

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Philosophy

2024

The Rules We Live By: Essays on Social Norms

Martinez, Camilo

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp01kd17cx17z>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

The Rules We Live By

Essays on Social Norms

Camilo Martinez

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR IN PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY
THE DEPARTMENT OF PHILOSOPHY

Adviser: Philip Pettit

January 2024

© Copyright by Camilo Martinez, 2024. All rights reserved.

Abstract

I study various puzzles arising from the idea that social norms are *rational equilibria*. This means that following these rules is rational for each of us, as long as we expect most other people to adhere to them. The puzzles concern (1) whether rule-following is rational, (2) whether we can be rationally motivated by our expectations of other people's attitudes and actions, (3) whether these expectations are stable enough to sustain an equilibrium, and (4) whether we see these equilibria as genuine norms, that is, as behaviors we ought to follow. I give a positive answer to all these questions based on the idea that norms are closely tied to *cultural practices*, that is, socially learned ways of acting that mediate our access to social goods such as reputation, esteem, and status. Rule-following is a rational way of partaking in certain practices, which we do expecting to receive approval from others. Moreover, these expectations endure among us because we teach them to newcomers to our groups, thereby establishing them as shared norms.

Contents

Abstract	iii
Introduction	vii
1 Rules and Rationalization	1
1.1 Acting for Reasons and Following Rules	2
1.2 The Dilemma	4
1.3 All Rules?	6
1.4 Biting the Bullet	8
1.5 Constrained Maximization	10
1.6 Sanctions and Internalization	12
1.7 The Practice Model	14
1.8 The Role of Rules	16
1.9 Look-Alikes	19
2 Two Problems for Bicchieri's Theory of Norms	22
2.1 Desiderata	23
2.2 Bicchieri's Theory	25
2.3 The Efficacy Problem	27
2.4 The Stability Problem	30
2.5 Esteem	33
2.6 Cultural Learning	35
2.7 Towards an Improved View	38
3 The Sanctions Paradox	41
3.1 The Motivational Question	43

3.2	The Sanctions Paradox	44
3.3	Altruistic Punishment	47
3.4	Cheap Sanctions	48
3.5	Esteem	50
3.6	Virtue Signaling	52
3.7	Cheap Signals	53
3.8	Social Sanctions as Cheap Signals	56
4	Taught Rules: Instruction and the Evolution of Norms	59
4.1	The Humean Program	61
4.2	Skyrms on Fairness	63
4.3	The Explanatory Deficit	65
4.4	Sympathy and Resentment	68
4.5	Population Replacement	71
4.6	Imitation and Myopia	74
4.7	Instruction and Natural Selection	75
4.8	The Emergence of Normativity	78
4.9	The Extended Humean Program	81
	Bibliography	85

List of Figures

2.1	Feedback loop between social expectations and norm-guided behaviors	39
4.1	Evolution of fairness without population replacement ($N=1000$; $\mu=0.001$; $t=200$)	73
4.2	Evolution of fairness with population replacement ($N=1000$; $\mu=0.1$; $t=200$)	73

Introduction

In the last two years, I've spent quite a lot of time in a coffee shop near my apartment. I like to get there early—the place opens at 7 a.m., seven days a week—sit down at a quiet table in a corner and write for 2 or 3 hours—or at least try to. From my table, I can see the front door and, a few steps away, the long wooden counter with the Italian espresso machine, the drip maker, and the industrial coffee grinder. Early mornings in the shop are typically quiet. Every once in a while, the odd customer will come in, greet the staff, order their drink, and get going. Between 8 a.m. and 9 a.m., however, it's rush hour. Customers start pouring into the shop, and very quickly, a long line forms, stretching from the register out to the storefront. Sometimes, two people standing in line will recognize and greet each other, but for the most part, the line consists of strangers, and there's no conversation. People patiently wait in line, often engrossed in their phones. Fortunately for all, the line moves quickly, and everyone gets their coffee within just a few minutes.

Scenes like this are so common in our lives that we usually take them for granted. If we ever stopped to think about them, however, they would surely fill us with awe. Here we have a group of strangers who manage to coordinate their actions subtly and effectively without any planning or communication. Coming into the shop, everyone wants to get their coffee as quickly as possible, but this tension or conflict in what different people want rarely leads to quarrels or fights breaking out between them. Instead, without having to be told what to do, each person knows exactly how to act so that everyone eventually gets what they want. Witnessing this feat of social coordination, the proverbial Martian scientist might be excused if he stands in puzzlement and wonder.

This dissertation is about the social tools we use to achieve these forms of collective, coordinated action. In the coffee shop and countless other social situations, we rely on *social norms* to help us figure out what to do. These unwritten, unspoken rules, such as the rule that one should stand in line and wait for one's turn to be served, are truly the bread and butter of social life: They're fundamental to our collective sustenance and flourishing, yet they're so commonplace as to be easily overlooked.

Social norms are fundamental to social life because they create a social *choreography* that allows us to live together peacefully and productively (Gintis, 2010). Thanks to these rules, we can get our coffee in the

morning by standing in line at the coffee shop, drive to work by following the traffic signs on the road, work together by each performing their assigned task or role, and enjoy the fruits of our joint labor by dividing them up fairly and equitably. Social norms are so central to our way of life that, even if some of them are outright unjust, destructive, or harmful, we all recognize that a life with norms is much better than a life without them—that a life without the social choreography they create would be chaos and mayhem. We want to get rid of bad norms, of course. But we wouldn't dream of getting rid of norms altogether.

Looked at closely, however, social norms turn out to be deeply puzzling phenomena. Norms, in a word, are mysterious beasts (Brennan et al., 2013, p. 2). Although they make our groups and societies tick, we don't fully grasp exactly what they are, where they come from, and why we follow them.

In the social sciences, social norms have long been regarded as a fundamental element of social order and a key determinant of human behavior. Indeed, norms are among the core resources—the “nuts and bolts,” as Jon Elster (1989b) calls them—that social scientists use to explain and predict human social action. For example, in the opening paragraphs of *The Rules of the Sociological Method*, Émile Durkheim insists on the importance of norms for our social behavior. As he writes, “[w]hen I fulfill my obligations as brother, husband, or citizen, when I execute my contracts, I perform duties which are defined, externally to myself and my acts, in law and in custom. Even if they conform to my own sentiments and I feel their reality subjectively, such reality is still objective, for I did not create them; I merely inherited them through my education” (1964, p. 57). Unsurprisingly, social norms crop up everywhere in the social sciences. Among many other things, social scientists invoke norms to explain why people treat strangers fairly (Henrich et al., 2005), why employers pay wages above market levels (Akerlof, 1982), why white men are more violent in the US South than in the North (Nisbett and Cohen, 1996), and why people vote in democratic elections (Knack, 1992).

But even in the social sciences, social norms are often met with perplexity, even skepticism. In an oft-quoted passage, for example, sociologist Jack Gibbs remarks that “no concept is invoked more often by social scientists in explanations of human behavior than ‘norm.’ However, despite the plethora of ‘normative’ explanations, the conceptual treatment of norms remains unsatisfactory” (1968, p. 208).

To begin with, talk of social norms often seems metaphysically suspect. As the above Durkheim quote might suggest, norms can easily look like “spooky” social entities existing over and above individuals and their interactions. Further, even if we grant that there's nothing metaphysically spooky about norms, it's still not clear exactly what individual-level facts are necessary and sufficient for a norm to exist in a group or how these facts first get established. Lastly, even accepting that social norms are often present in our lives, it's not clear why we are so wont to adhere to them, especially when doing so often seems to go against our

rational self-interest. As many economists have wondered, why do we bother doing such things as leaving a tip at a roadside diner we know we'll never visit again or cleaning up after ourselves after having a picnic at the park (Altman, 2012)?

In light of these worries, we might be tempted to conclude that social scientists are better off without employing the concept. That is, we might be tempted to become “eliminativists” about social norms, just as some philosophers are eliminativists about folk-psychological entities like beliefs, desires, and intentions. In the philosophy of mind, eliminativists argue that, since mind-body dualism is untenable and folk psychology is befuddled by conceptual and methodological problems, we should do without folk psychological notions. Likewise, we might be tempted to think that, since dualism about the social world is untenable—there are no “social substances” over and above individuals—, we should do without the notion of norms, seeing as it is plagued by conceptual and methodological difficulties.

But perhaps becoming eliminativists about norms would be too quick. For there is a way of understanding norms in philosophy that promises to assuage many of the worries just mentioned. This way of thinking about norms goes back at least to David Hume. In his famous discussion of the rules of justice and the institution of property in Book 3 of the *Treatise*, Hume writes:

I observe, that it will be for my interest to leave another in the possession of his goods, *provided* he will act in the same manner with regard to me. He is sensible of a like interest in the regulation of his conduct. When this common sense of interest is mutually express'd, and is known to both, it produces a suitable resolution and behaviour. And this may properly enough be call'd a convention or agreement betwixt us, tho' without the interposition of a promise; since the actions of each of us have a reference to those of the other, and are perform'd upon the supposition, that something is to be perform'd on the other part. (2007, p. 3.2.10; emphasis in the original)

As contemporary readers often point out, this passage contains an early formulation of a powerful and influential idea—namely, the idea that social norms can be understood as *rational equilibria*. Social norms are a particular type of social arrangement where each person in a group has a rational interest in following a certain rule provided she expects everyone else to follow it, too. For each person, it makes rational sense to stand in line, stop at traffic lights, do her part at work, and take only her fair share of the collective goods if she believes others will act in the same ways and expect her to act in these ways, too.

This *equilibrium theory of norms*, as I call it, is immensely influential in philosophy and adjacent disciplines. After Hume, it was taken up by David Lewis in his pioneering game-theoretic account of social conventions, and it has been articulated in different ways by different theorists since Lewis. Among contem-

porary writers, it features prominently in the work of Cristina Bicchieri, Ken Binmore, Philip Pettit, Robert Sugden, Edna Ullmann-Margalit, and Peyton Young, among others.

The equilibrium theory promises to dispel the puzzles surrounding the nature, origin, and efficacy of norms. Norms, it turns out, aren't spooky supra-individual social substances. At bottom, norms are nothing but arrangements of individuals whose beliefs, preferences, and plans are in equilibrium. These equilibria may come to exist in different ways. Hume himself believed that a norm like our rule of respect for each other's property can emerge out of a process of social evolution in which the rule "arises gradually, and acquires force by a slow progression, and by our repeated experience of the inconveniences of transgressing it" (2007, p. 3.2.10). Once a norm is established, however, people are willing to adhere to it simply because doing so is rational for them. Following the norm satisfies their preferences or desires, at least relative to their beliefs, which include beliefs about other people's desires and beliefs. By putting these ideas in place, the equilibrium theory of norms seems to defuse the threat of eliminativism about social norms.

Despite its promise, however, the equilibrium theory gives rise to a host of philosophical puzzles of its own. Indeed, over the last few decades, different problems for the view have been raised here and there by various critical commentators. One of the central tasks of my dissertation is identifying and properly formulating these different challenges. Across the four essays that follow, I identify and articulate four different puzzles or problems for the equilibrium theory of social norms:

(1) The *Guidance Problem* (aka *The Rule-following Dilemma*), discussed in Chapter 1—According to the equilibrium theory, norm-guided action is rational action. The Guidance Problem suggests that norm-guided action isn't rational because norms or rules can't ever *rationalize* one's actions. Whenever a rule applies to one's case, following it either goes against one's beliefs and desires or aligns with them. Hence, following the rule either makes *no* rational sense, or, if it does, this isn't *because* of the rule, but because one has a reason to promote what aligns with one's beliefs and desires.

(2) The *Efficacy Problem* (aka *The Sanctions Paradox*), discussed in Chapters 2 and 3—According to the equilibrium theory, people are motivated to follow norms by their beliefs or expectations regarding other people's attitudes and actions. The Efficacy Problem suggests that people can't be coherently motivated to follow a norm by these kinds of "social" expectations. In particular, they can't be motivated by their expectation that others will impose certain sanctions on them if they don't adhere to the norm and, *at the same time*, be motivated to impose sanctions on others themselves. If people are willing to impose sanctions on others, then they should be willing to follow the norm in the absence of sanctions.

(3) The *Stability Problem*, discussed in Chapters 2 and 4—According to the equilibrium theory, people's attitudes and actions concerning a norm constitute an equilibrium, in the sense of being self-sustaining and,

thus, stable. That is, people's expectations motivate them to act in ways that further reinforce these expectations. The Stability Problem suggests that people's expectations aren't sufficiently stable to constitute an equilibrium. In particular, people's beliefs or expectations about what other people think they ought to do aren't self-sustaining: They may motivate people to follow a norm, but they're not reinforced by widespread norm compliance.

(4) The *Normativity Problem* (aka *The Explanatory Deficit Challenge*), discussed in Chapter 4—According to the equilibrium theory, social norms can be understood as equilibria in people's beliefs, desires, and actions. The Normativity Problem suggests that this type of equilibrium doesn't yet amount to a proper social norm, insofar as people need not adopt any sort of normative attitude towards it. They may believe that others follow the equilibrium, and they may prefer to follow the equilibrium themselves. Yet, they need not think that one ought, in general, to follow the equilibrium, nor do they have to be disposed to censure or disapprove of people who don't follow it.

Taken together, these puzzles or problems pose a significant challenge to the equilibrium theory. They suggest that following a social norm can't be understood as rationally partaking in an equilibrium out of one's belief that others are partaking in that equilibrium, too. In contrast, perhaps following a norm consists in acting out of an intrinsic, internalized desire to follow a rule, independently of whether one thinks others are also following it. On this alternative picture, when we obey social norms, we're something like the "mindless plaything of social forces," (Elster, 1989a, p. 97) as opposed to agents who act out of an appreciation of what courses of action best promote their desires relative to their beliefs.

In the four essays that follow, I endorse, defend, and elaborate on the equilibrium theory of norms. I do so by offering solutions to the four puzzles or problems just introduced. Although they vary from case to case, my solutions have a common theme: In one way or another, they all rely on the idea that social norms are deeply *cultural* phenomena.

Norms are cultural phenomena because they're a part of *culture*, namely, the body of knowledge, skills, and information that we humans have managed to accumulate across generations. Our capacity for culture—for acquiring, pooling, and improving our knowledge over time—is a central part of what makes us a unique species in the animal kingdom. As evolutionary biologist Joseph Henrich (2015) puts it, culture is the "secret of our success." Culture allows us to adapt effectively and flexibly to changing environments because it gives each of us access to knowledge and skills that we could never dream of acquiring on our own. Thanks to culture, we know what plants are edible and which ones are poisonous, we know how to build buoyant ships and sturdy houses, and we know where to find potable water and arable land.

Social norms, too, are part of the body of knowledge and skills that we accumulate across generations. Norms, I suggest, are intimately tied to what I call *cultural practices*, namely, socially learned ways of acting that mediate our access to various social goods, such as approval, reputation, esteem, and status. Over time, different social groups develop different practices, often in response to certain adaptive problems, which are then passed on from one generation to the next. And, as they become part of a community's lore, practices also become a mark of its identity, a condition for membership to the group, and a collective expression of the sort of things group members value and care about. Social norms define, regulate, and sustain such practices—of fairness, reciprocity, respect, honor, etiquette, and so on. This idea, I argue, holds the key to solving the various puzzles surrounding the equilibrium theory of norms:

(1) In response to the Guidance Problem, norms or rules rationalize one's actions because accepting a norm involves wanting to participate in a certain practice and believing that the norm is constitutive of or helps define the practice. That is, one believes that participating in the practice *just is, inter alia*, following the relevant rule. As I argue in Chapter 1, when one has this belief/desire pair, one has a reason to *rule out* from one's deliberation any choice option whereby one fails to follow the rule.

(2) In response to the Efficacy Problem, one can be motivated to participate in a certain cultural practice by one's belief that one will be subject to other people's disapproval or disesteem if one doesn't. As I argue in Chapters 2 and 3, this motivation need not be incoherent with a desire to impose sanctions on others. The cost of disapproving of others and expressing this disapproval is much lower than the cost of being disapproved by them. Hence, one may be motivated to disapprove of and impose sanctions on others without thereby being motivated to adhere to a norm in the absence of disapproval or sanctions.

(3) In response to the Stability Problem, expectations about what other people think one ought to do can survive in a group if people are in the habit of transmitting social norms to others through teaching and instruction. In our groups, we transmit cultural practices by teaching them to children and newcomers. And this practice of instruction coordinates our expectations regarding how we expect each other to behave. As I argue in Chapter 2, social expectations survive because they motivate behaviors that are then taught to new group members.

(4) Finally, in response to the Normativity Problem, people can come to regard certain equilibria as proper norms in response to the need to teach these practices to children and newcomers to the group. As I argue in Chapter 4, although equilibria are self-sustaining social arrangements, they may not be robust enough to resist the destabilizing effect of new people coming into the group. In response to this problem, group members may come to see these practices not just as common behaviors, but as norms to which anyone in the group, particularly newcomers, must conform.

In the following four essays, I present, motivate, and articulate these different solutions to the puzzles arising from the equilibrium theory of norms. To end this short introduction, however, I want to briefly summarize the view of norms that seems to emerge from my four solutions, taken together. Perhaps the easiest way of seeing how all the elements hang together is *genealogical*. I will describe how, on my account, what we call social norms are slowly assembled by various social processes over time.

We may begin the genealogy with a community or group of people in social interaction. In the course of their lives, these people interact with one another, and inevitably some problems arise here and there in their interactions. Often, people in the group encounter the same type of problem time and again, perhaps because the problem arises from conflicts in what people in the group naturally tend to desire and intend. Thus, people may regularly find themselves faced with the problem of how to allocate scarce resources, how to divide labor in joint efforts, how to cooperate to achieve collective goals, and so on.

Over time, however, a practice emerges among group members as a solution to the recurrent problem. This practice may appear because at some point group members decide to get together and come up with a collective solution. But this need not be the case. Most likely, the practice emerges in the absence of much planning and communication. Some people initially adopt it and start doing surprisingly well whenever they encounter problems. And, given their success, other people start copying what they do, and adopt the practice, too. Soon, the practice spreads and conquers the group, in the sense that most people follow it because doing so seems like a good idea.

But then the group faces a new problem—namely, how to make sure that newcomers to the group also follow the practice. Every once in a while, new people arrive in the group; many of them come from groups where the practice doesn't exist yet. Likewise, over time, old group members are replaced by newer generations who may be unaware of the practice. In response to this problem, a new practice emerges in the group—namely, a practice of teaching the original practice to newcomers. And, in the context of this practice of instruction, people in the group come to see the original practice as more than just something they do. They see it as a rule that people must follow, at least if they wish to belong to the community.

Once group members are in the habit of teaching certain practices to newcomers, these practices become shared rules in the group. Because everyone has been taught the rules, everyone knows what they're supposed to do. They know that they're supposed to share equally, respect other people's property, contribute to joint efforts, and so on. And because everyone knows that everyone else has also been taught these rules, they expect everyone else to know what they're supposed to do, too. In short, everyone knows the rules and knows that everyone knows them, and so on.

Finally, because the rules are shared in this way, breaking them comes to be associated with a distinctive sort of social penalty or stigma. People in the group disapprove of and withhold esteem from rule breakers.

As a result, transgressors lose their standing in the group and the capacity to act as autonomous agents within it. They serve as examples to everyone else, who become motivated to adhere to the shared norms out of fear of experiencing the same consequences. Group members enforce the rules because they care about their peers contributing to the collective goods these rules serve to secure. Yet, they follow the rules at least in large part because they're afraid of being criticized or ostracized.

Here we have, then, a full-fledged social norm in operation. It is a shared rule of behavior sustained by social sanctions. And it is an equilibrium state in the group: People are rationally motivated to follow the rule provided they expect everyone else to follow it, and also to expect *them* to follow it, too.

Of course, this emerging view of norms is much more coherent than what's contained in the essays you're about to read. To some extent, this is inevitable. These essays reflect my thinking about social norms over the course of three years. Hence, they're bound to contain inconsistencies and tensions, which reflect the evolution of my thinking about these matters. Still, I think there is a unified picture of norms coming out of the essays. Moreover, I think this picture vindicates the traditional Humean equilibrium theory, while at the same time adding something new to it. In any case, I'll let you be the judge.

Chapter 1

Rules and Rationalization

Abstract

One follows a rule only if that rule explains one's action by *rationalizing* it, in the sense of articulating one's reason to so act. There's a puzzle, however, about whether rules can ever rationalize action. According to this puzzle, whenever a rule applies to one's case, following it either makes *no* rational sense or, if it does, this isn't *because* of the rule. I solve this puzzle by identifying the way whereby rules rationalize action. Rules make it rational for one to *rule out* acting against them. When one accepts a rule, one has a reason to exclude from one's deliberation the option of breaking it.

Introduction

Many of our actions are guided by rules. Or so, at least, it would seem both from the point of view of common sense and philosophical theorizing. We follow rules of various kinds: The rules of games like chess, mathematical operations like addition, languages like English, and social practices like etiquette and fashion. And the idea that we often follow rules like these is central to our self-conception as agents and to philosophical accounts of our various epistemic and practical capacities.

But the idea of rule-following gives rise to thorny philosophical puzzles. Perhaps the most famous of these was first formulated by Wittgenstein (1953) and later popularized by Saul Kripke (1982). It is a kind of skepticism about whether there's ever a determinate fact of the matter concerning what rule an agent is following at any given moment, given that her behavior can always be said to conform to infinitely many possible rules.

In this paper, I focus on a different philosophical puzzle about rule-following. This challenge arises in connection with another central idea about human action, namely, the view that intentional human action is always *rational* action, in the sense that it is done for reasons, whether these are, in an objective sense, good or bad reasons. According to this idea, whenever human beings act, and do so intentionally, their behavior is motivated by considerations that also *rationalize* what they do. These considerations justify or make rational sense of the person's behavior, at least in her eyes, that is, subjectively, if not objectively.

The puzzle I'll be discussing comes in the form of a dilemma. It purports to show that rules are invariably *misguiding* or *redundant* concerning the rational justification of action. That is, it is always the case that following a rule either makes *no* rational sense for the agent or, if it does, this is not *because* of the rule. Hence, considerations relating to rules can never rationalize one's actions. As we will see, the upshot of this dilemma seems to be that, insofar as rule-guided action is possible, it isn't a species of rational action.

In what follows, I offer a solution to this puzzle. I claim that, at least in many central cases, rules rationalize our actions in a distinctive way: They make it rational for us to *rule out* acting against them. That is, when we accept a rule as a standard of behavior, it makes rational sense for us to consider in our deliberation only those among our alternatives whereby we conform to the rule and exclude all others from consideration. On my view, then, following rules is, at bottom, a matter of ruling out certain options.

I proceed as follows. First, in section 1, I clarify what I mean by rational action and rule-guided action. Next, in sections 2 and 3, I present the rule-following dilemma and identify the subset of rules to which it applies most forcefully. These are what John Rawls (1955) calls *non-summary* or *constitutive* rules. In sections 4 to 6, I reject different ways of responding to the dilemma. Finally, in sections 7 to 9, I give my solution, draw some of its consequences for how we reason when we follow rules, and contrast it with other superficially similar proposals in the literature.

1.1 Acting for Reasons and Following Rules

Philosophers commonly hold two views about human action—that intentional human action is always rational action, namely, it is done for reasons, and that intentional human action is often rule-guided action, namely, it is done following rules. My question in this paper is how these two views of human action relate to each other. Before addressing this question, however, I need to describe each of them in more detail.

The first view is meant to be a thesis about human action in general. It says that *all* intentional human action is rational action. According to this idea, whenever someone acts, and does so intentionally, her behavior is motivated by considerations that also *rationalize* what she does, at least in her eyes (Anscombe, 1957; Davidson, 1963). These considerations justify or make rational sense of the person's action in light

of what she believes and particularly in light of what she finds valuable, worthy, enjoyable, or desirable. In brief, they reflect the agent's *reasons* for doing what she does.¹ If I spill my coffee because you jiggle my hand, then my spilling the coffee isn't something I do, it's merely something that happens to me. In contrast, if I spill my coffee wanting to see how it stains your rug, then my behavior stems from my reasons. It's not something that befalls me, but one of my actions (Davidson, 1971).²

What sort of considerations reflect the agent's reasons for action? According to one influential view commonly associated with Hume, someone's reasons are constituted by pairs of psychological attitudes. More precisely, for someone to have a reason to do ϕ just is for them to have a *desire* to do ψ and a *belief* that they can do ψ by doing ϕ (Davidson, 1963; Smith, 1987). Here, desire is taken to be a very broad type of psychological attitude. It includes states like wants, urges, and yearnings, but also states like finding something valuable, beneficial, or agreeable (Davidson, 1963). On this Humean view, then, someone acts for a reason insofar as they have some kind of positive attitude—a “pro-attitude”—towards actions of a certain type, a belief about how to act in the desired way, and these two attitudes cause their action. In spilling my coffee, for example, I act for a reason just in case my action is caused by a desire to do something, e.g., staining your rug, and a belief that I can do that thing by spilling my coffee, where these attitudes cause my action in the right sort of way.³

The second view about human action I want to discuss is meant to be a thesis about a specific type or mode of human action. It says that at least *sometimes*, intentional human action is rule-guided action. According to this idea, often, when someone acts, her behavior is motivated by a general *rule* or *principle* of action. This rule or principle demands or requires that the person acts in a specific way in the type of circumstances she finds herself in. And, in these circumstances, the agent counts as “following the rule” insofar as her actions are somehow guided by what the rule demands or requires.

What, exactly, is it to follow a rule in this sense? We can start with a negative observation: Following a rule isn't just *conforming* to it. Conforming to a rule is a matter of acting in the way the rule requires. Yet, this is plainly insufficient for rule-following. If I answer “19” to your question, “What's $10 + 9$?” then I conform to the rule of addition. But I also conform to the rule of “quaddition,” and, indeed, to an infinity of

¹In speaking of the agent's reasons in this paper, I'll always mean her *motivating* reasons, as opposed to her *normative* reasons (Smith, 1994). An agent's motivating reasons are the considerations that motivate her to act, independently of whether they're also normative reasons for doing what she does, that is, considerations that, in a more objective sense, justify her actions.

²Of course, not all actions are intentional. In turning on the light in my room, I may alert a prowler loitering outside my house. My alerting the prowler is something I do, but I don't do it intentionally. According to an influential view, however, all actions have a close connection to intentional action. Every action is intentional *under a description* (Davidson, 1963). My alerting the prowler counts as one of my actions because it can legitimately be redescribed as my turning on the light, which is something I do intentionally.

³This last qualification is needed to exclude cases where the relevant attitudes cause my behavior by somehow bypassing my rational agency, for example, by producing certain physiological effects in me. See Davidson (1980a).

other rules (Kripke, 1982).⁴ Among all these rules, however, addition is the only one I follow. Furthermore, conforming to a rule also seems unnecessary for rule-following. Perhaps, in answering your question, I slip and say “90.” Still, I count as following the rule of addition, in the sense that this is the rule I’m *trying* to conform to (Boghossian, 2015). My failure to conform to the rule is simply due to a performance mistake.

More positively, rule-following is intuitively associated with several features. According to Paul Boghossian’s (2015; 2008) “ACER model,” for instance, rule-guided action has the following four features:

- (A) Acceptance: The agent must somehow or other accept the rule.
- (C) Correctness: If an agent is following rule R, then there is a sense of correctness, according to which what he does is correct iff he conforms to R.
- (E) Explanation: Acceptance of R explains what the agent does.
- (R) Rationalization: Acceptance of R rationalizes or makes sense of what the agent does.

These conditions are meant to capture the intuitive idea that, when someone follows a rule, she conforms to it in a non-accidental way. Instead, she conforms to the rule—or at least tries to conform to it—*because* she bears some appropriate relation to it (Boghossian, 2008). The agent somehow accepts or endorses the rule, and her acceptance of it explains her behavior by rationalizing it in light of the standard of correctness set up by the rule. In answering “19” to your question, for example, I follow the rule of addition because my acceptance of this rule as a standard of correct behavior helps explain and rationalize my action.

1.2 The Dilemma

What’s the relationship between acting for reasons and following rules? Initially, it seems like these two modes of action are compatible with each other. As we saw, acting for reasons is about being motivated to act by considerations that rationalize one’s behavior, at least in one’s eyes. And, as we saw, when someone follows a rule, her acceptance of this rule explains her behavior by rationalizing it. Yet, the idea that rational action and rule-guided action are compatible faces a significant challenge. According to this challenge, rule-guided action is actually *incompatible* with rational action because one’s acceptance of a rule can never rationalize one’s behavior.⁵

⁴The rule of “quaddition” says that one should, when “quadding” two numbers, add them if they’re both less than 57, and otherwise answer “5.” The fact that conforming to a rule is insufficient for rule-following motivates an influential skeptical argument about rule-following. According to this argument, there may be no determinate fact of the matter about what rule an agent is following at any given moment (Kripke, 1982; Wittgenstein, 1953). I put this skeptical possibility aside here.

⁵For different versions of this challenge, see Shapiro (2005), McClennen (2004), and Shapiro (2001), who mostly focus on how the challenge arises concerning the rationality of following legal rules. The challenge is also similar to a traditional objection to rule consequentialism according to which it either collapses into act consequentialism or is incoherent (see Hooker, 2023, sect. 8). Finally, the challenge resembles a problem concerning the rationality of acting on our future-directed intentions, especially if these are understood as plan-like states *à la* Bratman (1987). See Bratman (2014).

