ROUTLEDGE FOCUS

NEUROFUNCTIONAL PRUDENCE AND MORALITY

A Philosophical Theory

Marcus Arvan

ROUTLEDGE Focus

Neurofunctional Prudence and Morality

A Philosophical Theory

Marcus Arvan

First published 2020

ISBN: 978-0-367-23015-9 (hbk)
ISBN: 978-0-429-27795-5 (ebk)

Chapter 3

Derivation of Morality from Prudence

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

3 Derivation of Morality from Prudence

I am far from the first to suggest that moral actions are prudent. As we have seen, this idea spans human history, from the teachings of major world religions, to ancient Western and non-Western philosophy, to common tropes in fiction, to how we typically raise our children and treat morality in our adult lives, and so on. However, Chapter 2's theory of prudence is novel. We saw that prudent individuals typically appear to internalize a specific form of *'moral risk-aversion'*: negative attitudes that treat immoral actions as risks categorically not worth taking due to the potential for severe regret, along with positive attitudes that treat moral actions as always having greater expected aggregate lifetime personal utility than immoral behavior. Chapter 2 theorized that prudent people tend to internalize these two sets of attitudes as standing constraints on first-order decision-making, thereby adopting a 'morally constrained utility-maximization' strategy in their everyday life decisions (see Figure 2.4).

This chapter argues that Chapter 2's theory of prudence entails a novel normative and descriptive theory of morality: a revised version of the theory I defended in *Rightness as Fairness: A Moral and Political Theory*. Section 3.1 provides an overview of Rightness as Fairness, along with a number of clarifying revisions. Section 3.2 then uses Chapter 2's theory of prudence to provide a new defense of Rightness as Fairness, showing how this book's theory of prudence can be used to defend the theory from a variety of criticisms. Finally, Section 3.3 shows how, when Chapter 2's theory of prudence is combined with this chapter's derivation of Rightness as Fairness, the result is a unified normative and descriptive neurofunctional theory of prudence, morality, and political morality.

1 *Rightness as Fairness*: An Overview

In *Rightness as Fairness*, I defended a broader psychological claim than I have here: namely, that virtually all of us (at least sometimes) want to know our future interests and weigh them against our interests in the present, so that we can be sure to act in ways we will not regret later.[1] Although I briefly drew connections between this claim and prudence,[2] I did not provide a detailed theory of prudence. Instead, I used a number of moral and non-moral cases to illustrate how we often appear to have these motives. Specifically, I argued that when the stakes of a decision seem high—such as when we are tempted to violate moral norms or are faced with a big life-decision (such as buying a home or getting married)—we sometimes want to know before making the decision whether we will regret it afterward.[3] I thought readers would find these cases persuasive. After all, people often talk about agonizing choices that 'keep us up at night', saying things such as, 'I wish I knew whether buying this home is the right choice', or 'I wish I knew whether I will regret lying to my boss'. Finally, I briefly suggested that these attitudes are prudent, as they keep us from making rash decisions such as buying a home on a whim or lying on impulse. The basic idea was that prudent people approach decisions with large stakes carefully, and with great forethought, in order to avoid decisions they might regret.[4]

I then argued that this kind of desire to know one's future interests generates a problem of diachronic instrumental rationality: the 'problem of possible future selves'—the problem of how an agent can possibly satisfy their desire to know their future interests before the future occurs.[5] I recognized that this problem is probably irresolvable in part, as the future can lead us to have involuntary and semi-voluntary interests (such as unexpected emotional reactions) that we cannot fully anticipate or control.[6] However, I then argued in Chapter 3 of *Rightness as Fairness* that it is possible for one's present and future selves to cooperate across time to partially resolve the problem, solving it as far as it can be. First, I argued that although one's present and future selves can both realize that one's present self cannot know which future will occur, both sets of selves can also realize that they each appear to have voluntary control over at least some of their interests: interests that they can *choose*.[7] Second, I argued that because both sets of selves can recognize that they can voluntarily choose at least some of their interests, it is

possible and rational for one's present and future selves to forge and uphold a diachronic contract (across time) on voluntary interests to share with each other, given one's ignorance of the future in the present. Finally, I argued that the only contract that can achieve this is an agreement to act on voluntary interests that one's present and *every* possible future self can agree to share, regardless of how the future might turn out. This enables one's present self to act in ways they know will satisfy their voluntary future interests no matter how the future turns out, provided they choose in the future to uphold the contract. The principle representing this contract is as follows (clarifying revisions in italics):

> **The Categorical-Instrumental Imperative:** voluntarily aim for its own sake, in every relevant action,[8] 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 other 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 of possible future selves—and then, when the future comes, voluntarily choose *to uphold* your having acted as such.[9]

On its face, this principle might seem implausible. First, it is very complex—which has led some to wonder how ordinary laypeople might understand or follow it.[10] Second, it may appear unclear why it is rational for one's present and future selves to forge and uphold such a contract, given how unlikely some of one's possible future selves are.[11] Third, it may seem unclear how one's future and every possible future self might arrive at such a universal agreement, given just how many possible future selves one has and the diversity of their potential interests.[12] Finally, several critics of *Rightness as Fairness* doubted the psychological claims the above argument is based upon. Why think that all of us sometimes want to know our future interests before the future occurs, so that we can avoid all possible regret?[13]

Shortly, I will argue that this book's theory of prudence helps resolve these issues. However, in order to do so, allow me to first present the rest of Rightness as Fairness as I previously defended it, along with some clarifying revisions.

1.1 The Categorical-Instrumental Imperative: Three Formulations

The Categorical-Instrumental Imperative holds that when one encounters the problem of possible future selves (viz. wanting to know one's future interests), one ought to act on voluntarily chosen interests that one's present and every possible future self can agree to share given their other interests. Let us now think about who one's possible future selves are. As Chapter 2 argued, prudent agents should regard their future as profoundly uncertain—as over the course of a complete life, one's life can turn in all kinds of unexpected directions. One related feature of human beings is that our involuntary and semi-voluntary inclinations to selfishness, other-regardingness, and altruism can change from moment to moment, week to week, or year to year. Nobody (besides Christ perhaps) always wants to behave altruistically. Sometimes—and science indicates it is more often than we might like to admit[14]—we are inclined to behave selfishly, often because of emotions that involuntarily or semi-voluntarily afflict us, such as anger or jealousy.[15] Conversely, we also sometimes have other-regarding interests, which may be rooted in emotions such as sympathy or compassion,[16] but also in reason.[17] Importantly, the extent to which we become more selfish, other-regarding, or altruistic may depend on our earlier choices and unexpected events. For example, a single choice in one's past—to pursue fame and career success over love—might alter the overall path of one's life to such an extent that, in one possible future, one becomes a narcissistic egoist, while in another possible future, one might have become more concerned for others.[18] Similarly, as St. Augustine's life famously illustrates, it is possible for a person to unexpectedly realize the error of their selfish ways, becoming a person committed to helping others.[19]

Because in different possible futures, one can involuntarily or semi-voluntarily end up with different levels of selfishness, other-regard, or even pure altruism, the Categorical-Instrumental Imperative requires one to act on voluntary interests that one's selfish and other-regarding possible future selves can voluntarily agree to share *despite* all of these possible differences. Consequently, the Categorical-Instrumental Imperative can be restated as follows (clarifying revisions in italics):

> **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 self-interested (viz. egoism) or concerned with the interests of other human or nonhuman sentient beings (viz. other-concern and altruism) along a vast continuum,* 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 to uphold your having acted as such.[20]

Now consider what such a universal agreement between one's present and every possible future self would involve. First, insofar as many of one's possible future selves care about their past—and many of one's possible past and future selves care in turn about their future—a truly universal agreement between one's present and every possible future self must be an agreement between every possible version of oneself one could possibly care about: past, present, and future. Such an agreement would, in other words, be a kind of 'categorical agreement': an agreement on how to act regardless of how the past or future might possibly turn out, and regardless of what one's interests in the past, present, or future could possibly be. Second, such an agreement would thus be one that advances the interests of one's egoistic possible selves but also the interests of one's other-regarding possible selves (including, as a limit-case, purely altruistic possible selves who care about the interests of all other human and nonhuman sentient beings equally).[21] Consequently, a truly universal agreement between one's present and every possible future self must strike a fair *intra*personal bargain between the interests of one's egoistic and other-regarding possible selves. Finally, to that extent, a universal agreement between all of one's possible selves will also constitute a fair *inter*personal bargain between one's own interests and the interests of other human and nonhuman sentient beings one might end up caring about. Consequently, I concluded that the Categorical-Instrumental Imperative requires acting in ways are fair to oneself (intrapersonally, to all of one's possible selves), where this is in turn identical to acting in ways that are fair to others (interpersonally, to all sentient beings). As a result, I concluded that the Categorical-Instrumental Imperative can thus be restated in a third way (clarifying revisions in italics):

> **The Kingdom-of-Human-and-Sentient-Ends Formulation:** voluntarily aim for its own sake, in every relevant action, to abstract away from the interests (or ends) of particular human or nonhuman sentient beings, acting instead on interests (or ends) it is instrumentally rational for all human and nonhuman sentient beings to universally agree to share given their different voluntary, involuntary, and semivoluntary interests—*including any and every possible combination of egoistic and other-regarding desires different beings might have*—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 to uphold your having acted as such.[22]

