Notes on *Reasons and Persons*

Posted by Jacob Williams on brokensandals.net. Last updated 2024-12-18.

Part One: “Self-Defeating Theories”

Chapter 1: “Theories That Are Indirectly Self-Defeating”

Summary: Parfit wants to figure out “the best theory” for answering the question “What do we have most reason to do?” (Parfit 3) The first part of the book is evaluating just one possible argument against certain theories: that they’re “self-defeating”. This chapter looks at two popular theories (Self-interest theory and Consequentialism) and whether they are self-defeating in one particular way (which he calls “indirectly”). He concludes that they are, but that this isn’t a problem for them.

Key ideas/terms:

- Theories about what we have reason to do
  - **Self-interest theory** (S for short): its “central claim” is: “For each person, there is one supremely rational ultimate aim: that his life go, for him, as well as possible.” (Parfit 4)
  - **Consequentialism** (C for short): its “central claim” is: “There is one ultimate moral aim: that outcomes be as good as possible.” (Parfit 24)
  - Both of these theories come in different varieties based on how you define “well” or “good”. Picking the best version of each theory is outside the scope of the book; Parfit tries to draw conclusions that apply to all versions.
- “Self-defeating” and related notions
  - **Self-defeating**: when “a theory fails even in its own terms, and thus condemns itself.” (Parfit 3) This is vague; Parfit will define four more precise forms of it,

categorized on two dimensions: *directly vs indirectly* and *individually vs collectively*.

  - **Indirectly individually** self-defeating: "when… if someone tries to achieve" the aims given to them by a particular theory, "these aims will be, on the whole, worse achieved." (Parfit 5)
    - This applies to Self-interest theory. Parfit gives examples related to psychological limitations of humans, and cases where we could benefit from making it clear to others that we will behave irrationally in the future.
      - We are less likely to be happy when we are explicitly trying to be happy (Parfit 6).
      - We may have difficulty making convincing promises when we need to, if we are actually willing to break promises whenever it would benefit us. (Parfit 7)
      - We can disincentivize others from threatening us or following through on their threats, if we can convince them that we will irrationally ignore their threats at any cost. (Parfit 13)
  - **Indirectly collectively** self-defeating: "when… if several people try to achieve" the aims given to them by a particular theory, "these aims will be worse achieved." (Parfit 27)
    - This applies to Consequentialism. Parfit's examples seem to be about psychological limitations of humans. (Parfit 27–28)
  - **Self-effacing**: when a theory implies that nobody should believe in that theory. (Parfit 24) This does *not* make the theory self-defeating (Parfit 27).
    - Example: Suppose Self-interest theory is true, but that believing in Self-interest theory makes your life go poorly; then Self-interest theory would say you should believe in whatever theory would make your life go better. Self-interest theory's main objective is for your life to go well, not for you to believe true things.
    - (Parfit allows that on *some* meta-ethical views, it might be a problem for Consequentialism to be self-effacing. He says here that he doesn't think it is fully self-effacing (Parfit 43). From his later book, we know he doesn't accept those meta-ethical views anyway—he was a moral realist.)
    - There's a difference between **"ought morally"** and **"ought intellectually"** (Parfit 43)
- **Rational irrationality**: cases where your behavior in the moment is irrational, but is a result of motives that were rational for you to adopt. (Parfit 13)

    - Example: Someone is using threats to coerce you; you become the sort of person who ignores threats, and once they realize this, they stop coercing you. Your change was rational. But it means that later, when a threat is made against your life, you ignore the threat and are killed. Ignoring that threat was irrational. (Parfit 22)

- **Moral immorality** / **blameless wrongdoing**: cases where your behavior in the moment is wrong, but is a result of motives you had good moral reasons to adopt. (Parfit 32)
    - Example: You love your child. But because of this love, on some occasion you choose to help your child in a smaller way rather than help a stranger in a bigger way. Depending on the precise facts, some versions of Consequentialism may imply that the love is good but that this particular act is bad. (Parfit 32)

- Against the idea "that rationality and rightness can be *inherited*, or *transferred*" (Parfit 40), Parfit says the following are **false**:
    - "…if it is rational for me to cause myself to have some disposition, it cannot be irrational to act upon this disposition" (Parfit 39)
    - "…if it is rational for me to cause myself to believe that some act is rational, this act *is* rational" (Parfit 39)
        - If this were true, it could be used to argue that Self-interest theory can provide a foundation for morality, since we may have self-interested reasons to *believe* that we should e.g. keep our promises (since believing it will make it easier to convince people we're trustworthy). Parfit rejects such arguments in section 8. (Parfit 19–23)
    - "…if there is some disposition that I ought to cause myself to have, and that it would be wrong for me to cause myself to lose, it cannot be wrong for me to act upon this disposition" (Parfit 39)
    - "…if I ought to cause myself to believe that some act would not be wrong, this act cannot be wrong" (Parfit 39)

- **Objectively vs subjectively right/wrong**: These are two perspectives—each useful for different purposes—for evaluating whether an action is right/wrong according to a given moral theory.
    - Objective: "what are or would have been the effects of what some person does or could have done" (Parfit 25)
    - Subjective: "what this person believes, or ought to believe, about these effects" (Parfit 25) (The inclusion of "or ought to believe" seems to me to make it less clear exactly how to apply this definition.)

