such a way from his or her rational credences in the original Units of Good Case. But if agent-neutral consequentialism directs each player to choose Option A in the one-player games that are derived from their rational credences, then

it also directs each player to choose Option A in the original Units of Good Case itself, because there is no normatively relevant difference between the choices that each individual would face in those one-player games and the corresponding choices that they face in the original Units of Good Case. As a result, initial defenders of the Implicit Analysis are forced either to abandon that analysis by admitting that agent-neutral consequentialism directs players to choose Option A in the original Units of Good Case, or else to bite an unacceptable bullet and insist that agent-neutral consequentialism almost never provides any guidance to our decisions at all, because physical indeterminacy always underlies our decisions.

The preceding discussion shows the importance of distinguishing between a normative theory's theory of objective value and its theory of choice. As we have just seen, a theory of objective value never directs us toward any outcomes itself – it is only in conjunction with a theory of choice that we are directed to make particular choices, and thereby directed toward particular outcomes. This shows that the analysis of direct collective self-defeat under consideration must be inadequate, because that analysis focuses only on satisfaction of a theory's theory of objective value, and not on satisfaction of its theory of choice. In other words, that analysis must be inadequate because the notion of direct collective self-defeat is about what a theory directs us toward, and without a theory of choice a theory never directs us toward anything at all.

For these reasons, an adequate full analysis of direct collectively self-defeat must be tied to a normative theory's theory of choice. Here is a proposal:

> A theory T is directly collectively self-defeating (DCSD) when: it is certain from the perspective of each of us that, if each of us successfully follows T's theory of choice, we will thereby cause our T-given aims to be worse achieved than they would have been if none of us successfully followed T's theory of choice.

If successfully following T's theory of choice is the same as doing what T requires, then this *New Analysis* is equivalent to the claim that: *A theory T is directly collectively self-defeating (DCSD) when: (from the perspective of each of us) it is certain that, if each of us does what T requires, we will thereby cause our T-given aims to be worse achieved than they would have been if none of us did what T requires*. This amounts to an analysis of direct collective self-defeat for cases in which everyone has the same two options. To test this analysis, we can consult our judgments about cases, and our judgments about the concept of direct collective self-defeat. Upon reflection, this New Analysis delivers the correct verdict on all of the cases that theorists have discussed in connection with direct collective self-defeat, and, unlike the Implicit Analysis discussed above, also fits our intuitive concept of direct collective self-defeat, according to which a theory is

DCSD when it directs each of us to act in a way that is certain to be worse than if everyone did not follow the theory's directions instead.

In response, a defender of the Implicit Analysis might raise the following objection: "Perhaps the New Analysis and/or (iii) captures the idea that a theory is DCSD when it directs us toward outcomes that are certain to be worse. But that idea is inconsistent with other more firmly held beliefs that we have about self-defeat, and so the notion of direct collective self-defeat must be regimented in a different way – most likely, in the way the Implicit Analysis suggests. That is because Donald Regan, Derek Parfit, and others have provided cases that show that normative theories sometimes direct us away from the best outcomes, but are not thereby self-defeating." What the objector has in mind are cases like the following:

> **Miners Case**
> Suppose that several miners are trapped, with floodwaters rising. Before we can find out where these miners are, we must decide which floodgate to close.

The possible outcomes of our decision are outlined in Table 1.

Table 1. Miners Case

	The miners are in Shaft A	The miners are in Shaft B
We close Gate 1	We save ten	All die
We close Gate 2	All die	We save ten
We close Gate 3	We save nine	We save nine

Assume that, on the evidence, the miners are equally likely to be in either shaft.[18]

In this case, we are required to close Gate 3, even though it is certain that we will thereby bring about an outcome that is not best; nonetheless, this does not show that normativity is directly collectively self-defeating. Does this undermine the idea that (iii) is a sufficient for direct collective self-defeat?