This challenge comes in the form of a dilemma. To formulate this dilemma, let me first introduce some basic terminology. Let's say that, whenever someone acts, her action can be modeled as a solution to a *decision problem*.⁶ This problem consists in choosing one among a set of *alternatives*, where these alternatives reflect the different courses of action that she might take in her situation, at least as things appear to her. When I leave my apartment in the morning, for example, I face the decision problem of whether or not to take an umbrella with me. My problem consists in choosing one of these two alternatives. And the way I end up acting constitutes my solution to this problem.⁷

Consider now a natural thought: For any decision problem someone may face, there's a subset of her alternatives that are supported by the *balance* of her belief/desire reasons for action.⁸ These are the things she has *most* reason to do in their situation in light of everything she desires and believes. To borrow a term from decision theory, these alternatives maximize the agent's expected utility. They give her the best prospect of satisfying her desires relative to her beliefs. If Ra is the subset of alternatives supported by the balance of the agent's reasons, then an alternative a belongs to Ra just in case for any other alternative b , a is at least as well supported by the agent's belief/desires reasons as b —or, to put it in decision-theoretic lingo, the expected utility of a is greater than or equal to the expected utility of b .

Now, rules also offer criteria for evaluating alternatives in a decision problem. As we saw, every rule sets up a standard of correctness, at least for the agents who accept it. Rules are *general* principles specifying what one should do in circumstances of a given type (Brennan et al., 2013).⁹ Whenever someone finds themselves in the type of situation covered by a rule she accepts, there's a subset of her alternatives that reflect what the rule requires her to do in that situation. These are the things it would be correct, appropriate, or acceptable for the agent to do in her case in light of the rule she accepts. Call this subset of alternatives Ru .

I'm now able to formulate the dilemma. Suppose someone's decision problem falls under a rule she accepts. That is, it is a situation of the type covered by the rule. Then, the following two possibilities are mutually exclusive and jointly exhaustive:

1. $Ra \cap Ru = \emptyset$

2. $Ra \cap Ru \neq \emptyset$

⁶For a brief exposition of the idea of a decision problem, as used in decision theory, see Bermúdez (2009, chap. 1).

⁷An action can be a solution to a decision problem even if it is not preceded by explicit and conscious deliberation or reflection. See Bermúdez (2009, chap. 1) for discussion.

⁸It could be that, given certain forms of uncertainty in the agent's beliefs or incommensurability in her desires and values, not every decision problem she faces has a clear subset of rational solutions. I ignore this possibility here.

⁹These general principles might take different forms. They could be normative propositions ("You shouldn't put your elbows on the table") or imperatives ("Whenever you receive an urgent email, answer it right away!"). See Boghossian (2015, sect. 1.1.2) for discussion.

In the first scenario, doing what the rule requires isn't supported by the balance of the agent's belief/desire reasons. It fails to maximize her expected utility. So, the rule would seem to be *misguiding*. It requires that the agent does something it makes no rational sense for her to do in light of everything she desires and believes.¹⁰ In the second scenario, doing what the rule requires is supported by the balance of the agent's reasons, but not because it is what the rule requires. Instead, the action is supported by the agent's reasons because it gives her the best prospect of satisfying her desires relative to her beliefs. So, the rule would seem to be *redundant*. It requires that the agent does something that it makes rational sense for her to do quite independently of the rule.

This dilemma would seem to establish the following conclusion: One's acceptance of a rule can never rationalize one's actions. In other words, it is never the case that, if a rule she accepts requires someone to do something, it is *thereby* rational for her to do that thing. Conforming to the rule either makes *no* rational sense for the agent; or, if it does, this is not *because* of the rule. Rules, it would seem, are invariably misguiding or redundant concerning the rational justification of action. Hence, they can't make sense of someone's behavior *qua* rational behavior.

This conclusion leaves open only two theoretical possibilities, both of which seem equally unpalatable. One possibility is that rule-following is a chimera. As we saw when discussing Boghossian's ACER model, part of our intuitive notion of rule-following is that, when someone follows a rule, her acceptance of this rule explains her behavior by rationalizing it. Yet, our dilemma would seem to show that rules can never rationalize one's actions. It follows that rule-following is impossible.

The other possibility is that rule-guided action is possible—maybe the ACER model is wrong—but it just isn't a species of rational action. When we follow rules, we're motivated to act by certain considerations. But these considerations fail to rationalize what we do. This would mean, however, that rule-following isn't a form of intentional action. It's less something we do than something that happens to us.

1.3 All Rules?

As I have formulated it, this dilemma about rule-following applies to rules and rule-guided action in general. Upon reflection, however, there seem to be types of rules for which the problem may not arise.

Consider, first, what we might call—following a recent trend in meta-ethics—*authoritative* or *robust* normative principles. What I have in mind is the sort of rules or principles that seem to have an *authority*

¹⁰This is not to suggest that it can never make rational sense, at least to some extent, for one to act against the balance of one's belief/desire reasons. When an agent gives in to akrasia, for example, she acts against the balance of her reasons, but she does so intentionally and *for a reason* (Davidson, 1980b). The point here, rather, is that it doesn't seem to make rational sense for one to act against the balance of one's belief/desire reasons in light of a consideration that isn't *itself* a belief/desire reason, such as a rule.

such that it would be a particularly serious type of mistake not to give them any weight in one's deliberation (Plunkett, 2019).¹¹ Moral principles seem to be a prime example of such authoritative normative rules (Finlay, 2019). Very plausibly, morality requires that one be loyal to one's friends, keep one's promises, and not torture babies. If so, then these are rules or principles one must give at least some weight in one's deliberation on pain of being somehow normatively at fault.

Authoritative normative principles may escape the rule-following dilemma on its second horn. They seem to be clearly relevant to what it makes rational sense, on balance, for one to do. Indeed, several philosophers argue that a form of *motivational internalism* is true about such principles. That is, whenever someone accepts such a principle, she also has a desire to conform to it, at least if she's rational (Smith, 1994).¹² If I come to accept that I should donate to famine relief, perhaps because you have convinced me that this is what morality requires of me, then I'll also come to have a desire to do so. At least, I'll come to have such a desire if I'm rational or coherent. It follows that moral principles are not redundant concerning rational justification. Instead, they issue in a belief/desire reason to act as morality dictates (Smith, 1998).

Consider, next, what John Rawls (1955) calls *summary rules*. These are rules or principles that summarize the outcome of past deliberation as applied to cases of a given type. They recommend that the agent acts in ways that have often been found to be best supported by reasons in the past, either by the agent herself or by others. "Always save 5% of your earnings" is a sound prudential rule of thumb. It reflects the fact that, very often, when you earn some money, saving at least 5% of it makes sense in light of the things you hope to achieve.¹³

Summary rules may escape the dilemma on its first horn. At least sometimes, following these rules seems rational, even if it involves acting against the balance of one's reasons. Of course, a rule of thumb can't carry any weight if the agent knows it to be misleading in her specific situation. If I know that I can't or shouldn't save up any money this month—I urgently need to buy a new laptop—then following the 5% savings rule makes no rational sense for me. But an agent doesn't always know what action is best supported by her belief/desire reasons. Moreover, figuring this out often takes time, effort, and resources she'd be unwise to spare. Think of how difficult it would be to decide *exactly* how much money you should save up each time you get a paycheck. Under these circumstances, it makes sense for her to follow a rule of thumb. The rule has a sort of *heuristic value*, in the sense of allowing the agent to find a good enough solution to her decision problem quickly and easily. Given this heuristic value, following the rule is a form of *bounded rationality*

¹¹For this idea of authoritative or robust normativity, as opposed to merely generic or formal normativity, see Plunkett (2019) and McPherson (2011).

¹²For a recent discussion of the idea of motivational internalism and its relevance to meta-ethics and moral psychology, see Asarnow and Taylor (2022).

¹³Boghossian's email rule ("Whenever you receive an urgent email, answer it right away!") seems to me like another example of a summary rule.

(Simon, 1957)—a good strategy for coping with one’s limitations—even if it occasionally leads one to act against the balance of one’s belief/desire reasons.¹⁴

There is, however, a subset of rules for which the rule-following dilemma arises in all forcefulness. These are what Rawls calls *non-summary* or *constitutive* rules.¹⁵ The central mark of constitutive rules, according to Rawls, is that they help constitute certain types of *practices*. Think of the rules of a game like chess. “The Knight only moves in an L-shape” states a rule that helps constitute the practice of chess playing. Playing the game of chess *just is, inter alia*, following this rule. Someone who moved the Knight differently wouldn’t be playing chess but a different, if closely related, game—call it chess*. The same applies, *mutatis mutandis*, to many other rules: Table manners help constitute the practice of dining etiquette; fashion rules, the practice of fashion; military honor codes, the practice of military honor; and the rule of addition, the practice of adding as opposed to “quadding.”

Constitutive rules or principles don’t have the authority we associate with moral requirements and other authoritative normative principles. Etiquette may well require that you wear a tie to a dinner party at the club, but you need not be normatively at fault if you desire to do no such thing (Foot, 1972). Further, constitutive rules don’t summarize the outcome of past deliberation. They’re not useful heuristics to use in cases where deliberating is hard and time-consuming. In short, constitutive rules don’t seem to escape the rule-following dilemma in the way other types of rules may be able to escape it. It is concerning *these* rules that the question of how they can ever rationalize our actions seems to arise most forcefully in light of the dilemma.

In what follows, I’ll focus on the case of constitutive rules. This is not because other types of rules don’t raise a problem. Authoritative rules and summary rules may not escape the dilemma, after all. Whether they do, however, has been much discussed already. The question of how moral requirements guide action is hotly debated in meta-ethics and moral psychology. And bounded rationality is a central topic in economics and the philosophy of action. Constitutive rules are especially interesting because some views that look plausible concerning other types of rules are just non-starters when it comes to rules of this kind.

1.4 Biting the Bullet

How, then, should we respond to the dilemma? If we accept it, then we must conclude that—at least when it comes to constitutive rules—rule-guided action is possible only if it isn’t a species of rational action. At first glance, this conclusion seems unpalatable. But perhaps it isn’t as unpalatable as it initially seems.

¹⁴Compare with the old act-consequentialist idea that following a general moral rule is justified whenever it’s too difficult for the agent to figure out what act maximizes the good in her situation (Smart, 1956; Sidgwick, 1874).

¹⁵See Kaplan (n.d.) for a recent discussion of the distinction between summary and non-summary rules.

One may be tempted to think this in light of what cognitive scientists have to say about how we usually follow rules. One common idea is that rules or norms often guide our behavior by bypassing our capacities for practical reasoning and deliberation (Kelly, 2020). The idea here is that human beings have an evolved *norm system* or *norm psychology*, namely, a host of psychological mechanisms allowing us to follow rules (Henrich, 2015). Although the precise contours of this system are still debated, we seem to have a clear idea of the tasks it is meant to accomplish (Sripada and Stich, 2006). The norm system is meant to allow people to detect, comply with, and enforce the rules and norms present in their social environment.¹⁶ But the system can achieve these tasks independently of our practical reasoning and reflection. The mechanisms comprising it have many of the features of so-called “System 1” cognition, like domain specificity and automaticity (Kelly, 2020).¹⁷

This idea might lead one to bite the bullet concerning the rule-following dilemma in the following way: Rule-following is possible, but it just isn’t a species of rational, i.e., reflective action. It falls under the domain of “System 1” as opposed to “System 2” cognition (Kahneman, 2011).¹⁸

However, the fact that much rule-following is unreflective and automatic is, by itself, neither here nor there concerning whether it is a form of rational behavior, in the sense of “rational” relevant here. Many rational actions are done by the agent in the absence of deliberation or reflection. A mother finds her house engulfed in flames with her children inside. Right away, she runs towards it intending to rescue them. Here, the mother’s actions are adequately explained and rationalized by her beliefs and desires, but not because she ever explicitly deliberates or reasons about what to do in light of such attitudes (Cavell, 1993). Similarly, in driving on the left side of the road, I might do so habitually and automatically. Yet, my action is explained and rationalized by my beliefs and desires, in the sense that, were my reasons to drive on the left suddenly to change—as happened to people in Sweden on Dagen H—my behavior would quickly follow suit (Lewis, 1975).¹⁹

These examples illustrate a general idea: The view that rational action is done for reasons isn’t meant to be a thesis about how people reason. In particular, it isn’t the view that all rational actions are the outcome of explicit, conscious reasoning. As far as the view goes, many rational actions may happen automatically and unreflectively. The crucial idea, instead, is that whenever someone acts intentionally, her behavior can

¹⁶For a philosophically savvy overview of the empirical research supporting the idea of a norm system, see Kelly (2020).

¹⁷According to one popular taxonomy, cognitive mechanisms can be divided in two broad categories. “System 1” mechanisms are supposed to be automatic, intuitive, and effortless, whereas “System 2” mechanisms are supposed to be deliberate, logical, and effortful. See Kahneman (2011).

¹⁸This, I think, reflects a common way of thinking about rules and rule-following in economics and other social sciences. For example, the shift toward “behavioral economics” that began several decades ago with the work of Kahneman and Tversky is often portrayed as relaxing the assumption of ideal rationality embodied in the model of “economic man” from neoclassical economics by, among other things, incorporating the idea that human beings make many decisions using “quick and dirty” rules and heuristics instead of doing utility calculus. See Camerer and Lowenstein (2004) for discussion.

¹⁹In this case, we might say that my beliefs and desires are *virtual causes* of my behavior, in the sense of being sufficient to produce or sustain it under various contingencies. See Pettit (1996) for this idea of a stand-by or virtual cause.

be adequately explained in light of her desires and beliefs. As Elizabeth Anscombe puts it, these attitudes reflect “an order which is there whenever actions are done with intentions” (1957, para. 42), even if the agent isn’t always explicitly or consciously aware of that order in a way that allows her to reason from her attitudes to the performance of an action.

In sum, the idea that rule-guided action isn’t a species of rational action shouldn’t be equated with the view that it happens unreflectively and automatically. Instead, it is the idea that, when someone follows a rule, we can never tell a rationalizing story about how she acts. Whatever rationalization we offer for her behavior is inevitably a *post hoc* rationalization, in the sense of failing to reflect the motives behind her action—because there are no such rational motives. This is quite a big pill to swallow. Or so, at least, it seems to me. It entails that, in following the rules constitutive of practices, we’re something like the “mindless playthings of social forces” (Elster, 1989a, p. 97). On this picture, rule-following isn’t something we do. It’s something that happens to us.

Of course, there is a form of rule-following that is *sub-personal* in this sense. When I catch a flyball in baseball, I manage to do so because I follow a rule—namely, I run toward the ball keeping my angle of sight constant (Alexander, 2007). However, this isn’t the type of rule-following we have in mind when we say that someone is following the rules of etiquette or the rules of a game like chess. With the flyball rule, it seems appropriate to say that it isn’t me but my brain that follows the rule (Boghossian, 2015). Indeed, I need not even be aware of the rule’s existence and content for it to guide my behavior. But this is plainly not the case when we follow the constitutive rules of a practice, like those of etiquette, fashion, or games. When I put on a black suit to attend a funeral, it is *me* who’s following the social rule, not merely my brain, even if I follow the rule automatically.

1.5 Constrained Maximization

Clearly, then, we shouldn’t bite the bullet concerning the rule-following dilemma. What should we say about it, then? One thing to note is that the dilemma arises from the Humean view that only beliefs and desires can rationalize action. It’s because of this assumption that rules appear to be invariably redundant or misleading. For, in every case where a rule applies, the action it requires either already has or lacks the right sort of connection with the agent’s beliefs and desires. Perhaps, then, to escape the dilemma we need to relax our ideas about what sort of considerations can rationalize action and say that sometimes a rule we accept can itself rationalize what we do. We may call this idea a form of *revisionism* insofar as it involves revising the widely influential Humean picture of rational motivation.²⁰ Taking a revisionist line may allow

²⁰The label for this view is due to McClennen (2004).

us to escape the dilemma on its first horn because it entails that it can make rational sense for one to follow a rule against the balance of one's belief/desire reasons.

This idea is at the center of David Gauthier's (1986) *constrained maximization* model of rational justification. According to this model, a rule can rationalize what we do independently of one's belief/desire reasons if a disposition to follow the rule is itself supported by the balance of one's belief/desire reasons. To see how this model is supposed to work, consider the well-known Prisoner's Dilemma (PD) game from game theory. In this game, each player can either cooperate or defect, where defecting is the dominant strategy for each, meaning it's what she prefers to do no matter what the other player does. The situation is a "dilemma," however, because mutual defection is collectively suboptimal. Indeed, this outcome is Pareto inefficient, meaning there's another outcome—mutual cooperation—where both players do better in terms of satisfying their preferences.

According to the standard analysis, in the PD game defecting is the best strategy in light of the agent's belief/desire reasons. This alternative maximizes her expected utility because she prefers to defect no matter what her partner does. According to Gauthier, however, at least under certain circumstances, the agent's belief/desire reasons also support becoming disposed to follow a rule of conditional cooperation in the PD game. An agent who follows this rule cooperates if she knows her partner will cooperate, and defects otherwise. As Gauthier argues, if players who follow this rule can confidently identify each other's dispositions, then they have an opportunity to cooperate that isn't available to those who aren't disposed to follow the rule. Under these conditions, a choice of becoming disposed to follow the rule maximizes their utility relative to the alternative of not being so disposed.²¹ The crucial idea of the constrained maximization model is that, when these conditions obtain, following the rule is rational, even if it leads the agent to act against the balance of her belief/desire reasons in the PD game.²²

There are at least two problems with this proposal, however. The first one has to do with the role of one's belief/desire reasons in one's deliberation. Gauthier's idea, recall, is that a rule can rationalize acting against the balance of one's belief/desire reasons just in case a disposition to follow this rule is itself supported by one's belief/desire reasons. Upon reflection, however, this seems problematic. How can a rule be one's rationale for failing to promote one's desires relative to one's beliefs when the rule itself is only supposed to be justified in terms of how it promotes one's desires relative to one's beliefs? An agent who follows a rule in this way would appear to be guilty of a kind of *incoherence*. She seems to be of two minds concerning her

²¹For example, imagine the players are *transparent* to one another. They can flawlessly identify each other's choice disposition. Then, the expected utility of being disposed to follow the cooperation rule is equal to $p(R) + (1 - p)(P)$, while the expected utility of not being so disposed is P (where p is the probability that the other player is disposed to follow the rule, R is the "reward" outcome associated with mutual cooperation, and P is the "penalty" outcome associated with mutual defection). Given that in the PD game $R > P$, $p(R) + (1 - p)(P) > P$.

²²It's called the "constrained maximization" model because an agent who's disposed to follow a rule accepts a constraint on her deliberation, as opposed to an agent who straightforwardly maximizes her expected utility, who accepts no such constraint.

belief/desire reasons, in that she both treats and fails to treat them as genuine reasons in her deliberation. She treats them as reasons in her choice of what action disposition to adopt, but not in her choice of what action to perform. But an agent who reasons like this seems to be criticizable as irrational.

This criticism is analogous to a traditional objection against rule consequentialism in ethics.²³ On one standard formulation, rule consequentialism is the view that an act is morally justified just in case it is required by a rule that is itself morally justified for a distinct reason, viz., because it maximizes the good (Rajczi, 2016). Hence, rule consequentialists claim that failing to maximize the good in one's situation can be morally justified if it is required by a rule that maximizes the good. A traditional objection against this claim is that following a rule under these conditions makes no rational sense. How can a rule make it rational for you to fail to maximize the good if your only reason to follow the rule is that it maximizes the good? At worst, this type of reasoning seems to be a form of *rule worship*, in that you act out of an irrational attachment to the rule, not out of an appreciation of how the rule serves to promote the things you value (Smart, 1956).

The second problem with Gauthier's model is that it has limited scope. It fails to explain why following some constitutive rules can be rational. As some theorists point out, many constitutive rules are meant to solve collective action problems with the structure of a PD game (Ullmann-Margalit, 1977; Bicchieri, 2006). The military code of honor, for instance, may partly exist to solve the prisoner's dilemma intrinsic to collectively fighting against an enemy (Ullmann-Margalit, 1977). The code of honor says that desertion is dishonorable. Hence, soldiers who follow the code will cooperate by standing their ground in battle. Maybe the constrained maximization view explains why following this and other such rules can be rational.

But not *all* constitutive rules can be modeled as solutions to prisoner's dilemmas or other problems of collective action (Elster, 1989a). For example, there's no clear sense in which the rules of chess, addition, or English grammar offer solutions to any collective action predicament. These rules simply don't govern situations of strategic interaction.²⁴

1.6 Sanctions and Internalization

It's time to start putting my cards on the table. I believe we should try to escape the rule-following dilemma on its second horn. That is, we should say that constitutive rules aren't redundant concerning what it makes rational sense for us to do. Instead, whenever we accept such rules, we have a belief/desire reason to follow the rule. We may call this strategy a form of *compatibilism*, insofar as it allows us to retain the idea

²³This is the so-called *incoherence objection* against rule consequentialism. See Rajczi (2016) for a recent discussion and vindication of this objection.

²⁴Of course, chess is often called a game of strategy, but its rules don't apply to a previously existing strategic situation.

that rule-following is both possible and a species of rational action without revising the influential Humean picture of rational motivation.²⁵ Before presenting my own compatibilist proposal, however, I will reject two other compatibilist alternatives one might be tempted to adopt concerning the dilemma.

Take, first, the idea that constitutive rules may carry weight in one's deliberation insofar as they're often backed by *sanctions*, i.e., penalties for breaking a rule.²⁶ When I cut in line in front of you, I can fully expect you to give me the stink eye. And if I go to a formal dinner party dressed as a clown, I know I'll be made the laughingstock (Lewis, 1969). As these examples illustrate, the type of sanctions associated with breaking rules of politeness, fashion, or etiquette aren't the sort of explicit, formal sanctions associated with breaking legal norms or institutional codes, such as fines and prison sentences (Elster, 1989a). Instead, they're often implicit, informal penalties. Still, insofar as we normally see these sanctions as unpleasant consequences to be avoided, we seem to have a reason to do what these rules require. Perhaps, then, constitutive rules rationalize what we do because we have a desire to avoid the sanctions associated with breaking them.

The problem with this idea, however, is that doing what a rule requires merely for the sake of avoiding a sanction isn't the same as being *guided* by the rule. As Edward McClellenn (1997) puts it, conforming to a rule in this way isn't a form of rule-guided action; It's a form of *sanction-goaded* behavior. We can understand this idea in light of the ACER model of rule-following from Section 1. According to this model, recall, following a rule involves conforming to it because one bears some appropriate relation to it. One must somehow accept or endorse the rule as a standard of correct behavior. And one's acceptance of the rule must explain one's behavior by rationalizing it in light of the standard set up by the rule. Yet, someone who conforms to a rule merely for the sake of avoiding a sanction need not accept or endorse the rule in any way. In playing chess with you, for instance, I may play by the rules as long as you're paying attention to what I do but then resort to cheating as soon as you become distracted. Here, I conform to the rules of chess because I'm afraid of your disapproval, but I'm not guided by them.

Consider, next, the second compatibilist idea I reject. This is the suggestion that constitutive rules often matter to an agent's deliberation because she has *internalized* them, meaning that she has acquired some kind of intrinsic, non-instrumental desire or motivation to follow them (Elster, 1989a; Sripada and Stich, 2006; Kelly, 2020).²⁷ Plausibly, people come to internalize a rule through some form of education or socialization. Once they have internalized it, however, they have a motive to follow the rule which operates independently of social influences. If I have internalized the rule that one must not slurp one's soup at the dinner table, then I'll be motivated to refrain from slurping even if there's no one around to monitor and possibly sanction

²⁵This label is also due to McClellenn (2004).

²⁶For the idea that social rules are often backed by sanctions, see Elster (1989a) and Bicchieri (2006).

²⁷For an overview of the idea of internalizing a norm, see Kelly (2020).

my behavior. One could say, then, that rules rationalize action because the agent intrinsically wants to do what the rule requires.

However, the problem with this formulation of the internalization idea is that it gives an inaccurate picture of how we usually follow constitutive rules. We seldom seem to follow these rules out of an intrinsic desire to do what they require. Instead, we often seem to have an *ulterior motive* in obeying them.²⁸ When I follow the rules of chess, for instance, I don't do this because I value conforming to them as an end in itself. I do it because I want to have fun, spend an enjoyable evening, show off my intelligence, etc. Likewise, when I refrain from slurping my soup or when I put on a black suit to attend a funeral, I don't do these things wanting to follow a social rule for its own sake. I do this because following these rules is a way of being respectful or polite, meaning that a desire to be respectful or polite is at least part of what motivates me to follow these rules.

To be clear, I don't think the problem with the sanctions and internalization views is that one's fear of sanctions and internalized motivations aren't often part of the reason why we follow constitutive rules. Surely, such motives often contribute to explaining why following these rules makes rational sense for us (more on this below). The problem, instead, is that these motives can't be what Donald Davidson (1963) calls our *primary reason* to act in conformity with a rule. An agent's primary reason articulates the consideration that immediately rationalizes her action, even if there are other considerations relevant to making sense of what she does. My desire to recover my stolen car helps make rational sense of why I call the police, but it isn't my primary reason for doing so. My primary reason is that I want to alert them to the theft and believe I can do this by calling them. One's fear of sanctions and internalized motivations aren't the primary reason why we follow constitutive rules. When we follow these rules, our immediate reason to do so isn't that we want to avoid a sanction or have an intrinsic motivation to do what they require, even if such sanctions and internalized motives are otherwise relevant to making sense of what we do.

1.7 The Practice Model

I turn now to articulating my own compatibilist proposal for escaping the rule-following dilemma. I call it the *practice model* of rule-following because it relies on the idea that there's a close, necessary connection between following constitutive rules and partaking in certain practices. In this section, I introduce the practice model. In the next, I explain how it solves the dilemma.

Let me start by going back to an observation about the nature of constitutive rules. The central mark of these rules, recall, is that they help constitute *practices*, by which I mean distinctive types of action or

²⁸This is the phenomenon that McClennen (2004) and Shapiro (2001) call "instrumental" rule-following.

activity. A rule of chess like “the Knight only moves in an L-shape,” for example, helps constitute the practice of playing chess, in the sense that partaking in this practice or activity *just is, inter alia*, following this rule. Someone who doesn’t follow this rule isn’t playing chess—but she may be playing chess*. Let’s make this more precise by saying that a rule R requiring that one does ϕ in circumstances C is a constitutive rule just in case there is a practice—a type of action or activity— ψ , such that one does ψ by, *inter alia*, doing ϕ in C.

The practice model of rule-following is, first and foremost, a model of what *accepting* a rule involves. As we saw with the ACER model, part of our intuitive idea of rule-following is the notion that an agent who follows a rule somehow accepts or endorses it. Further, an agent accepts a rule just in case she bears an appropriate relation to it, where this plausibly involves having some or other intentional state with the rule as its content (Boghossian, 2015). The practice model says, roughly, that an agent accepts a constitutive rule just in case she *wants* to partake in a practice and *believes* that the rule is constitutive of that practice, in the sense specified above. More precisely, she accepts a constitutive rule R requiring that one does ϕ in C just in case, if she believes she is in C, then she wants to do ϕ because she (1) desires to partake in practice ψ and (2) believes that one does ψ by, *inter alia*, doing ϕ in C.

Before discussing how the practice model solves the rule-following dilemma, let me say why I think it offers an independently plausible and attractive model of rule acceptance. First, the model makes sense of the difference between accepting a constitutive rule and *merely cognizing* it.²⁹ Everyone who cheats at a game like chess knows its rules. For example, the cheater knows that playing chess involves, among other things, moving the Knight in an L-shape. The difference between her and a fair player is motivational. The fair player wants to win *at chess* whereas the cheater doesn’t really want that. Otherwise, she wouldn’t break the rules. What the cheater wants, instead, is merely for it to *seem* as if she won. This is why she breaks the rules at the first opportunity. Accepting a constitutive rule, then, involves more than merely cognizing it, in the sense of having the right beliefs about it. It also involves wanting to partake in the practice the rule helps constitute.

Second, the practice model explains why someone who accepts a rule takes what H.L.A Hart (1961) calls “the internal point of view” toward it, in the sense of regarding it as a standard of correct behavior. If I want to play chess and believe that moving the Knight in an L-shape is part of what it takes to do so, then I must think that I get things *wrong* if I don’t move the Knight thus. I get things wrong *procedurally*, as it were, that is, in light of my own views about how to do what I’m trying to do. If someone points out to me that, in my last move, I mistook my Queen for one of my Knights, then I must conclude that I made a

²⁹For the idea of merely cognizing a rule, see Kelly (2020).

mistake given what I was attempting to do. I was trying to make a move in chess, but I failed: That's not how you move the Queen.

Can the practice model also explain why someone who accepts a rule often regards it as a standard not only for herself but for other people, too? It can: The desire to partake in a practice may be a desire that *we* partake in it. If I desire that we *both* play chess, then I'll regard the rules of chess as procedural standards applicable to your behavior as well as mine. And, if you break the rules, I'll be prone to criticizing you for not playing the game as you should.

Finally, the practice model explains why there seems to be an internal, necessary connection between accepting a rule and being *motivated* to follow it. As some theorists point out, accepting a constitutive rule has associated with it a characteristic motivational profile. It disposes the agent to guide and evaluate her conduct in light of the rule (Shapiro, 2006). The practice model explains this connection between rule acceptance and motivation. The model says that accepting a constitutive rule necessarily involves being motivated to participate in the practice the rule governs. Indeed, accepting a rule *just is, inter alia*, having a desire to partake in that practice.

1.8 The Role of Rules

The practice model solves the rule-following dilemma. It entails that constitutive rules play a distinctive role in the agent's deliberation. Hence, they aren't irrelevant or redundant to what it makes rational sense for her to do.

According to the practice model, whenever an agent accepts a constitutive rule like the rules of chess, fashion, or etiquette, she has a belief/desire reason to follow the rule. Her accepting the rule, recall, involves her wanting to partake in a practice and believing that following the rule is a (constitutive) means to do so. These attitudes give her a reason to follow the rule. Thus, if I accept the rules of chess, my acceptance of these rules gives me a belief/desire reason to, among other things, move the Knight in an L-shape.