1.2 Comparison to Kantian Ethics

Notice how similar the three formulations of this principle are to Immanuel Kant's moral principle, the Categorical Imperative. Kant's first formulation of the Categorical Imperative holds that morality requires acting on maxims one can will to be universal laws of nature.[23] The first formulation of my Categorical-Instrumental Imperative is similar, entailing that morality requires acting in ways that all of one's own possible selves (and by extension, every other being they could care about) could universally agree upon. Kant's second formulation of the Categorical Imperative holds that we are to treat the humanity in ourselves and others as an end-in-itself.[24] Although the meaning of this formula remains debated,[25] my second formulation of the Categorical-Instrumental Imperative is similar but more inclusive. It requires one to act in ways that respect and advance the interests of oneself and every *sentient* being one could possibly care about, given their possible interests.[26] Finally, my third formulation of the Categorical-Instrumental Imperative, which requires seeking a consensus agreement between one's own interests and the interests of all possible sentient beings one could care about, is similar to but more inclusive than Kant's Kingdom of Ends formula, which holds only that morality is a matter of acting on principles that bring the ends of all *rational* beings into unity.[27]

My Categorical-Instrumental Imperative thus bears many similarities to Kantian ethics. However, there are a number of critical

differences. First, my principle and its defense do not involve appeal to any special kind of imperative: namely, genuine categorical imperatives, or 'ought'-statements that are true without any condition.[28] Genuine categorical imperatives remain controversial.[29] My account does not deny that genuine categorical imperatives might be true. On the contrary, I argue elsewhere that if they are, they converge with the Categorical-Instrumental Imperative and Rightness as Fairness.[30] My account in *Rightness as Fairness* (and in the present book) merely holds that we can derive a robust form of 'categorical moral justification' from purely *instrumental* normative foundations.

This is important for a number of reasons. First, I contend it is an epistemically more certain foundation for moral philosophy. Whereas categorical normativity is controversial, virtually everyone (including children, criminals, and psychopaths) appears to recognize instrumental normativity.[31] Second, Rightness as Fairness appears more likely to motivate moral behavior than Kantianism, as my theory roots moral concern in mental time-travel, risk-aversion, and other-perspective-taking—all of which have all been consistently linked to moral cognition and moral motivation (see Chapter 1). Conversely, the motivational power of moral principles alone (*qua* Kantianism) has been empirically challenged.[32] Third, although some argue that Kantian ethics can include animals (through arguments similar to those in *Rightness as Fairness*[33]), my account includes animals more naturally, as my account roots morality itself in terms of what our possible future selves might end up caring about (including other creatures). Fourth, my account holds promise to unify a variety of competing moral frameworks—deontology, utilitarianism, contractualism, and virtue ethics—in a manner that Kantianism does not. As I will argue shortly, once we see that the Categorical-Instrumental Imperative is (typically) prudent for individuals, we can see that it also plausibly maximizes long-term expected social utility (viz. utilitarianism)—as it requires us to advance our own ends and the ends of others in a 'positive sum' manner via universal agreement (viz. contractualism). Further, as we will also see, the principles of fairness that I derive from the Categorical-Instrumental Imperative place constraints on how we can treat others (viz. deontology), while also requiring the development of standing dispositions of character to act on these principles (viz. virtue ethics). Consequently, this book's theory not only holds promise to unify prudence and morality but also the insights of four major moral frameworks.[34]

Finally, whereas Kant's formulas are thought to entail several different (and potentially divergent) moral tests,[35] my formulations of Categorical-Instrumental Imperative jointly entail a single moral test: a Moral Original Position. To see how, consider what it would be to act in ways that one's self and all possible human and nonhuman sentient beings could agree upon together, abstracting away from their differences (viz. my Kingdom-of-Human-and-Sentient-Ends Formula). John Rawls famously asks us to imagine all individuals in a given domestic society as represented behind a veil of ignorance,[36] a hypothetical device that precludes anyone from arbitrarily privileging themselves or their interests over others—which he then later extends to an international community of peoples.[37] Importantly, Rawls imposes a variety of stipulations on the original position, including that it only includes human beings[38] (in the first instance, representatives of citizens within a single domestic society assuming a closed society[39]), along with assumptions of reasonably favorable conditions,[40] strict-compliance with whichever principles are selected,[41] and a focus on major governmental institutions (viz. the 'basic structure' of society).[42] Many of these assumptions have been contested, with critics objecting to Rawls's strict-compliance assumption,[43] his anthropocentric focus on human beings,[44] his abstracting away from justice in families,[45] and focus on a closed society instead of a cosmopolitan concern for all human beings globally.[46]

My Categorical-Instrumental Imperative does not justify any of Rawls's controversial assumptions. The Categorical-Instrumental requires us to act in ways that all of our possible selves can agree upon *simpliciter*, irrespective of the conditions that they could possibly find themselves in, and irrespective of what they might possibly care about. Bearing this in mind, consider a Moral Original Position: a hypothetical situation where one deliberates from behind a veil of ignorance, but without any of Rawls's other controversial assumptions about justice.[47] In the Moral Original Position, we simply imagine a prudent agent behind a veil of ignorance whose task is to seek to advance their own interests, and the interests of every possible being they might end up caring about, without being able to bet on probabilities. This is what the Categorical-Instrumental Imperatives holds: that one should act on principles that *all* of their possible selves could accept as a consensus agreement, given all of the possible beings whose interests they might care about.

Although some critics have worried there is 'nothing moral' about the Moral Original Position, suggesting it is purely prudential and does not provide epistemic reasons to conform to whichever principles are rational from its standpoint,[48] we are now in a better position to understand precisely how the Moral Original Position is an epistemically justified moral model. First, we are epistemically justified in adopting the Moral Original Position—and in acting upon whichever principles are instrumentally rational from behind its veil of ignorance—to the extent that we have epistemic justification for believing that prudence normatively requires acting on the Categorical-Instrumental Imperative. I will provide this derivation shortly using Chapter 2's theory of prudence. Second, the Moral Original Position is a moral model insofar as it requires us to deliberate fairly about our own interests in relation to the interests of others. Its veil of ignorance not only requires us to care about the interests of all human and nonhuman sentient beings impartially, which is widely thought to be a defining feature of morality.[49] It also approximates Kant's influential notions of morality as universalizability[50] and of realizing a 'kingdom of ends', a harmony of ends arrived at via abstracting away from particular ends.[51] Although some readers may worry my account is still merely one of prudence (viz. instrumental rationality) rather than morality (viz. 'categorical imperatives'), Chapters 4 and 5 argue that principles of theory selection support reconceptualizing morality in these terms.

1.3 Four Principles of Fairness

Allow me to now summarize the Four Principles of Fairness that I derived from the Moral Original Position. In *Rightness as Fairness*, I argued that in the Moral Original Position, all one knows are general facts about sentience, interests, and agency.[52] First, one knows behind its veil of ignorance that all human and nonhuman sentient beings experience the world, having interests of their own. Second, one can recognize behind the veil that every being has higher-order interests: an interest in not having their first-order goals thwarted (viz. coercion), and an interest in receiving assistance when they cannot achieve their goals on their own and desire assistance (viz. mutual assistance). Because all human and nonhuman sentient beings have these two higher-order interests, I argued that it is rational to agree to at least two principles of fairness[53] from behind the Moral Original Position's veil of ignorance:

> **The Principle of Negative Fairness:** all of our morally relevant actions should have as a guiding ideal, setting all costs aside, avoiding and minimizing coercion in all its forms (coercion resulting from intentional acts, natural forces, false beliefs, and so on), for all human and nonhuman sentient beings, for its own sake.
>
> **The Principle of Positive Fairness:** all of our morally relevant actions should have as a guiding ideal, setting all costs aside, assisting all human and nonhuman sentient beings in achieving interests they cannot best achieve on their own and want assistance in achieving, for its own sake.[54]

I then argued, however, that these principles pose problems of costs and conflicts. First, the Principle of Negative Fairness can generate internal 'coercion conflicts'. In order to minimize coercion in the world, including what I have called natural coercion,[55] we sometimes need to coerce, as some degree of coercion (viz. police or military force) may be necessary to prevent greater amounts of coercion (viz. crimes or military invasion). Second, the Principles of Negative and Positive Fairness can conflict with each other. Sometimes the best means to help some is to coerce others (via taxation, etc.). Indeed, this is what libertarians and liberal-egalitarians disagree over in political philosophy: libertarians argue that it is wrong to coerce people to help others, whereas liberal-egalitarians think this is what justice requires.[56] Finally, the principles of Negative and Positive Fairness run up against issues of personal costs. Recall that although some of one's possible future selves may be altruistic—seeking to avoid coercing others and to help others selflessly—many of one's possible future selves are more self-interested, not wanting to bear particular costs for others' sake. This is often salient in discussions of applied ethics. For example, some theorists (such as Peter Singer) argue that people should be willing to endure great personal costs to help others in great need.[57] However, others argue such costs may require too much sacrifice for the sake of others.[58] The parties to the Moral Original Position should recognize these issues behind the veil. Specifically, although they possess *all-things-equal* grounds to avoid and minimize coercion (viz. Negative Fairness) and help others (viz. Positive Fairness), the various possible selves they might be also possess diverging interests in how to settle costs and conflicts generated by the Principles of Negative and Positive Fairness.

