BOOK REVIEWS

Arvan, Marcus. *Rightness as Fairness: A Moral and Political Theory.* Basingstoke: Palgrave Macmillan, 2016. Pp. xi+271. $100.00 (cloth).

Marcus Arvan's *Rightness as Fairness* is a highly ambitious book. In fewer than 230 pages, Arvan hopes to demonstrate that we ought to evaluate moral theories in a similar manner to the sciences, that existing moral theories fall short on that evaluation, that moral normativity reduces to instrumental rationality, and that a new theory of rightness as fairness meets the scientific evaluative standards better than any of the alternatives.

Arvan adapts seven principles of theory selection from the sciences (13–24): Firm Foundations, Internal Coherence, External Coherence, Explanatory Power, Unity, Parsimony, and Fruitfulness. Meeting Firm Foundations is Arvan's only necessary condition for theory selection; the other criteria are desiderata. Firm Foundations requires that "theories based on common human observation—or observations that are taken to be obvious, incontrovertible fact by all or almost all observers—should be preferred over theories based on controversial observations that may seem true to some investigators but not to others" (9). Arvan dismisses existing moral theories because they are not based on Firm Foundations (30–35).

Instrumental rationality, that "if one's motivational interests would be best satisfied by Φ-ing then . . . one instrumentally ought to Φ" (24), is universally recognized in everyday life and the history of moral philosophy (25–27). Thus, Instrumentalism satisfies Firm Foundations for Arvan's moral theory, a "Humean reduction" of the normative to the non-normative (28).

Arvan's most important and radical arguments take up chapters 2 and 3, together constituting almost half the book. Chapter 2 introduces the problem of possible future selves, starting with a discussion of uncertainty about the future. Arvan suggests that, at least sometimes, we don't merely want information about probable outcomes; we want certainty. In his example, we want certainty about the future housing market so that we can time a house purchase appropriately. Since Firm Foundations deals in the obvious and incontrovertible, we might think that all homeowners and future homeowners want this certainty about the housing market. Such foreknowledge would frequently be futile; if we all had knowledge of the future housing market, demand for properties would change, potentially making it harder to get the property we would want in the future. Importantly, we also frequently avoid information that would make our own futures more predictable, such as failing to have a symptom checked by a doctor or avoiding social media before watching a prerecorded game. In contrast to Arvan, I'd therefore suggest that we don't genuinely seek certainty; claims along such lines are more

accurately described as wishful thinking. We accept uncertainty as a fact of life. *Que sera, sera.* Unfortunately, Arvan's claim that we want certainty, not just probability, is essential for his theory of rightness as fairness. If this desire for certainty is neither as widespread nor as problematic as Arvan suggests, the theory would fail Arvan's own Firm Foundations. But perhaps these impossible cases are only illustrative of Arvan's experience of the phenomenon, so first we must consider cases Arvan finds more seriously problematic.

Even if I am right that we accept this uncertainty in most cases, Arvan thinks we will be particularly troubled by uncertainty in moral cases (48). If I am a student now deciding whether to cheat, I should worry whether I will get away with cheating or get caught. Will I feel guilty or elated? Will I lose my college place, pass my exam undetected, or become a habitual cheat, increasing my chances of being caught eventually? Arvan views this case as relevantly like the house buying case, claiming we want to know not whether we might get caught but whether we will get caught and what will follow. My response that we simply can't know about the future "puts the cart before the horse" (49). Arvan claims we are motivated to want to know the future, and since instrumental rationality "defines normative rationality in terms of things we are actually motivated to want," we rationally ought to pursue the possibility of gaining some certainty about relevant future events. The relevant events are those which present themselves as the "problem of possible future selves."

The problem of possible future selves is this: I am trying to decide right now whether to violate a moral norm. I rationally ought to do something that I won't regret in the future. Future regret depends on my current action thwarting the interests of my future self. But since I don't know who my future self will be, nor what interests my future self will have, I don't know whether my current action will thwart or further the interests of my future self. Resolving the problem of possible future selves is a task for chapter 3, but first, Arvan rejects an appealing solution to this "problem": why not take an educated guess about your future self and your future interests? If you're now a decent person, breaking moral norms will likely have some negative impact on your future self. If you're now a gangster, deception and coercion probably won't thwart the interests of your future self. This kind of thinking would be exactly right, suggests Arvan, were there no better option (71). This paves the way for Arvan's claim that "morality is the most instrumentally rational way to respond to the problem" (72).

