Review of *Rightness as Fairness*, by Marcus Arvan

François Jaquet

This is a penultimate draft.
Please cite the published version that can be found here.

In *Rightness as Fairness*, Marcus Arvan sets himself an extremely ambitious goal: finding a new foundation for morality, one that would outperform the usual contenders in moral philosophy.

The whole project is built on the following methodological stance: just as our best scientific theories do, our moral theories should respect seven principles that allow us to distinguish genuine truth from merely seeming truth. As far as possible, they should: be grounded in claims that are nearly universally accepted ("Firm Foundations"); be internally consistent ("Internal Coherence"); cohere with non-moral facts ("External Coherence"); explain some of our observations ("Explanatory Power"); unify apparently disparate phenomena ("Unity"); postulate few entities ("Parsimony"); and solve theoretical as well as practical problems ("Fruitfulness").

According to Arvan, the main contemporary attempts to found morality – intuitionism, reflective equilibrium, moral language analysis, constitutivism, and dialecticalism – all breach these principles. In particular, they do not satisfy Firm Foundations, as the claims in which they ground morality are invariably controversial. Hence, we should be suspicious of these views. Arvan proceeds to present his own account, which purports to better respect all seven principles: morality must be grounded in Instrumentalism, i.e. the claim that "if one's motivational interests would be best satisfied by Φ-ing, then it is instrumentally rational for one to Φ – that is, one instrumentally ought to Φ" (p. 24). His argument to this effect is straightforward: being essentially normative, morality must be founded on a norm; and Instrumentalism is the only norm that satisfies Firm Foundations.

Arvan does not spend much time defending Instrumentalism, which is quite all right since few people deny that one instrumentally ought to act so as to satisfy one's motivational interests. The rest of the book consists of an attempt to derive a moral theory from Instrumentalism. To begin with, Arvan notices that most of us have an interest in knowing the interests of our future self, an interest that we can hardly satisfy given the obvious fact that we do not know what the future will be, and in particular who our actual future self will be among all our possible future selves. In order to make an informed career choice, for instance, you might want to know what you will enjoy doing ten years ahead. Alas, this is impossible since, as far as you know, you might turn out to have very different tastes and aspirations.

Arvan then argues that the only way for us to tackle this issue (which he labels the "problem of possible future selves") is to agree with our actual future self to form the set of interests that it is instrumentally rational for all our possible future selves to agree on. Such an agreement would indeed allow us to know what our actual future self will want, and act accordingly. Suppose that all your possible future selves should agree on willing to be a philosopher. Then, this is what your present self and your actual future self should agree on. Provided that you know this, you will know what your actual future self has an interest in: being a philosopher.

In light of Instrumentalism, it follows that our present self and our actual future self *should* make such an agreement and thus form the corresponding interests. (Of course, this presupposes that we *could* form these interests, which in turn presupposes that we exert some sort of control over what we want. But Arvan contends that some of our interests are indeed malleable: our "voluntary interests" and, to a lesser extent, our "semivoluntary interests". These are the interests our present self and actual future self should consent to modify so that they converge to form a coherent set.)

But then, what exactly is it that our present and future selves should agree upon? What should be the upshot of their negotiation? According to Arvan, they should cooperate to act on the following Kant-inspired principle:

> **The Categorical-Instrumental Imperative**: 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 co-recognition 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. (pp. 92-93)

Then, from this principle, he derives two Kantian-ish "interpretations", the first of which is:

> **The Humanity and Sentience Formulation**: 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 co-recognition that one's voluntary, involuntary, and semivoluntary interests could be identical to those of any possible human or sentient being(s), 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 in cases where one's present self wants to know and advance their future interests – and then, when the future comes, voluntarily choose your having acted as such. (pp. 127-128)

(I spare you the second interpretation, which is very much in the same vein anyway.) This pair of formulations is supposed to follow from the Categorical-Instrumental Imperative on the reasonable assumption that some of our possible future selves care about the interests of others to the point that they come to share those interests. If your future self turns out to be upset by animal suffering, for instance, then they will share animals' interest in the end of factoring farming. Since our present self and our actual future self must agree on a set of interests on which all our possible future selves could agree, they will agree on a set that includes the interests of all sentient beings, in accordance with the Humanity and Sentience Formulation.

As Arvan next remarks, the set in question is precisely that which we would accept from behind an “Absolute Veil of Ignorance” – where “one is to assume that one’s interests could turn out to be identical with the interests of any possible human or nonhuman sentient being(s)” (p. 155). This inspires him to give yet another formulation of the Categorical-Instrumental Imperative:

> **The Moral Original Position**: voluntarily aim for its own sake, in every relevant action, to act on interests it is instrumentally rational to act upon from the standpoint of a ‘Moral Original Position’ in which you assume that your voluntary, involuntary, and semivoluntary interests could turn out to be identical to those of any human or nonhuman sentient being(s), where relevant actions are defined recursively as those it is instrumentally rational to treat as such from the standpoint of the Moral Original Position. (p. 149)

Then, from this, Arvan derives four principles of fairness – Negative Fairness, Positive Fairness, Fair Negotiation, and Virtues of Fairness –, which he summarizes later under the following criterion for moral rightness:

