Comments welcome at drkelly@purdue.edu, taylor.thiel.davis@gmail.com

To appear in *Social Policy & Philosophy*, Special Issue on Learning and Changing Norms

Last Changes: 3/26/17 DRK
Word Count: 7137, main body; 11,558 total

# Social Norms and Human Normative Psychology

By Daniel Kelly and Taylor Davis[1]

**Abstract**: Our primary aim in this paper is to sketch a cognitive evolutionary approach for developing explanations of social change that is anchored on the psychological mechanisms underlying normative cognition and the transmission of social norms. We throw the relevant features of this approach into relief by comparing it with the self-fulfilling social expectations account developed by Bicchieri and colleagues. After describing both accounts, we argue that the two approaches are largely compatible, but that the cognitive evolutionary approach is well-suited to encompass much of the social expectations view, whose focus on a narrow range of norms comes at the expense of the breadth the cognitive evolutionary approach can provide.

## I. Introduction

While research on norms spans the humanities and human sciences, Christina Bicchieri's social expectation account has recently risen to prominence in philosophy, and serves as a touchstone and focal point for much discussion.[2] As such it will serve as our jumping off point and stalking horse in this paper. In the following section we will motivate and explain the core ideas of her account, and note some points of interest. Our discussion here will be brief, in part because Bicchieri's view is well known, and in part because many of the other papers in this special issue also discuss it. Moreover, we go into more detail below, where relevant, when discussing those aspects of Bicchieri's account that contrast with the cognitive evolutionary perspective we endorse. Our third (and largest) section will describe the view of norms that emerges from this perspective, focusing on its picture of humans' distinctive normative psychology, and illustrating how a better understanding of the mechanisms underlying the acquisition, performance, and transmission of norms can help shed light on how norms influence behavior, and how they are susceptible to change. Our fourth and final section will compare and contrast Bicchieri's view with our own. We argue that while her account usefully and accurately explains the phenomena on which it focuses, it is too restricted in scope to function as a complete theory of norms and norm psychology. However, we believe that much of her social expectation account is compatible with the broader scope of the cognitive evolutionary approach, and that together they each make important contributions to a complete theory of human normative psychology.

## II. Social Norms as Self-fulfilling Social Expectations

In a number of papers and books, Christina Bicchieri has developed an impressive body of work on social norms.[3] On Bicchieri's picture, social norms are rules that govern the

behavior of individuals, in turn creating group-level regularities. Not all common behaviors, or group-level regularities, indicate the presence of a social norm, however, so the core idea of the theory is that a social norm is in place only if the relevant group-level regularity is sustained by a particular cluster of psychological states found within individual actors. The most important components of this complex are two types of social expectations, social because they are beliefs about other people. The first, which Bicchieri calls **empirical expectations**, are beliefs about how other people *will* act, specifically about how they are likely to behave in a particular type of situation. The second, which she calls **normative expectations**, are beliefs about how other people think one *should* act, specifically beliefs about what a person ought to do in a particular type of situation. These beliefs may often be accompanied by the belief that others will not just disapprove of, but actively sanction those who violate their normative expectations, punishing them for failing to act as they should (or that others will approve of and actively reward those who satisfy normative expectations). Importantly, then, this definition itself does not strictly imply the existence of social sanctions and rewards, although normative expectations often lead to sanctions and rewards, empirically.

A third important component of Bicchieri's social expectation theory speaks to the issue of motivation. **Conditional conformity** is another mark of social norms. On this view, in order for a group-level regularity to be properly explained as a social norm (rather than a moral norm, custom, practice, tradition, or any other group-level regularity), those engaged in the behavior must have a preference to follow the rule *only if* they believe those others will also follow the rule, *and* that those others believe that they ought to follow the rule—i.e., only if both kinds of social expectations are met. Bicchieri distinguishes social norms proper from what she calls descriptive norms, which are behavioral regularities marked by preferences conditional on empirical expectations, but not necessarily upon normative expectations; intuitively, a descriptive norm is one where people engage in a behavior because they believe everyone else engages in that behavior, even if no one believes people *should* engage in the behavior. Thus, the expectations and beliefs that underlie social norms and descriptive norms alike are social, and the preferences leading to conforming behavior are also social--i.e., dependent on how one expects others to behave. Patterns of behavior count as social norms only if they are brought about by a combination of beliefs about other people and motivations that depend upon other people. People prefer to comply with the social norm, but only on the condition that everyone else in the group is complying, and everyone else thinks it's the right thing to do, too.