Chapter 3 admits that there is only a partial solution to the problem of possible future selves. This partial solution requires our present self and all possible future selves to cooperate diachronically to (1) agree on a set of shared interests and (2) mutually act on these interests. Diachronic agreement and action by present and possible future selves involve considering these interests from each different perspective of all possible selves.

This is a complex maneuver, but Arvan claims that the motivation to cooperate comes from all selves recognizing problem cases for possible future selves, cases where we wish to avoid regret. Hence, each self should voluntarily decide how to respond to a desire for certainty about our future, despite our ignorance of it (80). Arvan finds this "commonsensical," in that this joint venture provides us with one element of certainty about our future and something over which we can have control. Return to Arvan's definition of rationality. If my goal is to avoid

regret, and cooperation with my possible future selves guarantees that I will avoid regret, then I rationally ought to cooperate with my possible future selves.

My future particular interests are unknowable. Hence, present and possible future selves must agree to act on "voluntary interests" which are instrumentally rational for all selves to agree to. Voluntary interests are those over which (phenomenologically, at least) we have control, such as choosing not to dwell on the past or more generally choosing not to fixate on things that are wholly or partially out of our control.

From this process, Arvan derives the Categorical-Instrumental Imperative (CII): "voluntarily aim for its own sake, in every relevant action, to best satisfy the motivational interests it is instrumentally rational for one's present and every possible future self to universally agree upon given their voluntary, involuntary and semivoluntary interests and corecognition of the problem of possible future selves, where relevant actions are determined recursively as actions it is instrumentally rational for one's present and possible future selves to universally agree upon as such when confronted by the problem—and then, when the future comes, voluntarily choose your having acted as such" (92–93).

Setting aside the tension involved in an imperative to do something voluntarily, I cannot imagine anyone acting on a maxim so difficult to parse. As I understand the CII, the underlying idea is this: Before I act, I should think about the fact that I don't know whether my actual future self will rejoice in or regret my action. If and only if I judge that this uncertainty is a problem, then I should take the interests of all my possible future selves into account. Then, no matter who my future self turns out to be, I can be certain—because she'll be cooperating with me for her own sake—that she'll agree I acted so as not to thwart interests I could now reasonably expect her to have. The consequence of this should be certainty that I won't later regret what I'm about to do now.

Arvan applies the CII to the case of cheating. In isolation, the action might seem rational, but it is not instrumentally rational to cheat in the broader context of a whole life (95). What we do now just might ruin our later life chances, and we learn from experience that it is unwise not to care about our future selves.

Arvan provides mathematical moves from decision theory to convince us that it is futile to focus on the outcomes of close possible selves (97–100). This move is essential for his method. When we encounter questions like "Should I cheat on this exam?" as problem cases, Arvan claims that decision theory requires us to sum up all relevant outcomes multiplied by their probability of occurring. Arvan states that in problem cases the CII requires us to consider an infinite number of possible future selves. Thus, all the possible positive outcomes of cheating (+infinity) added to all the possible negative outcomes of cheating (–infinity) add up to zero. So, there is no advantage in betting on probable outcomes, since an infinite number of very small possibilities that outcomes won't go as I betted will negate my bet.

Although Arvan denies that this is a mathematical trick (99), the process is unconvincing. First, conceivability doesn't necessarily entail possibility, so there may not be an infinite number of possible future selves. Next, I rationally ought to ignore the interests of certain possible future selves, such as all possible future selves who have an interest in jumping under a train next Monday morning. In

addition, betting on outcomes does not appear irrational once we consider that what I do right now is one of the biggest influencing factors over who I will become next; current actions stack the odds in favor of close possible selves.