Note, however, that this rule-based reason rationalizes the agent's actions in a distinctive way. On standard formulations of the Humean view of rational motivation, rationalization depends on a *promotion* relation holding between one's beliefs and desires, on the one hand, and one's actions, on the other. More precisely, the idea is that, if one desires to do ψ , then one has a reason to do ϕ just in case doing ϕ promotes one's doing ψ , in the sense of giving one a good prospect of satisfying this desire, given one's beliefs.³⁰

³⁰The precise conditions under which an action can be considered to offer one a good prospect of satisfying one's desires is a much-debated question. Evidentialists think that an action is a good prospect just in case performing it is good *evidence* that one will get what one wants (see, e.g., Ahmed, 2014). Causalists, on the other hand, believe that an action is a good prospect whenever it is likely to *cause* or *bring about* one's desired state of affairs (see, e.g., Joyce, 1999).

Spilling my coffee gives me a good prospect of staining your rug, at least given what I believe about the world. Hence, it is an action rationalized by my attitudes.

But this picture of rationalization doesn't fit the way rule-based reasons rationalize action. Promotion always comes in degrees. An alternative is always *better* or *worse* than others at promoting one's desires, given one's beliefs. Given what I believe, spilling my coffee gives me a better prospect of staining your rug than not spilling it. But it sounds confused to say that moving the Knight in an L-shape gives me a better prospect of playing chess than not moving the Knight thus. The relationship between these options isn't that one promotes my desire better than the other. Instead, it is that one *enables* me to act as I desire, whereas the other doesn't. In the case of constitutive rules, then, rationalization is not about promotion. It's about enabling. According to this idea, if one desires to do ψ , then one has a reason to do ϕ just in case doing ϕ enables one to do ψ , in the sense of making it possible for one act as one desires, given one's beliefs.

This distinction between promotion and enabling helps us identify the role of rule-based reasons in deliberation. These reasons structure one's alternatives in a distinctive way. Promotion rationalization structures the space of one's alternatives as a *ranking*. In this ranking, some options promote one's desires better than others, given one's beliefs. And responding rationally to this ranking involves selecting the option (or one of the options) at the top. Given my beliefs and desires, I have the ranking {spilling my coffee > not spilling my coffee}, and I'm rational if I choose the first alternative over the second.

Enabling rationalization, on the other hand, structures one's alternative set, not as a ranking, but as a *partition*. In this partition, some options enable one to act as one desires whereas others don't, given one's beliefs. And, plausibly, responding rationally to such a partition involves selecting an option in the enabling subset. Suppose my Knight sits at e5. Then, given my acceptance of the rules of chess, I have the partition {{moving the Knight to d3, moving it to c4, ...}, {moving the Knight to a1, moving it to a2, ...}}, and I'm rational if I choose an alternative in the first subset.

Now, given how rule-based reasons structure one's deliberative alternatives, how should one respond to these reasons vis-à-vis the rest of one's belief/desire reasons? Plausibly, one should give rule-based considerations *lexical priority* over other sorts of considerations. That is, in reasoning about what to do, one should prioritize responding to these reasons over other sorts of reasons. Otherwise, one risks being unable to coherently respond to one's reasons. If I form a ranking over *all* of my options, then it could be that the option at the top of my ranking falls on the wrong side of my rule-based partition. In that case, there's no coherent way for me to respond to my reasons. I'll choose either against my ranking or against my partition. In contrast, if I first exclude from consideration any alternative on the wrong side of the partition, then I can rank the remaining alternatives from best to worst. Thereby, I can coherently respond to both my rule-based and non-rule-based reasons.

So, rule-based reasons would seem to play the following role in reasoning: As opposed to other kinds of belief/desire reasons, which prompt us to choose the option that best promotes what we want, rule-based reasons prompt us to *rule out* those options that don't enable us to act as we want. That is, when considering what to do, these reasons lead us to consider only those options enabling us to act as we desire and to exclude from consideration all the rest. When I follow the rules of chess, for example, my acceptance of these rules makes it rational for me to only consider options that involve, *inter alia*, moving the Knight in an L-shape, while ruling out all other alternatives.

This picture suggests that there's an interesting interaction between one's rule-based reasons and the rest of one's belief/desire reasons in deliberation. Rule-based reasons determine the set of options we consider. In particular, they lead us to rule out any alternative that doesn't belong to *Ru*, that is, the subset of alternatives required by the rule. Once we have done this, however, we must still decide what alternative to choose among those we haven't ruled out. Here, the rest of our belief/desire reasons become relevant. They allow us to rank our remaining alternatives from best to worst.

The picture also explains why rule-following is often unreflective and automatic, as we saw in Section 4. Ruling out alternatives according to a partition will often be much less computationally demanding than ranking them from best to worst. Plausibly, then, this task may often fall under the domain of "System 1" cognition. At the same time, however, rule-following may also involve some reflection whenever the ruling out of options still leaves a few alternatives open. For example, in moving the Knight in chess, we unreflectively rule out forbidden squares, while reflectively choosing one among the permissible ones.

Having identified the role of constitutive rules in deliberation, we can now see a way out of the rule-following dilemma. The dilemma invites us to think as follows: Once you have ranked your alternatives in light of your belief/desire reasons, rule-based considerations wouldn't seem to carry any weight in your deliberation. Going by a rule in such circumstances—the idea goes—is like listening to the orders of the first mate when the captain is around. His orders either conflict or coincide with the captain's. So, you either *shouldn't* or *needn't* pay attention to them.

But, as we have seen, rule-based considerations aren't relevant to ranking our alternatives, or to choosing the best one according to the ranking. Instead, they're relevant to ruling out those alternatives whereby we fail to conform to the rule, thereby determining what alternatives we even consider for the purposes of ranking them. Rules, then, cannot, strictly speaking, either conflict or coincide with the balances of one's other reasons for action. They influence our deliberation *before* we do any balancing.

This, then, is my solution to the dilemma. It holds that one's acceptance of a constitutive rule rationalizes one's action because it gives one a belief/desire reason to follow the rule. The reason is that one wants to partake in a practice and believes that following the rule is part of what it takes to do so. And this reason

is grounds for excluding from one's deliberation those options whereby one fails to partake in the practice. It is a reason, for instance, to only consider options whereby one makes a move in chess, adds two numbers instead of quadding them, dresses fashionably, behaves politely at the dinner table, acts honorably, etc.

Let's put all this in the terminology of Section 2. My view entails that, whenever an agent is following a rule, Ra —the subset of her alternatives favored by her belief/desire reasons—is itself a subset of Ru —the subset of her alternatives favored by the rule. Hence, she always follows the rule to the extent that she acts rationally. In symbols:

1. $Ra \subseteq Ru$

2. $\therefore Ra \cap Ru = Ra$

Note, finally, that my solution avoids the pitfalls of the other compatibilist proposals we reviewed in Section 6. According to my view, when an agent follows a constitutive rule, she doesn't merely act out of a desire to avoid a sanction. Nor does she act out of an intrinsic desire to follow the rule as such. Her primary reason is that she wants to partake in a practice and believes that she can do so by following the rule. She wants to play chess, behave politely, or act respectfully, and believes that she can do these things by, *inter alia*, moving the Knight in an L-shape, not slurping her soup, or wearing black when mourning.

Still, as far as my practice model goes, sanctions and internalized motivations may still be relevant to making rational sense of the agent's behavior: They may be relevant to making sense of why she desires to partake in the practice in the first place. There are several reasons why someone may want to engage in a practice or activity. I may want to dress fashionably simply because I want to avoid the social ridicule reserved for those who, say, wear socks with sandals. Or I may strive to behave politely at the table lest my boss thinks I'm a rube. But I may also have an internalized, intrinsic motivation to participate in such practices. I may care for fashion, etiquette, and games intrinsically. Maybe I value fashion as a reflection of my identity and creativity, etiquette as an expression of our concern and respect for one another (Buss, 1999), and games as intrinsically enjoyable forms of activity (Nguyen, 2019).

1.9 Look-Alikes

I want to close this paper by discussing the differences between my practice model of rule-following and other similar models in the literature.

Take, first, Joseph Raz's (1999) influential *exclusionary reasons* model of rule-following. According to this model, accepting a rule gives one a second-order reason not to consider in one's deliberation certain first-order reasons, namely, those that would tip the balance of reasons against conforming to the rule. Thus,

my acceptance of the rules of chess gives me a reason to exclude from my deliberation any reason in favor of, say, not moving the Knight in an L-shape.

One worry about Raz's view is that it is unclear that we need to posit the existence of second-order, exclusionary reasons to account for rule-following.³¹ Perhaps following rules is about responding to first-order reasons of a distinctive type. My model of rule-following avoids positing second-order reasons. It says that accepting a rule gives one a reason to exclude from consideration, not some of one's first-order reasons, but some of one's options. This rule-based reason is a first-order reason, just one that has lexical priority over others. According to my model, then, it is not the case that rule-based reasons exclude other reasons. But they may exclude options that those reasons might otherwise favor.

Take, next, Scott Shapiro's (2008) *constraint model* of rule-following. Like the practice model, Shapiro's constraint model says that accepting a rule leads to certain options being ruled out from one's deliberation. However, according to him, this happens not because the agent has a reason to rule out such options, but because her acceptance of the rule makes these options *unfeasible*, in the sense of no longer being live options for her. When I accept the rules of chess, for example, not moving the Knight in an L-shape is no longer something I can do. It's not in my set of options.

Shapiro's model seems to have some paradoxical consequences. For starters, it entails that someone who accepts a rule isn't free to break it. Given her acceptance of the rule, acting against it is not among her options. Moreover, the model entails that, when the agent faces a dilemma arising from two rules she accepts, she has no alternatives whatsoever, in the sense that no course of action is feasible for her. Rule R1 tells her to ϕ , while rule R2 tells her not to ϕ . Since she accepts both rules, she's not free to either ϕ or not ϕ . My practice model avoids these paradoxical conclusions. It predicts, correctly, that someone who accepts a rule can still break it, although in doing so she acts irrationally. Similarly, it predicts, correctly again, that in cases of rule conflict, the agent is free to act in different ways. Her problem, rather, is that there's no way for her to act rationally in her situation, given that she accepts two conflicting rules.

Conclusion

The idea that we often follow rules is crucial to our common sense and philosophical self-conception as agents. However, it notably gives rise to difficult philosophical puzzles. In this paper, I have offered a solution to one of them. The challenge in question purports to show that rule-following isn't rational. I have shown, however, that it is—or, at any rate, that it can be, in the relevant sense of "rational." The key lies in realizing that following rules is, at bottom, a matter of ruling out options—namely, those whereby we fail

³¹See McClennen (2004) and references therein.

to conform to the rule. As it turns out, then, there's no deep incompatibility or tension between rule-guided and rational action. When we follow rules, especially those constitutive of practices like playing chess, polite dining, or fashion, we also act out of considerations that rationalize what we do, at least in our eyes. We're rule-followers, no doubt, and rational ones at that.

Chapter 2

Two Problems for Bicchieri's Theory of Norms

Abstract

I raise two problems for Cristina Bicchieri's influential game-theoretic view of social norms. First, the view gives an *incoherent* account of why social norms motivate behavior. Second, it offers an *incomplete* picture of why social norms are stable. Next, I argue that the problems with Bicchieri's view don't force us to abandon a rational choice approach to social norms. We can solve them by invoking further resources, namely, the idea that human beings are moved by a desire for esteem and the idea that they acquire norms from others through cultural learning. These ideas lay the basis for an improved theory of social norms that supplements and enriches Bicchieri's view.

Introduction

Cristina Bicchieri is one of the leading philosophical specialists in social norms. In two books and several articles, she has developed a comprehensive theory explaining what social norms are, how they change, and why people follow them (2017; 2014; 2006; Bicchieri and Mercier, 2014; Bicchieri et al., 2011). Bicchieri's theory is also significant methodologically. She provides a game-theoretic analysis of norms. If successful, her view illustrates the power of applying rational choice tools to questions in social theory.

As it stands, however, Bicchieri's theory of norms is problematic. As I will argue, her view faces two serious problems. On the one hand, it gives an incoherent account of why social norms motivate behavior.

On the other, it gives an incomplete picture of why social norms are stable. Yet, any adequate theory of norms must explain their motivational efficacy and stability.¹

Despite its problems, we shouldn't abandon Bicchieri's view or, indeed, a rational choice approach to social norms more generally. If anything, the issues I raise show that we need further resources to give an adequate account. In particular, we need to invoke two ideas. First, to solve the efficacy problem, we must assume that people follow norms out of a *desire for esteem*. Second, to solve the stability problem, we must assume that people acquire norms through *cultural learning*. As I will argue, these two ideas offer the key to an improved, rational choice theory of social norms.

The structure of the paper is this. First, I lay out three desiderata for an account of social norms. Next, I present Bicchieri's theory, showing how it claims to satisfy these desiderata. In Sections 3 and 4, I raise the two problems facing her view. Sections 5 and 6 are devoted to the ideas of esteem and cultural learning. Finally, Section 7 discusses how one might begin to articulate these resources in an improved theory of norms.

2.1 Desiderata

Let me begin by laying out three desiderata for an account of social norms. These desiderata come from the epistemic roles social norms play across several disciplines.

First, an account of social norms must explain their *motivational efficacy*, i.e., why they influence individual behavior. The idea that social norms motivate behavior comes from the social sciences, where they often appear in explanations of human action.² For example, social scientists have invoked norms to explain why people cooperate with strangers (Henrich et al., 2005), why employers pay wages above market levels (Akerlof, 1982), and why white US men are more violent in the South than in the North (Nisbett and Cohen, 1996).

According to the idea of motivational efficacy, social norms are genuine springs of action. Yet, theorists disagree about how exactly they motivate behavior. According to some, conformity to norms is mainly due to *social incentives*. People follow norms because they fear sanctions or experience other forms of social pressure (Pettit, 2019; Sugden, 2000; Coleman, 1994). According to others, norm compliance is mainly due to *internal motivations*. People have an intrinsic desire to follow norms. They may also feel guilt or shame at the idea of breaking them. Hence, they are motivated to obey them independently of social influences (Elster, 2015; Sripada and Stich, 2006; Anderson, 2000).

¹For other criticisms of Bicchieri, see (Brennan et al., 2013) and Kelly and Davis (2018). My criticism of Bicchieri's view is different from previous ones in being both more problematic for it and more sympathetic to it.

²For a discussion of the roles of norms in the social sciences, see Elster (1989b, chap. 4). See also Elster (2015, chap. 21).

Second, an account of social norms must explain their *stability*, i.e., why they tend to survive in a social group instead of dying out. The idea that social norms are stable comes from social theory, where they are taken to be fundamental to social order. According to several theorists, norms are crucial to explaining the stability and functioning of social systems (Horne and Mollborn, 2020; Coleman, 1994; Elster, 1989a; Parsons, 1951).

Social norms contribute to social order in at least two ways. First, norms facilitate *coordination* (Young, 2015; Binmore, 2010). They sustain regular patterns of behavior, thereby making social life predictable. Knowing what to expect from each other, people have an easier time harmonizing their actions. For example, if I know that you will stop at the red light, I can feel safe crossing the street.³ Second, norms promote *cooperation* (Horne and Mollborn, 2020; Ullmann-Margalit, 1977). They offset the temptation of free-riding on others' efforts. Hence, they enable people to work together in the service of joint goals. For example, in many neighborhoods, the norm against littering motivates people to do their part in keeping the local park clean.⁴

Finally, an account of social norms must explain their *normativity*, i.e., why group members regard them as somehow authoritative or obligatory (Anderson, 2000).⁵ The idea that social norms have normativity comes from philosophy, where it is seen as a conceptual constraint on any theory of norms that deserves the name (Hart, 1961; Gilbert, 1992). According to this idea, a social regularity can properly be called a social *norm* only if group members somehow consider themselves under an obligation or requirement to conform to it.

To appreciate the importance of normativity, note that a social entity may be efficacious and stable but fail to be a social norm. Take Lewisian conventions. According to Lewis (1969), conventions are behavioral regularities that solve coordination problems, i.e., situations where people want to harmonize their actions but can do so in more than one way. As such, conventions can be highly stable. People want to follow them if they believe others will follow them. Hence, once a convention gets started, everyone is motivated to go on behaving in the same way. But a Lewisian convention isn't a norm. Group members might not see it as authoritative or obligatory (Gilbert, 1992, pp. 40–56). They believe others will follow the convention, and they prefer that they do. Yet, they need not think that anyone is *required* to go along with the practice.⁶

³As we will see, Bicchieri distinguishes between social norms and descriptive norms. The latter include coordinating conventions, i.e., norms that solve coordination problems. I think she would agree, however, that social norms sustain behavioral regularities and make social life predictable, even if they can't be modeled as solutions to coordination problems.

⁴But see Elster (1989a, chap. 3) for skepticism about the role of norms in promoting cooperation.

⁵Of course, social norms need not be normative in a *robust* or *reason-giving* sense. They're normative because they're (taken to be) genuine requirements, not mere behavioral regularities. Still, people may have no normative reason to obey them.

⁶Still, a Lewisian convention can *evolve* into a norm. For some accounts of how this can happen, see Lewis (1969, pp. 97–100) and Sugden (2000).

To be sure, I don't claim that these three desiderata are exhaustive. An account of social norms might need to satisfy further constraints.⁷ However, efficacy, stability, and normativity are the core of any viable theory of social norms. They arise from the roles this concept plays in the social sciences, social theory, and philosophy. I turn now to discussing how Bicchieri's account claims to satisfy them.

2.2 Bicchieri's Theory

Bicchieri's central aim is explaining efficacy.⁸ Her view is that people are motivated to follow social norms because they have *social expectations*, i.e., beliefs about what other people think and do (2017, pp. 11–15). When a norm exists and is followed in a group, people expect others to follow it, and they expect them to approve of others who do the same. These expectations motivate them to obey the norm. Hence, norm-guided behaviors are *interdependent* because how each behaves depends on how they expect others to behave and what they expect others to approve of (2017, p. 2).

There are two kinds of social expectations. Both are necessary to motivate norm conformity. First, *empirical expectations* are beliefs about other people's behavior, i.e., what they usually do or will do in the future. People want to conform to a social norm partly because they expect others to comply with it. If they thought no one else followed the rule, they wouldn't want to follow it either. According to Bicchieri, this fact distinguishes social norms from moral principles. People's motivation to follow their principles doesn't depend on what they expect others to do. For example, most of us wouldn't want to commit murder even if we thought it was a common practice. In contrast, we wouldn't bother standing in a line if we thought most others were skipping it.

Second, *normative expectations* are beliefs about other people's normative attitudes, i.e., what they approve of and think should be done.⁹ People are motivated to conform to a social norm partly because they expect others to approve of it. If they didn't expect others to think they should conform, they'd prefer not to, even if they expected everyone else to follow the rule. According to Bicchieri, this fact distinguishes social norms from what she calls "descriptive norms." For example, in a Lewisian convention, people's empirical expectations are enough to motivate them to participate in the practice because they want to coordinate with others. If you came to expect everyone else to drive on the left, you'd want to use that side of the

⁷For some candidates, see Paternotte and Grose (2013).

⁸In reconstructing Bicchieri's theory, I rely exclusively on her two books on the subject (2017; 2006). My focus is on understanding how her account satisfies the desiderata introduced in Section 1. Hence, I leave out many interesting elements of the view.

⁹N.B.: On this use, normative expectations aren't normative attitudes. They are descriptive beliefs. Other authors use the term to refer to a kind of normative attitude whereby one expects others to act in a specific way on pain of becoming the target of one's disapproval. For this use, see, e.g., Sugden (2000).

road, too. In contrast, many people would not refrain from picking their noses if they didn't expect others to disapprove of this behavior.

For Bicchieri, a social norm can be said to exist just in case people are motivated to obey it in the presence of the right social expectations. More precisely:

A social norm is a rule of behavior such that individuals prefer to conform to it on condition that they believe that (a) most people in their [group] conform to it (*empirical expectation*), and (b) that most people in their [group] believe they ought to conform to it (*normative expectation*) (2017, p. 35; emphasis in original).¹⁰

Let's move to stability. Note that Bicchieri's view has the following consequence: When group members have the right normative expectations, mutual conformity to a social norm is a *Nash equilibrium* (2006, p. 22). It is a combination of strategies such that each person's action is the best reply to everyone else's. To see this, suppose that group members believe that condition (b) is satisfied. They expect most other group members to believe they should conform to a rule, *R*. In that case, group members prefer to follow *R* if they believe condition (a) is also satisfied. Namely, they want to follow *R* if they expect most others to follow it. If everyone conforms to the rule, then each does what they most prefer to do, given everyone else's actions.¹¹

Nash equilibria are relatively stable phenomena. In an equilibrium, expectations and behavior are mutually reinforcing. If people expect others to conform to an equilibrium, this belief motivates them to conform to it. No one has an incentive to adopt a different strategy. On the other hand, mutual conformity to the equilibrium confirms everyone's expectations. Seeing others conform to the equilibrium strengthens people's belief that others will generally conform to it (2006, p. 23; Ekeland, 2002). In short, a Nash equilibrium is *self-sustaining* (Sugden, 2000). It involves a feedback loop whereby expectations motivate behaviors that reinforce the original expectations.

Finally, let's turn to normativity. Note something interesting about Bicchieri's view: According to her, a social norm can exist and be followed even if it lacks *first-person normativity*. It could be that no one in the group personally approves of the norm, in that no one believes that one ought to comply with it. However, a norm must have *third-person normativity* to motivate behavior. Group members must have normative expectations to the effect that other people believe they should follow the norm. Otherwise, no

¹⁰In her 2006 book, gives a slightly more complicated account (2006, p. 11). To keep things simple, I stick with her more recent formulation, except that I phrase it in terms of norms existing in groups instead of *reference networks*, which is Bicchieri's most recent terminology.

¹¹This property of norms explains why they can solve cooperation problems, such as the Prisoner's Dilemma (2006, p. 3). For Bicchieri, social norms solve a cooperation problem by transforming it into a coordination game where mutual cooperation is a Nash equilibrium (2006, pp. 25–26).

one would want to follow it. Of course, social norms may have first-person normativity, too. Yet, third-person normativity is *always* needed for norm conformity.¹²

To appreciate how a social norm may lack first-person normativity and still be effective, consider norms held under *pluralistic ignorance*. In this state, people think that other people's attitudes are at odds with their own views, even when everyone's public behavior is identical (Bicchieri, 2017, p. 42). For example, in a group of college students, everyone may think that everyone but themselves approves of drinking heavily at college parties (Prentice and Miller, 1993). Under pluralistic ignorance, no one endorses the standing norm. Yet, they mistakenly believe that everyone else does. Hence, each is motivated to obey it.

Bicchieri's view has many virtues. It offers a simple and elegant explanation of the core desiderata for a theory of social norms. Yet, as I will argue in the next couple of sections, the view faces two serious problems.

2.3 The Efficacy Problem

Start with the first problem.¹³ As we have seen, Bicchieri claims that people obey social norms because they have social expectations. In particular, normative expectations are necessary to motivate compliance because people's personal attitudes are insufficiently motivating (Bicchieri, 2017, pp. 8–10). People may approve of a social norm. Yet, if they don't expect others to believe they should follow it, they prefer not to do it.¹⁴ The Efficacy Problem is that it seems incoherent to claim that people can be motivated to conform to a norm by other people's approval of it but not their own.

Let me pose this problem as a dilemma. Bicchieri mentions two general reasons why normative expectations may induce norm conformity (2006, pp. 23–25). First, one may fear the *sanctions* associated with acting against other people's normative views. Thus, I may want to obey a norm that others approve of because I'm afraid they will punish me, in the sense of expressing their disapproval of my behavior through some form of harsh treatment (Feinberg, 1965). For example, a Corsican man may be moved to take revenge

¹²However, normativity (first person or third person) isn't needed for a norm to merely *exist*. A norm can exist without being followed because people can have a conditional preference to obey it but lack the expectations that would trigger this preference (2006, p. 38). For a counterexample to Bicchieri's theory that exploits this feature of her view, see Brennan et al. (2013, p. 25).

¹³This section borrows heavily from Elizabeth Anderson's criticism of sanction-based theories of norms (2000, sec.4). Still, my contribution is original in two ways. First, I show that Bicchieri's view, while not a purely sanction-based view, is open to Anderson's objections. Second, I will later argue, *contra* Anderson, that the Efficacy Problem doesn't force us to give up a rational choice approach to norms.

¹⁴But this doesn't mean that people's normative attitudes are motivationally inert. Instead, what Bicchieri calls *sensitivity*, i.e., the extent to which someone endorses a norm's content, modulates the influence of social expectations over behavior. As sensitivity increases, the need for social expectations to motivate behavior decreases. Still, norm conformity is always conditional on social expectations. See (2017, p. 75).

because he dreads being shunned by his family and friends (Bicchieri, 2006, p. 42; Elster, 1990, p. 871).¹⁵ Second, normative expectations may induce norm conformity due to the *intrinsic merit* one ascribes to other people's normative beliefs. Thus, I may want to follow a norm because I recognize that other people's demands are justified, not only because they might punish me. For example, an employer might raise her workers' wages because she realizes that their salary expectation is only fair, given their output.

The Efficacy Problem is that, in either of these pictures, normative expectations can motivate behavior only if people's normative attitudes can motivate them, too. Hence, it is incoherent to claim that people are moved to follow a social norm by their expectation of other people's approval but not by their own acceptance of the norm.

Take the first idea: Normative expectations motivate norm conformity due to the fear of sanctions. If this is why people want to follow a social norm, then the sanctions must be *credible*. People must have good reason to think others will punish them for breaking the rules. Otherwise, they won't be motivated to obey them. Moreover, for sanctions to be credible, group members must be in the habit of punishing norm violators. Otherwise, people will quickly learn that breaking the rules has no consequences. Hence, normative expectations can induce norm conformity only if group members are prepared to enforce the norm through sanctions.

But enforcing a norm in this way is a *costly* activity. It takes time and effort to discover norm violations and punish offenders (Buchanan, 1975, p. 167). In doing so, one also risks provoking or alienating them (Elster, 1989a, p. 133). So, the question arises, why are group members motivated to engage in these activities? As Jon Elster puts it, what's in it for them? (1989a, p. 132).

One possibility is that people punish offenders out of self-interest. As Bicchieri points out, many social norms curb behaviors that produce *public negative externalities*, meaning collateral costs to third parties (Bicchieri, 2017, pp. 28–29). For example, when you litter the local park, you impose a cost on me and others. You cause our environment to become a bit more polluted. The norm against littering prevents people from imposing this cost on others. Hence, group members have a collective interest in seeing it enforced. Namely, everyone is better off if litterbugs are punished than if they aren't.

Yet, despite being in the group's collective interest, enforcing such prosocial norms needn't be in anyone's *individual* best interest. In this type of case, norm enforcement is a *public good*. People benefit from it even if they don't contribute to it. Hence, they lack an incentive to contribute to it. Indeed, not punishing others is the *dominant* strategy for each, meaning that it is the best choice for them no matter what other people do. If others enforce the norm, one is tempted to enjoy the benefits of enforcement without shouldering the

¹⁵Sanctions can be *positive* as well as negative. I may want to obey a norm out of "a desire to please others and thus be rewarded" (Bicchieri, 2006, p. 15). To keep things simple, I focus on the case of negative sanctions, keeping in mind that positive sanctions raise the same problems.

costs. If they don't, one knows that one's contribution isn't enough to make the norm effective. In either case, one is better off not helping enforce the rule.¹⁶

Another possibility is this: People want to punish offenders for the same reason they want to follow the norm. Namely, they expect others to believe that they should enforce it. Wanting to avoid being punished themselves, they have an incentive to punish others.

However, this suggestion is problematic. First, people don't always believe that others should enforce a social norm. Indeed, the contrary is often the case: Many social norms stand against people enforcing other social norms. For instance, pointing out someone's rude behavior is often considered impolite (Anderson, 2000, p. 181). Second, the suggestion invites a regress. What motivates people to punish those who fail to punish those who break a social norm? If the answer is further sanctions, what motivates group members to impose *them*? Someone must want to enforce the norm for reasons besides the fear of being punished. Otherwise, every norm would need infinite sanctions to be effective (Anderson, 2000, p. 181; Elster, 1989a, pp. 132–133).

The only alternative, it seems, is that group members enforce the norm because they approve of it. They believe people should follow the norm, so they are driven to punish offenders at a personal cost. But then they should want to follow the norm for the same reason. Someone willing to spend time and effort enforcing a norm should also be ready to conform to it. However, this runs contrary to Bicchieri's claim. Normative expectations, it turns out, aren't necessary to motivate norm conformity. Indeed, they can induce some people to follow a social norm only if others are independently motivated to follow and enforce it.

Now take the second idea: Normative expectations induce norm conformity due to the intrinsic merit one ascribes to other people's demands. This idea avoids the enforcement problem just described. It entails that social norms don't need to be enforced to be effective. People expect others to believe that they should follow the rules. Recognizing their demands as legitimate, they strive to satisfy them. Hence, there is no challenge to explain what motivates group members to dole out sanctions. Sanctions are superfluous to norm compliance.

However, this picture leads to an asymmetry problem. Suppose my motive to follow a social norm comes from recognizing that your demands are justified. You are right to think I should share the cake equally with you: It's the fair thing for me to do. In that case, shouldn't my own belief that I must share fairly with you get me to act? *Ex hypothesi*, the influence of your normative views over my behavior has nothing to do with the fact that they're *your* views. Instead, your views about my obligations induce me to act because I

¹⁶This observation leads some writers to be skeptical about the idea that social norms can solve cooperation problems. They can do this only if another cooperation problem related to norm enforcement can already be solved. See Heath (2008, pp. 42–63) for discussion.

recognize their intrinsic merit. If these beliefs motivate me to act when you hold them, they should also get me to act when I hold them.

Again, this runs contrary to Bicchieri's claim. According to her, there is an asymmetry between people's normative attitudes and their normative expectations regarding their influence on behavior. As it turns out, however, this asymmetry is untenable. People's normative expectations can motivate them only if their own approval of the norm can motivate them, too.