Consequently, I argued the parties to the Moral Original Position have rational grounds behind the veil to agree upon a third principle: a Principle of Fair Negotiation for reconciling conflicts within and between the Principles of Negative and Positive Fairness, and between the application of those principles and personal costs. The basic idea here is that negotiating a consensus compromise with others on how to apply the Principles of Negative and Positive Fairness is something that would advance one's own possible egoistic interests *and* the interests of any other human or nonhuman beings one could end up caring about. Although *Rightness as Fairness* could have been clearer about this, agents in the Moral Original Position should want fair negotiation to have a particular end: a consensus agreement between all those whom one's actions might affect (as again, the Categorical-Instrumental Imperative requires seeking universal agreement). *Rightness as Fairness* also could have been clearer about the nature of fair bargaining. Because the parties to the Moral Original Position are behind a veil ignorance—a situation in which they cannot arbitrarily favor any of their possible selves over others—they should want negotiated agreements not to arbitrarily favor anyone they might care about more above others. I will argue below that these clarifications have new and unexpected implications for Rightness as Fairness. For now, let us simply assume that it is rational for the parties to the Moral Original Position to agree upon the following principle for the above reasons (clarifying revisions in italics):

> **The Principle of Fair Negotiation:** whether an action is morally relevant, and how the Principles of Negative and Positive Fairness and Virtues of Fairness (see below) should be applied factoring in costs, should be settled *by a fairly negotiated agreement:* either an actual agreement predicated upon all human and nonhuman sentient beings potentially affected by the action *being motivated by the above principles but lacking any arbitrary bargaining advantages or disadvantages*, *or to the extent that such an actual agreement is infeasible* (*including because nonhuman beings affected may not be able to negotiate given their capacities*), through a hypothetical process approximating the same, for its own sake.[59]

Notice that this principle does not preclude negotiating agents from taking into account real-world probabilities. The *Categorical-*

Instrumental Imperative holds that morality is a matter of acting on principles that all of one's possible future selves can agree upon without betting on probabilities. The Moral Original Position then represents this rational requirement through its veil of ignorance, which prevents the agent from favoring any of their possible future selves over any others on the basis of probabilities. The Principle of Fair Negotiation, on the other hand, is an *outcome* of a rational agent's deliberations from behind the veil—and my contention here is that one's present and all of one's possible future selves can *agree* to this principle letting real-world agents negotiate on the basis of probabilities. In brief, all of one's selves can recognize that real-world choosers must take into probabilities when weighing the Principles of Negative and Positive Fairness against costs, since all of one's possible selves live in an uncertain world where choosing on the basis of probabilities cannot be ultimately avoided. This is important because the Principle of Fair Negotiation thus permits fair (moral) negotiation to incorporate our preferences for 'special relationships'—for the people (friends, family-members, fellow citizens, and so on) we ordinarily take ourselves to be more likely to care about (and care about more), compared to complete strangers. All the Principle of Fair Negotiation requires is that one does not act in ways that *unfairly* neglect anyone in negotiation, including complete strangers—as the Categorical-Instrumental Imperative requires us to recognize that it is possible for our future selves to unexpectedly care about strangers.[60] This account, I believe, is highly intuitive: morality does permit us to negotiate norms that protect and prioritize special relationships. It simply requires that we do so in ways that are not unfair to others and which we *might* regret later.

Next, I argued that individuals in the Moral Original Position have rational grounds to agree upon a fourth principle: a principle of virtue that requires developing standing psychological dispositions to act on the other principles just arrived at:

> **The Principle of Virtues of Fairness:** all of our morally relevant actions should aim to develop and express stable character traits to act in accordance with the first three principles of fairness, for its own sake.[61]

Further, insofar as things like recognizing and addressing coercion, recognizing and helping people and animals in need, and negotiating fairly with others are all skills, this principle implicitly involves cultivating such skills.[62]

Finally, I combined these Four Principles of Fairness into a single criterion of moral rightness, which I now simplify and revise as follows to reflect the clarifying revisions above:

> **Rightness as Fairness:** an action is morally right if and only if it is morally relevant and *pursues or conforms to a fairly negotiated agreement*, where such an agreement is the result of all human and nonhuman sentient beings potentially affected by the action either actually negotiating with each other in conformity with the Four Principles of Fairness, or else through a hypothetical process approximating the same, for its own sake; where, finally, moral relevance is determined recursively by the Four Principles of Fairness (coercion, mutual assistance, fair bargaining, and virtues of fairness demarcating morally-relevant cases).[63]

This criterion is admittedly complex—so much so that some readers might wonder whether it can possibly be correct. However, we will soon see that its complexities in fact appear to be borne out in how ordinary people think about morality, particularly the role that negotiation plays in social-political theory and social justice activism.

2 Rightness as Fairness: A Revised Defense

Critics of *Rightness as Fairness* questioned several of the claims summarized above. I will now argue that this book's theory of prudence can be used to defend Rightness as Fairness against these concerns in a manner that leads the theory in important new directions.

2.1 A Revised Defense of Rightness as Fairness's Moral Psychology

One common objection to *Rightness as Fairness* was that it is unclear whether all people at least sometimes have the motivations that produce the 'problem of possible future selves': namely, a desire to know one's own future interests.[64] However, Chapter 2 defended a much more qualified claim: that *prudent* people typically want to avoid regret in morally salient cases, wanting to make (moral) choices they know they will not regret (Figure 2.5).

Several points here are important. First, if my argument in Chapter 2 is correct, then critics may have been correct in that not

everyone may have the kinds of motivational interests that make Rightness as Fairness rational. Indeed, as Chapter 5 examines, there may be people who lack these motives. Chapter 2's point was more qualified: namely, that prudent people typically have motives that generate the 'problem of possible future selves', as well as an overall moral-prudential psychology that I argue below can be used to provide a newer, better derivation of Rightness as Fairness. Further, although I argued in Chapter 2 that 'moral risk-aversion' typically appears to be prudent, given life's many uncertainties and particular empirical regularities, Chapter 5 will discuss the possibility of (hopefully rare) counterexamples: individuals who can act prudently without moral risk-aversion, and for whom Rightness as Fairness may thus be irrational.

To some readers, it might seem unacceptable for a theory to contend that morality may not always be rational. However, two central points of this book (to be defended in Chapters 4 and 5) are that (1) we should select theories on the basis of rigorous principles of theory selection, and (2) this book's theory satisfies such principles across a wide variety of normative and descriptive phenomena better than alternative theories of morality and moral psychology. Consequently, my response to *Rightness as Fairness*'s prior critics is twofold. First, this book provides better normative foundations for the theory. Whereas *Rightness as Fairness* defended the overly-ambitious claim that virtually everyone has motivations that make its principles rational, Chapter 2 argued merely that prudence *typically* requires these attitudes—an argument further supported below in Sections 3.2.2 and 3.2.3. Second, I think we should take seriously the (admittedly disturbing) possibility that morality may not always be normatively binding, for precisely the reasons this book's theory identifies. For as Chapter 5 argues, some normative philosophical thought-experiments and empirical research both suggest that morality may not always be normatively binding—though I leave these matters for further inquiry.

2.2 A Revised Derivation of the Categorical-Instrumental Imperative

In *Rightness as Fairness*, I provided two derivations of the Categorical-Instrumental Imperative. The first derivation was direct. In cases where one's dominant motive is to know one's future interests, I argued there is only one way to satisfy this interest: one's

present and possible future selves must agree with one another on voluntary interests to share across time given the inability of one's present self to know which future self one will be.[65] I also provided a second derivation using decision-theory and the mathematics of infinity, as I argued we have an infinite variety of possible future selves.[66] Richard Dees argued that both derivations are unsuccessful, contending that I must show that the expected long-term value of moral action is greater than immoral action.[67] I will now address this concern.

To see how this book's theory of prudence can be used to show that the expected value of conforming to the Categorical-Instrumental Imperative is at least typically greater than alternatives, consider again the two moral-prudential lessons examined in Chapter 2. Chapter 2 argued that prudent individuals typically internalize 'morally risk-averse' attitudes: beliefs and desires that (1) immorality is not worth the risk of immense regret, and (2) morality has greater-expected personal outcomes in the long-term, at least if you are patient. As we have seen in this chapter, the first of these attitudes gives rise to the problem of possible future selves. In morally salient cases, the prudent agent does not want to risk outcomes their future selves might profoundly regret. Instead, they want to make a choice they know they will not regret. The direct derivation of the Categorical-Instrumental Imperative in *Rightness as Fairness* (the derivation not involving infinite expected utilities) showed that the only way to satisfy this negative motive is for one's present and possible future selves to universally agree on voluntary interests to share for their own sake, irrespective of what the future holds. The rest of *Rightness Fairness*—the derivation of the three formulations of the Categorical Instrumental Imperative, the derivation of the Moral Original Position as a moral test, and the Four Principles of Fairness—were then shown to be the only principles that achieve that kind of agreement. Consequently, the Categorical-Instrumental Imperative and moral content of Rightness as Fairness can be derived from this book's account of prudent individuals' internalized categorical negative attitudes that *violating moral norms is never worth the risk of immense regret*.

