1 Introduction

Derek Parfit (1942–2017) was one of the most important and influential moral philosophers of the late twentieth and early twenty-first centuries. This Element seeks to introduce the reader to his wide-ranging ethical thought, focussing especially on his two most significant works: *Reasons and Persons* (1984) and *On What Matters* (2011a).

Parfit was centrally concerned about objectivity in ethics and practical rationality. Section 2 of this Element discusses his arguments against commonplace "subjectivist" assumptions, and briefly touches on his meta-ethical views regarding the nature of objective morality.

The next three sections address Parfit's contributions to the consequentialist tradition within ethical theory. Consequentialists generally regard actions as morally significant insofar as they produce good or bad outcomes. *Act Consequentialism* directs us to maximize the good. *Utilitarianism* is a form of consequentialism which further specifies that the good consists in the well-being of sentient beings. *Act Utilitarianism* thus directs us to maximize aggregate well-being. This simple view has faced many pressing objections. Parfit's ethical theory can be understood in part as a reaction to these objections.

For example, egalitarians have objected that utilitarianism neglects the *distribution* of well-being across the population. Some object to allowing a single great harm to one to be outweighed by many small benefits to others. Section 3 relates Parfit's innovative response to such objections: to argue that the underlying intuitions are best accommodated by a modest revision to utilitarianism so as to give extra weight or priority to the well-being of the worse off. It also explores whether we can further improve upon Parfit's revisions here.

Others have raised concerns about the potentially self-effacing nature of consequentialist views. If believing some other view would have better consequences, does that suggest that consequentialism is self-defeating in any problematic sense? Is it always best to act according to the best rules? Section 4 discusses how Parfit's early work shed important light on such structural questions. Perhaps one of the most interesting results to emerge from Parfit's work here is an argument to the effect that it is common-sense morality, rather than impartial consequentialism, that faces the greatest risk of being (problematically) self-defeating.

A prominent objection to Act Consequentialism is that it too easily permits intuitively heinous acts such as the killing of innocent people (if so acting would save a larger number from similar harms). *Rule Consequentialism* directs us to follow, instead, the rules whose general acceptance would have the best consequences. This seems likely to include a rule against killing the innocent. Section

5 critically examines the prospects for Rule Consequentialism, alongside Parfit's ambitious arguments for the "Triple Theory", according to which the best forms of Kantianism, Contractualism, and Rule Consequentialism ultimately converge.

Our final two sections look at some of Parfit's most distinctive work. Section 6 explores questions of personal identity through time, and Parfit's arguments for the striking claim that identity is not what matters in survival. Section 7 offers a brief overview of key issues in population ethics – a new subfield of ethics that is largely built upon Parfit's seminal insights. In both cases, we find that incredible-seeming claims can be supported by arguments that seem almost inescapable. I find few philosophical puzzles to be as gripping – yet slippery! – and rewarding to grapple with as those contained within these pages.

Note: I have written this Element as an *opinionated guide* to Parfit's ethics, rather than attempting a neutral exegesis. Throughout, I try to explain what I find most valuable in his work, as well as what I think he may have been wrong about (and why). But the reader is encouraged to question my verdicts – and Parfit's too. We make progress in philosophy by questioning and probing each other's arguments and ideas. While I believe that this Element contains some important truths, my strongest hope is that it provokes readers to engage philosophically with Parfit's arguments and ideas, no matter whether they ultimately agree with them.

2 Rationality and Objectivity

Parfit was centrally concerned with questions about practical rationality and what we ought, all things considered, to do.[1] As Parfit uses these terms, they might come apart in cases of ignorance or misinformation. An agent might rationally act on mistaken beliefs, and thereby fail to do what they *ought* (given the facts) to do. But, in what follows, I will focus on cases in which the agent knows all the relevant facts, so that we can speak interchangeably of what they "ought" to do or what it would be "rational" for them to do.

It's common in our broader culture to implicitly equate practical rationality with self-interest. According to *Rational Egoism*, or the "Self-Interest Theory", what each person ultimately has most reason to do is whatever would make their own life go best, on the whole. A central concern of Parfit (1984), in the second of the book's four parts, is to undermine this common view. Parfit compellingly argues that rational egoism cannot sustain itself against simultaneous attacks

[1] *Practical* rationality concerns the rationality of choice and action – aspects of our agency that seek to *change* the world – in contrast to *theoretical* rationality which concerns the rationality of judgement and belief – aspects of agency that seek to *accurately represent* the world.

from two sides, and must ultimately give way to a competing view that is either more objective or else more subjective.