Thus, it seems that even if we care about our future selves, it's not irrational to care only about the interests of future selves in close possible worlds. A gangster isn't irrational if she decides to look after the interests of her possible future selves by striking the following deals with them: I'll keep my friends close and my enemies closer, I'll intimidate snitches, I'll coerce the weak but dangerous. I'll keep to the letter of the law as much as I can, if not the spirit of it, to minimize my chances of being caught. I'll use bribery and blackmail if I can to retain my freedom. If all this fails, my future self can agree that I took the most rational steps to increase my/our wealth, power, pleasure, and standing and minimize my/our chances of getting caught. Hence, I/we won't regret giving crime my/our best shot. Simply by acting like this, the gangster makes it considerably more likely that her future self will be someone who can shrug off her current immoral actions. There may be angelic future selves in remote possible worlds, but why should our gangster care about them? Worryingly, Arvan accepts that his view of morality is contingent on us having the kind of nature that worries about all possible future selves. Arvan could deny that the gangster experiences the possibility of different future selves as a relevant problem, but if he limits his CII to people who already do, or would be willing to, extend their concern to every possible self, Arvan is preaching to the choir.

The task of chapter 4 is to turn the CII into a moral principle. If, unlike me, you accept a CII that stretches to an infinite number of possible selves, then the extension of instrumental rationality beyond the self follows. This is because some of our possible future selves will identify their own interests with the interests of other sentient beings, both human and nonhuman. To reach agreement with all possible future selves, our current self should act on interests that it is instrumentally rational for all human and nonhuman sentient beings to agree on (128). One imagines that humans would act as advocates for the interests of nonhuman sentient beings. Of course, we might wonder whether or why we could or should give equal weight to the interests of all possible future selves, especially as this principle is becoming very costly. Arvan deals with costs in chapter 6.

In chapter 5, Arvan offers the Moral Original Position, in which moral agents deliberate behind an absolute veil of ignorance, treating the ends of all sentient human and nonhuman beings as if they were their own. But why is placing one's current self behind an absolute veil of ignorance the most instrumentally rational action, bearing in mind that some possible future selves are very unlikely to become actual future selves? What might have started as plausible, that we would be instrumentally rational to pursue certainty about the interests of our possible future selves, now seems anything but an observation "taken to be obvious, incontrovertible fact by all or almost all observers."

It is not until chapter 6 that Arvan takes up the topic of fairness. There is a thriving contemporary fairness literature, investigating the conceptual analysis of fairness, the conditions within which fairness is relevant, and the relation of fairness to morality. While the fair act is often the right act, importantly fairness and rightness can come apart. If a powerful parent threatens to inflict great

harms on humanity unless I give his undeserving daughter an A on her philosophy paper, then the right thing to do is to give the A grade, but it is nonetheless unfair to the other students. This view is common in the fairness literature: Brad Hooker notes that we "already have terms signifying the verdicts of all-things-considered moral reasoning . . . includ[ing] 'morally justified', 'morally legitimate', 'morally right', and 'morally best'. Don't we want 'fair' to have a distinctive and thus narrower meaning?" ("Fairness," *Ethical Theory and Moral Practice* 8 [2005]: 332). John Broome identifies instances of wrong but fair actions ("Fairness," *Proceedings of the Aristotelian Society* 91 [1990]: 87–101). I was particularly keen to see how *Rightness as Fairness* would buck the trend of this commonly held view among fairness theories. Sadly, Arvan doesn't engage with fairness literature on any level, but his arguments concerning fairness would have been stronger if he had.

I said earlier that the requirement to give equal weight to all possible future selves and all human and nonhuman sentient beings was not sufficiently widely supported to meet Firm Foundations. The contemporary literature offers a possible solution. Garrett Cullity's account of fairness trades on the close conceptual link between fairness and impartiality (see Cullity's *The Moral Demands of Affluence* [Oxford: Oxford University Press, 2004] and his "Public Goods and Fairness," *Australasian Journal of Philosophy* 86 [2008]: 1–21). Since impartiality comes in different forms and strengths, one form of impartiality will be more appropriate to a situation than another. If we fail to employ an appropriate form of impartiality, then we act unfairly. Arvan might have explored what kind of impartiality is most appropriate to the diachronic negotiation, rather than assuming that it requires a complete abstraction from our current interests.