When all of these conditions are met, a stable group-level regularity is sustained by the self-fulfilling interplay of the social expectations and social preferences of the individuals that make up the group, or network. Only group-level regularities thus stabilized, on Bicchieri's view, are properly identified as social norms. In light of this set of individually necessary and jointly sufficient conditions, we'll call the cluster of psychological states required for the presence of a social norm a *Bicchieri-cluster*, and the type of behavioral regularity it produces, which she simply calls a social norm, a *Bicchieri-norm*.

A couple more comments will be useful here. First, though Bicchieri's 2006 book is suggestively entitled *The Grammar of Society: the Nature and Dynamics of Social Norms*, its conceptual roots run through game theory and economics, and it draws on the resources of rational choice theory much more than linguistics (or cognitive science more broadly).

Indeed, while it rejects a purely behavioral or group-level definition of social norms, the types of psychological states at its core are essentially those of common belief/desire folk psychology, or the close counterparts (e.g. desires as preferences) typically represented in game theoretic models. That said, Bicchieri and her colleagues have recently developed the account in further detail, appealing to mental representations such as scripts and schemata,[4] social roles and positions in social networks, and traits of particular actors, such as perceived self-efficacy.[5] Related to this, she also qualifies the definition of a social norm by appeal to the useful concept of a reference network, the group of people one takes into account in one's social expectations and preferences, for any given norm. Finally, as she and her collaborators have emphasized recently, the self-fulfilling social expectations account has direct implications for crafting policies and other kinds of interventions designed to change these kinds of group level regularities. Bicchieri-norms are stabilized by expectations and preferences about what others are likely to do, and what they are likely to think. Thus, efforts at redirecting behavior governed by Bicchieri-norms are unlikely to succeed when they merely attempt to change personal preferences, or to correct false factual beliefs about health risks or other undesirable outcomes of the behavior in question. Rather, since a Bicchieri-norm is sustained by a Bicchieri-cluster, and the psychological states in a Bicchieri-cluster are social and other-oriented, successful interventions need to focus on what the members in the reference network believe about *what the other members in the reference network believe*.

## III. A Cognitive Evolutionary Approach to Norms

We will return to the Bicchieri view below, but will now set it aside in order to present a cognitive evolutionary approach. In this section we draw on a range recent work coming out of the cognitive sciences characterizing human capacities for cognizing and acquiring norms, and describe what are emerging as the key psychological properties of the mechanisms that underpin these capacities. Three points before we get started: First, while our account is informed by the insights and findings of a number of theorists, it is not identical to the view endorsed by any of them; our intent is to present a picture that represents the overlap and convergence of what we take to be the most promising work. For ease of exposition, we will refer to ours as the Minimal Account, which aims to capture what an emerging consensus sees as the core features of human norm psychology.[6] Second, while we appeal to work often discussed under the heading of "empirical moral psychology," the conception of normative psychology we work with is both broader and narrower that what might be considered the psychology of "morality". On one hand, there is a great deal of work in moral psychology that we will not consider here, because it is nominally not about norms or normative cognition at all, but rather is about such issues as, for instance, the identification of intentional versus unintentional behavior, character trait-based versus situation-based explanations of behavior, conceptions of the true self, how to assign responsibility and blame for implicit bias, etc. On the other hand, the Minimal Account aspires to accommodate not just putatively moral norms, but also nonmoral norms of all kinds, including norms of logic, language, epistemology, aesthetics, religion, etiquette or any other kind.[7] Finally, our characterization of the psychological underpinnings of normative behavior posited by the Minimal Account draws on a richer vocabulary than folk psychology provides, discussing mechanisms and subsystems sometimes in lieu of beliefs and desires (or belief-like and desire-like states such as expectations and preferences).[8] While departure from commonsensical origins often accompanies a research program's gains in conceptual

sophistication and explanatory power, we will address some particular implications of this fact when we compare the Minimal Account to the self-fulfilling social expectation view in section four.