It does not. What the Miners Case shows is that there is a crucial distinction between, on the one hand, *it being certain that an option will lead to an outcome that is not best* and, on the other hand, *it being certain that an option will lead to a worse outcome than some other antecedently identifiable particular option*, and that a theory is DCSD when it directs us to choose an option of the latter type, but not when, as in the Miners Case, it merely directs us to choose an option that is certain to be not best. In particular, if we close Gate 3 we bring about an outcome that is certain to be not best, but we do not bring about an outcome that is certain to be worse than the outcome

[18] This example is taken from Parfit, 1988, pp. 2-3, who follows Regan, 1980, pg. 265.

of any particular other option, because there is no other option that is antecedently *certain* to lead to a better outcome than closing Gate 3. This is in perfect tune with conditions (i), (ii), and (iii) above, because the natural way of extending those conditions to cases involving many options such as the Miners Case is by claiming that a theory is DCSD when it directs everyone to choose an option that is *certain* to lead to a worse outcome than an antecedently identifiable alternative option that it could have directed everyone to choose instead – but not when, as in the Miners Case, the theory merely directs everyone to choose an option that is certain to lead to an outcome that is not best.[19] As a result, the Miners Case does not ultimately raise a problem for the intuitive notion of direct collective self-defeat, and does not raise a problem for the view that conditions (i), (ii), and (iii) are each sufficient for direct collective self-defeat, and does not raise a problem for the New Analysis above.

In response to all of the preceding arguments, it might be objected that theories like agent-neutral consequentialism still imply that it is always metaphysically possible to bring about the outcome that is best without anyone acting in a way that is wrong – and that such a possibility of doing what is best without anyone doing wrong is how the notion of direct collective self-defeat is best understood. However, although optimal cooperative action is *metaphysically possible* in cases such as the units of good cases, each individual is also certain that such cooperation will not obtain, and as a result from the perspective of each individual *it is certain that the aims of morality will be worse achieved if each successfully follows morality than if everyone did not successfully follow morality instead* – which is of course just to say that morality is directly collectively self-defeating, because it would actually be wrong to act in accord with optimal cooperative action based on the full information of the case and what each knows about the morally flawless dispositions of others. This shows that agent-neutral consequentialism ultimately has no interesting advantage over other types of ethical theories with respect to direct collective self-defeat.

The arguments above also cannot be dismissed by simply insisting on an alternative definition of direct collective self-defeat on which the arguments above do not go through – for example, a stipulative definition on which (iii) is not a sufficient condition for direct collective self-defeat. In part, this is because direct collective self-defeat, like knowledge, is a notion that we track and care about prior to seeing any stipulative definition, as is illustrated by our interest in social dilemmas and other situations in

[19] Such an extension presumably must be restricted to cases in which everyone chooses between the 'same' options, where those options are individuated in a 'natural' way – and in other cases the notion of direct collective self-defeat seems to have no clear application. The Miners Case could also be redescribed as a two-option case, where Option One is to close Gate 3, and Option Two is to close one of the other gates. An analysis that includes the features advocated here also delivers the correct verdict given that description, because choosing Option One is *not certain* to lead to a worse outcome than choosing Option Two, and therefore such an analysis does not imply that morality is sometimes DCSD.

which self-interest is directly collectively self-defeating, and so direct collective self-defeat is not a notion that we are free to define however we like if the result is to have any interest to normative theory. More specifically, insofar as we should care whether a theory is sometimes directly collectively self-defeating, that is because having that property means that regrettable consequences are assured even in cases like those described in (iii) in which it is common knowledge that everyone knows the relevant facts and will successfully follow the theory. As a result, a definition on which satisfaction of (iii) is not sufficient for direct collective self-defeat has no practical or theoretical interest, not only because it does not track the important notion of 'directing us toward outcomes that are certain to be worse', but more importantly because it does not track the kind of collective self-defeat that it is regrettable for a theory to imply – because the most regrettable form of collective self-defeat is when a theory is collectively self-defeating in the sense of (iii), when it is collectively self-defeating even though it is common knowledge that everyone knows the relevant facts and will do what is required, and that regrettability is not mitigated in any interesting way when it is also true that if individuals had failed to do what they actually know they are required to do, the outcome could have been better. As a result, any discussion that rejects (iii) as a sufficient condition for direct collective self-defeat is doomed to reduce to a definitional exercise that has no connection to any property that we should care whether a normative theory has – whereas endorsing (iii) is essential to capturing the kind of collective self-defeat that is of central interest from both a practical and theoretical perspective.[20] So, such a stipulative definition has no chance of playing an interesting role in arguments about the nature of morality, and in particular has no chance of playing an interesting role in arguments against commonsense morality.