Subjectivist views hold that normative reasons are grounded in the agent's desires (which are not themselves rationally evaluable, at least on the most straightforward versions of the view). Practical rationality is thus limited to *instrumental rationality*, or the evaluation of means in terms of their effectiveness at achieving whatever the agent's chosen ends might be. This view is importantly distinct from Rational Egoism, as agents might care about more than just their own interests, and – most strikingly – they might fail to care about their own interests. While we will see that Parfit rejects Rational Egoism as unduly restrictive, subjectivism risks being too lax an account of practical rationality, as grossly imprudent behaviour tends to strike us as irrational. That is, a modicum of concern for your future interests strikes us as required by rationality, suggesting that – contra subjectivism – at least some of our ultimate ends are rationally evaluable after all.

Parfit is thus led to the view that some things objectively matter, in the sense that we all have normative reason to care about them, no matter what desires we happen to have to begin with. On this view, when we fail to care appropriately about the things that really matter, we are making a genuine *mistake*, and are rationally criticizable in much the same way that someone who fails to apportion their beliefs to the evidence is rationally criticizable.

While the central concern of this section is to explain the arguments outlined above, we will close by briefly exploring the metaphysics and epistemology of normativity that Parfit believed necessary to support objectivity in normative ethics.

2.1 Rational Egoism

In requiring agents to always prioritize their self-interest over any competing concerns, Rational Egoism is a strikingly restrictive theory. As Parfit (1984, chapter 6) observes, it surely seems like we can reasonably care about things other than just our own interests. We may, for example, care about other people, or about achieving some magnificent goal, more than we care about our own future happiness or overall interests.[2] Parfit thus proposes the following simple counterexample to Rational Egoism:

[2] Even if one thinks that these things would then count as being among one's interests, the point remains that we could reasonably care about them to a degree that is disproportionate in comparison to the amount that they contribute to our well-being. We may thus reasonably prefer an outcome in which we are overall worse off, but this special goal is better achieved, over an alternative in which we are personally better off but fail in this goal.

> *My Heroic Death.* I choose to die in a way that I know will be painful, but will save the lives of several other people. I am doing what, knowing the facts and thinking clearly, I most want to do. . . . I also know that I am doing what will be worse for me. If I did not sacrifice my life, to save these other people, I would not be haunted by remorse. The rest of my life would be well worth living. (Parfit 1984, 132)

Rational Egoism must condemn the agent's choice in *My Heroic Death* as irrational, as they knowingly go against what is in their self-interest. But given that the agent is fully informed, thinking clearly, and acting in a way that is morally admirable, it is difficult to see any fair, non-dogmatic basis for insisting that their choice is irrational, just because they chose to prioritize others' interests over their own. Unless supported by some incredibly compelling theoretical rationale, the implausibility of Rational Egoism's verdicts in cases like this gives us good grounds to reject the view in favour of some more permissive alternative.

Parfit (1984) goes on, in chapter 7, to undermine Rational Egoism at a more theoretical level. Compare the following three principles:

(A) No individual preference is intrinsically irrational (just in virtue of its content), not even preferring a lesser benefit over a much greater one.
(B) It's irrational to prefer a lesser benefit over a much greater benefit, merely on the grounds that the former occurs *now* whereas the latter occurs *later*.
(C) It's irrational to prefer a lesser benefit over a much greater benefit, merely on the grounds that the former accrues to *you* whereas the latter accrues to *another*.

Rational Egoists accept principle (B) but reject both (A) and (C). Parfit argues that this is an unstable position, as there are good theoretical grounds for treating (B) and (C) alike. Parfit's basic idea is that there is a kind of formal *analogy* between "I" and "now", or between *agent* relativity and *temporal* relativity. When Rational Egoism dictates that we must be *temporally neutral* (giving equal weight to our interests at all times) but *agent relative* (giving more weight to ourselves than to others), it reveals itself to be what Parfit calls an "incompletely relative" theory. A theory is on sounder structural ground, Parfit believes, when it is either fully relative or fully neutral, treating both these dimensions of variation alike.

Why does Parfit think this? One way to understand his core insight is to notice that choices are made not only by particular agents but also at particular times. (It may be helpful to think of the deliberating agent as a "momentary self", distinct from the various "future selves" that will replace them at later times.) Just as a deliberator may ask, 'Why should *I* sacrifice my interests just so that

some *others* may benefit?', so we may imagine them asking, 'Why should *I now* sacrifice my *current* interests just so that my *future* selves may benefit?'. If the former question is thought to raise a serious challenge to altruistic requirements, parity of reasoning would suggest that the latter question should be considered similarly challenging to requirements of prudence.