However, this is only half of the story we can now tell. The next question then becomes why satisfying this negative motive (viz. conformity to Categorical-Instrumental Imperative) is instrumentally rational. *Rightness as Fairness* argued that this can be established through the mathematics of infinity. However, we

can now see that the answer is more straightforward. Chapter 2 argued that prudent agents also typically internalize a positive set of categorical attitudes—attitudes that *morality always has better likely personal outcomes than immorality in the long-term*, at least if you are patient. Dees's challenge was that my instrumental derivation of the Categorical-Instrumental Imperative succeeds if and only if the expected value of acting according to it is higher than other alternatives. This book has not provided a proof of this. Instead, it has defended a narrower thesis: that *in real life, given various empirical regularities*, prudent people typically internalize attitudes which entail—for them, given their morally risk-averse attitudes—that adherence to the Categorical-Instrumental Imperative maximizes expected aggregate lifetime personal utility. If this is correct, then it is typically prudent to be the kind of person whose attitudes make it rational to obey the Categorical-Instrumental Imperative and the moral theory (Rightness as Fairness) that principle entails.

Some readers may consider this a serious capitulation, as it amounts to recognizing that the Categorical-Instrumental Imperative (and Rightness as Fairness) may not always be rational. However, my response to this concern is four-fold. First, proof that morality is always rational is a bar that no moral theory has ever clearly met (which is why there is still so much controversy over these issues[68]). Second, this book's theory of prudence identifies the rationality of morality as an empirical issue: namely, whether, and to what extent, moral risk-aversion (as I defended it) is rational for different individuals in different contexts. As we will see in Chapter 5, I believe this is a very important question for future research to pursue. Third, there are further reasons for optimism about morality's rationality, as recent empirical work suggests that moral behavior is beneficial for individuals' physical and psychological well-being across the human life span, from childhood through later life.[69] Finally, this book's theory of prudence provides an intuitive model of how moral behavior is typically a prudent long-term strategy.

We can see how plausible its account is by considering real-life cases. Consider wanton 'immoralists': the common criminal, hitman, mob boss, political dictator, and so on. Although individuals like these can get away with and benefit from immoral behavior, there are grounds for thinking that their willing to risk immoral behavior tends to be an imprudent life-strategy, at least in the long-run. Criminals tend to end up in trouble with the law,

often ending up in prison. Similarly, many political dictators throughout history (ranging from Adolf Hitler to Saddam Hussein, Muammar al-Gaddafi, and so on) have ended up deposed or killed. Finally, in the course of everyday life, although people sometimes seemingly get away with immoral behavior 'scot-free', we have repeatedly seen ourselves or others face severe social sanctions for immoral behavior, such as divorces for infidelity to loss of jobs for deception, and so on. Sometimes, as in the case of the #MeToo movement's allegations of sexual misconduct, we see people suffer profoundly negative consequences for immoral behavior committed many years earlier. For all of these reasons, prudent individuals typically learn to not be these kinds of individuals: to not be individuals willing to risk immoral behavior for personal gain. Prudent people like you and I have seen the regrets immoral behavior tends to give rise to—so we internalize dispositions to not risk such immense regrets, instead aiming to behave morally 'for its own sake'. We come to believe that morality is a better life-policy than immorality and want to behave morally rather than immorally.

Are there potential counterexamples—individuals for whom moral risk-aversion is not required by prudence? Perhaps. First, there may be 'high-stakes' counterexamples: cases of political dictators or business leaders who never suffer for their moral crimes. Second, there may be 'low-stakes' counterexamples: cases where people routinely get away with more minor moral infractions, such never returning favors, cutting in line, and so on. This book cannot refute all of these potential counterexamples, nor do I think it should. Although proof that morality is always rational is arguably the 'holy grail' of moral philosophy, no such proof is widely accepted. I now believe this is for good reason. For as Chapter 5 discusses, I believe the rationality of morality is an empirical issue. Chapter 2 merely suggested that you, I, and most other everyday prudent people tend to learn over time to treat immoral actions as not worth the risk, and to internalize attitudes that morality has greater long-term expected benefit. I believe Chapter 2 made a compelling case for this, and Chapter 4 will argue that this book's overall theory coheres well with behavioral neuroscience. At the same time, I will recognize in Chapter 5 that it is an empirical question how far my account of prudence generalizes, and hence, whether morality (*qua* Rightness as Fairness) always is rational. I will now move on to Dees's next objection.

2.3 Defending Rightness as Fairness's *(Revised) Principles of Fairness*

Dees also called into question *Rightness as Fairness*'s account of fair negotiation, its Four Principles of Fairness, and standard of moral rightness. Here again, I think there is much to learn.

2.3.1 Clarifying the Instrumental Value of Fair Negotiation

Dees's first concern about the content of Rightness as Fairness is that it is unclear what instrumental value negotiating with other people has (viz. the Principle of Fair Negotiation), and how negotiating with other people amounts to negotiating agreements between our own possible future selves.[70] We can now see how this book's account of prudence can help us understand the instrumental value of negotiating agreements fairly (viz. ideals of coercion-minimization, mutual assistance, and fair bargaining), as well as how negotiating agreements with others is akin to negotiating agreements between our own future selves.

As Chapter 2 argued, prudent individuals typically internalize the desire to avoid outcomes they might profoundly regret. This means that they want to act in ways that make as many of their possible future selves happy with their decisions as possible—in the limit, all of them. Now consider the possible future selves one has. Some of one's possible future selves may be relatively selfish, caring about their own interests to the exclusion of others' interests. However, other possible future selves are more strongly altruistic, caring about other people's (and other creatures') interests over their own. *Betting* on any of these selves is an inherently risky endeavor. We see this in the case of criminals, and most powerfully in morally and prudentially tragic cases such as Hitler, Nazism, the Holocaust, and World War II. Criminals bet on their selfish future selves, not caring about how they harm others. The problem for criminals, however, is that other people do not typically enjoy being treated unfairly. For notice again what tends to happen, at least as a broad empirical regularity, to immoralists including everyday criminals, spouses guilty of infidelity, people who cheat on exams, slaveholders, and dictators such as Hitler. They are typically (though again, not always) *punished* for their unfair behavior, at least in the longer run: criminals tend to end up in jail, cheating spouses are often found out and divorced, exam-cheaters often receive failing grades and academic suspensions, political dictators often end up dead or imprisoned, and so on.

Is it possible to commit immoral acts and never suffer severe consequences? Perhaps. Again, some people may indeed get away with immoral behavior in perpetuity. However, these individuals appear to be outliers. The vast majority of us typically learn across childhood, adolescence, and adulthood that immorality is a bad bet in an uncertain world. Conversely, why should we think that negotiating fair compromises with others is (instrumentally) a good bet? The answer again has to do with one's possible future selves. As Chapter 2 argued, prudent individuals have internalized attitudes that make them want to act, in morally salient cases, in ways they will not regret (Figure 2.5). On my account, it is this motive that generates the 'problem of possible future selves': the problem of how to *know* one's future interests before the future occurs. I argue only the Categorical-Instrumental Imperative resolves this problem.

Bearing this in mind, consider the ways one's future selves can end up regretting one's actions in the present. One way one can end up regretting one's actions is by behaving immorally, as this can result in social punishment or guilt. However, another way one can end up regretting one's actions is by sacrificing an unfair amount for others. Sometimes, when we give others more than is fair, we involuntarily or semivoluntarily end up feeling resentful, such as when we help friends who never help in return, or try to do nice things for others only to have them treat us 'like a doormat'. Because some of one's possible future selves are more selfish and others more altruistic, betting firmly on one set of selves over the others is a serious risk. The Categorical-Instrumental Imperative holds that the only instrumentally rational response to these possibilities is to act in ways that one's selfish and altruistic future selves can agree upon *despite* their differences—so that one's future selves will not regret one's actions in the present no matter how the future goes. Furthermore, I argue that this is precisely what fairly negotiated agreements accomplish. When you and I negotiate a fair agreement, we act in ways that advance our own egoistic interests while also advancing the interests of the other person—a person who our future selves might end up caring about, either for self-interested reasons (if you object to how I treated you and retaliate) or due to empathy. Conversely, when we fail to negotiate fair agreements with others, we run serious risks: risks of treating others in ways that they object to and punish us for, or ways that involve us sacrificing so much that we regret it because our future selves judge we sacrificed 'too

much'. Negotiated agreements enable one to act in ways such that one's future self can advance their interests without feeling like they sacrificed excessively.

I believe this is an intuitive idea. If I fail to negotiate fairly with you, treating you the way I (unilaterally) think is best, then you may object and seek to punish me for failing to care sufficiently for your interests as you understand them. Conversely, if you fail to negotiate fairly with me, then you may treat *me* in ways I think are unfair or expect too much. It is only by arriving at and upholding a fair agreement that we can both be sure to advance our interests, whether they are more self-concerned (viz. egoism) or other-concerned (viz. altruism).