*The Psychology of Normativity: A Minimal Account*

We begin with normativity in general. From a logical point of view, the defining feature of a normative proposition is that it says something about what is required, allowed, or forbidden. Normative statements either specify or imply what "should" or "shouldn't" be done, expressing a rule or making some prescriptive claim about how people ought to think, feel, judge or behave. This feature provides a useful point of departure for making sense of the relationship between the logic of norms, on the one hand, and the psychology of norms, on the other. First, it allows us to distinguish *normative* from *normal*. A behavior does not count as norm-governed, and a pattern of behavior does not count as realizing a norm, simply in virtue of the fact that it happens to be normal, statistically speaking, within a population. Second, it allows us to distinguish *good* and *bad* from *right* and *wrong*. As Nichols (2004, chapter 1) points out, toothaches and natural disasters are bad, but they aren't wrong; in such cases no rule has been violated or transgression committed. In contrast, if a person steals a few dollars from the tip jar at a coffee shop, or keeps showing up at dinner parties empty handed and without the requisite bottle of wine, they are not just acting badly, they are doing something they shouldn't do. In both cases the person's action breaks a rule, written or otherwise, and in virtue of this can properly be evaluated as wrong. Thus, in our sense, normative psychology is the psychology of oughts, and so part of the psychology of rules.[9]

Consequently, it is also the psychology of compliance and enforcement. Indeed, from a psychological point of view, a distinctive, defining feature of norms concerns approval and disapproval, and the use of punishments and rewards to influence behavior. Again, a behavior is not normative in our sense merely in virtue of being typical. If reward and punishment play no role in explaining why a given pattern of behavior is statistically normal within a population, this suggests the regularity occurs for reasons other than the presence of a norm prescribing it, and perhaps that the individuals in the population themselves are indifferent to whether anyone actually engages in the behavior. This amounts to the claim that in such a case there is no sense, at least within the confines of causal, naturalistic explanation, in which anyone ought or ought not to engage in that behavior. In short, norms imply "oughts," and "oughts" imply punishment and reward.

Guided by similar assumptions, researchers like Richerson and Boyd, Chudek and Henrich, and Gelfand and Jackson summarize bodies of evidence supporting the existence of "a suite of genetically evolved cognitive mechanisms for rapidly perceiving local norms and internalizing them".[10] We refer to this complex of traits as the *norm system*, a more-or-less integrated, hierarchically organized collection of functional capacities and subsystems. This work suggests that some parts of the system are genetic adaptations that develop reliably across a wide range of different cultural environments, while others are highly local, typically acquired via imitation and social learning, but also sometimes through individual trial and error, and or the process of personal deliberation and reasoning. The combination of both innate (i.e., genetically inherited) structure and social learning allows our Minimal Account of the norm system to accommodate the fact that while norms in general are a ubiquitous part

of human social life in all cultures, the content of particular norms varies greatly from culture to culture, and changes over time within cultures. An appeal to genetic adaptation is able to explain why human individuals are universally capable of identifying, acquiring, and performing *some* set of norms or other, but it falters when it comes to explaining diversity and variation of norms across cultures; innate capacities for learning do not by themselves explain the particular content that an individual comes to learn, even if those learning mechanisms are themselves targeted at a specific and well delineated domain.[11] In the same way, appeals to innate psychological machinery explain why humans universally share the ability to learn some language or other, but it does not explain why anyone speaks the particular languages she speaks, rather than others.[12]

Sripada and Stich provide a high-level model of the core psychological mechanisms comprising an individual's norm system.[13] That model depicts a key distinction between those mechanisms responsible for the identification and acquisition of local norms, on one hand, and the performance of norms that have been acquired, on the other. Mechanisms on both sides of this divide are likely to have aspects that are purely cognitive and representational, as well as other aspects that are more affective and motivational. And while motivational features include obvious and powerful passions of righteous anger and contemptuous disgust, they also include more subtle mechanisms. For instance, motivational capacities within the acquisition subsystem may influence attention, shaping what we find interesting, salient, and relevant in the behavior and interactions of other people. Even children do not need to be cajoled to notice social rules, and nor are they passive learners that have to be actively or explicitly taught them. Those as young as three years old appear motivated to attend to cues indicating that a behavior is normative, and to draw inferences about the rule that is governing it.[14]