Rational Egoists might seek to defend requirements of prudence by appealing to the objective features of normatively significant phenomena such as pain. Pain matters because of how it feels, and the felt badness of pain is not affected by mere differences in timing. This is, Parfit suggests, an excellent defence of (B). But it is not one that the Rational Egoist can comfortably appeal to, for analogous reasoning would equally support principle (C). After all, the felt badness of pain is likewise unaffected by mere differences in *who* feels it.

Rational Egoism is thus undermined on both intuitive and theoretical grounds. We should instead accept a theory of practical rationality that is either more subjective or more impartial. Parfit's arguments here provide a nice demonstration of the power of philosophy to force a rethinking of prevalent assumptions. As a result of such arguments, philosophers now overwhelmingly reject this view. The same cannot be said of Parfit's next target, however, which enjoys much greater philosophical influence.

2.2 Normative Subjectivism

Normative subjectivists claim that we have reason to do whatever will fulfil our ultimate (non-instrumental) desires. On the purest version of this view, agents may be susceptible to rational criticism when they fail to effectively pursue their goals, but the goals themselves are immune from rational criticism. As Hume (1739, 2.3.3.6) famously declared, ' 'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger.'

Parfit disagrees, as it seems to make perfect sense to criticize desires, and not just beliefs, as "crazy" or irrational. To illustrate, Parfit (1984, 124) imagines an agent with *Future-Tuesday Indifference*, who 'would choose a painful operation on the following Tuesday rather than a much less painful operation on the following Wednesday'. The imagined agent knows he will subsequently regret it, but simply doesn't care – about either his future agony or the associated regret. Such an agent seems less than perfectly rational. Many of us would probably describe such a pattern of concern as "senseless" or even "crazy". As Parfit sums up his case: 'Preferring the *worse* of two pains, for *no* reason, is irrational.'

Future-Tuesday Indifference shows us that there's more to practical rationality than just taking the effective means to whatever your ends may be. Our ends

themselves are open to rational evaluation. At a minimum, there's some rational pressure to treat like cases alike, or avoid arbitrary distinctions (Smith 1994): if pain is worth avoiding on other days, and it feels no different on those calendar days arbitrarily designated to be "Tuesdays", then we rationally ought to regard Tuesday pain as similarly worth avoiding.

This is to suggest a *structural* rational requirement – a requirement governing combinations of desires. Such structural requirements by themselves do not yet establish that any desire is *intrinsically* irrational; they just specify that certain combinations of desires cannot rationally be held together. Sophisticated subjectivists might happily insist, in this way, that whatever desires you have must cohere together and avoid arbitrary distinctions, while retaining their core commitment to the idea that any desire *could* be rationally held (in isolation, or with the right companion desires).

Parfit's objection to subjectivism can be pressed further: avoiding arbitrary distinctions by becoming indifferent to *all* future agony would simply compound the error of the Future-Tuesday-Indifferent agent. To restore rationality, it isn't enough to be consistent. If sufficiently wrong-headed, that might just make you more consistently irrational. To do better, we must respond to evaluatively significant features of the world in the ways that they actually merit.

Parfit (2011a, 76) thus affirms as a *categorical* requirement of reason that 'We all have a reason to want to avoid, and to try to avoid, all future agony.' You may wonder: What about masochists for whom some degree of pain can serve, instrumentally, to bring them pleasure? They can simply weigh their reason to seek pleasure against the reason to avoid pain, and see which is the greater. Parfit need not deny that there are possible cases in which the reasons to avoid pain are outweighed by sufficient instrumental benefits. But to simplify the discussion, it will help to focus on cases in which there are no such instrumental benefits in play. So let's interpret "agony" here as meaning a state that is experienced as *entirely* negative in valence. So understood, Parfit's datum – that all agents have reason to want to avoid future agony – seems difficult to deny.

Normative subjectivists have trouble accommodating Parfit's datum, however. For their view seems to imply that agents never really have *reason to want* anything: our wants are simply taken as given, and the subjectivist instead focusses on what we have reason to *do*, namely, effectively pursue whatever it is that we antecedently want.[3]

[3] This raises a puzzle: Why would we have reason to pursue some end that we have no reason to want? *Hypothetical imperatives* of the form, "If you want X, you should do Y", present relations of normative inherence: *given* that X is worth pursuing, then Y is too. But a view on which there

Returning to Parfit's counterexample: if someone presently happens not to care about future agony, then (subjectivism implies) they've no present reason to try to avoid such future agony. That seems wrong. So we have two grounds here for rejecting subjectivism: it falsely implies (i) that agents have no reason to *want* to avoid future agony, and (ii) that some possible agents have no reason to *act* so as to avoid future agony. This is Parfit's *Agony Argument*.