Finally, we can see the *disvalue* of failure to negotiate fairly by considering cases involving individuals or groups. Consider two spouses who disagree over how to raise their child. If one spouse insists upon their way, coercing the other spouse to go along with them, the other spouse may become resentful, blaming them for every bad decision the child makes and sparking constant arguments. And indeed, fairness violations generally appear to stimulate retribution.[71] Conversely, if the two spouses *fairly negotiate an agreement*, then they can raise the child in harmony, according to an agreement that both sides perceive as fair. Further, if the child-rearing strategy they agree upon appears to fall short of their mutual satisfaction, they can fairly renegotiate a new strategy. Similar considerations plausibly apply to groups, including nations. When one group treats another unfairly, abusing bargaining advantages to coerce the other group—as in slavery, racism, sexism, and so on—the dominated group may be (rightly) inclined to seek retribution. Further, although retribution may be impossible or ineffective in the short-term, it can *become* effective as time goes on—as illustrated by many social conflicts, ranging from the US Civil War to the Israeli-Palestinian conflict, and so on. Yet retribution and counter-retribution are plausibly in no one's interest. As historian and onetime US Diplomat to Germany William Dodd once said, 'It would be no sin if statesmen learned enough of history to realize that no system that implies control of society by privilege seekers has ended in any other way than collapse'.[72] It is only by fairly negotiating agreements with others, both individually and collectively, that we are able to realize a future that advances our own interests and the interests of others we *might* end up caring about.

As such, I believe Dees's concerns about the instrumental rationality of fair negotiation can be provisionally met. First, fairly negotiating with others appears to typically be instrumentally optimal given uncertainty about the future. Second, negotiating fairly with others is 'negotiating with one's future selves' in the sense that fairly negotiated agreements with other individuals and groups preserve the ability of one's own future selves to pursue their own goals and the goals of others they might end up caring about. However, Chapter 5 will explore the possibility that Rightness as Fairness (and hence, fair negotiation) is not always rational, ultimately leaving these as important empirical questions to be answered by future research.

2.3.2 Clarifying Coercion-Minimization, Mutual Assistance, and Fair Bargaining

Dees's second concern about the content of Rightness as Fairness concerns the *nature* of fair negotiation. First, Dees writes:

> [E]ven if we accept that the principle requires actual negotiations between separate people, approximating equal bargaining power only makes sense if we already know what counts as the requisite form of equality. But equality is a morally loaded notion, which is supposed to be the *product* of the negotiations.[73]

Second, Dees writes:

> …Arvan takes the actual negotiation clause very seriously, citing it as one of the great advantages of his view…. But to avoid the obvious problems with forcing people to negotiate for basic rights, he concedes that we do not need to negotiate with people who do not share a commitment to basic equality and to his principles of non-coercion and assistance to others (182–84). He somehow misses the fact that the most contentious debates—those about abortion, women's rights, LGBTQI rights, and even about welfare rights—are mostly about what is required to treat people equally and without coercion. On his grounds, then, these debates are not ones open to negotiations, but he thereby undermines the centrality of actual negotiations that are the hallmark of his theory.[74]

I argue that Chapter 2's account of prudence, and the revisions made to Rightness as Fairness in this chapter, resolve these problems in ways that lead the theory in important new directions.

I believe Dees is right: coercion-minimization, mutual assistance, and equal bargaining are all moralized notions. Abortion-opponents think abortion restrictions minimize coercion by protecting fetuses from death. Defenders of abortion rights, however, think abortion restrictions unfairly coerce women. And so on (for LGBTQIA+ rights, etc.). Let us think, then, about how a prudent individual—one who faces the problem of possible future selves—should consider this problem. Here the very same problem of uncertainty underlying Chapter 2's account of prudence recurs. Consider an anti-abortion activist or politician who opposes LGBTQIA+ rights. It may seem like a 'good bet' for this person (for instance, based on their personal religious convictions) to seek to impose their moral views on others—passing laws against abortion or LGBTQIA+ rights. However, what if—entirely unexpectedly—they come to have a family member who these policies would negatively impact? For example, what if, after passing an anti-abortion law, a politician has a daughter who loses her life seeking an illegal abortion? Or, what if, after passing a law against gay marriage, a politician learns their child is gay? In both cases, the person's moral beliefs and preferences may radically change—and indeed, there are notable cases in which people's moral views and preferences unexpectedly changed for precisely these kinds of reasons.[75] Do some people stick to their moral beliefs on particular issues (e.g. LGBTQIA+ rights) even after having a personal experience that challenges said belief? Almost certainly. The question, though, is whether it is rational to bet that one's moral beliefs on controversial issues will not change in an uncertain world. This is less obvious, given that people's beliefs often do seem to change over time and in unexpected ways. For example, cultural attitudes toward LGBTQIA+ issues changed quickly and dramatically in the US,[76] as they have for many other social moral issues, such as the morality of marijuana legalization,[77] the morality of particular wars,[78] and abortion.[79] If this book is correct, a prudent person should thus not assume that their beliefs and preferences on controversial moral issues will remain the same. They should instead recognize that life is profoundly uncertain over the long-term, and that their moral views on contentious moral issues *might* change over time in unpredictable ways.

Let us suppose, then, that prudent individuals will agree to Rightness as Fairness's principles (via my previous arguments), but also know that their own views and preferences about how to interpret and apply these principles might change dramatically over time. What should they then do? They should seek a *decision-procedure* that addresses that problem—one settling how people who are committed to ideals of coercion-minimization, mutual assistance, and fair (nonarbitrary) bargaining power should respond to the fact that their own views might change over time. Is there such a model? Indeed, there is: a series of Rawlsian Social-Political Original Positions.

To see how, consider the grounds that Rawls gives for the original position as a model of justice. Rawls supposes that the model represents citizens who (A) are committed to cooperating fairly for mutual advantage but (B) disagree over exactly what fairness involves, and also (C) recognize that their own views and preferences might change over time. In such a model, what is it to be committed to cooperating fairly for mutual advantage? As Rawls himself delineates, every person in the original position wants to advance their own preferences: they do not want to be coercively prevented from achieving things they want, and they may want to be helped by others (i.e. society) in pursuing their goals. The veil of ignorance then ensures that no one has arbitrary bargaining power over anyone else. Finally, the veil of ignorance requires the parties to recognize that their own preferences might change over time. Rawls's original position thus represents *all* of the preferences and uncertainty about the future that instrumentally rational agents should arrive at via the Moral Original Position. It models people committed to coercion-minimization (Negative Fairness), mutual assistance (Positive Fairness), nonarbitrary bargaining (Fair Negotiation), and the development of a standing sense of fairness (Virtues as Fairness).

Consequently, Rightness as Fairness should not be seen as entailing that its Four Principles of Fairness directly specify what is right or wrong for any situation or issue (e.g. abortion, global poverty). Rather, this book's theory of prudence can be seen as entailing a Prudential Original Position for deriving principles of prudence, a Moral Original Position for deriving moral principles, and a series of *Social-Political Original Positions* for interpreting and applying Rightness as Fairness under different possible social and political conditions, as people's views and preferences emerge or change over time. Allow me to explain.

2.3.3 Revising Rightness as Fairness's Content as a Moral and Political Theory

Dees's final critique is that I did not adequately circumscribe the moral limits of negotiation:

> Arvan emphasizes negotiations because he rightly observes that most human interactions are negotiated as we go…. But in practice the real-life negotiations that he promotes are either mere exercises in power or they are bounded by moral rules, rules that must be in place *before* the negotiations begin.[80]

Dees's concerns here are justified and have led me to see that Rightness as Fairness—as I previously developed it—is incomplete.

First, although Dees is correct that people rarely have equal/nonarbitrary bargaining power in real life, in many cases we do *approximate* it. We approximate equal bargaining power, for example, in friendships and relationships among equals, such as in marriages where neither spouse exploits unfair bargaining power over the other (such as financial power). In these cases, I think it is entirely intuitive to say that morality is the result of actual negotiation: in relationships among equals, the equal parties settle the moral terms of their interactions through forging and upholding fair agreements with each other.

Second, notice that there is an increasing realization in social and political theory that this is what justice requires more generally: eliminating arbitrary bargaining advantages. For example, consider evolving standards of consent in sexual relationships. One emerging ideal is that of equal bargaining power, such that the requirement not to abuse power is a requirement on consent itself. Similarly, consider conversations about white privilege: here, too, the idea is that it is wrong and unjust to not help those who have been (or are currently) oppressed, given one's own greater bargaining power as a white person. In other words, it is increasingly recognized that in order for our individual actions to be moral (viz. not wrongly coercing others through unfair bargaining power), social and political mechanisms must be in place to ensure there are no arbitrary inequalities in bargaining power.

Notice what this suggests. Rawls wrote in *A Theory of Justice* that the model of justice as fairness he presents is one that a

certain kind of person will find attractive: people who live in modern-democratic conditions. Why? I think we see the answer in the trends just mentioned. Many (though by no means all) people in modern democracies have, to a greater or lesser extent, *internalized* Rightness as Fairness's norms of fairness: ideals of non-coercion (e.g. free speech), mutual assistance (e.g. social security), and view that arbitrary bargaining advantages are unfair (no one should be disadvantaged on the basis of race, gender, etc.). More importantly, if this book is correct, there is a deeper truth here: namely, that *prudent* people will have these motives, wanting to organize their society on fair grounds—grounds that aim to reduce coercion, assist each other, and realize nonarbitrary bargaining power over contested issues (viz. Rightness as Fairness).