Arvan offers a negative principle of fairness, which requires us to avoid or minimize coercion of all sentient beings, and a positive principle of fairness to assist all sentient beings in achieving certain goals they cannot achieve better on their own, setting all costs aside. The negative principle is a widely accepted moral principle. The positive duty to assist is less widely accepted but nonetheless a moral principle. Are Arvan's negative and positive principles not only principles of morality but also principles of fairness?

On Broome's view, fairness is irrelevant unless people have claims to the negative and positive treatment identified in the first two principles. Arvan might respond that these two principles are simply a step toward the principle of fair negotiation. Fair negotiations, on Arvan's view, give all human and nonhuman sentient beings equal bargaining power concerning the costs they're willing to face in meeting the demands of the first two principles. The early idea of negotiating with all one's possible future selves required a vivid imagination. Negotiating with all sentient beings, whether human or nonhuman, looks impossible. What criteria must these negotiations satisfy to be fair? Arvan claims that fair negotiations are carried out from the Moral Original Position and with equal bargaining power. But to offer a counterexample, giving men an equal say in the abortion debate is not obviously a requirement of any common conception of fairness. Jonathan Wolff demonstrates that the ideals of equality and fairness are often in tension ("Fairness, Respect and the Egalitarian Ethos," *Philosophy and Public Affairs* 27 [1998]: 97–122). In contrast to Arvan's view, Broome's account of fairness renders the Moral Original Position redundant, since the business of fairness is to mediate between competing claims. We need to know about

the strength of claims and who holds them, but our identity need not be concealed from us.

While Arvan's work is highly innovative, stages of his argument fail to meet his own Firm Foundations requirement, and two of his principles of fairness are not obviously matters of fairness. I'm afraid I was unconvinced that we should abandon our preferred moral theories in favor of rightness as fairness.

Charlotte A. Newey
Cardiff University

Dorsey, Dale. *The Limits of Moral Authority.*
Oxford: Oxford University Press, 2016. Pp. 256. $74.00 (cloth).

In this book, Dorsey develops a novel account of the limits of moral authority—that is, an account of the ways in which, and the extent to which, moral considerations determine how we ought to live, all things considered. The book is carefully argued, and the view that Dorsey offers of the reason-providing force of moral and nonmoral considerations is in many ways appealing. It is an important contribution that anyone who aims to defend moral rationalism—that is, the view that rational justification entails moral justification—must engage with.

Dorsey has two main aims in the book. The first is to argue that moral rationalism is false (chaps. 2–4), and the second is to argue that, although complying with moral requirements is always justified, all things considered, at the default level (chap. 5), individuals can, by taking on certain kinds of commitments, make it the case that complying with at least some moral requirements would be wrong, all things considered (chap. 6).

Chapter 1 sets the stage for the book's central arguments by clarifying the questions that Dorsey aims to address and introducing a number of key concepts and claims. In chapter 2, Dorsey argues against a priori rationalism, which is the view that moral rationalism can be known to be true independent of first-order inquiry into the content of the moral and normative (i.e., all things considered) points of view. Chapters 3 and 4 offer independent arguments against moral rationalism. In chapter 3, Dorsey argues that the substantial appeal of the claim that the moral point of view is impartial in its content, along with the fact that prominent arguments against the impartiality of the moral point of view illicitly presuppose moral rationalism, gives us reason to reject moral rationalism, since it seems clear that it is at least sometimes rationally permissible to act in ways that are inconsistent with strict impartiality. Chapter 4's central claim is that there is no plausible account of the moral point of view that is consistent with both moral rationalism and the existence of supererogatory actions. Since there is, Dorsey thinks, good reason to believe that there are supererogatory actions, we should reject moral rationalism. In chapter 5, Dorsey argues that despite the fact that nonmoral considerations are sometimes sufficient to justify failing to comply with moral requirements, they will never, at the default level—that is, in the absence of special conditions that can affect the strength of nonmoral reasons—require noncompliance with moral requirements. Chapter 6 is a defense of the view that individuals can, by adopting a commitment to be guided by certain kinds of