The Efficacy Problem invites a very different picture of why people follow social norms. It suggests that norms influence behavior because people *internalize* them (Henrich and Ensminger, 2014; Sripada and Stich, 2006; Parsons, 1951). Due to socialization, they acquire an intrinsic desire to follow the rules of their community. Moreover, they are internally motivated to impose sanctions on offenders. Hence, norm-guided behaviors are *independent*, not interdependent. People don't need to have social expectations to conform to a social norm.

2.4 The Stability Problem

Let's turn to the second problem for Bicchieri's theory. As we saw, Bicchieri claims that mutual conformity to a social norm is a Nash equilibrium. In such a state, expectations and behaviors are mutually reinforcing. People's expectations motivate them to conform to the norm. In turn, mutual conformity to the norm confirms everyone's expectations. The Stability Problem is that, while this story explains the stability of *empirical* expectations, it fails to explain the stability of *normative* expectations. As we will see, these expectations are exogenous to the equilibrium.

The Stability Problem arises because the way people behave isn't a strong indication of what they think should be done. In particular, the mere fact that a group behaves in a specific way isn't strong evidence that they endorse a *norm* to the effect that one should act that way. For example, many people use an umbrella when it rains, but this doesn't mean they think others should also use one.¹⁷ Hence, while seeing others conform to a social norm may confirm one's expectations about their behavior, it wouldn't seem to confirm one's expectations about their normative views.

Let me explain in more detail. Suppose you have beliefs about what other people in your group are likely to think and do. In particular, you expect them to engage in behavior *b* when they are in circumstances *c*. Also, you expect them to believe that one should generally do *b* in *c*. Now, suppose you observe that most of your fellow group members do *b* in *c*. Their behavior conforms to this regularity. In that case, your evidence confirms your empirical expectations, but it doesn't confirm your normative ones. On its own, the

¹⁷Bicchieri openly agrees with this idea, hence her insistence that one can't infer the existence of a norm merely from a behavioral regularity (2006, pp. 8–9; 2017, p. 50).

regularity isn't strong evidence that people think others should conform to it. To be sure, your evidence is *consistent* with the hypothesis that people regard the practice as a norm. Yet, it is equally compatible with the behavior being a mere group habit or an optional tradition, like singing carols on Christmas.

To sharpen this problem, consider a further fact about social groups. Namely, many of them are *porous*. People flow through them, meaning that members are constantly joining and leaving them.¹⁸ Therefore, explaining the stability of social norms involves explaining how newcomers to a group come to learn and follow these rules.

Bicchieri's view explains how new group members come to have empirical expectations. Seeing others follow a rule, they come to expect them to act that way *as a rule*, i.e., regularly (Bicchieri, 2006, p. 13). For example, if you observe all your new neighbors mowing their lawns at dusk, you will believe that this behavior is common among them. Yet, Bicchieri doesn't explain how newcomers come to have normative expectations, i.e., why they expect others to think they should conform to the regularity. As we saw, the practice itself offers no definite evidence to that effect. For all you know, there might be no norm of lawn mowing at dusk among your new neighbors. Maybe all of them engage in this practice merely out of habit.¹⁹

One possible reply to my line of argument is this: Norm-guided behaviors go beyond norm compliance. They also involve norm enforcement. That is, people don't just follow social norms; they also discipline offenders. Norm enforcement allows new group members to form normative expectations. It gives them evidence that others hold certain normative views. For instance, if you see your new neighbors rebuking anyone who doesn't mow their lawn at dusk, you have a reason to think they expect you to participate in this practice, too.

Note, however, that this reply is unavailable to Bicchieri. It presupposes a solution to the Efficacy Problem. According to the reply, normative expectations are stable because they stand in a feedback loop with norm enforcement. Enforcement practices give rise to normative expectations, which in turn help sustain these practices. But, according to the first horn of the Efficacy Problem, normative expectations are insufficient to motivate norm enforcement. The suggestion that they are invites an infinite regress. Absent a solution to this problem, Bicchieri lacks the resources to explain the feedback loop between norm-guided behaviors and social expectations.

¹⁸In sociology and social psychology, this idea is sometimes expressed by saying that groups have *permeable boundaries*. For a recent conceptualization of this notion, see Armenta et al. (2017).

¹⁹In her 2006 book, Bicchieri says that normative expectations can be second-order empirical expectations (2006, p. 15). If they are, then the Stability Problem might not arise. This is because a behavioral regularity not only gives one a reason to think it exists. It also gives one a reason to think others have a reason to think it exists (see Lewis, 1969, pp. 57–58). Yet, Bicchieri does well to move away from this view of normative expectations. According to the normativity desideratum, norms involve more than reciprocal empirical expectations. They involve normative attitudes or at least the belief that others hold them.

The reply is also problematic for a second reason. As we saw above, many social norms stand against people enforcing other social norms. Hence, some social norms may seldom be enforced. For example, it could be that, while your neighbors do think others should mow their lawn at dusk, they wouldn't dare to rebuke those who don't do this. Doing so may go against a local norm of politeness. If norm enforcement is the only source of evidence about other people's normative attitudes, newcomers may never form the relevant normative expectations.

In sum, Bicchieri's view fails to explain the stability of normative expectations. Indeed, her view may also fail to explain the persistence of empirical expectations in some cases. As she herself emphasizes, not all social norms are prescriptive. Some are *proscriptive*, meaning that they rule certain behaviors *out* rather than ruling them *in* (2006, pp. 8–9, 8-9). Proscriptive norms sustain behavioral regularities. Yet, being omissive rather than commissive, these regularities are hard to detect from an outsider's perspective. They don't involve observable behaviors. Instead, they involve people who *avoid* acting in specific ways.

In the case of proscriptive norms, normative expectations seem to come before empirical ones. One first learns that others disapprove of, say, picking one's nose. Later, one forms the belief that others don't usually behave this way. Crucially, normative expectations make the corresponding regularity *salient*. They help newcomers see that group members don't usually pick their noses, while they may never notice that others don't tend to hum *La Marseillaise*. If Bicchieri's theory fails to explain how newcomers acquire normative expectations, then it fails to explain how they acquire the empirical expectations associated with proscriptive norms.

Let's take stock of the argument so far. Bicchieri's theory of norms faces two serious problems. As it turns out, both have to do with the notion of normative expectations.

On the one hand, Bicchieri gives an incoherent account of the efficacy of social norms. She claims that normative expectations are always needed for norm conformity because people's approval of a norm is insufficiently motivating. However, as we saw, normative expectations can influence people's behavior only if their own approval of the norm can also motivate them.

On the other hand, Bicchieri gives an incomplete account of the stability of social norms. She argues that social expectations are stable because they stand in a feedback loop with norm-guided behaviors. However, norm-guided behaviors can sustain people's empirical expectations but not their normative expectations.

At the same time, normative expectations are central to Bicchieri's view. She needs to invoke them to explain the normativity of social norms. Bicchieri cannot reformulate her view in a way that appeals only to, say, empirical expectations. Instead, normative expectations are crucial to distinguishing between mere descriptive norms, like Lewisian conventions, and social norms proper.

In the following two sections, I show how we can solve the two problems facing Bicchieri's view. Moreover, I show how we can do this while remaining broadly within a rational choice approach to social norms. In particular, we can retain two crucial game-theoretic insights of the theory. First, norm-guided behaviors are *interdependent* (instead of independent, as in the norm internalization view). Second, social norms are stable because they involve expectations and behaviors in *equilibrium*.

2.5 Esteem

Start with the Efficacy Problem. The key to solving it, I believe, is realizing that Bicchieri's taxonomy of reasons why normative expectations influence behavior isn't exhaustive. According to her, normative expectations induce conformity either because of the fear of sanctions or the intrinsic merit one ascribes to other people's demands. There is, however, a third possibility: Normative expectations motivate action due to the *esteem* one expects to get (or the *disesteem* one hopes to avoid) by doing what others believe one should do.²⁰

I take the idea that people desire esteem from Geoffrey Brennan and Philip Pettit (2004).²¹ According to them, desiring other people's esteem involves wanting them to hold a particular *attitude* towards one's person. This attitude has two features that are relevant for my purposes.

First, esteem is an *evaluative* attitude. It involves a rating of a person based on their performance in some evaluative dimension. For example, if I esteem you as a soccer player, my attitude towards you involves a positive evaluation of your soccer-playing skills. I must believe that you play the game fairly well. Otherwise, it would be impossible for me to esteem you on this front.

Second, esteem is a *comparative* attitude. It involves an assessment of how the person performs in the relevant dimension compared with others in a reference group. If I esteem you as a soccer player, I must believe that you are better at the game than other players. You may play the game very well. Yet, if you don't outperform others in that group, you won't enjoy my esteem.²²

As Brennan and Pettit point out, human beings seek other people's esteem. They long to be held in positive regard by others, and they dread being held in disesteem by them. On the one hand, this preference may be purely *instrumental*. I may want to be esteemed by others because of the good things their esteem will bring me. It will help me win their trust and access new opportunities for interaction. On the other hand, the desire for esteem may be *intrinsic*. I may value other people's esteem as something good independently

²⁰Bicchieri sometimes comes close to this idea. She claims, for example, that we can follow norms out of a desire to please others (2006, p. 24). Yet, she ties this desire with a further desire to be rewarded, that is, with a positive sanction (2006, p. 24).

²¹As Brennan and Pettit point out, the idea is accepted by a long list of writers, from Cicero to David Hume.

²²For Brennan and Pettit, esteem is also a *directive* attitude, in the sense that "it communicates the message that the sort of performance esteemed or disesteemed is one to be emulated or avoided" (2004, p. 22).

of its consequences. Indeed, human beings may be biologically primed to care about esteem intrinsically as this desire would appear to increase their inclusive fitness (2004, p. 30; see also Sugden, 2000, sect. 4)).²³

The desire for esteem can explain why normative expectations have the power to motivate behavior. If I expect others to believe that I should follow a social norm, then I can also expect them to hold me in disesteem if I fail to do it. I can expect them, that is, to hold a negative attitude of disapproval towards me based on my poor performance following our shared rule. Not wanting to become the target of their disesteem, I have an incentive to conform to the rule.

Note how this idea solves the Efficacy Problem. First, it avoids the first horn of the dilemma, i.e., the enforcement issue. When people value esteem intrinsically, their motivation to follow a norm doesn't depend on expecting to be punished. If they anticipate norm violations to attract disapproval, they'll obey the norm even if they don't think others will express their disapproval by sanctioning them. This means, in turn, that others don't have to enforce the norm through sanctions to ensure compliance. It is sufficient that they are around in enough numbers so that norm violations are likely to be detected. In that case, potential offenders will expect to be discovered, and they will be deterred (Pettit, 1990, p. 739).

Let me explain. According to the esteem picture, people are motivated to conform to social norms, but not because others might punish them. Instead, they follow the rules to avoid other people's disapproval. The prospect of disesteem deters them from transgressing. In this sense, disesteem functions like a sanction. Unlike a sanction, however, disesteem is an *automatic* deterrent (Pettit, 1990, p. 739). Group members don't need to express their disapproval of an offender for it to harm him. They just need to hold the attitude. But holding this attitude isn't something people do intentionally. They disapprove of others automatically based on perceived norm transgressions.²⁴

On this proposal, there is no challenge to explain what motivates group members to punish norm violators. Social norms can be effective in the absence of sanctions. Accordingly, there is no need to assume that people's approval of a norm can motivate them independently of their social expectations.

Second, the idea that people desire esteem avoids the second horn of the Efficacy Problem, i.e., the asymmetry issue. To see this, note that one might care a great deal about other people's esteem without caring about one's self-esteem nearly as much. First, self-esteem isn't as instrumentally valuable as social esteem. Having a good opinion of myself need not help me win other people's trust or access opportunities for interaction. Second, even when the desire for esteem is intrinsic, self-esteem may have little value

²³Plausibly, the desire for esteem is a part of what some social scientists call our *norm psychology*, viz., the host of evolved adaptations allowing us to learn and follow social norms (Henrich, 2015; Sripada and Stich, 2006). Human norm psychology might also include adaptations that make possible the cultural transmission of norms described in the next section (see Sterelny (2012, chap. 7).

²⁴Anderson (2000) describes this mechanism as an attitudinal or emotional sanction. Punishment, in contrast, is a behavioral sanction, in the sense that it operates through the enforcer's behavior. I avoid calling disesteem an attitudinal sanction because I wish to highlight its automatic, non-intentional aspect.

independently of social esteem. Often, a crucial source of our self-esteem is the esteem other people we care about grant us. In such cases, I must enjoy the good opinion of others to have a good opinion of myself.

It follows that normative expectations can induce norm conformity even if people's own approval of the norm fails to motivate them. When I follow a norm because I know you believe I should, the fact that *you* hold this normative attitude matters greatly to me. It entails that following the norm is a way of avoiding your disesteem. In contrast, the fact that I share this attitude need not motivate me nearly as much. I might not care about my self-esteem. Or, if I do, I may find it valuable only to the extent that you and others have esteem for me.

In brief, invoking the desire for esteem solves the Efficacy Problem. It solves the puzzle of how people can be motivated by other people's approval of a norm but not necessarily by their own approval of it. Other people's approval matters because it is a source of esteem, while one's own approval isn't—or, at least, it isn't a source of esteem to the same degree, or one that is independent of the esteem one enjoys from others.

To close this section, let me note that the desire for esteem also explains an interesting feature of Bicchieri's theory that she herself fails to account for. Namely, it makes sense of why people's preference to conform to a social norm is conditional on their *empirical* expectations, not only on their normative ones. As we have seen, esteem is a comparative attitude, meaning that I can hold you in disesteem for breaking a norm only if I believe you do worse at following it than others. Empirical expectations allow people to infer that they will do worse than others if they break the rules. They will be among the few offenders. Hence, they can expect others to hold them in disesteem for their transgression. In short, empirical expectations make it possible for normative expectations to motivate behavior.

2.6 Cultural Learning

Let's move, finally, to addressing the Stability Problem. To solve this problem, we must reject a background assumption in Bicchieri's account of how norm-guided behaviors reinforce a group's social expectations. According to this assumption, people's access to their own group's practices is *socially unmediated*. They encounter the local regularities in the same way they might discover social regularities elsewhere. Thus, they must infer other people's attitudes merely from their outward behavior. In reality, people's access to their own group's norms is always socially mediated. Social norms are transmitted through *cultural learning*: We acquire them by interacting with other people in our community.²⁵

The idea that social norms are transmitted through cultural learning is widely accepted in the social sciences (Tomasello, 2016; Henrich and Ensminger, 2014; Boyd and Richerson, 2005; Rogoff, 2003; Tomasello

²⁵For an influential theory of cultural learning in general, see Tomasello et al. (1993).

et al., 1993). Moreover, some theorists argue that the cultural learning of norms takes a specific form. It is structured like an apprenticeship, that is, it happens through *guided participation* in social activities (Sterelny, 2012; Rogoff, 1990). Apprentices in a craft learn the trade by working under an expert craftsman. In the same way, newcomers to a group learn its rules by partaking in the local practices with the guidance of more experienced group members.

This apprenticeship in social norms has several features. First, it often relies on *explicit* instruction. A parent, for example, may explicitly teach their child that they shouldn't have their elbows on the table. Very frequently, however, the guidance provided by group members to novices is only *implicit*. Group members can pass along norms to a newcomer by teasing or shaming her if she falls short (Rogoff, 2003, 217-218). Furthermore, group members structure the newcomer's learning environment, exposing her to examples of approved and forbidden behaviors through chitchat, gossip, stories, fables, and even games (Sterelny, 2012, 164).

Second, the cultural transmission of norms is often *vertical*. Norms are handed down from parents to children and from one generation to the next. Thus, we learn many of our group's codes from our parents and elders. But the transmission of norms can also be *horizontal*. It can happen between social peers. For example, the unwritten rules of the workplace seem to spread in this way. New employees learn them from senior ones, who act as provisional mentors.

Finally, social norms are often transmitted within *formal* institutions, such as the school or the workplace. Yet, much cultural learning of norms is *informal*. It takes place outside institutions, relying instead on unplanned, decentralized interactions between group members. Even when norms spread within institutions, this is often not because they're deliberately taught. Instead, places like the school and the workplace are hotbeds of norm transmission because they offer individuals increased opportunities for interaction.

In the social sciences, the idea that norms are culturally learned is often wedded to the view that people internalize them. People follow social norms out of an intrinsic desire to follow them, and they acquire this desire because they are *socialized* into following these rules.²⁶ However, the cultural learning of norms can play a very different role in a social group, which I now want to highlight. Namely, it can give rise to *mutual normative expectations* between group members. Thanks to shared cultural learning practices, group members form expectations about each other's normative attitudes.

²⁶The connection between socialization and norm internalization features prominently in classic sociological accounts of norms such as Durkheim's and Parsons's. See Heath (2008, chap. 2) for an overview. Bicchieri is a staunch critic of the socialization/internalization picture (see, e.g., 2014).

To see this, let me first introduce the idea of a *public indicator*,²⁷ by which I mean a state of affairs, p , with the following features:

1. The fact that p holds is public.
2. p supports an inference to a further fact, q .
3. The fact that p supports an inference to q is public.

Condition (1) says that p is a public state of affairs. It is out in the open between the members of a social group. Condition (2) says that p is an indicator of a further fact, q . It is a cause, a concomitant, or a symptom of q . Hence, someone who knows that p holds is licensed to infer that q holds. Finally, condition (3) says that p 's status as an indicator of q is itself out in the open between group members. Take the fact that the clouds in the sky are dark. This is a public indicator that it will rain. It is a public fact that publicly indicates the prospect of rain. When there are dark clouds in the sky, everyone knows they should carry an umbrella.

I now wish to argue that, within a social group, the cultural learning of the local norms by newcomers is a public indicator. It publicly indicates to group members that they all know these rules and use them to evaluate each other's behavior.

First, the cultural learning of norms is a public state of affairs. As we have seen, within a group, everyone undergoes an apprenticeship in the local rules. Thus, everyone knows firsthand that newcomers are generally trained to learn these rules. Second, the cultural learning of norms is an indicator that group members know the rules. Indeed, group members can be expected to know them precisely because they're all inducted into them. Finally, this fact itself is public in the group. It shows up in the things we say to each other when someone breaks the local norms. "I guess his parents didn't teach him to say please and thank you!" "Didn't anybody tell you? You're supposed to pass the Port to the left," "Weren't you taught not to touch things that aren't yours?" As statements like these show, we publicly accept cultural learning as the default explanation for why people can be expected to know the local standards.

In sum, the cultural learning of norms publicly indicates to a group that everyone is fluent in the local norms. As such, it gives rise to mutual normative expectations between them. Given shared cultural learning practices, anyone can expect others to hold her behavior to the shared standards. She can expect them to believe that she should share equally with others, tip the waiter, or whatever. As we have seen, these expectations can motivate people to obey these norms.

²⁷This notion is inspired by Cubitt and Sugden's (2003) idea of a reflexive common indicator. One advantage of my terminology is that it doesn't presuppose an analysis of publicity in terms of common knowledge, which has come to be controversial (see Lederman, 2018). Further, unlike Cubitt and Sugden, I'm not interested in giving a formal account of public indication. An informal description is enough for my purposes

I'm now able to state my solution to the Stability Problem. The problem, recall, was to explain how normative expectations are stable. This involves explaining how newcomers to a social group come to acquire them. The answer is that new group members form normative expectations due to cultural learning. Group members introduce them to the local rules. For this reason, they come to learn these rules, but they also come to expect others to learn them. Indeed, cultural learning creates a *public normative framework*: a shared stock of rules that people use to evaluate each other's conduct. Combined with the desire for esteem, these shared rules can sustain social regularities, which then become the training ground for the next generation of group members. Thus, we have a feedback loop whereby expectations motivate behaviors, which in turn sustain the original expectations.

In closing, note that cultural learning also explains how people acquire the empirical expectations associated with proscriptive social norms. Thanks to cultural learning, people come to know what they're *not* supposed to do. Hence, they also learn what actions others expect them *not* to perform. This expectation can bring to salience the fact that others don't usually do such things. They don't pick their noses or cut in line. Noticing these facts allows newcomers to form the corresponding empirical expectations.

2.7 Towards an Improved View

Let's take stock of what has been achieved so far. I have identified two problems facing Cristina Bicchieri's theory of social norms. I have also shown how these problems can be solved by invoking the ideas of esteem and cultural learning. Crucially, these ideas offer an improved account of social norms broadly within the rational choice approach Bicchieri and others favor.

The ideas of esteem and cultural learning allow us to retain two central insights of Bicchieri's game-theoretic analysis of norms. First, the idea that people seek others' esteem helps us preserve the view that norm-guided behaviors are *interdependent*. According to the esteem picture developed in Section 5, people follow social norms because of a kind of social incentive. Namely, they expect others to hold them in disesteem if they don't follow the local rules. Thus, we don't need to think that norm-guided behaviors are independent in that people are internally motivated to follow social norms.

Second, the idea that people acquire norms through cultural learning preserves the view that social norms involve expectations and behaviors in *equilibrium*. According to the account developed in Section 6, norm-guided behaviors can sustain social expectations because people learn norms from others. Indeed, cultural learning gives rise to mutual normative expectations between group members. Thus, we have a feedback loop between social expectations and norm-guided behaviors (Fig. 2.1). Social expectations motivate norm

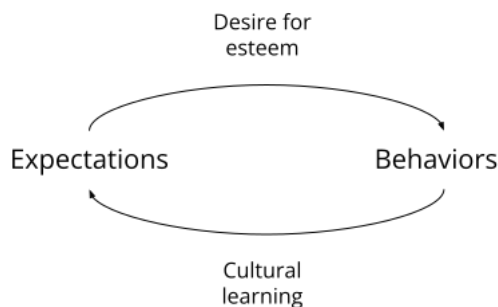


Figure 2.1: Feedback loop between social expectations and norm-guided behaviors

conformity due to the desire for esteem. In turn, norm conformity sustains the expectations due to cultural learning.

As we can see in Fig. 1, when we put them together, the ideas of esteem and cultural learning offer the skeleton of an improved theory of social norms. These ideas explain why social norms are motivationally efficacious and stable. At the same time, like Bicchieri’s view, the emerging picture satisfies the normativity desideratum. According to the proposal, social norms are normative because group members have mutual normative expectations. They expect each other to know and apply certain behavioral rules.

Of course, much work remains to be done to turn this schematic view of norms into a full-fledged theory. On the esteem front, we need to know under what conditions the desire for esteem influences behavior, how it interacts with other motives like self-interest and morality, what adaptations lie at its basis, and where they come from. On the cultural learning front, we need to know under what conditions it operates, to what extent it can reliably transmit norms, what capacities lie at its basis, and where they come from. As is perhaps apparent, we won’t solve most of these questions from the armchair. Solving them will require significant empirical work.

Let me conclude this paper with some brief comments on the prospects of applying rational choice ideas to questions in social theory. Bicchieri’s theory is a prime example of how game-theoretic notions can be used to impose conceptual clarity and give empirical content to our talk of social norms. Indeed, Bicchieri sees herself as offering a *rational reconstruction* of this concept (2006, pp. 3–4).²⁸ She uses the machinery of rational choice theory (notions such as expectations, conditional preferences, and equilibrium) to substitute a precise concept for an imprecise one.²⁹ Bicchieri’s analysis of norms allows us to distinguish between social

²⁸The idea of using rational reconstruction to elucidate our talk of norms is due to Edna Ullmann-Margalit (1977).

²⁹Much like Ullmann-Margalit, Bicchieri aims to give an *explication*, in Carnap’s sense, of the notion of a social norm. Carnap describes explication as “transforming a given more or less inexact concept into an exact one or, rather, in replacing

norms and other social entities. Moreover, it has empirical, testable content. Thus, it can underwrite efforts to measure and change social norms.

At the same time, the problems with her view speak to the need to combine rational choice rigor with social-psychological insight. As we have seen, these problems can only be solved by invoking empirically informed claims about why people follow social norms and how they acquire them. In short, a bare-bones rational reconstruction of norms is insufficient to meet the desiderata for an adequate theory of the phenomenon. Our account must put social-psychological flesh on the rational choice bone to make sense of the roles social norms play in the social sciences, social theory, and philosophy.³⁰

This eclectic approach, by the way, is not alien to the rational choice tradition in social theory. It is already present in David Lewis's trailblazing account of social conventions (Guala, 2007). In developing his account, Lewis appeals both to the game-theoretic idea of equilibria in a coordination game and the social-psychological concept of salience. I see no reason to be less eclectic in theorizing about social norms.

the first by the second" (1962, p. 3). Unlike conceptual analysis, the primary goal of explication is not to describe how we, as a community of speakers, use a term. Instead, it aims to introduce an analogous new concept that is fruitful for some theoretical enterprise, even if it distorts the pre-theoretic notion.

³⁰Of course, this is not to suggest that Bicchieri's account isn't informed by social-scientific insight. On the contrary, her theory invokes many important ideas from a wide range of disciplines to explain how norms emerge, how they change, and how people follow them in context. Still, her core analysis of what social norms are is highly abstract.

Chapter 3

The Sanctions Paradox

Abstract

According to a natural idea, people in a group are motivated to adhere to social norms because they are willing to impose social sanctions on each other in case anyone breaks them. However, this view faces the *sanctions paradox*, according to which people who are willing to sanction others for breaking a norm should also be motivated to follow the norm in the absence of sanctions. I offer a solution to the paradox based on the idea that social sanctions can serve a communicative function in a group. If people use sanctions as *signals* of wrongdoing, the cost of sanctioning others is *cheap* compared to the cost of being sanctioned by them. Hence, one may be motivated to sanction others without thereby being motivated to obey the norm absent sanctions.

Introduction

Why do we follow *social norms*, i.e., the unwritten rules of conduct that exist in every human group? Consider some examples. When you have a picnic at the park, you're supposed to take your trash away with you. And when you want to get coffee in the morning, you're supposed to stand in line at the coffee shop. The question is, why do you bother doing these things? Taking your trash with you is a hassle; standing in line takes time.

Of course, cleaning up after yourself helps keep the park clean. But leaving a couple of napkins behind when you leave won't make the park a mess. Likewise, standing in line helps everyone get their coffee. But if others are already standing in line, you can save yourself a lot of time by skipping it.

These considerations motivate the following thought: To the extent that we follow these social rules, we do it because *other people* care if we do and may behave unpleasantly toward us if we don't. They may gossip about us to their friends or refuse to greet us when we cross them in the hallway. At worst, they may call us out on our behavior, much to our embarrassment. In a word, they may impose various *social sanctions* on us for disobeying these norms.

You may be thinking that this idea paints an excessively unflattering picture of our motives. We often see some value in behaviors like cleaning up after oneself and respecting one's turn in line independently of their social consequences. For example, even if leaving some bits of trash behind won't make the park a mess, you may still care about being a clean and neat person. Hence, the idea goes, we often prefer to do these things independently of whether other people care if we do them.

But here a moment of self-examination is called for. Ask yourself: If nobody minded trashing the park, would you still want to clean up after yourself before leaving? And if everyone else was jumping the line, would you still want to stand in it? For many of us, the answer to these questions seems to be "no" (Bicchieri, 2006). Even if we see some value in these actions, this may only be because other people see value in them too, and may sanction us if we don't act accordingly.

In this paper, I study the idea that we follow social norms mainly because of the sanctions associated with them. As we will see, this view is held by many social theorists, who see it as the key to explaining why social norms are so effective in motivating action. Social norms are effective because they are backed by a system of social sanctions.

Yet, this view faces a challenge. According to some critics, the view is *self-defeating*: If it's true, then it's false. As opposed to what happens with legal norms, there's no official authority in charge of enforcing social norms. We enforce these norms ourselves by sanctioning each other when we break them. But if we care enough about a norm to bother sanctioning each other for breaking it, then it seems we must also care enough about it to follow it *in the absence of sanctions*. This is the *sanctions paradox*.

In what follows, I offer a solution to this paradox. As we will see, my solution integrates many ideas for how to solve the paradox but also adds something new to them. I argue that social sanctions are a way for people in a group to communicate with each other about violations of their shared norms. Sanctions, in a word, are a kind of *signal*. Yet, given people's shared interests in disapproving of norm violators, these signals can be reliable while still being quite *cheap*. Since the cost of sanctions is low compared to the cost of norm compliance, we may care enough about a norm to bother sanctioning others, yet not enough about it to follow the norm in the absence of such sanctions.

I proceed as follows. In Section 1, I explain why social norms give rise to the question of why we're reliably motivated to follow them, even when this question fails to arise for other kinds of social rules. In

Section 2, I formulate the sanctions paradox in detail. Sections 3 to 6 are devoted to discussing different ideas for solving the paradox. In each case, I argue that the idea adds an important element to the solution, even if it doesn't fully solve the paradox. Finally, in Sections 7 and 8, I develop my own solution.

3.1 The Motivational Question

Social norms raise a question about motivation. When a social norm exists in a group, people are reliably motivated to adhere to it. While some people may break the rule sometimes, most tend to follow it most of the time.¹ The question that arises is: What explains their motivation to do so?

Let me explain how this motivational question regarding social norms arises. Consider, first, the fact that social norms, or at least many of them, are *collectively beneficial* rules (Horne and Mollborn, 2020; Coleman, 1994; Ullmann-Margalit, 1977).² Norms are collectively beneficial in the sense that we're all better off with them than without them (Pettit, 2019). As many social theorists have pointed out, social life would be extremely challenging, if not impossible, without a framework of social rules that include truth-telling, promise-keeping, respect for each other's property, and cooperation. Social norms are a fundamental component of social order: A world without such rules would resemble the bleak Hobbesian vision of a state of nature involving the war of all against all (Horne and Mollborn, 2020).