We can now see that Rawls gave us a model of exactly this: a model of free and equal individuals seeking to apply Rightness as Fairness to social-political conditions. By representing every individual in society behind a veil of ignorance, every party to the original position is given (A) an equal, nonarbitrary say over the extent to which they are (B) free from coercion and (C) able to seek and receive assistance by others—which is what Rightness as Fairness's Four Principles of Fairness entitle everyone to. Hence, Rawlsian Social-Political Original Positions *just are* models of what Rightness as Fairness requires, viz. prudence and morality. But there are further implications. If, as I argued previously, prudence and morality are a matter of acting in ways that treat one's future self as though their interests could be identical to any possible human or nonhuman sentient beings, agents motivated by Rightness as Fairness should want to realize a fair *world*: a world that treats every person and nonhuman sentient being they might care about as fairly as possible, including future generations. This suggests that prudence—which in the broadest sense means realizing a world that is rational for one's future selves, given life's uncertainty—requires settling how to apply Rightness as Fairness through a Cosmopolitan Original Position, which is something Rawls's cosmopolitan critics have long maintained.[81]

If this is correct, then Rightness as Fairness is not the end of social-political theory or applied ethics: it is the beginning. Prudence is a matter of acting on principles that are rational from a Prudential Original Position, morality is a matter of acting on principles (Rightness as Fairness) that are rational from a Moral

Original Position, and interpreting and applying Rightness as Fairness requires adopting a series of Social-Political Original Positions—beginning with a Cosmopolitan Original Position that includes all persons and sentient beings as *entitled to fairness*. Allow me to elaborate.

First, there is a question of which principles individuals in a Cosmopolitan Original Position would agree to at the level of ideal theory, that is, for defining a world in which all persons and sentient beings would be treated fairly. Many have suggested these principles are ones that would afford every human being in the world equal rights and liberties, fair equality of opportunity, and maximize the income and wealth of the world's global poor[82]—and others have argued for including animals in Rawlsian original positions.[83] However, another possibility is that individuals in a Cosmopolitan Original Position would recognize the importance of nation-states. In that case, it might be rational for individuals in the Cosmopolitan Original Position to agree to adopt an International Original Position to model fairness to nation-states of the sort Rawls defended in *The Law of Peoples*.[84] Next, individuals in a Cosmopolitan Original Position plausibly have grounds to seek an agreement on what fairness requires within nation-states—which would require them to adopt a Domestic Original Position of the sort Rawls defended in *A Theory of Justice* and *Political Liberalism*.[85]

Finally, individuals at each level should be concerned not only with understanding what Rightness as Fairness requires under ideal conditions (viz. assumptions of strict-compliance), but also how to interpret and apply Rightness as Fairness's principles fairly in our presently unfair world. In recent work, I have argued that this is best understood in terms of another type of iteration of the original position: a Nonideal Original Position that can be adapted to different kinds of unfair conditions (global unfairness, unfairness in modern-democracies, etc.) to specify what is fair in a given social-political domain given unfairness.[86] Although this work is still in its early stages, I have suggested that the Nonideal Original Position results in a variety of plausible principles for redressing injustice and unequal bargaining power.

The result we have ended up at is this. Some philosophers have criticized Rawls's outsized influence on modern political philosophy—arguing that Rawls's overall approach to theorizing about justice is misguided.[87] Our conclusions here, however, suggest otherwise. If this book's theory of prudence and

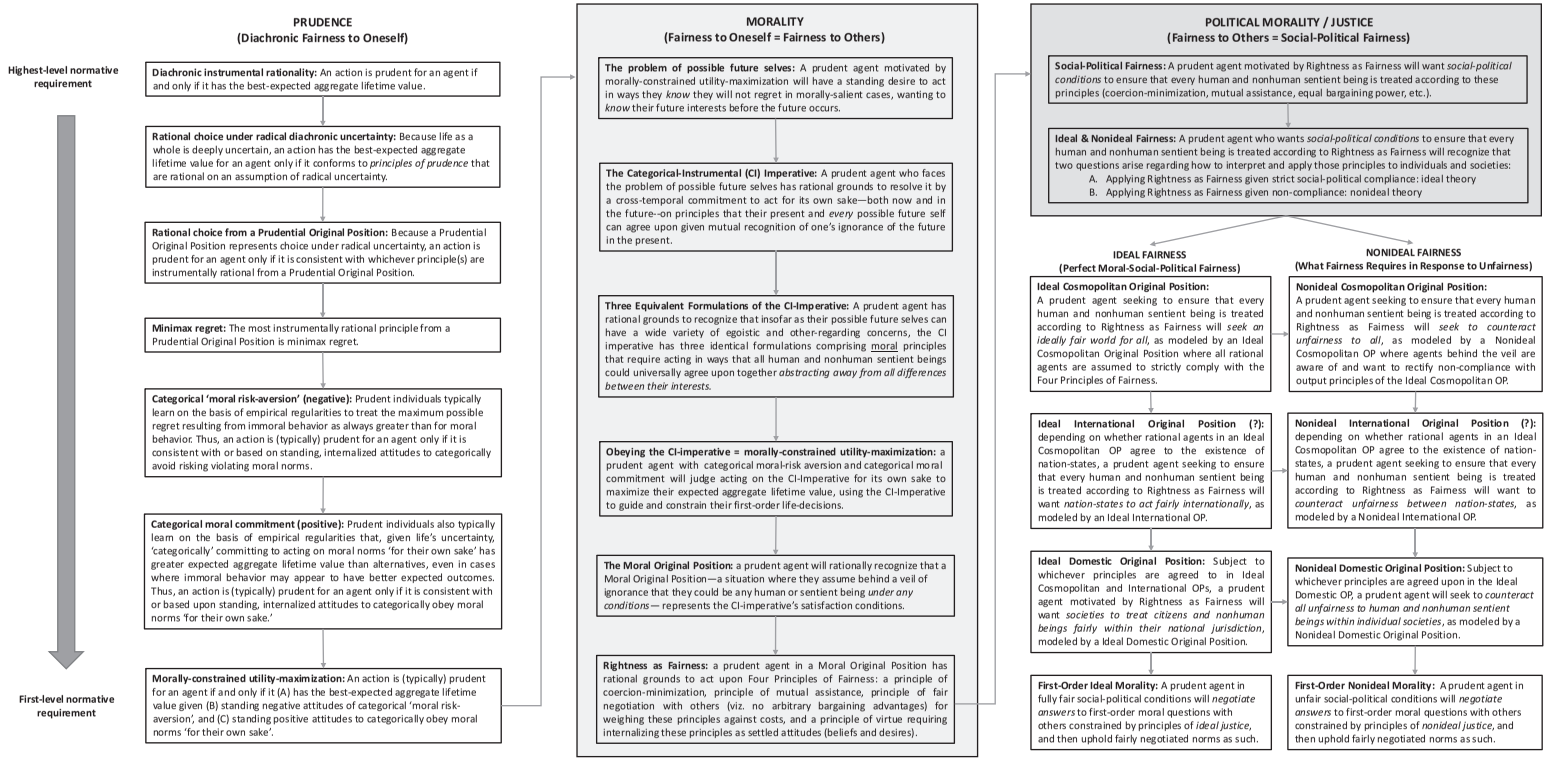

Figure 3.1 Outline of a Unified Normative Theory of Prudence, Morality, and Justice.

morality is indeed the best explanation for a wide variety of relevant phenomena—as Chapter 4 will argue—then the importance of Rawls's original position has been *under*estimated: a variety of iterations of the original position are necessary for understanding the normative nature of prudence, morality, *and* justice (Figure 3.1).

3 A Unified Descriptive Theory of Prudential, Moral, and Social-Political Psychology

Finally, my account entails the following unified descriptive theory of prudential, moral, and social-political psychology (Figure 3.2). Prudent agents learn to worry about the future, wanting to categorically avoid immoral behavior (as too risky) and instead categorically commit themselves to acting morally even if it appears that moral behavior might not maximize personal utility—as prudent agents learn to believe that moral behavior is likely to maximize long-term utility even when it appears it might not. These motivations make it rational for prudent individuals to encounter the 'problem of possible future selves'—that is, to want to know their future interests before the future comes, so that they can avoid possible regret. This leads prudent agents to want to justify their actions to all their possible future selves, but also, in turn, to all possible human and nonhuman sentient beings. This leads prudent agents to (at least implicitly) adopt a Moral Original Position, whereby they ask which principles of action all beings could agree upon from a standpoint of perfect fairness. This in turn leads prudent agents to recognize and aim to conform to Four Principles of Fairness. These Four Principles must then be applied, leading prudent agents to investigate what nonarbitrary bargaining power (viz. the Principle of Fair Negotiation) involves at a social and political level, both at a global level (viz. world affairs) and within particular societies. Because a variety of Social-Political Original Positions represent precisely this, the prudent individual will (at least implicitly) adopt Social-Political Original Positions to arrive at answers to questions of justice, and a Nonideal Original Position to determine what is right and just at a first-order level in an unjust world. Although imperfectly prudent agents (e.g. you, I, and every other human being) may only conform to this psychology imperfectly and perhaps at a very incomplete or implicit level, elements of this descriptive model should show up in the cognition and motivation of people who have internalized prudential motives.