The preceding discussion suggests the following evaluation of Parfit's main argument for never-directly-collectively-self-defeating moral theories:

Parfit's Main Argument
To be plausible, a moral theory must be never DCSD.
So, we must reject common-sense morality and other theories that are sometimes DCSD, and instead endorse a version of never-DCSD moral theory.

The premise is false, because morality is sometimes DCSD. As a result, not only is it consistent to deny the conclusion, but there is decisive reason for thinking that the conclusion is false, because a moral theory is false if it is never DCSD. If we were to

[20] An additional consideration is that many theorists, including Parfit, take facts about what would be wrong when agents know the relevant facts as explanatorily fundamental – which provides decisive reason to think that the sense of direct collective self-defeat captured by (iii) is the sense that must have the greatest theoretical importance, because (iii) is explicitly concerned with whether a theory is collectively self-defeating when everyone knows the relevant facts. (See Parfit, 2011, Section 21.)

follow Parfit in thinking that consequentialists, contractualists, Kantian theorists, and most others have been "climbing the same mountain" toward the goal of developing the most plausible version of never-DCSD moral theory, then this would mean that those theorists have all been climbing the wrong mountain.[21]

This is not to denigrate Parfit's work, which has the highest virtues of clarity, testability, originality, and importance. Because Parfit's work has such virtues, identifying a clear objection to his arguments leads to important progress in normative theory.

In sum, morality and all other interesting forms of normativity are sometimes dramatically directly collectively self-defeating, which means that many influential normative theories are either false, or at least don't have the consequences that their adherents take them to have. In particular, morality and other forms of normativity cannot be relied upon to solve collective action problems even in a world of normatively flawless agents. A practical upshot is that many of the most important questions about modern moral life cannot be answered by asking 'But what if everyone did that?', or by a more sophisticated appeal to a form of 'universalizability'.

Appendix: The Equilibrium Objection

In "Group Morality", Frank Jackson uses an example that bears some similarity to the Stampede Case to argue that it is possible to "have a group action which is wrong, yet every constituent act is right; and a group action which is right yet every constituent act is wrong" (1987, 102). Parfit accepts Jackson's conclusions in later work, but neither Parfit nor Jackson take these conclusions to show that morality is sometimes directly collectively self-defeating.[22] This appendix shows that Jackson's conclusions do not clearly follow from the example he discusses, and that his discussion cannot be extended to show that morality is sometimes directly collectively self-defeating (DCSD) – but that such conclusions are vindicated by the examples discussed above, despite an important objection that is suggested by reflection on Jackson's discussion.

[21] For this metaphor and a summary of Parfit's arguments that the most plausible versions of consequentialism, contractualism, and Kantian ethics all imply that morality is never DCSD, see Parfit (2011, 25-26). Parfit endorses Parfit's Main Argument in Parfit (2011, 306) : "In [social dilemmas], in acting on common sense moral principles, we are acting in ways that are directly collectively self-defeating. If we were Rational Egoists, that would be no objection to our view, since this form of Egoism is a theory about individual rationality and reasons. But moral principles or theories are intended to answer questions about what all of us ought to do. So such principles or theories clearly fail, and condemn themselves, when they are directly self-defeating at the collective level". See Parfit (2011, 111 and 113) for an earlier discussion and more explicit presentation of the argument.

[22] Jackson does not claim that his conclusions show that morality is sometimes DCSD, and in later work Parfit continues to rely on the premise that morality is never DCSD despite Parfit's endorsement of Jackson's conclusions in Parfit, 1988.

Here is Jackson's example:

> [Suppose that] There is a steady stream of traffic going to work. Everyone is driving at 80 kilometres per hour. It would be safer if everyone was driving at 60. The right group action is for everyone together to drive at 60. But what about each person, should he or she drive at 60? The answer may well be no; for it may well be the case that if he or she were to drive at 60, everyone else would still drive at 80, and so a lot of dangerous overtaking would result. For each individual the right action is to keep driving at 80, so avoid dangerously disrupting the traffic flow; yet the right group action is for everyone to drive at 60. Thus, we have in this example a right group action – everyone together driving at 60 – with each and every constituent individual action – each action of a person driving at 60 – wrong. And also we have a wrong group action – everyone together driving at 80 – with each and every constituent action – each action of a person driving at 80 – right. We see, therefore, that not even the attractive-sounding principle that if a group action is right, at least one of its constituent acts is right, is valid. (1987, 102-3)

This case presupposes some initial wrongdoing by some individuals – in particular, the initial drivers who break the speed limit – which means that the case does not show that a morally suboptimal outcome would result if *everyone* followed morality, which means that the case does not show that morality is sometimes DCSD. Furthermore, even if we imagine a group of morally flawless agents somehow 'thrown into' the case Jackson describes as in a stampede, the case still does not clearly show that morality is sometimes DCSD, and for similar reasons does not support Jackson's own conclusions.