But the affective and motivational aspects involved in following and enforcing norms are particularly important for our purposes, largely because they give acquired social rules their distinctive normative motivational force, or *normative force* for short. Our Minimal Account incorporates the claims that this normative force is (1) intrinsic, as opposed to instrumental, that it is (2) two-pronged—both self- and other-oriented—and that it (3) can be quite strong. More fully, when an individual genuinely acquires a norm, rather than merely becoming aware of or simply cognitively grasping a social rule, the rule is represented in the database of her norm system, and thereby inherits the motivational features associated with the performance of norms. Thus, to fully *acquire* a norm is to develop an intrinsic motivation to perform it, rather than following or enforcing it out of instrumental motivation to avoid sanctions or attain rewards.[15] And normative motivation in this sense is not just non-instrumental, but also two-pronged: a person who has acquired a norm is thereby motivated to obey simply because it is "the right thing to do", but she is also motivated to enforce the norm on others, punishing those who violate it, and often forming longer-standing reactive attitudes toward transgressors. This is not to say that a person's *behavior* is always in accord with the norm, or that she actually enacts the punishments or rewards she deems appropriate, but merely that this piece of her psychology generates an impulse to do so, which may be overriden. Intrinsic motivation is not irresistible motivation, and both of these types of normative motivations, like any others, can be suppressed or superseded by other, more powerful motivations. That said, the force of compliance and punishment motivations can be, and often is, quite powerful, sometimes overriding self-interested goals, leading to examples of sacrifice, commitment to a group or cause, and in extreme cases even

martyrdom.[16] Aside from strength, the quality of normative motivation can differ along other dimensions as well, some of which have to do with details of situations to which a norm might apply, some having to do with the norm itself, and some perhaps with which emotion the norm draws on, for instance anger, disgust, guilt, shame, outrage, spite, etc.[17]

Appeal to the norm system, and the normative force it confers on acquired norms, surely will not explain every instance of cooperation or collective action, of people acting in accordance with social rules, or of people sanctioning those who disobey them. One person may follow a social rule out of an instrumental desire to avoid the punishments that come from violating it, while another person may follow the same rule for its own sake, regardless of the instrumental, practical value of doing so. But we maintain that this Minimal Account of normative psychology, and the research on which it draws, provides indispensible explanatory purchase in understanding how individuals dole out social rewards and punishments and how they respond to the rewards and punishments doled out by others. It also thus fits into a larger picture about the group level characteristics of norms, and factors that influence how those change over time. We turn to this next.[18]

*Norm Psychology, Cultural Evolution, and Social Change*

To clarify how this Minimal Account might contribute to a broader picture of social change, we begin by showing how the norm system helps bind individuals to groups, getting them in sync with the social arrangements that structure life within any given group. The Minimal Account appeals to social learning to explain how an individual comes to have the particular norms represented in her norm system, suggesting that, as in the case of language, the acquisition mechanism of an individual's norm system intuitively and automatically "soaks up" norms from her social environment.[19] The plausibility of this claim comes from the commonalities in the developmental trajectory of normative capacities that appear to hold across cultures, as well as from patterns in the distribution of particular norms throughout populations. Like languages, norms exhibit common patterns of within-group similarity and between-group difference.[20]

The Minimal Account thus depicts human normative psychology as "expecting" certain kinds of cues and regularities in an individual's social world, from which it will be able to glean information about the particular norms that prevail locally. These cues manifest in other people and their interactions with each other, as well as with the individual herself, and the most salient behaviors will be those group-level regularities stabilized by reward and punishment. These pockets of the social world provide information about what kinds of actions are forbidden, permissible, and required, and, in the episodes of social learning that fix on them, the acquisition machinery of a person's norm system makes (perhaps innately constrained) inferences about what the rule is that governs the observed interaction (as well, perhaps, as what follows from that rule).[21] When she acquires those norms, she becomes attuned to her culture and its social arrangements, her own normative sensibility harmonizes with the general normative framework of the group of which she is a member, with perhaps some fine-grained calibration determined by the particular set of social roles she occupies in it. Due to the normative force exerted by the norms she has acquired, she not only typically acts in accordance with them, but typically enforces them as well. Thus, by obeying rules and punishing transgressors, she makes her individual contribution to the stability of the

collective, supporting the durable structure of the often long-standing social arrangements of her culture.