Notice, however, that these types of social rules can be collectively beneficial only at the expense of being *individually burdensome* (Pettit, 2019). Norms require each of us to do things we often would prefer not to do. They require that we refrain from telling lies, breaking promises, seizing others' property, and free-riding on others' efforts, even when acting in these ways might seem to serve our rational self-interest. Adhering to social norms is individually burdensome or costly because it involves sacrificing our pursuit of individual interests for the sake of the group's interests.

This property of social norms can be more precisely defined. Indeed, many social theorists contrast the rules involved in social norms with other types of social rules, such as coordinating conventions (Pettit, 2019; Bicchieri, 2017, 2006; Ullmann-Margalit, 1977). These different kinds of social rules contribute to social order in different ways. Coordinating conventions solve coordination problems, that is, situations where people want to coordinate with each other but there are multiple ways to do so (Lewis, 1969). Coordinating conventions solve these problems by allowing people to know what to expect from each other. For instance, if there is a convention of driving on the right side of the road, I can expect you to drive on the right, and I can plan to do the same.

¹This fact about norms is what distinguishes them from mere social standards, that is, principles people endorse but need not live up to. For this distinction between norms and standards, see Pettit (2019).

²See, however, Elster (1989a, chap. 3) for a critique of this idea.

However, when it comes to social norms solving collective action problems, these aren't merely coordination problems. Instead, they are what Cristina Bicchieri (2006) calls *mixed-motive games*—situations where people share a common interest in acting together, but they also have conflicting interests over how to do so.³ The classic example is the free-riding problem, where people aim to achieve a collective good, such as a litter-free environment, yet each person is tempted to let others make all the effort to secure that collective good (Hardin and Cullity, 2020). In this scenario, people have a shared interest in maintaining a clean environment, but they have conflicting interests in deciding who will contribute to keeping it clean. Social norms resolve this problem by requiring everyone to contribute, even if they'd prefer not to.

However, this distinction between social norms and conventions leads to a motivational question that doesn't arise in the case of conventions. Once a convention is established in a group, it is clear why everyone should be motivated to follow it consistently. Following the convention aligns with each person's rational self-interest, provided they expect everyone else to do the same. Conventions are *Nash equilibria* in the underlying coordination game, meaning that each person benefits from following these rules whenever everyone else does (Lewis, 1969).⁴

However, this isn't the case for social norms. Social norms aren't equilibria of the underlying games they solve.⁵ Therefore, simply expecting others to follow these rules is insufficient to motivate people to abide by them. For example, if I can expect other people to refrain from littering the park, this might give me even more reason to litter because I can enjoy the collective benefit without incurring any individual cost. The question then becomes: What motivates me and others to consistently adhere to such rules?

3.2 The Sanctions Paradox

But maybe this motivational question regarding social norms has a simple answer. As many social theorists point out, social norms are typically backed by *social sanctions*, namely, social penalties associated with breaking such rules (Horne and Mollborn, 2020; Elster, 2017, 1989a; Bicchieri, 2006; Anderson, 2000). It is natural to think, then, that people are motivated to adhere to social norms because they wish to avoid being the object of such sanctions.

Of course, social norms don't trigger the same type of sanctions associated with breaking more formal rules and codes, such as the law (Elster, 1989a). Firstly, when it comes to social norms, there is no official authority responsible for penalizing deviations from these rules. When you mix stripes and plaids, there's no

³Bicchieri's notion of a "mixed-motive" game is slightly different from the traditional understanding of this idea in game theory, c.f., Schelling (1980, p. 89).

⁴Strictly speaking, conventions are what Lewis calls "coordination equilibria," meaning Nash equilibria such that departing from them not only makes the individual worse off, it also makes other players worse off.

⁵According to Bicchieri, social norms *transform* the games they solve into games where following the norm is an equilibrium (2006, p. 26). Yet, this transformation presupposes that sanctions are available and effective in the group.

literal fashion police around to administer a punishment fitting for your “crime.” Instead, with social norms, people administer the sanctions themselves by monitoring and penalizing each other’s behavior.

Secondly, the type of sanctions associated with social norms aren’t the sort of formal, material punishments and penalties associated with breaking the law, such as fines and prison sentences. Instead, they are informal penalties related to people’s ability to enjoy social goods like approval, reputation, and status. They include such things as criticism, disapproval, gossip, snubbing, and ostracism (Pettit, 2019; Horne, 2007; Elster, 1989a).⁶ Still, for social creatures like us, these kinds of social penalties are genuine penalties. Hence, most people are motivated to avoid them.

Although it may seem compelling at first, the idea that people follow social norms because they wish to avoid sanctions encounters a significant challenge. As several critics have argued, this idea would seem to be *self-defeating*—that is, if the idea is true, then it would seem to be false (Anderson, 2000; Elster, 1989a). I call this the *sanctions paradox*.

To see how this paradox arises, consider, first, that—as things seem—sanctions are always *costly* to impose, even when they’re only the sort of informal social penalties associated with disobeying social norms (Elster, 1989a; Buchanan, 1975). For one thing, imposing such sanctions on others takes time and effort: One must monitor their behavior and respond appropriately to deviations by expressing one’s disapproval in criticism, gossip, or whatever. Moreover, in sanctioning others, one also risks provoking or alienating them. For example, when you confront someone for littering the park, you waste a good amount of time and energy, all while risking conflict or retaliation. Thus, social sanctions give rise to a *second* motivational question, that is, if sanctions are individually costly, why are people reliably motivated to sanction norm transgressors?⁷ As Jon Elster poignantly puts this question, what’s in it for them? (1989a, p. 132).

The answer to this second motivational question can’t be that people sanction others because they’re afraid of being sanctioned themselves (Elster, 1989a). First, as a matter of empirical fact, people don’t typically sanction people who fail to sanction norm transgressors (Anderson, 2000).⁸ Second, this suggestion invites a regress. What motivates people to impose sanctions on those who fail to sanction others? If the answer is further sanctions, where does the cycle end? Sanctions can’t go on indefinitely. Somewhere or other, people must want to sanction norm transgressors for a reason other than wanting to avoid a sanction.

What might this reason be? One natural thought is that people sanction norm transgressors simply because they *approve* of the norm (Anderson, 2000; Elster, 1989a). That is, they think everyone should

⁶As will become clear below, another difference between social sanction and formal/legal sanctions is that the latter are typically intentionally imposed, while the former need not be imposed intentionally on others.

⁷It is sometimes said that norms that solve collective action problems give rise to a *second-order* collective action problem having to do with the imposition of sanctions (Buchanan, 1975).

⁸Indeed, as Anderson (2000) points out, sometimes punishing norm violations is itself regarded as a norm violation. Pointing out other people’s impoliteness is sometimes considered impolite.

follow the norm, so they're motivated to sanction offenders, even at a personal cost. I may, for instance, think that trashing the park is the wrong thing to do. And this belief may drive me to chastise any litterbug I happen to encounter.

But now we come to the crux of the paradox. For if norm approval is enough to motivate people to sanction norm transgressors, then it seems it should also be enough to motivate them to obey the norm *in the absence of sanctions*. In other words, if norm approval is a good answer to the second motivational question—i.e., why do people sanction norm transgressors?—then it should also be a good answer to the first motivational question—i.e., why do people follow the norm?

To see this more clearly, let's say that when someone approves of a norm, they incur a cost whenever this rule is broken (either by themselves or by others)—call this cost a . Also, following the norm has a cost coming from the individual burden of compliance—call it b . Finally, let c be the cost of sanctions. Then, people in a group are generally motivated to sanction norm transgressors only if, for most of them:

$$(1) a > c$$

Likewise, sanctions motivate norm compliance in the group only if, for most people:

$$(2) pc > b \text{ (where } p \text{ is the probability of transgressors getting caught)}$$

But (1) and (2) jointly entail that, for most people in the group:

$$(3) a > b$$

That is, if norm approval is enough to get people in the group to pay the cost of effective sanctions against norm transgressors, it should also get them to endure the burden of adhering to the norm.

It follows that the idea that people obey social norms because they want to avoid sanctions is self-defeating. This idea presupposes that sanctions are (1) *available*, in the sense that people are generally motivated to sanction norm transgressors. The idea also presupposes that sanctions are (2) *effective*, in the sense that the threat of sanctions deters people from transgressing. However, as the paradox seems to show, whenever effective sanctions are available, they're also (3) *redundant*. Under these circumstances, people are motivated to comply with the norm independently of sanctions.⁹

⁹For other formulations of this sanctions paradox, see Buchanan (1975, pp. 132–133) and Brennan and Pettit (2004, chap. 14).

The sanctions paradox suggests a different view on why people are motivated to follow social norms. The idea is that people follow norms because they have *internalized* them, meaning that they have acquired an intrinsic desire to obey and enforce them (Anderson, 2000; Elster, 1989a).¹⁰ For example, they cooperate with others, not because this suits their individual interests, but because they have been socialized to value cooperation in and of itself. At best, sanctions play only a secondary role in norm compliance. They're an important backup motive in cases where one's intrinsic motivation falters. If I'm ever tempted to litter the park, for example, the thought that you might catch me may dissuade me from doing so. Most of the time, however, sanctions are redundant to norm compliance.

3.3 Altruistic Punishment

Can the sanctions paradox be solved? In the next four sections, I'll discuss various attempts to solve it. Although, as we shall see, these proposed solutions are incomplete, each contributes an important element to the solution I'll develop starting in Section 7.

Consider, first, the idea that the paradox can be solved by appealing to the phenomenon of *altruistic punishment*. This is the phenomenon whereby people exhibit a tendency or disposition to punish others at a cost to themselves and without clear gains (Fehr and Gächter, 2002; Van Lange et al., 2014). In cooperative interactions, altruistic punishers are willing to impose costly sanctions on non-cooperators, even when doing so has no strategic or material benefits for them.

Various empirical studies have shown that people tend to behave like altruistic punishers. For example, in a now-classic study, Fehr and Gächter (2002) had subjects play a public goods game¹¹ with the option of punishing each other. Subjects were given tokens that they could use to make contributions to a public pot. Moreover, they could observe other people's contributions and punish them by having some of their tokens deducted from them for a fee. Using this experimental paradigm, Fehr and Gächter found that subjects were willing to punish non-cooperators even when (1) doing so was costly to them, (2) there were no repeated interactions and, hence, punishment didn't deter future non-cooperation, and (3) participants were unknown to each other and, hence, punishment didn't improve one's reputation as a cooperator.

One might think that altruistic punishment offers a solution to the sanctions paradox. Given this phenomenon, there seems to be no special need to account for people's motivation to enforce social norms through sanctions. Even if sanctions are costly, they're also readily available, at least in the case of social norms that promote cooperation. Norm compliance, on the other hand, isn't altruistic. On the contrary,

¹⁰For a detailed discussion of the idea of internalization, see Kelly (2020).

¹¹In a public good game, players have the choice of making contributions to a public pot. The amount in the pot is then multiplied and divided between the players. For each player, it is individually rational to let others contribute to the pot, while he keeps his resources for himself. Yet, when everyone does this, they all do worse than if they had all contributed to the pot.

the prospect of sanctions is crucial to explain compliance with cooperative norms. In altruistic punishment studies, for example, introducing the option of punishing others substantially increased the overall level of contributions to the public pot. These studies suggest that, in general, people know that others will likely punish them if they don't cooperate, and this knowledge motivates them to cooperate.¹²

Altruistic punishment explains various important features of social sanctions. First, as just mentioned, the idea explains why social sanctions are widely *available*; that is, it explains why, in general, norm transgressions are associated with social penalties like disapproval and criticism. Second, the idea accounts for why social sanctions tend to be *sincere*; that is, it explains why, in general, people don't impose sanctions on others merely to further their own self-interest, but rather, they do so out of a sincere commitment to cooperative norms. Finally, the idea explains why social sanctions can often involve an *emotional* element. Altruistic punishment studies indicate that punishment is often associated with strong negative emotional reactions.

However, considered in and of itself, altruistic punishment doesn't fully solve the sanctions paradox. This is because, recall, solving the paradox not only involves explaining how sanctions can be widely available in a group but also explaining why, in being widely available, sanctions don't also become *redundant*. The idea of altruistic punishment doesn't account for why people's motivation to punish non-cooperators doesn't also translate into a motivation to cooperate in the absence of punishment.

Indeed, altruistic punishment only raises the same puzzle in a slightly different form. With altruistic punishment, we might ask what sort of process is responsible for our tendency to punish others altruistically. As with other psychological propensities or dispositions, the possibilities here range from natural selection to cultural evolution and individual learning (or a combination of these processes) (O'Connor, 2022). But then the worry is that any process that selects for a capacity for costly punishment of norm transgressors should also select for a capacity for costly norm-following. If the process sustains a disposition to punish others at a cost to oneself because of the benefits associated with cooperative behaviors, why shouldn't it directly sustain a disposition to engage in such behaviors? (Boyd et al., 2003). So here, again, it seems we're right back in the paradox.

3.4 Cheap Sanctions

As we have seen, the sanctions paradox suggests that anyone motivated to impose costly sanctions against norm transgressors should also be motivated to adhere to the norm in the absence of sanctions. However, this idea holds only if the cost of sanctions is comparable to the cost of norm compliance. Elliot Sober and

¹²For a review of the various findings of altruistic punishment studies, see Van Lange et al. (2014).

David Sloan Wilson (1998) question this assumption. They argue that, while sanctions are costly, their cost is generally lower than the cost of compliance. Sanctions are *cheap* compared to norm conformity.

According to Sober and Wilson, obeying cooperative social norms has a significant cost given the nature of the social problems these rules are meant to solve. As we saw in Section 1, in promoting social goods such as honesty, fidelity, respect for property and reciprocity, social norms impose a significant burden on individuals. As Sober and Wilson point out, other cooperative norms of division of labor and fair sharing demand large expenditures of time, energy, and resources. Indeed, these costs can be so high that individuals might prefer not to incur them were it not for the threat of social rewards and sanctions.

In contrast, imposing sanctions on others can be done at a much lower cost to the individual. Sober and Wilson illustrate this idea with the following example:

[In 13th-century Spanish Jewish society] a law was... in place that a man who did not pay his taxes would have a blot placed upon his genealogy... A blot on a person's genealogy could prevent a marriage. Thus, failure to pay taxes could have severe costs, not only for the individual but also for his descendants... It costs an individual a great deal to fork over a large fraction of his income. It costs the keeper of the genealogy almost nothing to make a blot. (1998, pp. 144–145)

Supposing this kind of asymmetry exists between the cost of norm compliance and the cost of sanctions, the sanctions paradox fails to arise. This asymmetry presupposes the existence of another asymmetry, at least if sanctions are effective—namely, it entails that the cost of *sanctioning others* is significantly lower than the cost of *being sanctioned by others*. This is reflected in Sober and Wilson's example. The cost of making a blot of ink on a piece of paper is negligible, but the cost of being deprived of the kind of social goods that typically come with marriage is quite significant, especially in a society where being married is central to being able to access other social goods such as reputation and status.

With this asymmetry in mind, it's easy to see where the sanctions paradox falls apart. As I have reconstructed it in Section 2, the paradox relies on the unexpressed assumption that the cost of sanctioning is equal to or greater than the cost of being sanctioned. If this assumption is incorrect, however, then (1) and (2) don't jointly entail (3). If the cost of sanctioning, c^* , is lower than the cost of being sanctioned, c^{**} , then $a > c^*$ and $pc^{**} > b$ need not entail $a > b$. That is to say, social sanctions can be available and effective without thereby being redundant. People's approval of a norm motivates them to altruistically sanction norm transgressors, and these sanctions are sufficiently severe to motivate norm compliance. Yet, people aren't motivated to follow the norm by their own approval of it. The threat of sanctions is needed to motivate norm compliance.

The idea of cheap sanctions also solves the puzzle arising from the phenomenon of altruistic punishment. Since punishment is cheaper than norm conformity, a process that selects for a capacity for altruistic punishment need not also select for a capacity for obeying norms altruistically. On the contrary, cheap sanctions make it possible for processes like natural or cultural evolution to stabilize cooperative norms because it makes it possible for such processes to stabilize practices of punishing deviations from such norms.¹³ When this happens, people follow norms to avoid being punished, yet they punish others altruistically.

As we can see, then, the idea of cheap sanctions promises to solve the sanctions paradox. Nevertheless, a complete solution is still lacking. For we don't yet have an explanation for why, in general, sanctions are so cheap. Regarding this point, Sober and Wilson write that "[i]t is an important fact of human life that individuals can often greatly increase (reward) or decrease (punish) the fitness of others at trivial costs to themselves" (1998, p. 144). But what accounts for this "important fact of human life"? In the next few sections, I suggest that the explanation has to do with our desire for approval or esteem.

3.5 Esteem

To explain why social sanctions are cheaper than norm conformity, we can start by observing that sanctions are often associated with attitudes of approval and disapproval. That is, when someone breaks a norm, others not only impose sanctions on him in the form of behaviors like criticism, gossip, or ostracism, they also disapprove of his conduct. Indeed, these sanctioning behaviors are just various more or less conventional ways of *expressing* this attitude of disapproval. Likewise, when someone honors a norm, this behavior might trigger certain social rewards, but also attitudes of approval and admiration on the part of others.

The attitudes of approval and disapproval behind social sanctions are important because people not only care about how others *act* towards them but also about what they *think* about them. They have a desire for social approval or—as some theorists call it—*esteem* (Pettit, 1990; Sugden, 2000; Brennan and Pettit, 2004).¹⁴ That is to say, they want others to have a good opinion of them, particularly one that casts them in a better light than relevant others (Brennan and Pettit, 2004). For instance, as a philosopher, I desire the esteem of my peers, hoping they see me as a proficient philosopher, perhaps even surpassing others in the field.

The desire for esteem is partly instrumental. Enjoying the good opinion of others is crucial to being able to access a host of social goods. When others have us in good esteem, they're more likely to trust us and cooperate with us. But the desire or preference for approval is also importantly intrinsic. That is, most of

¹³Sober and Wilson call this process the *amplification* of altruism. Given this amplification, processes like group selection may select for behaviors that may otherwise be too individually costly to be selected.

¹⁴For a detailed discussion of the nature of the attitude of esteem, see Brennan and Pettit (2004, chap. 1).

us value social approval for its own sake, independently of the good consequences it might bring us. We care about what others think of us, and we are terrified at the idea that they might hold a bad opinion of us, even if, for some reason or other, this has no upstream consequences in terms of how they behave toward us.

Because people care deeply about disapproval or disesteem, the cost of being sanctioned by others can be quite high. As previously mentioned, social sanctions convey others' disapproval of one's behavior. Furthermore, because this disapproval is based on standards and rules shared within a group, it carries a sort of collective validation which makes it particularly costly for the recipient. Social sanctions express one's group's disapproval of one's behavior, over and above each individual's personal assessment of it (Bicchieri et al., 2018). In Sober and Wilson's example, for instance, the blot of ink upon one's genealogy is a socially sanctioned expression of condemnation against one's refusal to pay one's taxes.

However, in contrast to the high cost of being the object of social disapproval or disesteem, the cost of disapproving of others can be quite low. Indeed, as Philip Pettit has argued, disapproving of others can be *costless*, insofar as people don't disapprove of others intentionally (Pettit, 2023, 1990; see also Brennan and Pettit, 2004, 2000). Unlike the *expression* of disapproval through criticism and other sanctions, the *attitude* of disapproval isn't a kind of intentional response. Instead, one forms this attitude more or less automatically based on perceived norm transgressions. When you see me littering the park, for example, you form a bad opinion of me, but this isn't something you do intentionally. Hence, your reaction doesn't carry the costs in time and energy necessarily associated with intentional behavior.¹⁵

If people desire social approval or esteem, then they suffer a cost whenever they know others disapprove of them. Even if such others never express their disapproval, the mere fact that they hold this attitude is a kind of penalty; It frustrates one's desire for esteem. This means that disapproval can itself be a sanction, whether or not it gets expressed through criticism. If I know that others will think poorly of me for breaking a norm, then I know I'll suffer a penalty for breaking it, even if they never say a word to me.

According to Pettit, the possibility of non-intentional disapproval offers a solution to the sanctions paradox. Often, it is a matter of common awareness between group members that certain behaviors will trigger other people's disapproval. If I cut in line at the coffee shop, it will be a matter of common awareness between me and the people standing in line that my behavior is censurable in light of our shared norms. The prospect of being the target of this social disapproval may be sufficient to deter me from cutting in line, but others don't have to do as much as lift a finger to enforce this rule.

¹⁵Of course, even if disapproval isn't costly in the sense of being intentionally imposed on others, it's still costly in another sense. Disapproval—at least of the kind associated with norm transgressions—is a *psychologically taxing* attitude (Shoemaker and Vargas, 2021). When I disapprove of you, holding this attitude carries various mental costs for me. It affects my “equanimity, time, energy, self-control, and self-governance” (2021, p. 582). Just think about how unpleasant and unsettling it can be to see someone cutting in line at the coffee shop. However, this is not a cost the agent incurs intentionally. Because disapproval is a non-intentional or automatic response, it doesn't give rise to what I have called the “second motivational question.”

The idea of non-intentional disapproval captures one crucial way whereby social norms are enforced. However, it isn't a full response to the sanctions paradox. Considered in and of itself, the idea doesn't explain why, when social norms are broken, people *do* typically express their disapproval of violators through criticism, gossip, and ostracism. As we saw, part of the observation that social norms are backed by sanctions is the idea that norm transgressions typically trigger not only sanctioning attitudes like disapproval but also various sanctioning behaviors (Elster, 1989a; Nolan, 2017). However, if these behaviors are costly enough, then this raises the paradox once more. In that case, we would need to conclude that whatever motivates people to express their disapproval of norm transgressions should also motivate them to adhere to the norm in the absence of disapproval.

3.6 Virtue Signaling

Perhaps to explain why people express their disapproval of norm violators through criticism, gossip, and other social sanctions, we must appeal to the idea that these behaviors can operate as a kind of *signal* within a group. In particular, sanctioning norm transgressors might be a form of *virtue signaling*, in that people do it to communicate to each other their commitment to abiding by the norm.

Again, the starting point for this suggestion is the idea that people generally seek the approval or esteem of others. Naturally, this desire may translate into a concern for maintaining a good reputation among them (Pettit, 2023). If people know that others generally approve of a norm, they'll want to cultivate a reputation for being a reliable norm follower—someone committed to honoring the norm. For instance, if I know that people around me care about recycling, then I'll want to cultivate a reputation as a zealous recycler because this will improve my esteem and status within the group.

The desire to have a good reputation might motivate people to sanction norm violators as a way of advertising their commitment to the norm to others. In such cases, people engage in virtue signaling: They participate in public moral discourse to improve their moral reputation (Westra, 2021; Tosi and Warmke, 2016). For example, to showcase my commitment to recycling, I may publicly condemn anyone whom I catch not sorting plastic and paper into different bins.

This kind of virtue signaling is possible because, in condemning norm offenders, one opens oneself up to a cost if one ever breaks the norm oneself (Pettit, 2023). Namely, if this happens, others may rightfully accuse one not only of breaking the norm but also of being a hypocrite. Condemning norm offenders, then, is only rational if one is truly motivated to honoring the norm. Hence, it serves to indicate to others that one indeed has this motivation. In other words, virtue signaling through sanctions can be understood as a kind of *costly signaling* (Zahavi, 1975). That is, it is a form of signaling whose cost carries reliable information

about the presence of an underlying trait. Just as only the fittest peacocks can afford a big and flashy tail to attract peahens, only those people truly committed to a norm can afford to engage in costly displays of sanctioning behavior.

Like non-intentional disapproval, virtue signaling is an important way by which norms are enforced. However, this idea can't explain why sanctions tend to be widely available in a group. Costly signals only work well if they're not accessible to all. If most individuals can afford the signal, then the signal lacks informational value because it fails to expose outliers (Sterelny, 2012, p. 110).¹⁶ For example, if any peacock could afford a big and flashy tail, peahen wouldn't be able to distinguish high-quality from low-quality males based on their tails. When understood as a form of costly signaling, then, virtue signaling may explain why people engage in relatively costly and infrequent forms of sanctioning, such as very public and ostentatious forms of condemnation. Yet, it fails to explain why they engage in less costly and more frequent forms of sanctioning.

3.7 Cheap Signals

In this section, I'll start developing my solution to the sanctions paradox. My solution draws on the various ideas for solving the paradox we have discussed so far, particularly the esteem account of Section 5. I'll argue that sanctions are cheap signals of norm violations. That is, when someone breaks a norm, behaviors like criticism, gossip, and ostracism serve to communicate this transgression to the group in a way that need not involve a high cost. It follows that sanctioning others is generally cheaper than being sanctioned by others.

As we saw when discussing the esteem view in Section 5, under certain conditions, social sanctions can be costless, in the sense that people don't intentionally spend time or energy imposing them on others. In particular, norm enforcement through attitudinal sanctions of disapproval can be costless under what economists call conditions of *perfect monitoring* (see e.g., Mailath and Samuelson, 2006), that is, when people's choices, including their norm violations, are a matter of common awareness among group members.

Under conditions of perfect monitoring, people don't spend time or energy being on the lookout for norm transgressions. Instead, whenever someone breaks a norm, everyone knows about it, and this as a matter of common knowledge. Moreover, upon discovering norm violations, people don't spend time or energy punishing the transgressors. They enforce the norm by forming an attitude of disapproval toward them, which they don't do intentionally. An example of a situation that approximates these conditions is standing in a line. When someone jumps a line, this fact is immediately clear to everyone who's standing in it. Further, people in line enforce it by coming to disapprove of the line jumper.

¹⁶For an insightful discussion of why costly signaling theory doesn't apply well to social traits that are widespread in a group, see Sterelny (2012, chap. 5)

However, many, if not most, cases where norms are broken aren't like this. Instead, they are cases of *imperfect monitoring*, in the sense that norm violations aren't a matter of common awareness among group members. Under conditions of imperfect monitoring, only some but not all people in the group know that a shared rule has been broken. If Jones cheats me on a business deal, you may not know anything about it. If his cattle trespass my property, you—who live on the other side of the hill—may remain ignorant of this fact. If he fails to share fairly with me, he might still share fairly with you.

Under conditions of imperfect monitoring, there's room for communicating about norm violations. Situations of imperfect monitoring are situations of imperfect information: In these situations, when someone breaks a norm, some people know about this, while others don't. Hence, there's an opportunity for the former to inform the latter about the transgression. If you don't know about Jones's wayward behavior, I have an opportunity to inform you about it. Clearly, this opportunity doesn't arise under conditions of perfect monitoring where norm violations are a matter of common awareness in a group.

Assuming that under conditions of imperfect monitoring there's room for communication about norm transgressions, what can we say about the cost of this communication? That is, if group members send and receive signals concerning the violation of certain shared rules, what can we expect the cost of these signals to be? We can approach this question from the perspective of *costly signaling theory* (Spence, 2002, 1973; Veblen, 1953).¹⁷ One of the central insights of this theory is that the cost of a signal is a function of the interests of the communicative parties. In particular, if the parties have conflicting interests in communicating, then the signals they use to do so can be honest or reliable only if they're sufficiently costly. The parties have conflicting interests whenever successful deceit would benefit the signaler at the expense of the receiver (Bird and Smith, 2005). For example, peacocks benefit at the expense of peahens if they're able to deceive them about their quality as males. Hence, between peacocks and peahens, honest communication is only possible through costly signals, like a big and flashy tail.

In contrast, if the communicative parties have a common or shared interest in communicating honestly, then the signals they use to do so can be quite cheap (Maynard Smith, 1991). That is, if the signaler wouldn't profit from deceiving the receiver at her expense, then both parties can communicate honestly or reliably at no great cost to each.

In the case of norm violations, what can we say about the interests of group members in communicating about these events? Clearly, people typically will want to deceive others about their own norm-following behavior. For example, deceiving you and others about my cheating ways is clearly beneficial for me. If I deceive you successfully, I avoid your disapproval for breaking our shared rules. Moreover, this deceit benefits me at your expense because you continue cooperating with me when you otherwise wouldn't. I get

¹⁷See Bird and Smith (2005) for a review of costly signaling theory and its central tenets.

to enjoy the benefit of being a cheater, an exploiter, or a free rider, while you and others have to put up with the costs. Hence, people's signals about their own norm-following behavior can only be trusted if they're sufficiently costly. They must "put their money where their mouth is" by, for example, being prepared to publicly condemn norm violations, as happens in virtue signaling.

Note, however, that this conflict of interests all but disappears whenever people are communicating about *other people's* norm following behavior. People in a group have no interest in deceiving each other about whether third parties break the shared rules.¹⁸ If I happen to catch someone breaking a norm, I have no interest in covering for them. For one thing, I don't benefit from them avoiding the group's disapproval. If anything, this goes against my interests, for in that case, I have to put up with the costs of their non-cooperation. I have to put up with the costs of their cheating, exploiting, or free-riding.

To see this more clearly, consider how social approval or esteem normally operates, particularly in the sort of collective action problems that social norms are meant to solve, as we discussed in Section 1.¹⁹ When people in a group are pursuing some collective goal, they share a rational interest in social approval or esteem being granted to group members based on their contribution to that collective goal. They all want reliable cooperators to be socially rewarded because this increases the group's chances of success (Ridgeway, 2019). On the other hand, each individual wants as much approval or esteem as possible for themselves independently of their level of contribution to the group effort (ibid.). This is because people generally desire esteem, both instrumentally, given its effects on social trust, power, and influence, but also intrinsically.

If this is how our desire for approval or esteem operates in the context of collective action, then there will be a conflict of interest between group members in communicating about their own adherence to a norm. People will generally want to lie about their behavior as a way of securing or maintaining the approval of others. However, there will be no conflict of interest between group members in communicating about the transgressions of third parties. As we just saw, group members share a rational interest in allocating approval or esteem to people based on their contribution to the group's goals. If Jones breaks our shared cooperative norms, for example, both you and I have an interest in him being deprived of social approval or esteem.