Sobel (2011, 63) responds that subjectivists might yet accept a *Reasons Transfer Principle* according to which: 'If one will later have a reason to get O, then one now has a reason to facilitate the later getting of O.' If so, the agent's future reason to avoid concurrent agony provides the present agent with a reason to avoid that future agony. It's an interesting question whether we should consider the Reasons Transfer Principle to be compatible with the spirit of subjectivism. (It requires positing a kind of normative authority that goes beyond the agent's present deliberative perspective, thus conflicting with the traditional "internalist" strain of subjectivism associated with Williams (1981).) But even if (some) subjectivists can in this way avoid the problematic verdict about our reasons for action, they still face the first part of the objection: that their view appears to be incompatible with our having *reasons to want* to avoid future agony in the first place.

Perhaps it's psychologically inevitable that future agony will entail some thwarted future desires (assuming that agony necessarily either involves or generates a concurrent desire for the agonizing experience to cease). By subjectivist lights, those future desires may generate future reasons to avoid being in agony, and by the Reasons Transfer Principle, those future reasons may likewise give the present agent reason to avoid the future agony (if they can). But what is the status of the future desires that started all this? For subjectivists, they generate reasons just in virtue of being desires that the agent has – their specific content is irrelevant to their reason-giving force. So the agent may have equally strong desires to experience agony (without enjoying it in any way), or to robotically count blades of grass, any of which would end up having the same normative significance as the desire to avoid agony. This seems a troubling verdict: many of us, at least, would be inclined to think that the desire to avoid agony is *warranted* in a way that a gratuitous desire to experience agony, by contrast, is not. Such considerations may help to push us towards a more objective normative view.

Subjectivists like to point out that we often have reason to do what we desire. If desires ground reasons, that would certainly explain the correlation. But it is

are *only* hypothetical imperatives is effectively a form of normative nihilism – no more productive than an irrigation system without any liquid to flow through it. Or so it seems to me.

not the only available explanation. Parfit instead explains away the correlation: first, our desires might indirectly affect our reasons, for example, by making it the case that we would *enjoy* some activity (or else be unhappy without it). On any plausible objective view, happiness is one of the things that objectively matters, so it is to be expected that we will typically have reason to fulfil our desires if this would make us happier. Second, our desires may often *track* the things that really matter, or are objectively good (in much the same way that our beliefs track the truth). Candidate objective goods include things such as happiness, achievement, success in one's central life goals, friendship and loving relationships, and helping others in need. It should come as no surprise that reasonable people tend to desire and pursue such ends, if (as many objectivists believe) they are genuinely good things that *merit* our attraction and pursuit.

To properly test our intuitions about subjectivism, then, we must consider special cases in which desire-satisfaction diverges from happiness and other candidate objective goods. In such cases, it no longer seems so plausible that desire-satisfaction is the only thing that matters. A major remaining challenge for the Parfitian objectivist, however, is to assuage our theoretical misgivings about how anything *could* really matter.

2.3 Objective Normativity

Objections to normative realism (the idea that some things really matter) come in two broad flavours: metaphysical and epistemic. The former concern the nature of *mattering*, or how normative properties could really *exist*. Next, assuming that objective normative truths are somehow "out there", epistemic objections remain about how we could possibly come to *know* them.

Mackie (1977, 38) famously objected that 'If there were objective values, then they would be entities . . . of a very strange sort, utterly different from anything else in the universe.'[4] Parfit (2011b, chapter 31) seeks to defang such metaphysical qualms by denying that objective values (or normative properties more generally) would have to exist 'in the universe' at all. Nor do they exist in some separate, ghostly Platonic realm. That is still to treat them too much on the model of concrete objects that exist in space and time. Instead, Parfit suggests, abstract entities like numbers and objective values exist in a 'non-ontological' sense. True claims about numbers and values are as true as true can be, but – Parfit insists – these truths 'have no positive

[4] As Kirchin (2010) argues, it's not so clear just what Mackie's target is. I focus here on objective values, broadly speaking, and ignore Mackie's misguided assumption that these would necessarily have a magnetic pull on our motivation.

ontological implications' (Parfit (2011b, 479). This is Parfit's *Non-Metaphysical Cognitivism* in a nutshell.[5]