Or so it might go in a "perfect" world. On the highly idealized picture just described, the norm system would successfully bind individuals to their groups, but seamlessly. The (somewhat unsettling?) vision just sketched suggests individuals who are perfect learners getting totally enculturated into a society, flawlessly performing its norms, making them (the individuals) pristine models for the next wave of norm learners, and turning them (the individuals again) into impeccable members of a society that is almost completely static, and whose norms thus remain, but for the occasional exogenous perturbation, largely unchanged as they are passed from one generation of individuals to the next. The reality is of course (happily?) messier, but this picture, like any good idealization, gets some key things right. Many group-level regularities are far from ephemeral; social arrangements and the interconnected clusters of norms that govern them can be quite stable over time, even doggedly resistant to change. Moreover, modeling work in evolutionary game theory strongly suggests that one of the most important stabilizing factors is punishment.[22] Via the kind of feedback loops just described, when the members of a group enforce norms and punish transgressors, the aggregated effect can render interconnected clusters of norms endogenously stable. Situations that fit this description are in equilibrium, or in an evolutionarily stable state, and the mechanisms that keep them in homeostatis also make them robust in the face of a range of external influences—including deliberate attempts to change them.[23]

Two points will allow us to begin complicating this picture. First, as noted above, part of the messiness of the actual world is that there is cultural variation in norms, social arrangements, and group-level behavioral regularities. The norms and arrangements that organize one group can differ dramatically from the norms and arrangements that organize another group, and yet both can be endogenously stable. A second way in which reality departs from the idealized picture is that the norms and social arrangements of groups do, in fact, change over time, shifting from one stable state to another. Together these points show that there is more than one way to be stable (some of which are more adaptive, collectively efficient and cooperative, some of which are less so). Stability is a property that can be realized by many different configurations of norms; there are multiple stable equilibria, and punishment and the operation of human normative psychology can in principle stabilize any of them. And since there are so many ways to be stable, appeal to norms, punishment, and the operation of human normative psychology alone will be inadequate to fully explain either variation or change. While it will be a key, perhaps necessary, factor in any viable explanation, the Minimal Account can't by itself explain why those stable configuration of norms that have actually been realized *have* actually been realized, while other possible stable configurations have not. Nor can it fully explain, by itself, the fact that norms and social arrangements change over time, or how. This requires further resources.[24]

We maintain, however, that theories of cultural evolution in general *can* deliver these goods. This shouldn't be controversial; "evolution" means "change," and so theories of cultural evolution are theories of cultural change.[25] In general, such theories use the concepts and models of biological evolution and evolutionary game theory to understand changes in the frequencies and distribution of *cultural variants*, where cultural variants are understood as behavior-affecting bits of information that are acquired socially.[26] Variants are typically

different kinds of socially transmitted ideas, values, beliefs, preferences, skills, habits, norms, etc. These are subject to competition in the face of cultural selective pressures of differing strengths and from assorted origins. Cultural variants are understood to have varying levels of *cultural fitness* depending on how likely they are to be copied and transmitted to others, and thus spread across populations and between generations.[27] When quantified, these properties can be represented in the models, which are used to calculate the changes in the frequencies of cultural variants that will occur under various conditions. This literature is thriving and continues to grow, but for now we will briefly mention a few representative and important ways that the psychological features of individuals, including features of their norm systems, can influence the transmission and spread of cultural variants. Since norms are cultural variants themselves, these psychological features are relevant for attempting to influence social change and contribute to the evolution of social arrangements.

**Transmission and Learning Heuristics**

Unlike in our idealized story, humans are not perfect learners, and cultural transmission is reliable but not extremely so, and it does not always lend itself to precise, high-fidelity copying. This produces noise and variation, though some of this is counterbalanced by the fact that humans are not equal opportunities learners. Rather, social learning capacities, including those responsible for the acquisition of norms, are guided by a number of transmission and learning heuristics. When added to humans' hypertrophied (compared to other social animals) tendency to imitate each other, these heuristics are features that facilitate a groups' collective ability to more effectively and efficiently produce adaptive cultural variants. These heuristics go some way in "correcting" for the effects of noisy transmission or sloppy individual learning.[28] For example, a prestige heuristic makes us more likely to adopt norms and other cultural variants from those with the greatest success and status within our group, or within some subculture or reference group with which we are concerned.[29] Similarly, a conformity bias makes us more likely to imitate the most common behaviors, and, in the case of norms, acquire those performed by most of our peers, or aspirational peers in the in-group we wish to join.[30] Together, these exert systematic influence on how the frequencies of norms in a population can change over time. It is worth noting that conformity and prestige heuristics give norms and other cultural variants a cultural fitness boost independently of the content of those variants, and independently of whether or not the norms are just, fair, utility maximizing, etc. Intuitively, the idea is that messengers matter, and these heuristics exert an influence on the spread of norms based solely on who performs them. Nevertheless, they still directly affect how norms are transmitted, influencing which are more likely to be copied and acquired by other members of the population.