Thanks to this coincidence of interests, communication about norm violations in a group can be quite cheap. Under conditions of imperfect monitoring, people can use cheap signals to inform one another about the norm violations committed by third parties. In sum, people's shared desire to allocate esteem to others based on their contribution to joint efforts drives down the cost of communication concerning people's failures to contribute to these efforts.

¹⁸I'm here abstracting away from cases where the communicative parties are somehow related to the norm violators, through ties of family or friendship or whatever.

¹⁹For a general discussion of how the desire for esteem or status operates in the context of collective action, see Ridgeway (2019, chap. 2).

3.8 Social Sanctions as Cheap Signals

The sanctions paradox can be solved if we appreciate the fact social sanctions can serve as signals of the violation of certain shared rules. In this section, I'll argue that social sanctions can operate as signals, and show how this idea solves the sanctions paradox in a way that ties together the different proposed solutions that we reviewed in Sections 3 to 6.

As we have seen, sanctioning behaviors such as criticism, gossip, and ostracism are expressions of social disapproval triggered by the violation of social norms. When I criticize or confront you for cutting in line or littering the park, my actions express my disapproval of your behavior, where this attitude is based on our shared standards concerning proper behavior when standing in line or going on a picnic. Moreover, in the case of social norms, this kind of sanctioning behavior is often public. I display my disapproval in ways that not only you but also other people in the group can see. If you cut in line in front of me, I might loudly voice my complaints about this. And if you run a stop sign, I might honk my horn at you.

Expressions of disapproval like these can serve a communicative role or function in the group. They may serve to indicate both to the transgressor and to other group members that a shared rule or standard has been broken. According to various moral philosophers, reactive attitudes such as blame have a communicative point (Fricker, 2016; Macnamara, 2015; Bennett, 2012). When I blame you for something you did to me, my attitude is a way for me to (non-verbally) communicate to you that, as I see things, you have broken an important norm governing our interaction. Likewise, social sanctions can serve a similar function. They can be a way for people in a group to (verbally or non-verbally) communicate to others that, as things seem to them, an important shared norm has been breached, perhaps in a particularly serious way. In short, social sanctions can operate as signals of wrongdoing. In expressing my disapproval of your conduct, I may be communicating to you and others that our norms have been compromised.

Of course, some people may sanction others precisely because they recognize that social sanctions have a communicative role and wish to communicate to others that a norm has been broken. That is to say, the communicative role of sanctions may be the motivating reason why some people engage in sanctioning. However, sanctions can serve a signaling function even if the sanctioner doesn't intend that they do, or isn't even aware of this fact. Blame plays a communicative role, yet, in blaming you, I need not intend to communicate anything to you. Likewise, sanctions may play a communicative role even if people who impose them don't intend that they do. Here, it seems enough that other group members regard sanctions as reliable signals of wrongdoing, independently of the intentions of the signaler.

In any case, the important point is that, as we saw in the last section, my sanctioning behaviors can operate as honest, reliable signals with minimal associated costs. Of course, my expression of disapproval

already involves some costs in time, effort, and energy. However, my disapproval need not carry any further costs for it to be taken as a credible signal that a norm has been broken. It need not involve the sort of costly display that we observe in the peacock's tail or in the ostentatious grandstanding of the virtue signaler. Often, simply voicing my disapproval will be enough to communicate the relevant information.

We can see these features of social sanctions reflected in Sober and Wilson's example from Section 4. Making a blot of ink upon someone's genealogy is a kind of punishment. As such, it expresses the group's disapproval of the tax evader's actions. At the same time, the blot clearly serves a communicative function. The blot serves as an indication to other people in the group, particularly potential mates, that the person so marked has transgressed an important shared cooperative norm. At the same time, the blot is also an incredibly cheap sanction to implement. Because people in the community have a common interest in tax evaders being punished and chastised, the blot can operate as a reliable signal of tax evasion without involving high costs.

If all this is correct, then we have a complete solution to the sanctions paradox. This view of sanctions entails that there is a substantial asymmetry between the cost of sanctioning others and the cost of being sanctioned by others. The cost of sanctioning others is, simply, the relatively low cost of expressing one's disapproval of their behavior. If sanctions are signals of wrongdoing, however, the cost of being the object of sanctions can be quite steep. In being sanctioned, one comes to be disapproved, not only by the person imposing the sanction but also by other group members. And, as we discussed in Section 5, being the object of this kind of concerted social disapproval can be very costly for creatures like us who care about and seek the approval and esteem of others.

In light of this asymmetry between sanctioning and being sanctioned, it's just not the case that one's motivation to sanction others must necessarily translate into a motivation to follow the relevant norm in the absence of sanctions. In particular, one's motivation regarding social norms might be *Janus-faced*. As potential enforcers of these norms, our motivation to impose sanctions on others might be altruistic, as emphasized by the altruistic punishment literature. It could be that, upon observing a norm being violated, we're motivated to punish or otherwise sanction the transgressor, not because this brings some strategic advantage to us, but merely because we value the norm intrinsically. Hence, we're willing to endure some personal cost to have the norm enforced. On the other hand, as potential norm abiders, our motivation to follow the norm depends on our expectation of being sanctioned if we don't. We don't follow the norm altruistically. Instead, we do so to avoid the loss of approval or esteem concomitant with being caught breaking it. In sum, looking at others, we enforce the norm altruistically; looking at ourselves, we follow the norm out of a fear of sanctions.

This view also clarifies why the process responsible for our tendency to enforce norms altruistically need not also endow us with a tendency to follow norms altruistically. Because sanctions have a low cost, a process of natural selection or cultural evolution may stabilize a practice of enforcing deviations from cooperative norms, even if the same process cannot stabilize a practice of following such norms. Yet, once the enforcement practice is in place, the norm is widely followed because group members have a rational interest in doing so.

Conclusion

Why do we follow social norms? According to a natural and plausible view, we follow norms because we want to avoid the sanctions that come with breaking them.

In groups where the people in charge of sanctioning are the same people who must follow these rules, this natural idea seems to give rise to a paradox. Namely, if people care enough about a norm to bother sanctioning others for breaking it, then it seems they should also want to follow the norm in the absence of sanctions.

We have seen how this paradox can be dealt with. This is a generalization of the esteem view discussed in Section 5. When people care about esteem, they share a common interest in allocating esteem to others based on their contribution to the group's collective efforts. In light of this common interest, expressions of disapproval triggered by norm transgressions can operate as cheap signals that a common, shared norm has been broken. Under these conditions, sanctioning others is much less costly than being sanctioned by others. Hence, one's motivation to sanction others need not translate into a motivation to follow the norm in the absence of sanctions.

The emerging picture is that our motivation concerning social norms is Janus-faced. When we look at other people's behavior, we care about norms intrinsically and are willing to enforce them altruistically, even at a personal cost. When we look at our own behavior, we care about norms instrumentally, in the sense that we see norm-following as a way of securing or maintaining our esteem or status in a group. These two contrasting aspects of our motivation aren't in tension. On the contrary, they explain why we can live peaceful and productive lives based on shared rules that govern our interactions. We follow these mutual covenants because we are afraid of the swords that we ourselves wield.

Chapter 4

Taught Rules: Instruction and the Evolution of Norms

Abstract

Why do we have social norms—of fairness, cooperation, trust, property, or gender? Modern-day Humeans, as I call them, believe these norms are best accounted for in cultural evolutionary terms, as adaptive solutions to recurrent problems of social interaction. In this paper, I discuss a challenge to this “Humean Program.” Social norms involve widespread behaviors, but also distinctive psychological attitudes and dispositions. According to the challenge, Humean accounts of norms leave their psychological side unexplained. They explain, say, why we share equally, but not why we disapprove of those who don’t. I defend the Humean Program against this challenge. In particular, I suggest an idea for how to extend the Program to account for the psychological side of norms. Socially adaptive behaviors aren’t just likely to emerge in a group; They’re also likely to be widely *taught* within it. The transmission of these behaviors through instruction explains why they’re associated with distinctive normative attitudes and dispositions. These attitudes play a pedagogical role in helping transmit these behaviors to children and newcomers.

Introduction¹

Having agreed to go hunting together, the Lion, the Fox, the Jackal, and the Wolf manage to catch a stag. If each of them contributed equally to the hunt, how should they divide up the kill?

¹This paper is accepted for publication in *Philosophical Studies*. Please cite the published version.

For most of us, this isn't a hard question: Into four equal parts, of course! That's what *fairness* requires. Indeed, absent more information about the case, any other division seems unfair. Thus, we resent the Lion when we learn he's taken the lion's share of the kill.

A harder question is why this answer comes so naturally to us. Why do we tend to think that, in this situation and others like it, one *should* share alike? From where comes the idea that doing otherwise is *wrong*?

According to many philosophers, these ideas are best accounted for in *evolutionary* terms. We disapprove of unequal division because we have somehow evolved to think so. But the kind of evolution at stake here isn't biological. It is *cultural* evolution, which means, roughly, a change in belief over time (Alexander, 2007, p. 19).

I call the project of explaining our commitment to fairness and other social norms in cultural evolutionary terms "The Humean Program." The name comes from David Hume's account of the origin of the rules of justice, particularly the institution of property, in Book 3 of *A Treatise of Human Nature*. There, Hume argues that these rules aren't the outcome of an explicit agreement between rational parties. Instead, they arise out of a gradual process of social evolution. As he writes:

Nor is the rule concerning the stability of possession the less deriv'd from human conventions, that it arises gradually, and acquires force by a slow progression, and by our repeated experience of the inconveniences of transgressing it. (2007, sect. 3.2.10)

Modern-day Humeans apply this insight to explain the origin of various social norms. Besides norms of fairness, they have studied norms of cooperation and coordination, trust, property, and gender.² In each case, Humeans aim to offer a naturalistically plausible account of how these norms first emerged.

Yet, despite its promise, the Humean Program faces a significant challenge. As some critics have argued, the program incurs an *explanatory deficit*. Humeans aim to explain our commitment to social norms by showing that certain *behaviors* associated with them are highly likely to emerge and stabilize in a group through cultural evolution. However, explaining why people tend to behave in certain ways isn't yet to explain why they endorse a *norm* to the effect that one should act that way. It doesn't explain, for example, why they consider deviations from such behaviors to be *wrong*, not just unexpected. Insofar as social norms are more than mere behavioral propensities, the Humean Program must be found lacking.

In this paper, I defend the Humean Program against this challenge. In particular, I put forward a promising idea for how to extend the Program to account for the "normativity" of social norms. To be clear,

²For fairness, see Skyrms (2014, chap. 1), Alexander (2007, chap. 5), Zollman (2008), and Binmore (2005). For cooperation and coordinating, see Axelrod (1984), Alexander (2007, chap. 3), Young (1993), and Sugden (2004). For trust, see Skyrms (2014, chap. 3) and Alexander (2007, chap. 4). For property, see Gintis (2007). For gender, see O'Connor (2019).

I don't offer an actual extension; As we will see, doing so will involve substantial theoretical and empirical work.

Put briefly, the idea is this: Socially adaptive behaviors like equal sharing aren't just likely to evolve in a group. They're also likely to be widely *taught* within it. Once they emerge, group members are likely to transmit these behaviors to newcomers by monitoring and correcting how they act. SUGDENut this type of instruction profoundly transforms the nature of these practices. Monitoring and correction imply the type of psychological attitudes and dispositions in virtue of which a behavioral regularity counts as a social norm proper. Given instruction, practices like equal sharing become shared *rules*, not just common behaviors.

The structure of the paper is this. In Sections 1 and 2, I describe the Humean Program, focusing on one specific application of it: SUGDENrian Skyrms' influential account of the origin of our norm of fair division. Next, in Sections 3 and 4, I raise the explanatory challenge and argue against some ways Humeans have responded to it. Finally, in Sections 5 to 8, I develop my idea for how to meet the challenge, focusing again on the case of fairness.

4.1 The Humean Program

The guiding hypothesis of the Humean Program is that social norms are culturally evolved solutions to recurrent problems of social interaction (Alexander, 2007, p. 23; Skyrms and Zollman, 2010, p. 266). Humeans aim to explain the presence of a social norm in a group in terms of that norm's cultural success in promoting socially adaptive behaviors. Confirming this hypothesis for any given social norm N present in group G involves establishing three different claims.

First, Humeans must establish *Recurrence*. They must show that there is a type of social situation, S , such that members of G regularly find themselves in it. In other words, S is a recurrent social situation in the group.³ To describe the central features of S , Humeans usually model it as a *game* in the sense of game theory. A game in this sense refers to a set of two or more agents or "players," each of whom can adopt one of several different action strategies. In a game, each player's outcome or "payoff" depends on the combination of strategies adopted by all the players. Humeans model different types of social situations using different games.

Second, Humeans must establish *Nuisance*. They must show that whenever members of G find themselves in S , they face a social interaction problem, p . In other words, p is a problem inherent to the social situation. Usually, p involves some tension between individual rationality and collective action. It is a situation where

³ S might refer to a specific type of social situation or a broad *class* of social situations. Here, I focus on the former case for simplicity's sake. However, the latter approach promises to make sense of why there is often ambiguity and conflict in the application of social norms. See Skyrms and Zollman (2010, p. 266) for discussion.

each person acting rationally doesn't ensure that they act together successfully. In every such case, group members must find some way of coordinating or cooperating.^{4,5}

Finally, Humeans must establish *Resilience*. They must show that norm N has culturally evolved among the members of G because it allows them to respond to problem p in situation S . In other words, N is a culturally adaptive response to the social interaction problem. To show this, Humeans usually offer an *evolutionary model* of the population of G when they are repeatedly faced with cases of S . At the minimum, such a model must include a representation of the state of the population at any given time, as well as a specification of the dynamical laws describing how that population state changes over time (Alexander, 2007, p. 25). With such a model in hand, Humeans aim to show that norm N —represented in the model as a particular action strategy—is *evolutionarily robust*, meaning that it is very likely to emerge and stabilize in the population given plausible assumptions. Ideally, one wants to show that N emerges as a stable evolutionary equilibrium in the model under a maximally wide range of initial conditions.⁶

Humeans use different types of evolutionary models.⁷ Although some of them admit a biological interpretation, they are meant to be models of cultural evolution. This term means, roughly, a change in cultural traits (i.e., beliefs, behaviors, skills) over time (Alexander, 2007, p. 19). Cultural evolution is possible because some cultural traits are more adaptive than others, allowing people to cope better with social life. Furthermore, adaptive cultural traits are transmitted more often between group members, thus spreading in the population. The transmission of cultural traits relies on different forms of *social learning*, namely, learning where people acquire beliefs or behaviors from others (Laland and Hoppitt, 2013; Henrich and McElreath, 2003). Cultural traits spread in a group because people acquire them from others through mechanisms such as imitating those who are more successful than them.

For Humeans, establishing Recurrence, Nuisance, and Resilience goes a long way toward explaining a social norm's presence in a group. Taken together, these three claims offer an *equilibrium explanation* of the norm (Sober, 1983). This type of explanation is both *less* and *more* informative than a standard causal explanation, where an event (like the emergence of a norm in a group) is explained by citing its cause. Unlike a causal explanation, an equilibrium explanation doesn't specify the causal history of the explanandum. Yet, it makes up for this lack of detail by showing that a specific causal history isn't needed to account for the explanandum's presence in a given setting. Insofar as the explanandum is a robustly emergent equilibrium

⁴In other cases, S may be problematic not because it involves a tension between individual rationality and collective action but because group members lack the required information or willpower to implement the uniquely rational course of action for everyone.

⁵For an accessible introduction to some familiar games used to model problematic social situations, see O'Connor (2022, sect. 4).

⁶My notion of an evolutionarily robust norm or strategy is similar to Sober's (1983) notion of a *global equilibrium* and Alexander's (2000) notion of a *stochastically robust strategy*.

⁷See Alexander (2007, chap. 2) for an overview.

in that setting, it is highly likely to be brought about by some cause or another. In other words, equilibrium explanations present a *disjunction* of causal scenarios (Sober, 1983, p. 204). Thus, they are more informative than explanations that focus on a single causal history.⁸

4.2 Skyrms on Fairness

Moving forward, it will be useful to see how one can apply the Humean Program to a specific case. I'll focus on Brian Skyrms' (2014; 1996) account of a widespread norm of fair division according to which people with equal claims to some good should receive equal shares of it. Skyrms' account of this norm is one of the best developed and most discussed applications of the Humean Program. As such, it will be my focus for the remainder of the paper.

Skyrms' account of fair division starts with Recurrence. He describes a type of social situation that people in any human group are likely to encounter. These are cases where two or more individuals must decide how to divide a good between them on pain of losing it. Moreover, the parties to this interaction are all symmetrically positioned in that all have a similar level of strength, experience, speed, etc. (Kitcher, 1999, p. 223). Imagine, for instance, a group of hunters who, having caught a stag together, must decide how to divide it between them before it goes to waste.⁹

Skyrms introduces a version of the *Nash bargaining game* as a formal model of this simple distribution problem. This game involves two players who must figure out how to divide a chocolate cake between them. They must do so by each independently writing a final claim to a percentage of the cake on a piece of paper and handing it to a referee. If both claims add up to more than 100%, the referee eats the cake. Otherwise, each player gets what they claimed.

Moving to Nuisance, Skyrms notes that players in this divide-the-cake game face an *equilibrium selection problem*. Without communicating, they must somehow agree on a way of dividing up the cake. If they ask for more than 100%, they get nothing. If they ask for less, they fail to get as much cake as they can. The problem is that there are many ways for them to jointly ask for *exactly* 100% of the cake. They could both ask for half, or one could ask for 60% and the other for 40%, etc. Each such combination of strategies is a *Nash equilibrium* of the game, in the sense that each player does as well as they can do given what their partner is doing. The players' problem consists in coordinating to select one among many such equilibria.

⁸Even if Humeans models fall short of explaining how a social norm potentially evolved, in the sense of specifying the conditions under which it likely emerged and stabilized in a group, they might still offer other types of valuable information. For instance, they may offer insight into how the norm possibly evolved or about the minimal conditions required for it to evolve. For this distinction between “how-potentially,” “how-possibly,” and “how-minimally” uses of evolutionary models, see O'Connor (2019, sect. 0.2).

⁹For more contemporary examples of situations with this structure, see Alexander (2007, p. 150).

Turning to Resilience, Skyrms wants to show that in a group regularly facing the divide-the-cake game the strategy of demanding half of the cake is highly likely to evolve. He argues for this claim by offering an evolutionary model. In particular, he models the population according to the *replicator dynamics*. This model assumes that different game strategies (Demand Half, Demand 60%, etc.) are initially present in the population with different frequencies. Moreover, the frequency of any given strategy increases or decreases depending on how the expected fitness of those who follow it compares to the average fitness of the population (Alexander, 2007, p. 28). Here, one's expected fitness is measured in terms of one's expected payoff in the game. So, if a strategy is such that its players can expect to do better in the game than the group average, then the proportion of people playing it increases accordingly.¹⁰

The replicator dynamics can be interpreted in strictly biological terms (Taylor and Jonker, 1978) However, the model also admits a cultural interpretation (Weibull, 1995). Indeed, one can show that when a group uses a particular social learning strategy, their cultural evolution approximates the replicator dynamics (Schlag, 1998). This learning strategy consists in imitating others with a probability proportional to their success in the game. Someone who follows this learning rule randomly selects another player from the population and then compares her own payoff to theirs. If the other person's payoff is higher, the player adopts their strategy with a probability proportional to the payoff difference (Alexander, 2021).

Having introduced this model, Skyrms shows that demanding half of the cake is an evolutionarily robust strategy in the population under the assumption that strategies in the divide-the-cake game are weakly and positively correlated. According to this assumption, people playing the same strategy are slightly more likely to partner up with each other than with people who follow a different strategy.¹¹ There are several ways of defending this assumption in Skyrms' model. Social interactions may happen within the same family or clan or in social networks involving "neighborhoods" of like-minded individuals (Alexander, 2000). In any case, under these conditions, the strategy of demanding 50% of the cake emerges as a stable evolutionary equilibrium under virtually all the initial conditions of the model. Thus, Skyrms concludes that:

¹⁰More formally, the replicator dynamics represents the state of the population at any given time using a state vector $\vec{s} = (s_1, \dots, s_m)$, where s_i denotes the proportion of group members who follow strategy i . The expected fitness of an agent following i ($F(i|\vec{s})$) is their expected payoff, namely, the sum of the payoffs they would obtain playing every other strategy in the population, weighted by their probabilities. The crucial dynamical assumption of the model is that the instantaneous rate of change of s_i is a function of the difference between the expected fitness of agents who play that strategy and the average fitness of the population. In mathematical form:

$$\frac{\partial s_i}{\partial t} = s_i(F(i|\vec{s}) - F(\vec{s}|\vec{s}))$$

In this equation,

$$F(\vec{s}|\vec{s}) = \sum_{i=1}^m s_i F(i|\vec{s}).$$

For how to derive this equation, see Alexander (2007, p. 29).

¹¹In Skyrms' model, correlation works through a function that inflates the likelihood that a strategy meets itself and deflates the likelihood that it meets a different strategy. For details, see Skyrms (1996, chap. 1, fn. 30).

In a finite population,¹² in a finite time, where there is some random element in evolution, some reasonable amount of divisibility of the good and some correlation, we can say that it is likely that something close to share and share alike should evolve in dividing-the-cake situations. This is, perhaps, a beginning of an explanation of the origin of our concept of justice. (1996, p. 21)

4.3 The Explanatory Deficit

Humean accounts of the origin of social norms can fail in several ways. First, regarding Recurrence and Nuisance, any such account might fail to be *representative* (D'Arms et al., 1998, p. 89). It might posit a situation S and a problem p that the group is unlikely to have ever encountered. Hence, norm N can't be explained as an evolved response to situations with that structure. For example, some critics argue that Skyrms' divide-the-cake game is an unrealistic model of distribution problems. Although these problems are frequent in social life, they seldom involve a referee or near-perfect symmetry between the parties (Kitcher, 1999, p. 223; D'Arms et al., 1998, pp. 89–90).¹³

Further, regarding Resilience, a Humean model might fail to be sufficiently *robust* (D'Arms et al., 1998, p. 90). The model might fail to show that norm N is likely to emerge and stabilize in the population across a sufficiently wide range of conditions. For instance, D'Arms et al. (1998) argue that Skyrms' model of fair division is insufficiently robust. As we saw, this model assumes that action strategies in the divide-the-cake game are weakly correlated. Yet, people might rationally want to play this game in a way that introduces anti-correlation between their behaviors. Someone who demands 60% of the cake might seek a partner who asks for 40%, not one who asks for 60%. Assuming anti-correlation, Demand Half is much less likely to evolve in the replicator dynamics.

Even if Humean accounts are representative and robust, they face a deeper problem. Indeed, this problem affects the Humean Program in general, not this or that specific model. As some critics have pointed out, the program seems to incur an *explanatory deficit*. Humeans purport to explain our commitment to social norms (e.g., our norm of fair division) by showing that certain action strategies associated with them (e.g., sharing equally) are evolutionarily robust. However, explaining why people tend to act in a certain way isn't yet to explain why they endorse a *norm* to the effect that one should act that way. Social norms aren't mere behavioral regularities. Hence, fully accounting for them requires more than explaining the behaviors associated with them.¹⁴

¹²Strictly speaking, Skyrms' model posits an infinite population, so this summary of his view seems to be inaccurate in this respect. This point is made by Alexander (2007, p. 159).

¹³For a response to these criticisms of Skyrms' model, see Alexander (2000, p. 150).

¹⁴For different versions of this critique, see Kitcher (1999), D'Arms (2000), and Anderson (2000).

Social norms go beyond mere behavioral regularities in at least two ways. First, when a social norm exists in a group, people adopt a special attitude towards deviations from the regularity in behavior. They consider such deviations to be *wrong*, not just unexpected (Hart, 1961; Anderson, 2000). Second, because people think of deviations this way, they're disposed to punish or otherwise sanction those who deviate (D'Arms, 2000). Consider our norm of fair division. When someone takes more than their fair share of a good—like the Lion in Aesop's fable—we're not only surprised by their behavior. We also think they have acted wrongly. Moreover, we usually resent them and may even call them out on their behavior. In sum, social norms have a *psychological* dimension. They are behavioral regularities undergirded by distinctive attitudes and dispositions.

Humean models of the evolution of norms leave their psychological side unexplained. These models address the evolution of behavior, not psychology (O'Connor, 2022). That is, while they show that certain behaviors are likely to culturally spread in a population on account of being socially adaptive, they're silent about the psychological mechanisms behind them. As far as the models go, the behaviors might be caused in a variety of ways. A behavior like sharing equally *could* be produced by the type of attitudes and dispositions characteristic of social norms, but it could also be caused by something else. It might, for instance, result from a “fast and frugal,” System 1 heuristic, like the rules of thumb commonly used by chess players (Alexander, 2007, pp. 22–23).¹⁵ This insensitivity to the psychology behind culturally evolved behaviors impacts the explanatory power of Humean models. As Justin D'Arms puts the point, “a model that is avowedly insensitive to whether the behavior it explains is (regarded as) [normatively] significant cannot be an explanation of that significance” (2000, p. 299).

One may wonder whether Humeans really incur an explanatory deficit. As we saw, Humean models of the origin of norms are meant to be models of *cultural* evolution. But these models represent interactions between cultural agents, that is, agents whose behavior is often guided by attitudes such as normative beliefs, values, and emotional tendencies. Given this assumption about the kind of agents they apply to, it seems we can interpret Humean models as describing the evolution of behaviors to which people attach certain normative significance from the outset. In short, we can interpret them as models of the evolution of full-fledged norms, not mere regularities of behavior.

However, even if we can attribute normative attitudes and dispositions to the agents who figure in Humean models, the issue is whether these models accord *any* role to such attitudes in the emergence of the relevant practices. The answer would seem to be “no” (D'Arms, 2000). In Humean models, cultural fitness is *exclusively* a function of how agents behave in social interaction problems. Hence, whatever selection pressures these models target, they act primarily on behaviors. Due to these pressures, some behaviors

¹⁵Some Humeans explicitly compare social norms with such heuristics. See Alexander (2007, pp. 22–23).

proliferate, others die out, irrespective of whether group members have normative attitudes towards them or not. What guarantees, then, that they will have such attitudes?¹⁶

Maybe this criticism of the Humean Program is too quick. Perhaps Humean accounts of norms can be extended to account for their psychological dimension. As we saw, Skyrms only claims to offer “a *beginning* of an explanation of the origin of our concept of justice” (1996, p. 21; emphasis added). This claim suggests that he thinks more work is needed to give a complete account of how fairness evolves.

However, the problem with the Humean Program isn’t just that it incurs an explanatory deficit but that it makes it hard to see how this deficit could be bridged. As we saw, Humeans aim to show that strategies like equal sharing are evolutionarily robust. But this means that these behaviors must be highly *stable* in the face of deviations. In Skyrms’ model, for example, any “mutant” or innovator who decided to adopt a strategy other than Demand Half would be driven into extinction pretty quickly. The reason for this is simple: Once Demand Half has taken over the group, playing a different strategy simply doesn’t pay. Anyone who deviates from the regularity is guaranteed to have a lower expected payoff than the population average.¹⁷

Under these circumstances, it is hard to see what would be the *point* of the type of attitudes and dispositions characteristic of social norms (D’Arms, 2000). If cultural evolution already weeds out deviations from the regularity in behavior, why would group members ever come to disapprove of such deviations and be disposed to sanction those who deviate? Having such attitudes and dispositions would be *costly* for them. Blame and disapproval are psychologically taxing (Shoemaker and Vargas, 2021). Imposing sanctions on others takes time and effort (Elster, 1989a; Buchanan, 1975). Yet, these attitudes and dispositions would seem to play no useful role in the group. They’re not needed to sustain the practice.

One could argue that normative attitudes and dispositions play a role in the evolutionary dynamics described by Humean models. According to this idea, if playing a strategy different from Demand Half doesn’t pay in a group, then this is partly because one will suffer other people’s disapproval and sanctions. So, plausibly, part of the reason why cultural evolution weeds out deviations from fairness has to do with these attitudes and dispositions.

¹⁶As critics of the Humean Program point out, this problem need not arise for all evolutionary models of the origin of norms. Some models may accord a relevant role to normative attitudes in the emergence and stabilization of socially adaptive behaviors. See, e.g., Gibbard (1982)

¹⁷Indeed, in the divide-the-cake game, Demand Half is the only *evolutionarily stable strategy* in the sense of Maynard Smith and Price (1973). Roughly, a strategy is evolutionarily stable just in case it is able to withstand invasion by mutants once it has taken over a population. This doesn’t mean, however, that Demand Half is the only evolutionarily stable state in the game. As Skyrms shows, there are “polymorphic” states involving more than one strategy that are evolutionary stable. Part of his account of the evolution of fairness focuses on specifying the conditions under which a population can avoid such “polymorphic traps.” For a generalization of the idea of an evolutionarily stable strategy, see the notion of an *evolutionarily stable set*, i.e., a set of strategies that remains stable in a population even if there is some drift between the different strategies in the set (Thomas, 1984).

However, this just isn't so. In Skyrms' model, strategies other than Demand Half go extinct due to the replicator dynamics. Yet, this dynamics makes absolutely no reference to disapproval or sanctions, only to the agent's expected payoff in the divide-the-cake game. If a strategy isn't copied as much as others, this isn't because people who play it get punished more often, but because they get less cake on average. As far as the model goes, disapproval and sanctions are irrelevant to the cultural evolution of fairness.

In sum, the Humean Program faces a dual challenge. It incurs an explanatory deficit because it fails to explain the psychological dimension of social norms. Moreover, it makes it hard to see how this deficit could be bridged because it seems to leave no space for normative attitudes and dispositions to play a social role or function.