- Note, IIUC, it's only the person's beliefs about the *effects* that matter, not the person's beliefs about right/wrong. Suppose someone believes that their action will cause an innocent person to suffer; and also believes that causing innocent people to suffer is right; but in reality their action will cause lots of happiness and no suffering. We would say that according to utilitarianism, their action is objectively right and subjectively wrong.

Thoughts: Most of this made sense to me, but I have some nagging discomfort around Parfit's discussion of "ought implies can". He thinks that if we're going to deny that I *ought* to have done something, on the grounds that I *couldn't* have done it, then the relevant notion of *couldn't* is: "acting in this way would have been impossible, even if my desires and dispositions had been different" (Parfit 15) (regardless of whether there was any way, given the state of the universe and the laws of nature, for my desires/dispositions to actually be different). I agree, but I wonder if this undermines his grounds for saying that adopting certain motives which lead to "rational irrationality" or "moral immorality" are rational/moral in the first place. Why, for example, should we say that it's moral for a parent to love their child in a way that may lead to immoral behavior in certain circumstances? (Parfit 32–33) Wouldn't it be more moral to both love the child *and* have a disposition to disregard that love when the love would lead one to act immorally? Parfit seems to reject this on the grounds that it's psychologically impossible, but if he's not going to allow that as a defense against saying we ought to have performed some act, it seems inconsistent to use it as a defense against saying we ought to have adopted some disposition.

Chapter 2: "Practical Dilemmas"

Summary: Parfit now defines another way of being self-defeating: "directly". He does not think this applies to Consequentialism, but he does think Self-interest theory can be "directly collectively self-defeating". The two-person Prisoner's Dilemma is a simple example of this, though he doesn't think it arises much in the real world. But the Many-Person Dilemma comes up frequently in the real world: situations where each person doing what's in their own best interests results in everyone being worse off. This is at least a "practical problem" for Self-interest theory. Parfit lists five general approaches for addressing this practical problem.

Key ideas/terms:

- **Agent-neutral vs agent-relative** theories: Consequentialism is agent-neutral because it gives everyone the same goal (promote the general good). Self-interest theory is agent-relative because it gives everyone a different goal (my goal is to promote my welfare, your goal is to promote your welfare).

- **Directly self-defeating**

  - Parfit gives multiple definitions (Parfit 54–55) and it's not clear if any are meant "to cover all cases" (Parfit 54)

    - The key difference from “indirectly self-defeating” seems to be that a theory is *directly* self-defeating if people “successfully follow[ing]” (Parfit 54) it leads to worse outcomes (according to the theory’s own way of evaluating); whereas when a theory is *indirectly* self-defeating, the worse outcome stems from people merely “*tr[ying]* to achieve” (Parfit 5, emphasis added) its aims.

    - **Directly individually** self-defeating: “when it is certain that, if someone successfully follows [some theory] T, he will thereby cause his own T-given aims to be worse achieved than they would have been if he had not successfully followed T…” (Parfit 55)

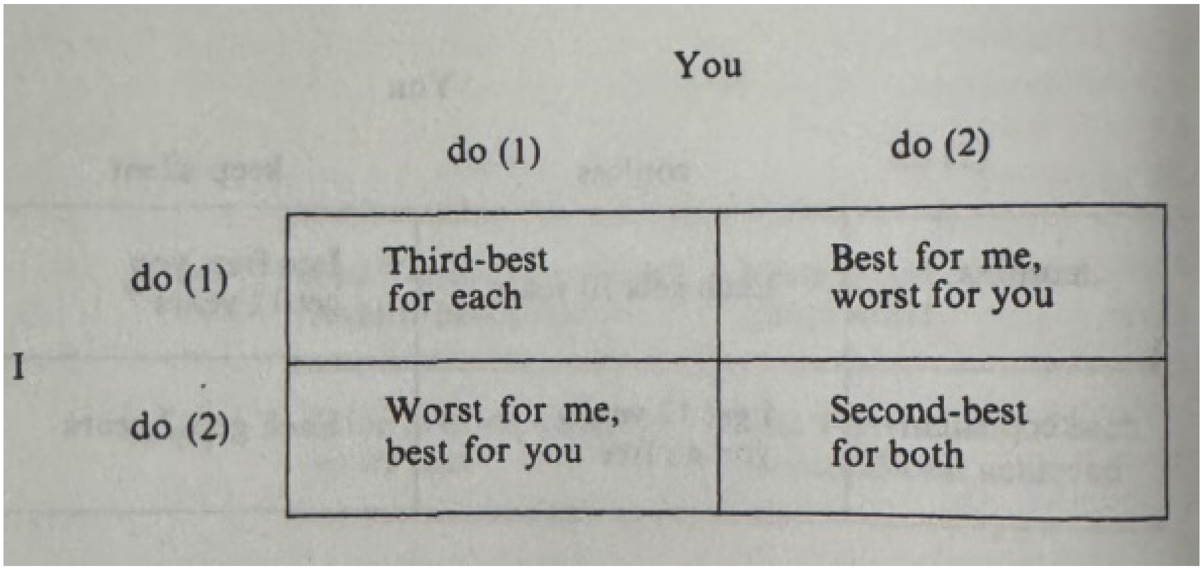
        - I don’t see how this could be true of any theory; it seems like basically a contradiction in terms (at least if to “follow” a theory means to try to achieve its aims). Which is, IIUC, essentially why Parfit says it isn’t true of Self-interest theory. (Parfit 55)