**Epidemiology-Inspired Psychological Stickiness**

Another psychological factor that can influence the transmission and cultural fitness of norms—and one that also responds in part to features other than the content of the variants in question—is what we will call *psychological stickiness*. This is the degree to which a variant easily "meshes" with a variety of features of cognitive machinery other than capacities dedicated specifically to social learning. The general epidemiological approach to culture has been pioneered and developed by, and fruitfully applied to the study of religion.[31] The idea can be extended to norms as well: to the extent that a norm is salient, easy to identify, and

easy to remember, it will also be more easily transmitted to others.[32] Moreover, a norm might engage psychological mechanisms in addition to those core components of the norm system, and in doing so it may get a further fitness boost. Other psychological mechanisms that have been shown to boost the cultural fitness of a norm include particular emotions, such as disgust[33], narrative capacities, and the way a norm is embedded in a recognizable kind of plot or narrative[34], close association with attention-grabbing, button-pushing supernormal stimuli[35], or really anything else that increases the norm's salience, ease of comprehension, retention, and transmission by making it attractive and intuitive to some piece of our psychological repertoire. Indeed, meshing with the norms that an individual has already acquired could itself make a new candidate norm psychologically stickier for that person, and if other individuals in the group are similarly normatively attuned, the new candidate norm will be stickier throughout the population, thereby receiving a fitness boost. In this way, norms influence the selective environment for other norms.[36]

**Ancient Sociality and Tribal Sociality**

While transmission and learning heuristics and psychological stickiness are general factors that can affect the spread of any kind of cultural variant, recent research suggests an important division between two broad families of psychological systems humans have for navigating different kinds of social exchanges. In virtue of this, these systems are more directly relevant to norms and normative behavior, but in different and important ways. The idea is that human capacities for interacting with each other have two distinct strata, and while both help facilitate cooperation and coordination, each has been differently shaped by its own distinct evolutionary history, and those differences remain visible in the ways each influences social interaction today. The ancient social instincts[37] are those that we share with many other social animals, which are responsible for our interactions with family and friends—conspecifics with whom one shares blood relations or regular patterns of interaction. These more ancient capacities operate according to principles associated with kin selection and (often reputation-based) reciprocal altruistic solutions to cooperative dilemmas. This complex of psychological traits also includes the elements of our status psychology that are based on dominance.

Tribal social instincts, by contrast, are evolutionarily recent and uniquely human. As their name suggests, they are responsible for our interactions with members of tribal-sized groups, whose size far outstrips that of a circle of family members, friends and acquaintances (even acquaintances merely by reputation). These include many features of human normative psychology, as described above, but other components as well. One such is a set of capacities related to monitoring tribal membership. A mark of these is a heightened sensitivities to tribal boundaries, or to the symbolic markings that people use to signal what groups they belong to, as well as what station or roles they occupy within those groups, and thus what norms and beliefs they have likely acquired. These capacities also have motivational features, typically shaping differences in the way individuals behave toward ingroup and outgroup members.[38] Indeed, there has been a great deal of work recently exploring how ingroup favoritism and outgroup bias manifest, especially with respect to other aspects of social instincts and cooperation. Some of it suggests these heuristics emerge rather early in development, driving noticeable differences in norm enforcement and reputation management.[39] In a series of papers, Carsten De Dreu and colleagues have shown that this kind of tribalism runs deep, and that the effects of oxytocin on social interactions

bifurcate along ingroup and outgroup lines along a number of different dimensions, including empathy, conformity, and cooperative and competitive tendencies.[40]

Another aspect of our tribal psychology relevant to social change and the transmission of norms is the human disposition to give deference based on prestige, rather than just physical dominance.[41] This can manifest in many behavioral ways associated with status and status-related behavior, which includes, as mentioned above, differentially imitating highly prestigious individuals.[42]