4.4 Sympathy and Resentment

Humeans haven't failed to notice this challenge to their project. Indeed, Hume himself was aware of the need to account for the attitudes of approval and disapproval behind social norms. In the *Treatise*, he distinguishes between two questions we can ask about the origin of justice. First, how do the rules (viz., regularities) of justice get established among us? And second, why do we come to "attribute to the observance or neglect of these rules a moral beauty and deformity"? (2007, sect. 3.2.1). In other words, why do we come to think of acts of justice and injustice in terms of *right* and *wrong*? As we saw, under the Humean Program this second question arises for social norms in general. Humeans have proposed different answers to it.

Hume's own answer appeals to the emotion of *sympathy*. We disapprove of deviations from regularities like equal sharing because we feel for the people who are harmed by them. Once these regularities exist in a group, deviating from them hurts other people's interests. In a population where Demand Half is the statistical norm, a "greedy" mutant who asks for 60% of the cake will disadvantage most people who interact with him. The mutant asks for more cake than is customary, thus causing others to get no cake at all. In Hume's view, group members condemn greedy behavior out of sympathy for those negatively affected by it.

In the latest edition of *Evolution of the Social Contract*, Skyrms suggests a similar explanation for why the practice of equal sharing is likely to become a norm. He writes: "If the equal split is a convention in dividing-the-cake situations, it is no surprise that greedy players should be despised or ostracized, since they spoil things for those with whom they interact" (2014, p. 22).

However, this account doesn't quite work. In divide-the-cake games where one party acts fairly and the other acts greedily, *both* end up empty-handed. They jointly ask for more than 100% of the cake. Yet, we're only inclined to feel sympathy for the fair one. The reason for this, I take it, is that we think she's the one who's been wronged. In other words, sympathy for the fair party presupposes a belief that her interests

aren't just frustrated, but *wrongfully* so. Otherwise, we would feel sympathy for the greedy party, too. But then people who sympathize with others in this type of case must already think that it is wrong to be greedy.

Robert Sugden (1998) proposes a similar view. When we have good reason to expect that others will satisfy our preferences, we tend to *resent* them if they don't. Sugden calls these expectations that trigger resentment *normative expectations*. They shouldn't be confused, he says, with beliefs to the effect that someone should do something.¹⁸ For Sugden, normative expectations are just reasonable beliefs that someone will do something where one prefers that they do that thing. Practices like equal sharing give rise to such expectations. When equal sharing takes over a population, people can reasonably expect that others will partake in this behavior. Moreover, they prefer that they do so because it makes it easier to coordinate with them. Hence, they will resent those who don't share equally because they act against their normative expectations.

Sugden's view is similar to David Lewis's (1969) account of how coordinating conventions can turn into norms. Lewis writes:

...if [other people] see me fail to conform [to a convention], not only have I gone against their expectations; they will probably be in a position to infer that I have knowingly acted contrary to my own preferences, and contrary to their preferences and their reasonable expectations. They will be surprised, and they will tend to explain my conduct discredibly. The poor opinions they form of me, and their reproaches, punishment and distrust are the unfavorable responses I have evoked by my failure to conform to the convention. (1969, p. 99)

This view falls prey to the same problem as the sympathy view. Suppose that in previous years you've always bought me a fancy present for my birthday. Accordingly, I now believe that you'll do the same this year. If you act against my expectation, I may feel frustrated and displeased. Yet, I wouldn't resent you for it unless you also *promised* to get me a fancy gift. In short, to trigger disapproval, expectations require a belief that others are under some obligation to satisfy them. But this means that people who resent others in the divide-the-cake game must already think that they should share equally.

The problem with these two views is that they try to account for the attitudes and dispositions associated with social norms in terms of emotions and emotional responses that presuppose their existence. Hence, they fail to meet the first part of the challenge to the Humean Program. They offer an explanation for the psychological side of norms, but it is viciously circular.

Further, these views also fail to meet the second part of the challenge, viz., that of elucidating the role or function of normative attitudes and dispositions. The views depart from the spirit of the Humean Program.

¹⁸Nor with beliefs that others believe that one should do something, as the term is used by Bicchieri (2006).

The guiding hypothesis of the program, recall, is that social norms are culturally evolved solutions to social problems. According to these accounts, however, disapproval and sanctions don't arise in a group in response to a problem. Instead, they are additions to an independently evolved solution. The idea is that once a behavior like equal sharing evolves, it interacts with people's emotional tendencies in a way that leads them to regard it as normatively significant. But this interaction isn't guided by the type of selection pressures that explain the initial emergence of the behavior. The normativity of the practice "evolves" in the sense of developing gradually, but it isn't *selected for*.

Is there a way for Humeans to discharge their explanatory burden? Can the Humean Program be extended to account for the type of normative attitudes and dispositions behind social norms? In the remainder of the paper, I suggest a possible extension. The key to the extension, I believe, lies in the relationship between cultural evolution and social learning.

As we saw above, cultural evolution is possible because human beings are social learners—we have a capacity to acquire beliefs and behaviors from each other. Indeed, according to various theorists, our capacities for social learning evolved through natural selection precisely because of the role they play in cultural evolution. These capacities allow us to accumulate a body of adaptive knowledge across generations—a *culture*. And having access to this knowledge increases our fitness (Henrich and McElreath, 2003). For example, it gives us access to skills and information that no individual could acquire on her own.

The crucial insight behind the Humean Program is that we should think of norms as *cultural products*—as a part of this body of adaptive knowledge. Norms encode *socially* adaptive information, i.e., information allowing us to cope better with social life. As such, they come to be transmitted between people through social learning. Eventually, they become part of our common lore.

My idea for how to extend the Humean Program has to do with the specific way norms are socially transmitted. As we saw, Humean models like Skyrms' tend to assume that social norms are transmitted through some form of imitation. People learn to share equally because they copy other people who share equally. My suggestion is that social norms are also likely to be transmitted through *instruction*. People learn to share equally, not just because they copy others, but because those others *teach* them how to play fair.

The transmission of behaviors like equal sharing through instruction profoundly transforms their nature. Or so, at least, I wish to suggest. Teaching someone how to play fair involves monitoring and correcting how she acts. But these activities imply normative attitudes and dispositions. These attitudes and dispositions play a *pedagogical role* in helping transmit the practice.

In the next few sections, I sketch how my suggested extension to the Humean Program might look like, focusing on the case of our norm of fair division. My goal isn't to work out all the details. Instead, I want to offer reasons to think that the details *can* be worked out. You may take what follows as a proof of concept.

4.5 Population Replacement

In Humean fashion, I'll start with a version of Recurrence. I'll describe a type of situation that human groups are likely to encounter once they have evolved practices like equal sharing. As I'll argue, groups are prone to facing forces that destabilize the evolutionary equilibria behind these practices

Consider one way Humean models fail to be representative of real human groups. These models usually abstract away from the phenomenon of *population replacement*, namely, the process whereby people in a group are replaced by others over time. Population replacement can happen for several reasons. For example, it can result from *migration*, with some people immigrating to and others emigrating from the group. Or it can happen due to *generational change*, with newer generations gradually succeeding older ones.

Plausibly, most real human groups experience some degree of population replacement over time. People often move in and out of neighborhoods. Families and clans frequently gain and lose members. Yet, Humean models of the evolution of social norms don't usually take these processes into account. For instance, in Skyrms' model of the origin of fairness, group practices change over time because group members update their strategy depending on their expected payoff, not because anyone gets replaced by a different player.

Some Humean models involve a mutation parameter, μ , representing the probability that any given group member switches strategy after each round of play. In principle, this parameter can be used to represent the random substitution of group members by others who play different strategies (Kandori et al., 1993). In practice, however, the parameter is mostly used to model processes like innovation or experimentation with new strategies. Hence, it is assigned a small value. In J. McKenzie Alexander's (2000) agent-based model¹⁹ of the evolution of fairness, for instance, the value of μ is as low as 0.001. This means that in a population of 1,000 people, only about one person will change strategy after each round of the divide-the-cake game. We should expect the effect of population replacement to be greater than this. Migration and generational replacement seem like more significant sources of behavioral variation within a group than innovation or experimentation.²⁰

¹⁹Unlike the replicator dynamics, which models a population using an aggregate state vector, agent-based models represent a population discretely. They include information about each individual group member, such as what strategy they use and how they're spatially or socially positioned. See Alexander (2007, chap. 2) for discussion of the different types of evolutionary models.

²⁰Of course, if we want to use μ to model population replacement, we cannot estimate an adequate value for it simply from the armchair. We need a better understanding of the phenomenon. For example, we need to estimate how often population replacement happens relative to how often group members face the relevant social interaction problem. This already points to

Population replacement can lead to the introduction and elimination of game strategies, thus changing their relative frequency in a group. It should be clear how emigration and deaths lead to the elimination of strategies from a group. But how can immigration and new births introduce game strategies? On the one hand, new arrivals may come from groups that haven't (yet) evolved the relevant practices. Or, upon migrating, they might switch strategies thinking that circumstances in the new group warrant a change. Someone who used to behave fairly might turn greedy because they hope to get more cake than they used to. Or they might switch to being modest, asking for only 40% of the cake, wanting to curry favor with their new peers. Children, on the other hand, are unfamiliar with the type of social situation the practice solves. Hence, they may approach it using strategies that make sense in other circumstances. Faced with the divide-the-cake game, they might use the otherwise sensible strategy of trying to get as much of the good as possible.

By eliminating and introducing game strategies, processes like migration and generational change can affect the stability of evolutionary equilibria. As we saw, according to Humeans, evolutionarily robust strategies like equal sharing are highly likely to emerge and stabilize in a group. Yet, migration and generational replacement can act as external forces that cause perturbations in such equilibria. A population might reach an equilibrium only to be displaced away from it by the inflow and outflow of group members. Being an evolutionarily robust state, the equilibrium is likely to withstand such perturbations. But it might be subject to further perturbations. In the long run, a practice can undergo cycles of stabilization and destabilization.²¹

As an illustration of this phenomenon, consider Figures 4.1 and 4.2, both of which show the evolution of a population of 1,000 agents who play 200 rounds of Skyrms' divide-the-cake game.²² For simplicity, only three strategies are represented in the population: The *Fair* strategy (Demand Half), the *Modest* strategy (Demand 40%), and the *Greedy* strategy (Demand 60%). Initially, every strategy has the same frequency. In each round of this model, agents randomly play the game with others and then update their strategy using the imitative learning rule described in Section 2, namely, they imitate others who have a higher payoff than them with a probability proportional to the difference between their payoffs.

an observation I'll make below about how extending the Humean Program in the direction I suggest will require substantial theoretical and empirical work.

²¹Also, the way population replacement affects the stability of a practice may be influenced by the specific type of social interaction problem the practice evolved to solve. Situations with the structure of a Stag Hunt may be destabilized differently than situations with the structure of a Prisoners' Dilemma.

²²These models were created using Abed-1pop, an agent-based modeling framework developed by Izquierdo et al. (2019). You can download Abed-1pop here: <https://luis-r-izquierdo.github.io/abed-1pop/>. Please e-mail me for a file with the specific parameters I used in my models.



Figure 4.1: Evolution of fairness without population replacement ($N=1000$; $\mu=0.001$; $t=200$)

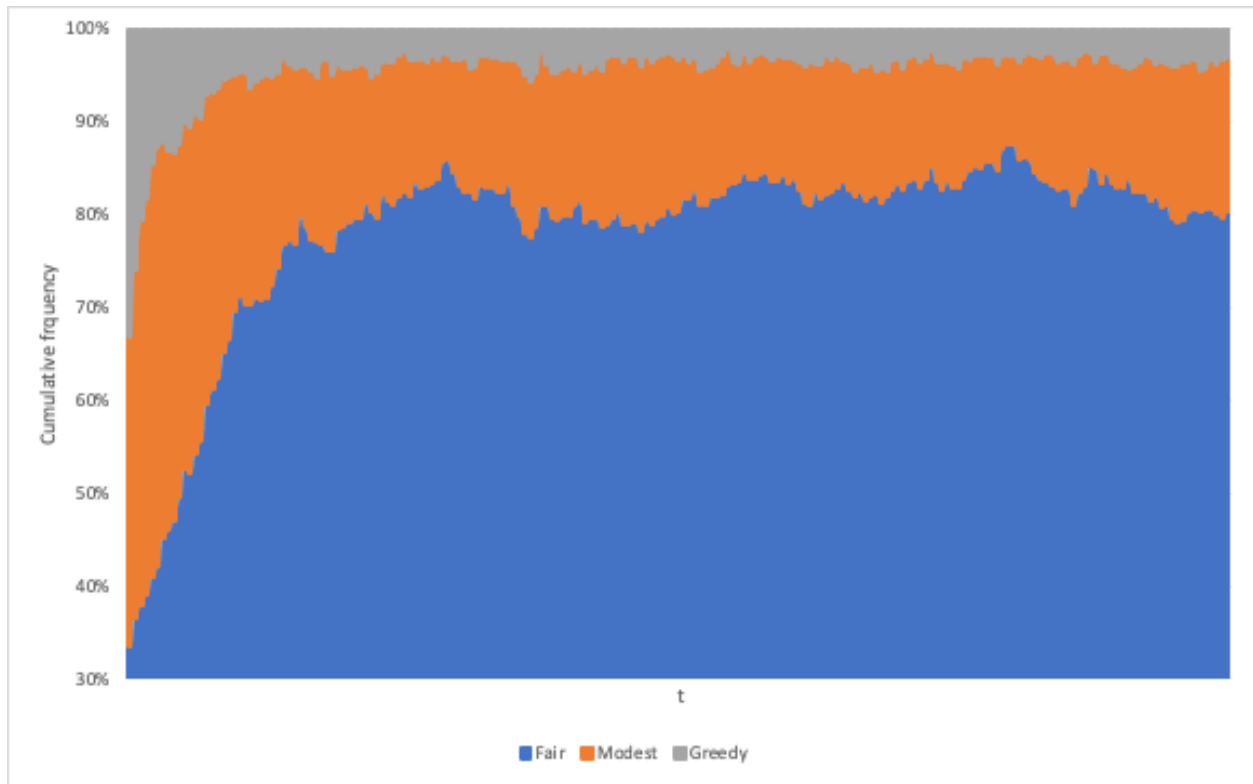


Figure 4.2: Evolution of fairness with population replacement ($N=1000$; $\mu=0.1$; $t=200$)

Figure 1 shows how the population evolves without much population replacement ($\mu = 0.001$).²³ As we see, under these conditions, fairness quickly emerges and stabilizes in the group, following Skyrms' results. The population *converges* to fairness in the sense of Alexander (2000), that is, in the long run, everyone except $N \times \mu$ group members behaves fairly. Figure 2, in contrast, shows how the population behaves under conditions of population replacement. More precisely, 10% of the population is replaced by others with random strategies after each “generation” or round of play ($\mu = 0.1$). Under these conditions, the population never converges to fairness: The proportion of people who behave fairly is always below 90%. Moreover, the population seems to go through cycles of stabilization and destabilization: It starts to converge to the equilibrium only to be displaced away from it later.

4.6 Imitation and Myopia

I'll now move to a version of Nuisance. I'll argue that, given frequent population replacement, groups with practices like equal sharing may face a problem concerning how they readapt to these practices after they have been destabilized. This problem has its roots in the social learning strategy group members use to acquire behaviors from each other.

In the divide-the-cake game, Demand Half is the *optimal* strategy for everyone to adopt in the long run (Alexander, 2000). Given the evolutionary dynamics at play, everyone is better off playing fair: Under other states, fair players can expect to get less than 50% of the cake because they are likely to meet a few greedy players; Modest players can't expect to get more than 40%, come what may; And, as the population approaches fairness, the payoff of greedy players goes well below 50%. Hence, when a population is displaced away from the state of widespread fairness, it's in everyone's best interest to go back to that state as quickly as possible.

But a group may not readapt to fairness as quickly as desirable. The reason comes down to the type of social learning rule people use to update their strategy. As we saw, in Humean models people are usually assumed to use some sort of imitation rule to learn from others. One common rule of this sort consists in imitating others who are more successful than them with a probability proportional to the difference between their payoffs. Yet, this kind of rule can be *myopic* in the following sense: It can lead some of the population to shift towards a strategy whose fitness benefits are only transient. This is because a strategy may have an

²³N.B. In these models, I use parameter μ to represent different rates of population replacement instead of mutation, in line with Kandori et al.'s (1993) suggestion that this parameter can be used to model the random substitution of group members with others who play different strategies, such as immigrants. Of course, random replacement is an idealizing assumption. Very plausibly, what strategies new group members use isn't purely a random matter. For example, the strategies of children are likely influenced by their parents. More sophisticated models may represent population replacement differently. They may, for instance, explicitly represent the flow of individuals between different interacting populations. Here I only want to provide an illustration of the effect that population replacement might have over a practice like equal sharing, so I abstract away from these complications. More realistic models of the phenomenon must take these details into account.

advantage over others simply because of current peculiarities of the population, not because it is the best strategy in the long run (Alexander, 2021). Still, for a while, the strategy gets copied, thus spreading in the group.

Given population replacement, imitative social learning can lead some group members to shift to modesty, even if they'll eventually return to fairness. This process might work as follows.

At first, population replacement introduces new greedy players to the group. This increase in the frequency of greedy players gives modesty a slight advantage over fairness. Modest players do better against greedy players than fair players do. Playing against a greedy player, a modest player gets 40% of the cake. In contrast, a fair player gets no cake. In the short term, some of the population moves towards modesty. For example, if I'm a fair player who happens to encounter a greedy player, I'll fail to get any cake. If I then compare my payoff to that of a modest player, the imitative learning rule will lead me to switch from fairness to modesty.

Eventually, however, greedy players go extinct. Although they do well against modest players, they do very poorly against fair players and other greedy players. Once this happens, fairness becomes the best strategy to play in the game. No matter who they play, fair players get half of the cake while modest players only get 40% of it. So, the population shifts back to fairness.

In brief, once a population is displaced away from fairness due to population replacement, its road back to it can have a "detour" through modesty. Some group members may temporarily switch to modesty because this strategy does better against greediness. But this detour gives rise to a kind of social problem. Because of it, the population readapts to fairness more slowly than it otherwise could, to everyone's disadvantage.²⁴

4.7 Instruction and Natural Selection

Finally, I turn to a version of Resilience. I'll argue that faced with the problem described in the last section, human groups may develop a way of responding to it, namely, they may come to use a different social learning strategy. Instead of acquiring behaviors like sharing equally merely through imitation, human populations may evolve a practice of acquiring them through instruction.

Human groups may respond in several ways to forces that threaten to unravel socially beneficial practices like equal sharing. Indeed, a large literature in the social sciences and game theory studies various mechanisms that communities might use to stabilize cooperation and other socially beneficial arrangements. For example,

²⁴Imitation can also give rise to other problems. For example, the learning rule of imitating the best may favor adopting the strategy of the best agent in one's local neighborhood, who may turn out to have a very poor strategy from the global point of view. Imitation may also be problematic if individuals are prone to making mistakes when copying the actions of others (Castro and Toro, 2014). As I discuss in Section 9, one question that arises under my proposed extension to the Humean Program is what exactly are the limitations of imitative social learning that give rise to a need for instruction.

in response to the proliferation of non-cooperators in a group, group members may respond by directly withholding cooperation from defectors (Axelrod, 1984; Axelrod and Hamilton, 1981; Trivers, 1971), keeping track of people's reputations to better decide who to cooperate with (Nowak and Sigmund, 2005; Leimar and Hammerstein, 2001; Boyd and Richerson, 1989), and developing community enforcement practices, such as various forms of punishment and exclusion (McAdams, 1997; Hirshleifer and Rasmusen, 1989; Fudenberg and Maskin, 1986).

Of course, a group may respond in all these ways to the sort of destabilization of its practices brought about by population replacement. I believe, however, that the problem described in the last section creates a sustained pressure for groups to develop a new and distinct solution. This problem, recall, specifically had to do with how behaviors are socially transmitted in a group. In particular, the problem arises due to the myopic character of social learning through imitation. Given the recurrent nature of this problem, we might expect groups to develop a solution that addresses it at its root by somehow improving how behaviors are socially transmitted. This solution, I want to suggest, is instruction.

As discussed in Section 1, Humean accounts of the origin of norms rely on the idea that people are social learners: They have a capacity for acquiring beliefs and behaviors from each other. As we saw, one important type of social learning is imitation. In it, people acquire a behavior by copying the actions of another person, the “model.” This type of behavior transmission doesn't require the model to do anything to facilitate imitation. Indeed, she need not be aware that she's being emulated. But social learning can also happen through *instruction* (Tomasello et al., 1993). Here, the model of the behavior takes steps to facilitate its transmission to others. She may do things like demonstrating the behavior to them, giving them instructions on how to perform it, and correcting them if they get it wrong.

Under what conditions should we expect a behavior to be taught, rather than merely copied? On the one hand, instruction has various benefits relative to imitation. First, it can increase the fidelity with which behaviors are transmitted. With imitation, there's always a risk that one might get what the model is doing wrong. This risk decreases, however, if that person is involved in the learning process. Moreover, instruction can pass along behaviors that are impossible to acquire otherwise. For example, “negative” behaviors, such as refraining from picking one's nose, can't be copied. One needs to be taught not to engage in them. Finally—and crucially for our purposes—instruction can transmit behaviors more effectively than imitation. Many behaviors are acquired more easily and quickly if they're taught rather than copied. Think, for instance, of how hard it would've been for you to learn how to tie your shoes without anyone's assistance.

On the other hand, instruction also has various costs. In particular, it is costly for the model, who must take actions aimed at handing down the practice. Indeed, instruction is a form of *cooperation*, in that one

individual engages in a costly behavior that benefits others with no immediate benefit to herself (Thornton and Raihani, 2008, p. 1825).

One natural idea is this: Instruction is favored by selection pressures only if the long-term benefits teachers can expect to obtain from teaching outweigh the costs to them of doing so (Thornton and Raihani, 2008, p. 1826). The expected benefits of teaching depend on the difference that teaching makes to how effectively pupils acquire a behavior. In particular, instruction has a high utility only if it increases the likelihood and ease with which pupils acquire the behavior relative to other forms of learning. For example, adult meerkats obtain long-term fitness benefits in terms of kin selection if pups learn how to handle prey effectively. Moreover, pups would have a very hard time picking up this skill through imitation. Hence, in meerkat populations, natural selection likely favors the transmission of prey handling through a form of teaching where adult meerkats let pups handle some of the prey they catch (Thornton and Raihani, 2008, p. 1827).

Now, consider culturally evolved practices like equal sharing. Once these practices evolve, population replacement seems to create a pressure for them to be passed down through instruction, not just imitation. First, group members obtain fitness benefits if newcomers to the group learn how to share equally, especially if they pick up this behavior quickly. In particular, if new arrivals and children learn how to play fair soon after joining the group, then everyone avoids the “detour” through modesty in the road back to fairness. The practice of fairness is destabilized by the inflow of new people, but the group quickly readapts to the optimal equilibrium. Moreover, new group members are unlikely to pick up this behavior quickly enough through imitation. Indeed, imitation is what causes the detour due to its myopic nature. Under these conditions, the benefits group members can expect to obtain from instructing newcomers in the practice may well outweigh the costs they incur in doing so.

But we need to tread carefully here. Upon reflection, there’s an important disanalogy between instruction as it happens in meerkats and how a practice like equal sharing might be taught among humans. In meerkats, natural selection favors the teaching of prey handling because adult meerkats who instruct their pups in this skill have higher expected fitness than those who don’t (plausibly because adults with competent children have a higher expected number of grandchildren). Among humans, however, instructing others in practices like equal sharing seems to be a *public good*. If I teach you how to play fair, then whatever benefits this brings me will also be shared by other group members. If so, then it’s hard to see how natural selection might favor a disposition to teach such behaviors. Teachers incur a cost, but this cost doesn’t give them a comparative advantage over others.

However, in human beings, natural selection may favor instruction through the process of *cultural group selection*. The idea that cultural groups are subject to group selection pressures is emphasized by several

recent theorists of cultural evolution (Richerson et al., 2016; Henrich, 2004; Henrich and McElreath, 2003).²⁵ According to this idea, groups with certain cultural traits are more likely to grow, resolve internal conflicts, overcome resource constraints, win wars, and replace other groups. Because of these factors, such groups are more likely to survive and flourish than others. Given this type of cultural selection acting on groups, natural selection may slowly favor traits that contribute to group success (Richerson et al., 2016, p. 5). The presence of these traits makes it more likely that an individual's group will do well. Hence, it indirectly increases their expected fitness. As we just saw, social learning through instruction seems to be one such trait. It contributes to a group's success by allowing its members to respond to the destabilizing effect of population replacement more effectively. If so, then natural selection might favor a capacity for this type of social learning.

To recap, the picture so far is this: The cultural evolution of practices like equal sharing creates a selection pressure favoring a capacity to transmit these practices through instruction. Natural selection favors social learning via instruction because this capacity allows populations to readapt more effectively to socially beneficial equilibria which have been destabilized by forces like migration and generational change.

This general picture fits well with *culture-gene coevolutionary theory*, one prominent recent approach to explaining the evolution of human sociality (Boyd and Richerson, 2005; Henrich and McElreath, 2003; Feldman and Laland, 1996). According to this theory, many pro-social human traits are best explained as the outcome of two processes working in tandem: cultural and biological evolution. More precisely, the idea is that cultural and biological adaptations co-evolve: Culturally evolved innovations bring about changes in human social environments; These changes give rise to new selection pressures; And these selection pressures shape our minds, making us better equipped for sociality (Kelly and Setman, 2021). My suggestion is that culturally evolved practices like equal sharing produce new selection pressures favoring social learning through teaching and instruction.

4.8 The Emergence of Normativity

But the transmission of equal sharing and similar behaviors through instruction deeply transforms their nature. Instruction brings on stream the type of attitudes and dispositions that distinguish social norms from mere regularities in behavior. As I will argue in this section, instruction involves certain activities on the part of teachers, these activities require them to have certain capacities, and these capacities necessitate normative attitudes and dispositions.

²⁵The idea of cultural group selection shouldn't be confused with the idea that natural selection can act directly on groups. This latter idea is highly controversial in biology, with many theorists rejecting it. See Henrich (2004, sect. 4) for discussion. More recently, some theorists have attempted to rehabilitate the notion that natural selection may sometimes act on groups. See Sober and Wilson (1998) and Okasha (2006).

In order to instruct newcomers to a group in a practice like equal sharing, group members must engage in at least two types of activities. First, in their role as teachers, group members must reliably *monitor* the behavior of newcomers. That is, they must keep track of how new arrivals act in divide-the-cake situations, and they must register whenever they deviate from the established practice of fairness. Such monitoring need not require that group members actively seek out transgressors. But it does require that, whenever they witness or participate in interactions with newcomers, they pay increased attention to their behavior and be alert to deviations.

Second, in cases where they detect deviations on the part of newcomers, group members must respond by taking appropriate measures. In particular, they must attempt to *correct* their pupils' behavior, in the sense of intervening to make it less likely that they will deviate again from the practice in the future. Such correction may involve various kinds of intervention. Among other things, group members can indicate that there has been a deviation, impose appropriate penalties, and offer guidance or advice.

The activities of monitoring and correction require group members to have a host of different capacities. On the one hand, monitoring plausibly requires being able to identify who the newcomers are, determine how they behave in the relevant social situations, and keep a record of their past and current behavior. Correction, on the other hand, plausibly requires being able to determine what interventions are appropriate in each particular case, track their effect on pupils, and make adjustments to them as needed.

Besides these general capacities, however, monitoring and correction crucially depend on some other abilities. By focusing on these abilities, we can appreciate how instruction gives rise to normativity.

In particular, monitoring and correction presuppose two types of capacity. Monitoring requires a capacity to make comparisons between *observed* and *expected* behaviors. That is, when engaging in monitoring, group members must be able to compare how newcomers actually behave in divide-the-cake situations with how they are *supposed* to behave in such situations. Otherwise, it wouldn't be possible for them to detect deviations from the practice. Correction, in turn, requires a capacity to modify one's own behavior should the behavior one observes in others be *unexpected*. That is, having detected a deviation, group members must be able to reorient their behavior towards the transgressor. Otherwise, it wouldn't be possible for them to administer the necessary correctives.

However, to have these capacities, group members must have attitudes and dispositions of a distinctively normative kind. Or so, at least, I wish to argue. First, in order to make comparisons between observed and expected behaviors, teachers must have *expectations* regarding how people in the group are supposed to behave in a certain class of situations. These expectations can't simply be beliefs about what others are likely to do in the future. They must be *normative* expectations, in the sense of expectations the frustration

of which one regards as a kind of failure that calls for an appropriate response.²⁶ Further, in order to modify their behavior to respond to such failures, teachers must have various behavioral dispositions. They must be disposed to intervene in relevant ways whenever their pupils don't behave as expected.

As we saw in Section 3, when group members have these types of attitudes and dispositions concerning a practice, this practice isn't a mere regularity of behavior. Instead, group members adopt a special attitude towards it: They effectively treat the practice as a *rule*, that is, as a standard of correct behavior against which the actions of others are to be measured and deviations from which call for some sort of response (Hart, 1961). In teaching newcomers the practice, group members consider deviations from it to be wrong, not just unexpected, and they're prepared to punish or otherwise sanction those who deviate.

This line of argument may raise two worries. First, one could think that teaching a behavior like equal sharing need not involve monitoring and correction. After all, not all forms of teaching involve such activities. Among meerkats, as we saw, adults train pups in how to handle prey simply by letting them play with it. And, among human beings, some forms of teaching rely on "social tolerance" and "opportunity provisioning" (Kline, 2017). Here, teachers don't monitor and correct how pupils act; They simply tolerate them being around and offer them opportunities to learn that they may otherwise not have.