    - **Directly collectively** self-defeating: “when it is certain that, if we all successfully follow T, we will thereby cause the T-given aims *of each* to be worse achieved than they would have been if none of us had successfully followed T.” (Parfit 55)

        - This does *not* apply to Consequentialism or any agent-neutral theory. (Parfit 54–55)

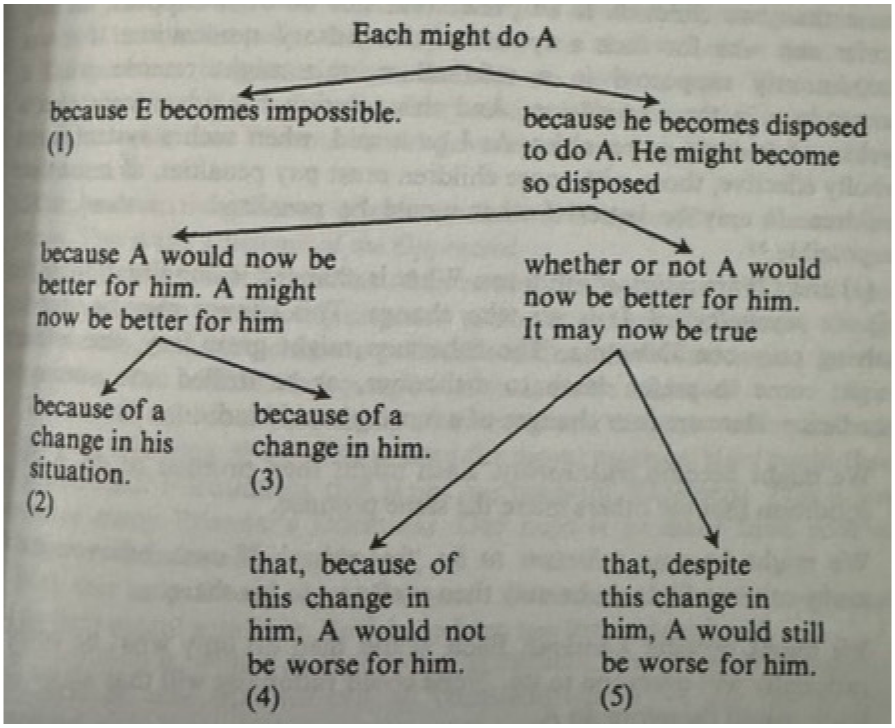
        - But it *does* apply to Self-interest theory. The fact that different people have different aims makes the Prisoner’s Dilemma and Many-Person Dilemmas possible. (Parfit 56)

- **Prisoner’s Dilemma**: a two-person, non-repeating situation like the following, in which neither person’s decision can affect the other’s: (Parfit 58)

    - 

		You	
		do (1)	do (2)
I	do (1)	Third-best for each	Best for me, worst for you
	do (2)	Worst for me, best for you	Second-best for both

- **Many-Person Dilemma**: a situation “when it is certain that, if each rather than none of us does what will be better for himself, this will be worse for everyone.” (Parfit 59)

- This comes up in decisions about whether to help strangers (Parfit 60) or contribute to public goods (Parfit 61).
- We would like to change the situation or people involved in order to avoid the worse outcome. Parfit gives the following diagram of possible types of intervention (where “A” refers to the altruistic action, and “E” to the self-interested/egoistic action). (Parfit 63) “(1) and (2) are *political* solutions. … (3) to (5) are *psychological*.” (Parfit 64)
    - 

      Each might do A

      because E becomes impossible. (1)

      because he becomes disposed to do A. He might become so disposed

      because A would now be better for him. A might now be better for him

      because of a change in his situation. (2)

      because of a change in him. (3)

      whether or not A would now be better for him. It may now be true

      that, because of this change in him, A would not be worse for him. (4)

      that, despite this change in him, A would still be worse for him. (5)

References

Parfit, Derek. *Reasons and persons*. 1. issued in paperback (with corr.), reprinted with further corr, Clarendon Press, 1987.