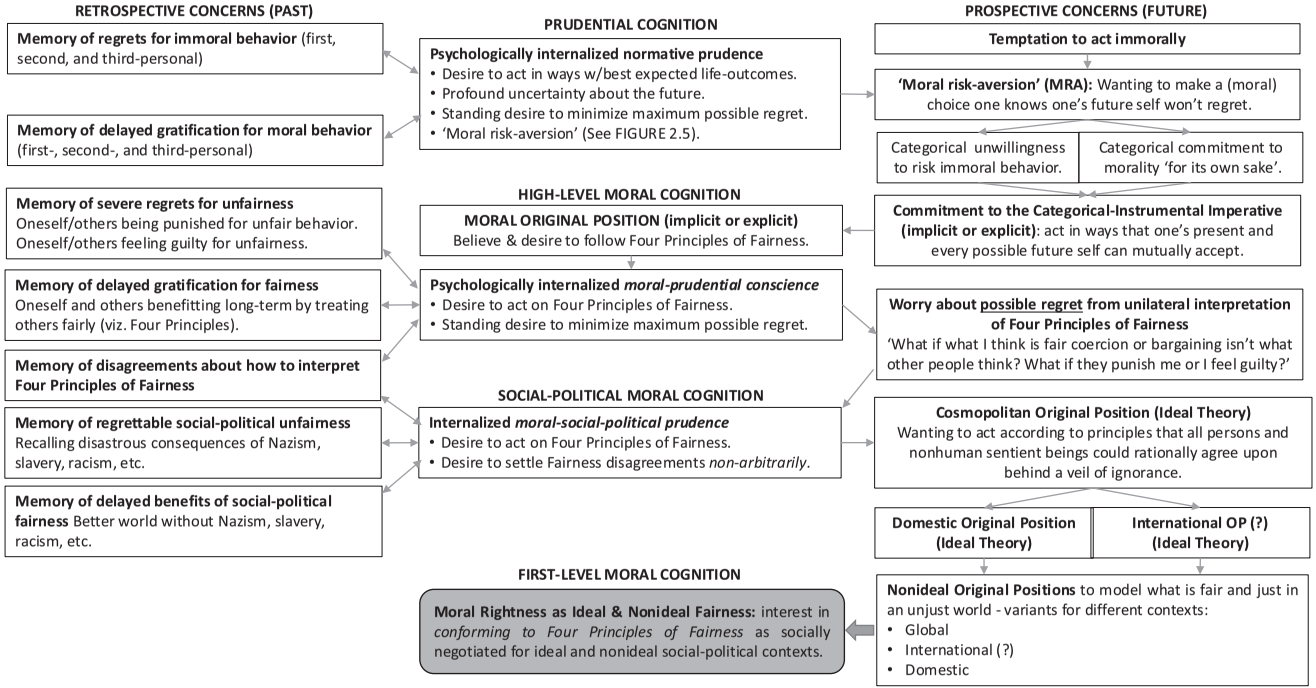

Figure 3.2 A Unified Descriptive Model of Prudential and Moral Psychology.

4 Conclusion

Some readers will undoubtedly find the theory of morality just defended, based on Chapter 2's findings, to be controversial. However, as we will now see in Chapter 4, it and the theory of prudence it is based upon both cohere with and explain the neurobehavioral phenomena summarized in Chapter 1 better than existing alternatives.

Notes

1 Arvan (2016a): 47–51.
2 Ibid.: 93–7.
3 Ibid.: Chapter 2, §2.4.
4 Ibid.: Chapter 2, §2 and Chapter 3, §5.
5 Ibid.: Chapter 2, §2.
6 For an in-depth discussion of these different types of interests, see Ibid.: Chapter 3, §2 and Chapter 6: §1.3.
7 Ibid.: 90–91, 56–64.
8 An important idea in *Rightness as Fairness* is that agents should not always encounter the problem of possible future selves, the normative problem that makes morality rational. Instead, prudent agents face the problem only when they worry about the future so much that they want to *know* how to act in a way that they will not regret (viz. the prudential psychology defended in Chapter 2 of this book). On my account, rational agents then use the Categorical-Instrumental Imperative to recursively define which other cases they should worry about in a similar fashion, thus delineating the limits of morality from within. Because I argue the Categorical-Instrumental Imperative entails Four Principles of Fairness—one of which is a Principle of Fair Negotiation—this involves socially negotiating fair standards regarding when we should worry about this future in this way, and hence, which of our actions are morally relevant. This is an important result, as one major objection to many ethical theories is that they are too demanding (Wolf 1982, 2012). Rightness as Fairness avoids this charge in that it permits us to socially negotiate how demanding morality is, including when we should or should not engage in moral reasoning.
9 Arvan (2016a): 76, reproduced with permission of Palgrave MacMillan.
10 Dees (2017), Newey (2017): 232.
11 Moore (2017): 526–7, Spencer (2018): 796–7.
12 Spencer (2018): 796.
13 Dees (2017), Moore (2017): 526–7, Newey (2017): 230–1.
14 Batson (2015).
15 See Arvan (2016a): 85–90.
16 Ibid.: 118–27.
17 Bentham [1780]: Chapter 1.
18 *The Family Man* (2000).
19 Augustine [400].
20 Arvan (2016a): 117, reproduced with permission of Palgrave MacMillan.
21 Ibid.: Chapter 4, §2.

22 Ibid.: 117, reproduced with permission of Palgrave MacMillan.
23 Kant [1785]: 4:421.
24 Ibid.: 4:429.
25 Arvan (2012).
26 I do not hold here that all sentient beings can actually rationally agree to things, as nonhuman animals in general do not appear capable of forging agreements. As I argued in Arvan (2016a): 156-57, the interests of sentient nonhuman animals are to be represented *by us* 'by proxy'—by us acting in ways we see would be rational for them to agree to if they possessed capacities for rational agreement.
27 Kant [1785]: 4:433–4.
28 Ibid.: 4:389.
29 Bukoski (2018), Joyce (2007).
30 Arvan (unpublished manuscript).
31 Arvan (2016a): Chapter 1.
32 See Batson (2015): esp. Chapter 4, though May (2018): Chapter 7, §7.3.3–7.4.1 defends the motivational power of moral principles.
33 Compare Korsgaard (2018), Arvan (2016a): Chapters 4–6.
34 Cf. Parfit (2011): Parts 2–5, Darwall (2014), Nebel (2012).
35 See Johnston and Cureton (2018): §9 for an overview. For an argument that Kant's formulas are identical (and hence co-extensive), see Arvan (2012).
36 Rawls (1999a).
37 Rawls (1999b).
38 Rawls (1999a): 475 and §3.
39 Ibid.: 7.
40 Rawls (1993): 297.
41 Rawls (1999a): 4–5, 216–7.
42 Ibid.: 6–10.
43 Farelly (2007), Mills (1997, 2017).
44 See Garner (2012, 2013).
45 Okin (1992).
46 Kuper (2000), Moellendorf (2002).
47 Arvan (2016a): Chapter 5.
48 Jaquet (2017): 318–9.
49 See Baier (1965), Hare (1952), Parfit (2011), and others.
50 Kant [1785]: 4:421.
51 Ibid.: 4:433. Cf. Arvan (2016a): 130, Rawls (1999a): §40.
52 Arvan (2016a): Chapter 6.
53 Newey (2018): 233–4 suggests it is unclear how these are principles of fairness *per se*. Although space constraints prevent protracted discussion, the short answer is that they are properly considered principles of fairness because they are justified by a perfectly fair procedure (the Moral Original Position).
54 Arvan (2016a, 2016b): 6, 153, reproduced with permission of Palgrave MacMillan.
55 Arvan (2016a): 161, 204, 207.
56 See Nozick (1974): Chapters 3 and 5.
57 Singer (1972).
58 Arthur [1981]. Cf. Wenar (2003).

59 Arvan (2016a): 154, reproduced with permission of Palgrave MacMillan.
60 Ibid.: Chapter 4.
61 Ibid.: 7, 154, 176, reproduced with permission of Palgrave MacMillan.
62 Stichter (2018).
63 Compare to Arvan (2016a): 7, 154, 178.
64 Dees (2017), Moore (2017): 526–7, Newey (2017): 230–1.
65 Arvan (2016a): 92–3.
66 Ibid.: 93–109.
67 Dees (2017).
68 See Bukoski (2016, 2017, 2018), Joyce (2007).
69 Post (2008).
70 Dees (2017).
71 Skarlicki and Folger (1997).
72 Nagorski (2012): 138.
73 Dees (2017).
74 Ibid.
75 See Lynch and Palmer (2013).
76 Schmidt (2019).
77 Felson *et al.* (2019).
78 Rosentiel (2008).
79 Fredericks (2019).
80 Dees (2017).
81 Caney (2005), Kuper (2000), Moellendorf (2002).
82 Ibid.
83 Elliot (1984), Garner (2013).
84 Rawls (1999b). Cf. Beitz [1979].
85 Rawls (1993, 1999a).
86 Arvan (2019).
87 Freiman (2017), Gaus (2016), Mills (2005).