However, these non-corrective forms of teaching might not be sufficient to transmit behaviors like equal sharing in the way required to avoid the detour through modesty described in Section 6. Social tolerance and opportunity provisioning would work by offering pupils increased chances to observe and imitate experts in divide-the-cake situations. Recall, however, that in these situations imitation can be myopic, in the sense of leading pupils to acquire behaviors whose fitness benefits are only transient. What is needed is a form of teaching where sub-optimal behaviors are quickly identified and purged. This form of teaching, I think, will likely involve activities of monitoring and correction.

Second, one could argue that, while monitoring and correction do require certain normative attitudes and dispositions, these need not be of the kind that characterizes social norms. These attitudes may be purely *prudential*, in the sense of reflecting ideas about what behaviors are most beneficial or expedient. Teachers may monitor and correct newcomers merely because they think that they must share equally if they hope to do well in the group. They need not think that pupils are under a non-prudential requirement to be fair. In contrast, we think of practices like fair division as involving non-prudential requirements, in the sense that if people don't act fairly, then they're on the hook independently of whether being fair advances their interests.

²⁶N.B. These expectations are different from what Sugden (1998), as we saw, calls normative expectations. For him, acting against normative expectations is seen as a failure because it frustrates one's personal preferences, not because it indicates lack of competence in a practice.

Note, however, that instruction is likely to be unsuccessful if it relies merely on prudential normativity. As we saw, correction involves imposing sanctions on newcomers when they deviate from the established practice. Yet, prudential normativity is not a good basis for sanctions. Someone who acts against a prudential requirement may always claim that this requirement no longer applies to her because her interests have changed. To be effective, instruction must be associated with a non-prudential form of normativity. Group members must think that not being fair is wrong, but not simply in the sense of being imprudent. They may, for instance, consider it wrong because it goes against local custom: It's at odds with "how we do things around here" (Brennan et al., 2013; Pettit, 2019).

In sum, effective social learning through instruction involves the type of attitudes and dispositions characteristic of social norms. These attitudes play a *pedagogical* role in helping us teach practices like equal sharing to others. If natural selection favors a capacity to transmit some behaviors through instruction, then it also likely favors a capacity to hold such attitudes and dispositions. Evolution likely shaped our minds to make us capable of treating some behaviors as rules that people must follow, and be prepared to enforce such rules with sanctions. And this capacity may be the basis for our *norm psychology*, that is, the host of adaptations allowing us to "do norms" by detecting, complying, and enforcing them (Kelly and Setman, 2021).

4.9 The Extended Humean Program

So far, I have only offered the sketch of an extension to the Humean Program in response to the explanatory deficit challenge. In this section, I'll discuss how the sketch might be completed. The question is: How can my proposed extension be developed into a serious candidate explanation of the normativity of social norms?

Consider my two central claims:

1. Socially adaptive behaviors, such as equal sharing, aren't just likely to emerge in a group through imitation; They're also likely to stabilize in it through instruction.
2. Instruction brings on stream the type of normative attitudes and dispositions that are characteristic of social norms and distinguish them from mere regularities of behavior.

Developing my proposed extension to the Humean Program will require substantiating these two claims. How can this be done? On the one hand, supporting the first claim will require gaining a better understanding of the evolution of teaching. We need to determine under what conditions groups are likely to develop a practice of teaching certain behaviors. And we need to show that these conditions are met in the case of socially adaptive behaviors like equal sharing. On the other hand, supporting the second claim will require gaining

a better understanding of the relationship between instruction and normativity/normative psychology. We need to produce evidence to the effect that these phenomena are, indeed, intimately connected.

The evolution of teaching can be studied through modeling work. Evolutionary models can offer insight into the conditions under which teaching is favored by selective pressures. What sort of models could these be? One approach is suggested by work on the evolution of social learning more generally. Several theorists use evolutionary game theoretic models to study the conditions under which evolution favors social learning over individual learning or innate mechanisms when it comes to acquiring adaptive behaviors (Wakano and Aoki, 2006; Kameda and Nakanishi, 2003; Rogers, 1988). One big lesson from this body of work is that evolution favors social learning when the environment where organisms learn changes at a moderate rate. When the environment changes slowly, social learning isn't worth it: Other learning strategies are equally effective for acquiring adaptive behaviors while being less costly. When the environment changes quickly, social learning isn't useful: There's not enough time for adaptive knowledge to accumulate in the population.

The evolution of teaching could be studied using a similar framework. For example, we might consider models where teaching is compared to imitation across different rates of population replacement. (The idea is that population replacement causes changes in the *social* learning environment individuals face). Intuitively, we should expect these models to throw results similar to those obtained in social learning models. When the rate of population replacement is low, teaching wouldn't seem to be worth it: The situation is like that of a group facing the odd mutant or innovator who quickly goes extinct. When the rate of population replacement is high, teaching wouldn't seem to be useful—indeed, it may even be prejudicial: The people doing the teaching are unlikely to possess adaptive social knowledge.

Evolutionary models can also help us better characterize the problems associated with imitative learning that drive the evolution of instruction. In this paper, I have focused on one such problem: the myopic character of imitative learning rules. But imitation may give rise to other problems, too. Recent work in evolutionary biology studies the conditions under which various problems associated with imitation might create a selection pressure in favor of teaching and instruction. For example, imitation may be problematic when the behaviors that need to be transmitted in a group are complex and difficult to acquire (Fogarty et al., 2011) or when pupils are likely to make mistakes in copying the actions of models (Castro and Toro, 2014). My proposed extension to the Humean Program is compatible with this work and may draw from it to better describe the pitfalls of imitation which give rise to a need for teaching as a new, improved mechanism of cultural transmission.

At any rate, once modeling work allows us to spell out the conditions favoring the emergence of teaching, we need to show that these conditions are likely to be met in the case of socially adaptive behaviors like

equal sharing. Ideally, we would want to show that, just as such behaviors are likely to emerge across a wide range of conditions, so they are likely to be taught across a similarly wide range of circumstances.

Finally, the relationship between teaching or instruction and normativity might be best explored through empirical research. In particular, my proposal predicts various correlations between our normative tendencies and our pedagogical practices. One of my claims is that normative attitudes and dispositions emerge to serve a pedagogical role. If I'm right, then we should observe correlations between the utility of teaching some behaviors and our tendency to adopt normative attitudes towards them. That is, the normativity of a behavior should be roughly proportional to the utility of teaching it.

This prediction can be tested empirically. For example, the “altruistic punishment” paradigm (Fehr and Gächter, 2002) could be used to determine if people's tendency to punish or impose sanctions on others co-varies with whether these others are likely to modify their ways as a result of being punished. In this experimental paradigm, players in a social dilemma have the choice of punishing others at some cost to themselves. Studies using this paradigm have shown that people tend to punish non-cooperators even when doing so is highly costly to them and can have no benefits in terms of reciprocity or reputation (that is, the punishment is “altruistic”) (see Van Lange et al., 2014, pp. 6–8) . One could test if this tendency varies as a function of whether non-cooperators are “corrigible,” in the sense of being capable of learning to cooperate. If my proposal is on the right track, then we should observe less punishment when non-cooperators are perceived to be less corrigible.

Conclusion

Let me briefly summarize our discussion so far. Modern-day Humeans, as I have called them, believe that social norms, like our norm of fair division, are best accounted for as culturally evolved solutions to recurrent problems of social interaction. This explanatory project, however, faces a challenge. The challenge is showing why culturally evolved solutions to social problems should take a distinctive shape—why they should be *norms*, as opposed to mere behavioral regularities.

In this paper, I have suggested a way for Humeans to meet this challenge. Norms, as Humeans claim, encode socially adaptive information. This is why they're likely to emerge and stabilize in a group through a form of social learning that relies on imitation. But a group's capacity to access and use this socially adaptive information is compromised by forces like migration and generational change. Hence, there's a pressure for norms to be transmitted through instruction, not only imitation. Normativity emerges in connection to such instruction. Normative attitudes and dispositions play a pedagogical role in the context of activities like guiding and correcting others' behavior.

I believe my suggestion offers a promising way of extending the Humean Program to account for the psychological side of norms. But the suggestion has to be made good. In arguing for versions of their view, Humeans normally advocate for a combination of theoretical and empirical work. In particular, abstract, formal models like Skyrms' replicator dynamics help us formulate the precise conditions under which a practice like equal sharing is likely to emerge and stabilize in a group. In turn, empirical work helps us determine whether these conditions actually obtain in the relevant social groups. Likewise, extending the Humean Program in my suggested direction will require a similar combination of theoretical and empirical contributions. We need to develop abstract, formal models allowing us to formulate the precise conditions under which a practice like equal sharing is likely to be transmitted through instruction. And, in addition to this, we need empirical evidence that helps us determine whether these conditions are representative of real human populations and whether instruction is generally correlated with normativity.

In this paper, I have tried to offer reasons for thinking that the extension might be achieved. Whether we can successfully do so, however, remains to be seen. Here, as with many other questions, the proof is in the pudding.

Bibliography

- Ahmed, A. (2014). *Evidence, decision, and causality*. Cambridge University Press.
- Akerlof, G. (1982). Labor contracts as partial gift exchange. *The Quarterly Journal of Economics*, 97(4), 543–569.
- Alexander, J. M. (2000). Evolutionary explanations of distributive justice. *Philosophy of Science*, 67(3), 490–516.
- Alexander, J. M. (2007). *The structural evolution of morality*. Cambridge University Press.
- Alexander, J. M. (2021). Evolutionary game theory. In E. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/sum2021/entries/game-evolutionary/>
- Altman, M. (2012). *Behavioral economics for dummies* (1st edition). Wiley.
- Anderson, E. (2000). Beyond homo economicus: New developments in theories of social norms. *Philosophy & Public Affairs*, 29(2), 170–200.
- Anscombe, G. E. M. (1957). *Intention*. Blackwell.
- Armenta, B. M., Stroebe, K., Scheibe, S., Van Yperen, N., Stegeman, A., & Postmes, T. (2017). Permeability of group boundaries: Development of the concept and a scale. *Personality and Social Psychology Bulletin*, 43(3), 418–433.
- Asarnow, S., & Taylor, D. (2022). Judgment internalism. In M. Vargas & J. Doris (Eds.), *The Oxford handbook of moral psychology* (pp. 139–166). Oxford University Press.
- Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.
- Axelrod, R., & Hamilton, W. (1981). The evolution of cooperation. *Science*, (4489), 1390–1396.
- Bennett, C. (2012). The expressive function of blame. In J. Coates & N. Tognazzini (Eds.), *Blame: Its nature and norms* (pp. 66–83). Oxford University Press.
- Bermúdez, J. L. (2009). *Decision theory and rationality*. Oxford University Press.
- Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

- Bicchieri, C. (2014). Norms, conventions, and the power of expectations. In N. Cartwright & E. Montuschi (Eds.), *Philosophy of social science: A new introduction* (First edition., pp. 208–229). Oxford University Press.
- Bicchieri, C. (2017). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.
- Bicchieri, C., & Mercier, H. (2014). Norms and beliefs: How change occurs. In M. Xenitidou & B. Edmonds (Eds.), *The complexity of social norms* (pp. 37–54). Springer International Publishing.
- Bicchieri, C., Muldoon, R., & Sontuoso, A. (2018). Social norms. In E. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2018). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2018/entries/social-norms/>
- Bicchieri, C., Xiao, E., & Muldoon, R. (2011). Trustworthiness is a social norm, but trusting is not. *Politics, Philosophy & Economics*, 10(2), 170–187.
- Binmore, K. (2005). *Natural justice*. Oxford University Press.
- Binmore, K. (2010). Social norms or social preferences? *Mind & Society*, 9(2), 139–157.
- Bird, R., & Smith, E. (2005). Signaling theory, strategic interaction, and symbolic capital. *Current Anthropology*, 46(2), 221–248.
- Boghossian, P. (2008). Epistemic rules. *The Journal of Philosophy*, 105(9), 472–500.
- Boghossian, P. (2015). Rules, norms and principles: A conceptual framework. In M. Araszkiewicz, P. Banaś, T. Gizbert-Studnicki, & K. Pleszka (Eds.), *Problems of normativity, rules and rule-following* (pp. 3–11). Springer Cham.
- Boyd, R., Gintis, H., Bowles, S., & Richerson, P. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, 100(6), 3531–3535.
- Boyd, R., & Richerson, P. (1989). The evolution of indirect reciprocity. *Social Networks*, 11(3), 213–236.
- Boyd, R., & Richerson, P. (2005). *Not by genes alone: How culture transformed human evolution*. University of Chicago Press.
- Bratman, M. (1987). *Intention, plans, and practical reason*. Harvard University Press.
- Bratman, M. (2014). Temptation and the agent’s standpoint. *Inquiry*, 57(3), 293–310.
- Brennan, G., Eriksson, L., Goodin, R., & Southwood, N. (2013). *Explaining norms*. Oxford University Press.
- Brennan, G., & Pettit, P. (2000). The hidden economy of esteem. *Economics & Philosophy*, 16(1), 77–98.
- Brennan, G., & Pettit, P. (2004). *The economy of esteem: An essay on civil and political society*. Oxford University Press.
- Buchanan, J. (1975). *The limits of liberty: Between anarchy and Leviathan*. University of Chicago Press.
- Buss, S. (1999). Appearing respectful: The moral significance of manners. *Ethics*, 109(4), 795–826.

- Camerer, C., & Loewenstein, G. (2004). Behavioral economics: Past, present, future. In C. Camerer, G. Loewenstein, & M. Rabin (Eds.), *Advances in behavioral economics* (pp. 3–51). Princeton University Press.
- Carnap, R. (1962). *Logical foundations of probability* (2d ed.). University of Chicago Press.
- Castro, L., & Toro, M. (2014). Cumulative cultural evolution: The role of teaching. *Journal of Theoretical Biology*, 347, 74–83.
- Cavell, M. (1993). *The psychoanalytic mind: From Freud to philosophy*. Harvard University Press.
- Coleman, J. (1994). *Foundations of social theory*. Harvard University Press.
- Cubitt, R., & Sugden, R. (2003). Common knowledge, salience, and convention: A reconstruction of David Lewis’ game theory. *Economics & Philosophy*, 19(2), 175–210.
- D’Arms, J. (2000). When evolutionary game theory explains morality, what does it explain? *Journal of Consciousness Studies*, 7(1), 296–299.
- D’Arms, J., Batterman, R., & Gorny, K. (1998). Game theoretic explanations and the evolution of justice. *Philosophy of Science*, 65(1), 76–102.
- Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, 60(23), 685–700.
- Davidson, D. (1971). Agency. In R. Binkley, R. Branaugh, & A. Marras (Eds.), *Agent, action, and reason* (pp. 1–37). University of Toronto Press.
- Davidson, D. (1980a). Freedom to act. In *Essays on actions and events* (pp. 63–81). Oxford University Press.
- Davidson, D. (1980b). How is weakness of will possible? In *Essays on actions and events* (pp. 21–43). Oxford University Press.
- Durkheim, É. (1964). *The rules of sociological method*. Free Press.
- Ekeland, I. (2002). Equilibrium: In the balance. *Nature*, 417(6887), 385–385.
- Elster, J. (1989a). *The cement of society: A survey of social order*. Cambridge University Press.
- Elster, J. (1989b). *Nuts and bolts for the social sciences*. Cambridge University Press.
- Elster, J. (1990). Norms of revenge. *Ethics*, 100(4), 862–885.
- Elster, J. (2015). *Explaining social behavior: More nuts and bolts for the social sciences*. Cambridge University Press.
- Elster, J. (2017). Norms. In P. Hedström & P. Bearman (Eds.), *The Oxford handbook of analytical sociology* (pp. 195–217). Oxford University Press.
- Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868), 137–140.
- Feinberg, J. (1965). The expressive function of punishment. *The Monist*, 49(3), 397–423.
- Feldman, M., & Laland, K. (1996). Gene-culture coevolutionary theory. *Trends in Ecology & Evolution*, 11(11), 453–457.

- Finlay, S. (2019). Defining normativity. In T. Kevin, P. David, & S. Scott (Eds.), *Dimensions of normativity: New essays on metaethics and jurisprudence* (pp. 62–104). Oxford University Press.
- Fogarty, L., Strimling, P., & Laland, K. (2011). The evolution of teaching. *Evolution*, 65(10), 2760–2770.
- Foot, P. (1972). Morality as a system of hypothetical imperatives. *Philosophical Review*, 81(3), 305–316.
- Fricke, M. (2016). What’s the point of blame? a paradigm based explanation. *Noûs*, 50(1), 165–183.
- Fudenberg, D., & Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3), 533–554.
- Gauthier, D. (1986). *Morals by agreement*. Oxford University Press.
- Gibbard, A. (1982). Human evolution and the sense of justice. *Midwest Studies in Philosophy*, 7(1), 31–46.
- Gilbert, M. (1992). *On social facts*. Princeton University Press.
- Gintis, H. (2007). The evolution of private property. *Journal of Economic Behavior & Organization*, 64(1), 1–16.
- Gintis, H. (2010). Social norms as choreography. *Politics, Philosophy & Economics*, 9(3), 251–264.
- Guala, F. (2007). The philosophy of social science: Metaphysical and empirical. *Philosophy Compass*, 2(6), 954–980.
- Hardin, R., & Cullity, G. (2020). The free rider problem. In E. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2020). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2020/entries/free-rider/>
- Hart, H. L. A. (1961). *The concept of law*. Oxford University Press.
- Heath, J. (2008). *Following the rules: Practical reasoning and deontic constraint*. Oxford University Press.
- Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization*, 53(1), 3–35.
- Henrich, J. (2015). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.
- Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q., & Tracer, D. (2005). “Economic man” in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28(6), 795–815.
- Henrich, J., & Ensminger, J. (2014). Theoretical foundations: The coevolution of social norms, intrinsic motivation, markets, and the institutions of complex societies. In J. Ensminger & J. Henrich (Eds.), *Experimenting with social norms: Fairness and punishment in cross-cultural perspective* (pp. 19–44). Russell Sage Foundation.

- Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, 12(3), 123–135.
- Hirshleifer, D., & Rasmusen, E. (1989). Cooperation in a repeated prisoners' dilemma with ostracism. *Journal of Economic Behavior & Organization*, 12(1), 87–106.
- Hooker, B. (2023). Rule consequentialism. In E. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Spring 2023). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2023/entries/consequentialism-rule/>
- Horne, C. (2007). Explaining norm enforcement. *Rationality and Society*, 19(2), 139–170.
- Horne, C., & Mollborn, S. (2020). Norms: An integrated framework. *Annual Review of Sociology*, 46(1), 467–487.
- Hume, D. (2007). *A treatise of human nature: A critical edition*. Oxford University Press.
- Izquierdo, L., Izquierdo, S., & Sandholm, W. (2019). An introduction to ABED: Agent-based simulation of evolutionary game dynamics. *Games and Economic Behavior*, 118, 434–462.
- Joyce, J. (1999). *The foundations of causal decision theory*. Cambridge University Press.
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus; Giroux.
- Kameda, T., & Nakanishi, D. (2003). Does social/cultural learning increase human adaptability?: Rogers's question revisited. *Evolution and Human Behavior*, 24(4), 242–260.
- Kandori, M., Mailath, G., & Rob, R. (1993). Learning, mutation, and long run equilibria in games. *Econometrica*, 61(1), 29–56.
- Kaplan, J. (n.d.). A new problem for rules. *Philosophy and Phenomenological Research*.
- Kelly, D. (2020). Internalized norms and intrinsic motivations: Are normative motivations psychologically primitive? *Emotion Researcher*, 1, 36–45.
- Kelly, D., & Davis, T. (2018). Social norms and human normative psychology. *Social Philosophy and Policy*, 35(1), 54–76.
- Kelly, D., & Setman, S. (2021). The psychology of normative cognition. In E. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2021). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2021/entries/psychology-normative-cognition/>
- Kitcher, P. (1999). Games social animals play: Commentary on Brian Skyrms's Evolution of the social contract. *Philosophy and Phenomenological Research*, 59(1), 221–228.
- Kline, M. A. (2017). Teach: An ethogram-based method to observe and record teaching behavior. *Field Methods*, 29(3), 205–220.
- Knack, S. (1992). Civic norms, social sanctions, and voter turnout. *Rationality and Society*, 4(2), 133–156.

- Kripke, S. (1982). *Wittgenstein on rules and private language: An elementary exposition*. Harvard University Press.
- Laland, K., & Hoppitt, W. (2013). *Social learning: An introduction to mechanisms, methods, and models*. Princeton University Press.
- Lederman, H. (2018). Uncommon knowledge. *Mind*, 127(508), 1069–1105.
- Leimar, O., & Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society B: Biological Sciences*, 268(1468), 745–753.
- Lewis, D. (1969). *Convention: A philosophical study*. Wiley-Blackwell.
- Lewis, D. (1975). Languages and language. In K. Gunderson (Ed.), *Minnesota studies in the philosophy of science* (pp. 3–35). University of Minnesota Press.
- Macnamara, C. (2015). Reactive attitudes as communicative entities. *Philosophy and Phenomenological Research*, 90(3), 546–569.
- Mailath, G., & Samuelson, L. (2006). *Repeated games and reputations: Long-run relationships*. Oxford University Press.
- Maynard Smith, J., & Price, G. R. (1973). The logic of animal conflict. *Nature*, 246(5427), 15–18.
- Maynard Smith, J. (1991). Honest signalling: The Philip Sidney game. *Animal Behaviour*, 42(6), 1034–1035.
- McAdams, R. (1997). The origin, development, and regulation of norms. *Michigan Law Review*, 96(2), 338–433.
- McClennen, E. (1997). Pragmatic rationality and rules. *Philosophy and Public Affairs*, 26(3), 210–258.
- McClennen, E. (2004). The rationality of being guided by rules. In A. R. Mele & P. Rawling (Eds.), *The Oxford handbook of rationality*. Oxford University Press.
- McPherson, T. (2011). Against quietist normative realism. *Philosophical Studies*, 154(2), 223–240.
- Nguyen, C. T. (2019). Games and the art of agency. *Philosophical Review*, 128(4), 423–462.
- Nisbett, R., & Cohen, D. (1996). *Culture of honor: The psychology of violence in the South*. Westview Press.
- Nolan, J. (2017). Social norms and their enforcement. In S. Harkins, K. Williams, & J. Burger (Eds.), *The Oxford handbook of social influence* (pp. 147–164). Oxford University Press.
- Nowak, M., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437(7063), 1291–1298.
- O'Connor, C. (2019). *The origins of unfairness: Social categories and cultural evolution*. Oxford University Press.
- O'Connor, C. (2022). Methods, models, and the evolution of moral psychology. In M. Vargas & J. M. Doris (Eds.), *The Oxford handbook of moral psychology* (pp. 441–464). Oxford University Press.
- Okasha, S. (2006). *Evolution and the levels of selection*. Clarendon.
- Parsons, T. (1951). *The social system*. Free Press.

- Paternotte, C., & Grose, J. (2013). Social norms and game theory: Harmony or discord? *The British Journal for the Philosophy of Science*, 64(3), 551–587.
- Pettit, P. (1990). Virtus normativa: Rational choice perspectives. *Ethics*, 100(4), 725–755.
- Pettit, P. (1996). Functional explanation and virtual selection. *British Journal for the Philosophy of Science*, 47(2), 291–302.
- Pettit, P. (2019). Social norms and the internal point of view: An elaboration of hart’s genealogy of law. *Oxford Journal of Legal Studies*, 39(2), 229–258.
- Pettit, P. (2023). *The state*. Princeton University Press.
- Plunkett, D. (2019). Robust normativity, morality, and legal positivism. In D. Plunkett, S. Shapiro, & K. Toh (Eds.), *Dimensions of normativity: New essays on metaethics and jurisprudence* (pp. 105–136). Oxford University Press.
- Prentice, D., & Miller, D. (1993). Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology*, 64(2), 243–256.
- Rajczi, A. (2016). On the incoherence objection to rule-utilitarianism. *Ethical Theory and Moral Practice*, 19(4), 857–876.
- Rawls, J. (1955). Two concepts of rules. *The Philosophical Review*, 64(1), 3–32.
- Raz, J. (1999). *Practical reason and norms* ([New ed.]). Oxford University Press.
- Richerson, P., Baldini, R., Bell, A. V., Demps, K., Frost, K., Hillis, V., Mathew, S., Newton, E. K., Naar, N., Newson, L., Ross, C., Smaldino, P. E., Waring, T. M., & Zefferman, M. (2016). Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. *Behavioral and Brain Sciences*, 39, e30.
- Ridgeway, C. (2019). *Status: Why is it everywhere? why does it matter?* The Russell Sage Foundation.
- Rogers, A. (1988). Does biology constrain culture? *American Anthropologist*, 90(4), 819–831.
- Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*. Oxford University Press.
- Rogoff, B. (2003). *The cultural nature of human development*. Oxford University Press.
- Schelling, T. (1980). *The strategy of conflict* ([Rev. ed.]). Harvard University Press.
- Schlag, K. (1998). Why imitate, and if so, how?: A boundedly rational approach to multi-armed bandits. *Journal of Economic Theory*, 78(1), 130–156.
- Shapiro, S. (2001). Judicial can’t. *Noûs*, 35, 530–557.
- Shapiro, S. (2005). The rationality of rule-guided behavior: A statement of the problem. *San Diego Law Review*, 42(1), 55–60.
- Shapiro, S. (2006). What is the internal point of view? *Fordham Law Review*, 75(3), 1157–1170.

- Shapiro, S. (2008). How rules affect practical reasoning. In B. Verbeek (Ed.), *Reasons and intentions*. Routledge.
- Shoemaker, D., & Vargas, M. (2021). Moral torch fishing: A signaling theory of blame. *Noûs*, 55(3), 581–602.
- Sidgwick, H. (1874). *The methods of ethics*. Thoemmes Press.
- Simon, H. (1957). *Models of man*. Wiley.
- Skyrms, B. (1996). *Evolution of the social contract*. Cambridge University Press.
- Skyrms, B. (2014). *Evolution of the social contract* (Second edition.). Cambridge University Press.
- Skyrms, B., & Zollman, K. (2010). Evolutionary considerations in the framing of social norms. *Politics, Philosophy & Economics*, 9(3), 265–273.
- Smart, J. J. C. (1956). Extreme and restricted utilitarianism. *Philosophical Quarterly*, 6(25), 344–354.
- Smith, M. (1987). The humean theory of motivation. *Mind*, 96(381), 36–61.
- Smith, M. (1994). *The moral problem*. Blackwell.
- Smith, M. (1998). The possibility of philosophy of action. In J. Bransen & S. Cuypers (Eds.), *Human action, deliberation and causation* (pp. 17–41). Kluwer Academic Publishers.
- Sober, E. (1983). Equilibrium explanation. *Philosophical Studies*, 43(2), 201–210.
- Sober, E., & Wilson, D. S. (1998). *Unto others: The evolution and psychology of unselfish behavior*. Harvard University Press.
- Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3), 355–374.
- Spence, M. (2002). Signaling in retrospect and the informational structure of markets. *The American Economic Review*, 92(3), 434–459.
- Sripada, C., & Stich, S. (2006). A framework for the psychology of norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind, volume 2: Culture and cognition* (pp. 280–301). Oxford University Press.
- Sterelny, K. (2012). *The evolved apprentice: How evolution made humans unique*. The MIT Press.
- Sugden, R. (1998). Normative expectations: The simultaneous evolution of institutions and norms. In A. Ben-Ner & L. Putterman (Eds.), *Economics, values, and organization* (pp. 73–100). Cambridge University Press.
- Sugden, R. (2000). The motivating power of expectations. In J. Nida-Rümelin & W. Spohn (Eds.), *Rationality, rules, and structure* (pp. 103–129). Kluwer Academic Publishers.
- Sugden, R. (2004). *The economics of rights, co-operation, and welfare* (2nd ed.). Palgrave Macmillan.
- Taylor, P., & Jonker, L. (1978). Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1), 145–156.

- Thomas, B. (1984). Evolutionary stability: States and strategies. *Theoretical Population Biology*, 26(1), 49–67.
- Thornton, A., & Raihani, N. (2008). The evolution of teaching. *Animal Behaviour*, 75(6), 1823–1836.
- Tomasello, M. (2016). Cultural learning redux. *Child Development*, 87(3), 643–653.
- Tomasello, M., Kruger, A., & Ratner, H. (1993). Cultural learning. *Behavioral and Brain Sciences*, 16(3), 495–511.
- Tosi, J., & Warmke, B. (2016). Moral grandstanding. *Philosophy and Public Affairs*, 44(3), 197–217.
- Trivers, R. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46(1), 35–57.
- Ullmann-Margalit, E. (1977). *The emergence of norms*. Oxford University Press.
- Van Lange, P., Rockenbach, B., & Yamagishi, T. (2014). Reward and punishment in social dilemmas: An introduction. In *Reward and punishment in social dilemmas* (pp. 1–14). Oxford University Press.
- Veblen, T. (1953). *The theory of the leisure class; an economic study of institutions*. New American Library.
- Wakano, J. Y., & Aoki, K. (2006). A mixed strategy model for the emergence and intensification of social learning in a periodically changing natural environment. *Theoretical Population Biology*, 70(4), 486–497.
- Weibull, J. (1995). *Evolutionary game theory*. The MIT Press.
- Westra, E. (2021). Virtue signaling and moral progress. *Philosophy & Public Affairs*, 49(2), 156–178.
- Williams, R. M., & Gibbs, J. P. (1968). Norms. In D. L. Sills (Ed.), *International encyclopedia of the social sciences* (pp. 204–213, Vol. 11). Macmillan.
- Wittgenstein, L. (1953). *Philosophical investigations*. Blackwell.
- Young, H. P. (1993). The evolution of conventions. *Econometrica*, 61(1), 57–84.
- Young, H. P. (2015). The evolution of social norms. *Annual Review of Economics*, 7, 359–387.
- Zahavi, A. (1975). Mate selection—a selection for a handicap. *Journal of Theoretical Biology*, 53(1), 205–214.
- Zollman, K. (2008). Explaining fairness in complex environments. *Politics, Philosophy & Economics*, 7(1), 81–97.