Parfit thus hopes to secure the best of both worlds: the objectivity of robust normative realism, without the ontological costs. Whether this is a coherent position is, unfortunately, less clear.[6] Parfit (2011b, 479) claims that abstract entities 'are not a kind of entity about which it is a clear enough question whether, in some ontological sense, they exist, or are real, though they are not in space and time'. He seems to draw from this the conclusion that we can comfortably rely upon abstract objects at no theoretical cost. I wonder if a better conclusion would be that the theoretical costs of positing abstract objects are, as yet, *unclear*. But even this more moderate conclusion may be consoling in its own way. For it undermines the suggestion that there is anything *obviously* objectionable (or theoretically costly) about positing objective values.

Some sceptics have thought that objective values would be more problematic than other abstract objects. Mackie (1977, 40) supposed that they must be imbued with a kind of magical motivating force, claiming that '[a]n objective good would be sought by anyone who was acquainted with it'. Parfit (2011b, 268), by contrast, takes great care to distinguish motivating and normative reasons. We are substantively irrational when we fail to be moved by (known) normative reasons. But there is no force in the universe that prevents us from being irrational. Normativity is causally inert, on Parfit's view: it marks what truly ought to be done, but it cannot push us to do it. Their causal inefficacy makes Parfit's non-natural properties more metaphysically innocent (being compatible with the principle that physical effects can only stem from physical causes), but perhaps more epistemically puzzling.

If abstract objects cannot causally influence physical objects such as our brains, how can we possibly know anything about them? Parfit (2011b, chapter 32) suggests that causally responsive perceptual faculties are only required for detecting *contingent* truths, which could have been otherwise. Following Sidgwick (1907), Parfit suggests that the necessary truths of logic, mathematics, and philosophy are *self-evident* in the sense that full rational understanding of the claim in question gives one sufficient justification for believing it: no causal interaction or external evidence is required.[7]

[5] Parfit (2016) seeks to develop this meta-ethical view, together with Railton's naturalism and Gibbard's expressivism, so that all three converge. We haven't space to explore this here, but interested readers may look to reviews of the volume such as (Roojen 2017).

[6] Cf. Suikkanen (2017) and Mintz-Woo (2018). Related views are defended in Scanlon (2014) and Skorupski (2010).

[7] Indeed, the a priori nature of fundamental moral truths can be used to argue against metaethical naturalism, as per Howard and Laskowski (2019) and (Chappell n.d.a).

To appreciate that 2+2=4, or that pain is bad, you don't need to run a scientific experiment to better reveal the causal structure of the world. Once you've acquired the relevant concepts, you just need to think clearly. Not all self-evident truths are so obvious as these examples, and we are all fallible, imperfectly rational beings. So people may disagree about what is truly self-evident, and sometimes get it wrong. But the core suggestion is nonetheless that careful thinking *may* see us right, and at any rate is the only hope we have, so we might as well give it our best shot.[8]

3 Distributive Justice

Traditional consequentialist views (such as utilitarianism) are commonly criticized for neglecting *distributional* concerns. The most straightforward of these concerns involves the value of equality: Would it not seem better to have everyone content than to have half the population ecstatic while the other half is miserable, even if global net happiness is the same either way? Others object to aggregating different people's interests together, so that small benefits to sufficiently many might together outweigh great harms to a few. Finally, some have raised concerns about whether consequentialism can adequately account for obligations not to contribute to collective harms (such as pollution or climate change). In this section, we will examine Parfit's contributions to addressing these challenges.

3.1 Equality and Priority

Many people are drawn to the *egalitarian* idea that it is in itself bad if some people are worse off than others.[9] Parfit (1997) invites us to imagine a *Divided World*, where each half of the population lives in complete isolation from, and ignorance of, the other half. This stipulation allows us to bracket any merely *instrumental* effects of inequality, and focus instead on whether inequality is bad *in itself*, even apart from any bad effects it might typically have. Now compare the following two states of affairs:

(1) half at 100 units of well-being; half at 200
(2) everyone at 145.

Many people are drawn to the view that (2) is better than (1), even though it contains less well-being in total. If we take this evaluative claim to be a moral

[8] I further defend a version of Parfit's moral epistemology against sceptical worries in (Chappell 2017a).

[9] For simplicity, I focus here on the view that Parfit calls 'Telic Egalitarianism'. There is an alternative view, 'Deontic Egalitarianism', which directs us to remedy unjust inequalities, but does not count inequality as making outcomes *worse*. See Parfit (1997, 207–10) for more detail.