The problem is that, contrary to what Jackson tacitly assumes, each individual driver can choose among a wide range of possible speeds. This detail undermines Jackson's argument, because although no individual driver is required to reduce his or her speed instantaneously to the morally ideal speed of 60, nonetheless at each moment each individual is required to reduce his or her speed *slightly* – which means that if everyone in the group follows morality, the morally ideal speed of 60 will be reached by the group *in the morally optimal way given the group's starting point*. (Upon reflection, it seems clear that this is what morality would require in such a case, on the assumption that it is common knowledge that morality will be universally followed.) As a result, if everyone follows morality, this leads to the morally optimal outcome of everyone driving 60, and it leads to that outcome along a path that is also morally optimal given the relevant starting point – which arguably means that if each person does follow morality along that ideal path, then the group itself also acts rightly at each moment along that path, given its suboptimal starting point. As a result, Jackson's case does not clearly support his conclusions that it is possible to

"have a group action which is wrong, yet every constituent act is right; and a group action which is right yet every constituent act is wrong". Furthermore, even if 'is wrong' is stipulated to mean 'has a suboptimal instantaneous outcome' (as Jackson intends),[23] Jackson's case is still consistent with the idea, and might even seem to illustrate the truth of the idea, that the optimal course of action for a group is in perfect harmony with the optimal course of action for each of its constituent individuals *whenever a stable equilibrium develops as a result of every individual following morality*.

It could be claimed that this *equilibrium objection* also undermines the force of the Stampede Case discussed above. However, a crucial difference is that in a stampede, in contrast to highway traffic, individuals have only two real options: continue stampeding at the dictated rate, or else be trampled – and if everyone continues stampeding at the dictated rate, then all individuals will continue to have only those two options, ensuring that the ultimate outcome never tends toward an equilibrium that is morally desirable, given realistic assumptions.[24]

More importantly, even if such an equilibrium explanation were available for the stampede cases, such an explanation is not available regarding the units of good cases discussed above, because those latter cases involve a 'one-shot' decision situation in which it is simply impossible for a desirable equilibrium to develop in the way the equilibrium objection assumes. As a result, those cases provide a decisive demonstration that morality and all other forms of normativity are sometimes dramatically DCSD, and a decisive demonstration that the best course of action for a group can radically come apart from the best course of action for each of its constituent individuals, even when a stable equilibrium develops as a result of each individual following morality.

References

Adams, R. 1997. "Should Ethics be More Impersonal?", in Dancy, J. (ed.) *Reading Parfit*. Wiley.

Gibbard, A. 1971. *Utilitarianism and Coordination*. Harvard U. Reprinted 1990, Garland.

[23] See Jackson's discussion of 'objectively right' (1987, 92).

[24] Another crucial difference is that stampedes arise without any wrongdoing by any individual, unlike Jackson's example involving high-speed highway traffic.

Jackson, F. 1987. "Group morality". In Smart, Pettit, Sylvan, and Norman (eds.), *Metaphysics and Morality: Essays in Honour of J. J. C. Smart*. New York, NY, USA: Blackwell.

Parfit, D. 1984. *Reasons and Persons*. Oxford UP.

Parfit, D. 1986. "Comments", *Ethics*.

Parfit, D. 2011. *On What Matters*, Volume One. Oxford UP.

Parfit, D. 1988. "What we together do". Unpublished m.s.

Rabinowicz, W. 1989. "Act-utilitarian prisoner's dilemmas", *Theoria*.

Rachels, S. 1989. "Counterexamples to the Transitivity of Better Than", *Australasian Journal of Philosophy.*

Regan, D. 1980. *Utilitarianism and Cooperation*. Oxford UP.

Schelling, T. 1957. "Bargaining, Communication, and Limited War", *Journal of Conflict Resolution*.

Temkin, L. 2011. *Rethinking the Good*. Oxford UP.