Finally, tribal social instincts are thought to include a suite of uniquely human emotions that evolved in tandem with our normative, tribal membership, and prestige capacities. Examples of emotions that have been hypothesized to fit this description these include empathy, which allows us to share feelings and cognitive states with others, shame and guilt, which can be seen to function as internalized enforcement mechanisms that make a person more likely to obey norms she has acquired, and loyalty and pride, which can emotionally bind one to her tribe. Alternatively, Kelly argues that disgust, an emotion that is uniquely human but relatively ancient and not initially social, nevertheless came to play several roles in our more modern tribal social psychology, including producing stigmatizing, dehumanizing aversion towards the members, norms, and values of particularly loathed outgroups.[43]

An important claim here is that, like the newer emotions, and the addition of prestige to our status psychology, neither the human normative capacity nor the sensitivities to group membership and tribal boundaries are merely refinements to, or elaborations of, the ancient social instincts we share with other animals. They are a new thing under the sun. One happier result of this is that humans can cooperate with each other in many different ways, on a number of different scales. Ancient and tribal instincts provide different tricks to identify others in the group as likely cooperators and, more importantly, to detect and sanction defectors.[44] It has been posited that together these can collectively act as a 'moral hidden hand,' or a source of pro-social behavior and psychological stickiness that influences the spread and evolution of norms. Recall that on its own punishment can stabilize any norm: the useful and the pointless alike, as well as the just and unfair, the cruel and the kind. The moral hidden hand, however, can act as one of the pressures that drive social change towards the better—or at least the more pro-social—by giving a cultural fitness boost to norms that lead individuals to act for the good of the group, paying personal costs for the sake of others. Thus, norms that activate our feelings of empathy, our sense of fairness, or our aversion to gratuitous harm receive a transmission advantage over those that do not. This, in turn, acts as a gentle but persistent selection pressure favoring more equitable and compassionate social arrangements over the long run, because the norms that prescribe such arrangements are more likely to 'mesh' well with the range of human cooperative instincts.[45]

A less happy but equally interesting upshot is that these two sets of instincts can also be at odds with each other, and the resulting struggle can manifest both within individuals, and, collectively, at the level of groups. In other words "These new tribal social instincts were superimposed onto human psychology without eliminating ancient ones favoring friends and kin. This resulted in an inherent conflict built into human social life."[46]

Finally, a controversial but compelling hypothesis posits that another distinct and important process driving the cultural evolution norms and social change is *cultural group*

*selection*.[47] Darwin himself famously endorsed the basic idea: "A tribe including many members who, from possessing in a high degree the spirit of patriotism, fidelity, obedience, courage, and sympathy, were always ready to aid one another, and to sacrifice themselves for the common good, would be victorious over most other tribes; and this would be natural selection."[48] Correctly understood, this is a macro form of cultural evolution, and concerns competition and cultural selection between different *clusters* of norms, social arrangements and other cultural variants. When one 'tribe' – a group of people bound together by shared norms, values and other cultural institutions – competes with another, the tribe equipped with more efficient, effective, and advantageous cluster of variants will typically win, and its technology, norms, and social arrangements will spread at the expense of the defeated tribe's.[49]

## IV. The Social Expectations Account Meets The Cognitive Evolutionary Approach: Two Ways of Putting the "Social" in Social Norms

Both Bicchieri's account and the cognitive evolutionary account distinguish normative behavior from merely normal behavior, and both accept a common broad notion of what a norm is, namely, a rule of behavior that has both individual and group level properties.[50] Moreover, while group- or population-level regularities are central to each account, both reject the idea that norms are to be accounted for purely at the group or population level. Rather, both identify norms by appeal to psychological characteristics of the individuals that make up the group, but each does so by appeal to different kinds of psychological structure, resulting in different explanations for the stability of the relevant group-level regularities. The social expectations account defines norms in terms of what we called the Bicchieri-cluster: empirical expectations about what others will do, normative expectations about what they think should be done, and conditional preferences to comply if others do. The cognitive evolutionary account appeals to a suite of systems that includes mechanisms for the social learning of practices of *enforcement* (including rewards), along with the specialized psychological mechanisms that make up human norm system, as specified by the Minimal Account.