> **Rightness as Fairness**: an action is morally right if and only if … it is [sic] (A) is morally relevant, (B) has coercion-avoidance and minimization, assisting human and nonhuman sentient beings to achieve interests they cannot best achieve on their own and want assistance in achieving, and the development and expression of settled dispositions to have these ends, as at least tacit ideals, and (C) is in conformity with the outcome of an actual process of fair negotiation approximating all human and sentient beings affected by the action being motivated by the above ideals and having equal bargaining power over how those ideals should be applied factoring in costs, or, if such a process is impossible, the outcome of a hypothetical process approximating the same, where moral relevance is determined recursively, by applying (B) and (C) to the question of whether the action is morally relevant. (p. 178)

(If that was not clear so far, both Arvan’s philosophical views and his writing style definitely place his work in the Rawlsian tradition.) At this stage, he takes himself to have established morality on the basis of Instrumentalism. In the remaining chapters, he maintains that his account reconciles theories as antagonistic as consequentialism, Kantianism, contractarianism, virtue ethics, libertarianism, egalitarianism, and communitarianism while faring better than each of them with respect to the seven principles of theory selection listed above.

Hereafter, I will accept Arvan’s arguments for the Categorical-Instrumental Imperative, the Humanity and Sentience Formulation, and ultimately the Moral Original Position – if only because I found them hard to grasp at times. Rather, I will take issue with his derivation of Rightness as Fairness.

Apart from its name, there is nothing moral to the Moral Original Position – or to the idea that it correctly models the way our present and future selves should deliberate, for that matter. For the sake of argument, I am prepared to grant that the actions to which we would assent in the Moral Original Position are those that satisfy Rightness as Fairness. What I contest, however, is that this provides us with a criterion for moral rightness, indeed that this has any ethical implication. What is striking in this respect is that Arvan does not seem to advance any argument to bridge that gap.

All he does, it seems to me, is insist that it is *instrumentally* rational to accept Rightness as Fairness from behind the Absolute Veil of Ignorance, and *a fortiori* at all (p. 178). Be that as it may, what we want to know is whether it is *epistemically* rational to do so, whether we have epistemic grounds for believing in Rightness as Fairness, or – and this amounts to the same thing really – whether Rightness as Fairness is true. Arvan remains silent on that matter. Now, maybe he would object that this is a distinction without a difference. After all, on more than one occasion he claims (in line with Unity) that it is a virtue of his theory that it unifies the normative realm, by making both prudential and moral reasons instrumental. Maybe, therefore, he believes that epistemic reasons are instrumental too because all reasons are instrumental in the end. This interpretation is uncharitable, though: Instrumentalism about epistemic reasons is no trivial claim, yet nowhere in *Rightness as Fairness* does Arvan explicitly defend it.

A claim he does defend, relying on empirical evidence, is that prudence and morality are one and the same thing (p. 222). A number of studies thus suggest that people who do not care much about their own future – children, adolescents, and psychopaths – tend not to care about morality either. Our instrumental reason to accept Rightness as Fairness would then also be a moral reason to act accordingly. But this argument is unconvincing for two reasons. First, these empirical data seem perfectly compatible with the view that morality and prudence are distinct normative systems. Since these people do not “appreciate the consequences of their actions” (p. 46), it should come as no surprise that they are interested neither in prudence nor in morality, for both prudential and moral norms depend on the consequences of our actions. Besides, one can imagine that these people are uninterested in prudential and moral norms simply because they are uninterested in norms as such. Second, the data in question would be unimpressive anyway once put in front of the wide consensus according to which morality and prudence are very different beasts.

Alternatively, Arvan could contend that the principles that stem from the Moral Original Position, whose form is not clearly that of *moral* principles, are nonetheless moral in virtue of their very content. Consider the Principle of Positive Fairness by way of illustration. This principle urges us to assist all sentient beings “in achieving interests they cannot best achieve on their own and want assistance in achieving” (p. 168). Taking into account the interests and wants of all sentient beings as it does, this looks very much like a moral injunction indeed. All things considered, it thus appears that the Moral Original Position generates moral principles. Unfortunately, this rejoinder raises the same issue we met earlier: it does not satisfy Firm Foundations since it rests on a contentious assertion. As a matter of fact, many moral philosophers would deny that moral principles are moral in virtue of their content, arguing instead that moral principles are moral because they generate reasons of a peculiar nature: categorical non-conventional reasons.

Although he does not address the present objection, Arvan anticipates a related concern. The concern is that the Categorical-Instrumental Imperative cannot ground morality because it provides us with the wrong kind of reasons: the reasons it gives rise to are hypothetical (i.e, dependent on our desires), whereas moral reasons are categorical (i.e, independent of our desires). Arvan's reply is as should be expected given his background methodology: while it may seem obvious to some that moral reasons are categorical, this view is far from universally accepted; many people, including philosophers, believe that moral reasons are hypothetical. This objection does not respect Firm Foundations and can be dismissed as a result (pp. 38-39).

Still, this will not suffice to rebut the present criticism, which does not presuppose that moral reasons are categorical but merely asks for a justification: why should we infer a moral claim (Rightness as Fairness) from a claim that is manifestly non-moral (the Moral Original Position)? To be clear, it is Arvan who bears the burden of proof in this context. For all he has shown, it may still be that moral judgments are uniformly false, even though it is instrumentally rational for us to believe in some. Until proven otherwise, it therefore looks like Arvan has failed to ground morality in Instrumentalism. At best, he has established that we have non-moral reasons to act in ways that are generally taken to be morally good – which would already be quite an achievement. But this is not enough to uncover morality's foundation. In retrospect, *Rightness as Fairness* was perhaps too ambitious an endeavour.