References

Arthur, J. [1981]. World Hunger and Moral Obligation: The Case against Singer. In S.M. Cahn (ed.), *Exploring Philosophy: An Introductory Anthology*. Oxford: Oxford University Press. 2009, 142–5.

Arvan, M. (2019). Nonideal Justice as Nonideal Fairness. *Journal of the American Philosophical Association*, *5*(2), 208–28.

——— (2016a). *Rightness as Fairness: A Moral and Political Theory*. New York: Palgrave MacMillan.

——— (2016b). *Errata – Rightness as Fairness: A Moral and Political Theory*, https://philpapers.org/rec/ARVER, retrieved 24 July 2019.

——— (2012). Unifying the Categorical Imperative. *Southwest Philosophy Review*, *28*(1), 217–25.

——— (unpublished manuscript). Reformulating the Categorical Imperative.

Augustine, St. [400]. *The Confessions of St. Augustine*. J.K. Ryan (trans.), New York: Doubleday.

Baier, K. (1965). *The Moral Point of View*. New York: Random House.
Batson, D. (2015). *What's Wrong with Morality? A Social-Psychological Perspective*. Oxford: Oxford University Press.
Beitz, C. [1979]. *Political Theory and International Relations*. Princeton: Princeton University Press, 1999.
Bentham, J. [1780]. *The Collected Works of Jeremy Bentham: An Introduction to the Principles of Morals and Legislation*. Oxford: Clarendon Press, 1996.
Bruckner, D.W. (2003). A Contractarian Account of (Part of) Prudence. *American Philosophical Quarterly*, *40*(1), 33–46.
Bukoski, M. (2018). Korsgaard's Arguments for the Value of Humanity. *Philosophical Review*, *127*(2), 197–224.
——— (2017). Self-Validation and Internalism in Velleman's Constitutivism. *Philosophical Studies*, *174*(11), 2667–86.
——— (2016). A Critique of Smith's Constitutivism. *Ethics*, *127*(1), 116–46.
Caney, S. (2005). *Justice Beyond Borders: A Global Political Theory*. Oxford: Oxford University Press.
Darwall, S. (2014). Agreement Matters: Critical Notice of Derek Parfit, On What Matters. *Philosophical Review*, *123*(1), 79–105.
Dees, R. (2017). *Review of Rightness as Fairness: A Moral and Political Theory*. *Notre Dame Philosophical Reviews*, https://ndpr.nd.edu/news/rightness-as-fairness-a-moral-and-political-theory/, retrieved 5 June 2019.
Elliot, R. (1984). Rawlsian Justice and non-Human Animals. *Journal of Applied Philosophy*, *1*(1), 95–106.
Farelly, C. (2007). Justice in Ideal Theory: A Refutation. *Political Studies*, *55*(4), 844–64.
Felson, J., Adamczyk, A., & Thomas, C. (2019). How and Why Have Attitudes about Cannabis Legalization Changed So Much? *Social Science Research*, *78*, 12–27.
Fredericks, B. (2019). Poll Shows 'Dramatic Shift' in Americans' Attitude Toward Abortion. *New York Post*. https://nypost.com/2019/02/25/poll-shows-dramatic-shift-in-americans-attitude-toward-abortion/, retrieved 15 July 2019.
Freiman, C. (2017). *Unequivocal Justice*. New York: Routledge.
Garner, R. (2013). *A Theory of Justice for Animals: Animal Rights in a Non-ideal World*. Oxford: Oxford University Press.
——— (2012). Rawls, Animals and Justice: New Literature, Same Response. *Res Publica*, *18*(2), 159–72.
Gaus, G. (2016). *The Tyranny of the Ideal: Justice in a Diverse Society*. Princeton: Princeton University Press.
Hare, R.M. (1952). *The Language of Morals*. Oxford: Clarendon Press.
Jaquet, F. (2018). Marcus Arvan, Rightness as Fairness: A Moral and Political Theory. *Dialectica*, *72*(2), 315–20.
Johnson, R. & Cureton, A. (2018). Kant's Moral Philosophy. *The Stanford Encyclopedia of Philosophy*, E.N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/kant-moral/.

Joyce, R. (2007). *The Myth of Morality*. Cambridge, UK: Cambridge University Press.

Kant, I. [1788]. *Critique of Practical Reason*, In M.J. Gregor (ed.), *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy*. Cambridge: Cambridge University Press, 1996, 133–271.

——— [1785]. *Groundwork of the Metaphysics of Morals*, in Ibid., 38–108.

Korsgaard, C.M. (2018). *Fellow Creatures: Our Obligations to Other Animals*. Oxford: Oxford University Press.

Kuper, A. (2000). Rawlsian Global Justice: Beyond the Law of Peoples to a Cosmopolitan Law of Persons. *Political Theory*, *28*(5), 640–74.

Lynch, S.N. & Palmer, K. (2013). Republican Senator with Gay Son now Backs Gay Marriage. *Reuters*. https://www.reuters.com/article/us-usa-portman-gaymarriage/republican-senator-with-gay-son-now-backs-gay-marriage-idUSBRE92E0G020130315, retrieved 2 July 2019.

May, J. (2018). *Regard for Reason in the Moral Mind*. Oxford: Oxford University Press.

Mills, C. (2017). *Black Rights / White Wrongs: The Critique of Racial Liberalism*. Oxford: Oxford University Press.

——— (2005). "Ideal Theory" as Ideology. *Hypatia*, *20*(3), 165–84.

——— (1997). *The Racial Contract*. Ithaca: Cornell University Press.

Moellendorf, D. (2002). *Cosmopolitan Justice*. Oxford: Oxford University Press.

Moore, L. (2017). A Critical Review of Rightness as Fairness: A Moral and Political Theory. *Res Publica*, *23*(4), 523–9.

Nagorski, A. (2012). *Hitlerland: American Eyewitnesses to the Nazi Rise to Power*. New York: Simon & Schuster.

Nebel, J. (2012). A Counterexample to Parfit's Rule Consequentialism. *Journal of Ethics and Social Philosophy*, *6*(2), 1–10.

Newey, C.A. (2017). Marcus Arvan, Rightness as Fairness: A Moral and Political Theory. *Ethics*, *128*(1), 230–5.

Nozick, R. (1974). *Anarchy, State, and Utopia*. New York: Basic Books.

Okin, S.M. (1991). Justice, Gender, and the Family. *Philosophy and Public Affairs*, *20*(1), 77–97.

Parfit, D. (2011). *On What Matters*, *Vols. 1&2*. Oxford: Oxford University Press.

Post, S.G. [Ed.] (2008). *Altruism and Health: Perspectives from Empirical Research*. Oxford: Oxford University Press.

Pronin, E., Olivola, C.Y., & Kennedy, K.A. (2008). Doing Unto Future Selves as You Would Do Unto Others: Psychological Distance and Decision Making. *Personality and Social Psychology Bulletin*, *34*(2), 224–36.

Rawls, J. (1999a). *A Theory of Justice: Revised Edition*. Cambridge, MA: The Belknap Press of Harvard University Press.

——— (1999b). *The Law of Peoples, with 'The Idea of Public Reason Revisited'*. Cambridge, MA: Harvard University Press.

——— (1993). *Political Liberalism*. New York: Columbia University Press.

Rosentiel, T. (2008). Public Attitudes toward the War in Iraq: 2003–2008. *Pew Research Center*. https://www.pewresearch.org/2008/03/19/public-attitudes-toward-the-war-in-iraq-20032008/, retrieved 15 July 2019.

Schmidt, S. (2019). Americans' Views Flipped on Gay Rights. How Did Minds Change So Quickly? *The Washington Post*. https://www.washingtonpost.com/local/social-issues/americans-views-flipped-on-gay-rights-how-did-minds-change-so-quickly/2019/06/07/ae256016-8720-11e9-98c1-e945ae5db8fb_story.html?noredirect=on&utm_term=.91629d24a229, retrieved 2 July 2019.

Singer, P. (1972). Famine, Affluence, and Morality, *Philosophy and Public Affairs*, *1*(3), 229–43.

Skarlicki, D.P. & Folger, R. (1997). Retaliation in the Workplace: The Roles of Distributive, Procedural, and Interactional Justice. *Journal of Applied Psychology*, *82*(3), 434–44.

Spencer, E. (2018). Rightness as Fairness: A Moral and Political Theory, Written by Marcus Arvan. *Journal of Moral Philosophy*, *15*(6), 795–8.

Stichter, M. (2018). *Ethical Expertise and Virtuous Skills*. Cambridge: Cambridge University Press.

The Family Man (2000). https://www.imdb.com/title/tt0218967/, retrieved 5 June 2019.

Wenar, L. (2003). What We Owe to Distant Others. *Politics, Philosophy and Economics*, *2*(3), 283–304.

Wolf, S. (2012). 'One Thought Too Many': Love, Morality, and the Ordering of Luck, Value, and Commitment. In U. Heuer & G. Lang (eds.), *Themes from the Ethics of Bernard Williams*. Oxford: Oxford University Press, 71–94.

——— (1982). Moral Saints. *Journal of Philosophy*, *79*(8), 419–39.