Other than the expected terminological differences (which we suspect are reconcilable), the key contrasts between the two views flow from what they have to say about motivation and acquisition. On our view, human psychology is equipped with a distinctive, specifically *normative* kind of motivation. This motivation is intrinsic, driving an individual to act in accordance with those behavioral rules represented in her norm box for their own sake, regardless of instrumental or conditional reasons. On this account, the motivation to comply with or enforce an acquired norm is, in the relevant sense, *asocial*, or *independent* of social reasons. This seems to be in direct contrast to the social expectations account, which defines norms in terms of an individual's social, other-oriented beliefs and preferences. In this way, it defines norms in contrast to other behavioral regularities like customs, practices, and traditions. This strikes us as an odd feature of the account, since it seems obvious to us, and natural to say, that customs, practices and traditions are *themselves governed by norms*, i.e. by culturally acquired, intrinsically motivating, socially enforced, behavior-guiding rules that specify how to participate in the relevant activity.

On the cognitive evolutionary account, what makes a social norm *social* isn't necessarily the expectations or preferences that cause anyone to conform to it or enforce it, but rather

the way in which it is acquired by individuals: norms are *socially learned* from cultural peers and parents. This is why we have emphasized the psychological machinery dedicated to norm acquisition, and the ways in which the Minimal Accounts shows this machinery to be both automatic and intuitive, but also exquisitely sensitive to cues of group membership and social status. The cultural transmission and acquisition of norms is of enormous importance, especially since, once a norm has been acquired and represented in an individual's norm system, that norm becomes imbued with intrinsic motivations of substantial force.

On the evolutionary cognitive account, the following state of affairs is the paradigm for norms and human normativity: social acquisition of a behavioral rule that in turn leads to intrinsic motivation to comply and sanction. However, motivation is not compliance, and intrinsic motivation is not unconditional compliance. We understand Bicchieri's temptation to cordon off these cases as extreme, to save the hard cases for later, and to dub them "moral norms" in order to distinguish them from the "social norms" on which her account focuses. There is a large and difficult conversation to have here, but the most pressing problem is not that we are without any way to delineate morality from other forms of normativity, or to cleanly distinguish moral norms from social norms, conventions, prudential rules, or the rest. The problem is that there are *too many* plausible ways to delineate the domain of morality, but no two seem to slice the pie in the same way.[51] For instance, Bicchieri's account of social norms places them within a larger taxonomy of behavioral phenomena.[52] In it, she separates out what she calls customs, legal injunctions, and moral rules from what she calls norms (descriptive and social). Her conception of moral norms defines them as those wherein compliance is unconditional, not dependent either on social expectations or social preferences. She does not say state it in these terms but this seems to entail that individuals conform to moral norms because they are intrinsically motivated to do so, and also that such intrinsic motivation *only* accompanies moral norms, that it is a unique and essential mark of morality. On our view, this mistakes a feature common to all norms represented in a person's norm system (normative force) for one that is distinctive of morality, and only morality. In doing so, it also restricts her account so that it has no conceptual resources for explaining the cases that emerge as *paradigmatic* from the point of view of the cognitive evolutionary account, and the fact that culturally acquired intrinsic motivations are also conferred all kinds of (intuitively nonmoral) norms, including epistemic norms, aesthetic norms, norms of logic, language, religion, etiquette, etc.

That said, on our view, the selective focus of the social expectations account is quite understandable and even justifiable from the point of view of the practical aims of providing actionable policy advice and diagnosing, measuring and changing many important norms.[53] But what is good for practical purposes is not necessarily what is good from the point of view of full theoretical understanding. Bicchieri-norms bracket off what the cognitive evolutionary account identifies as the most interesting, central and important aspects of human normativity, including the roles they play in generating cooperative behavior and collective action, and the psychological adaptations that evolved specifically for negotiating social environments in which norms are prevalent. Bicchieri-norms are an important subset of norm-related social phenomena at the intersection of individual psychologies and collective level regularities, but the idea that Bicchieri-norms count as 'social norms' proper, or that they make up the core subject matter of the study of norms and normativity in general, is quite implausible.

Fortunately, we see no reason why Bicchieri-norms cannot be incorporated into a theoretical perspective that is much broader in scope, which does address the study of norms and normativity in general. In many important cases of normative behavior, individuals do conform conditionally, on the basis of instrumental motivations derived from social expectations. We see no reason to reject or exclude the accurate and detailed account of those cases that Bicchieri has developed.

## V. Conclusion

We end on an ecumenical note. The types of resources we have discussed in this paper will all eventually contribute to our nascent but growing understanding of norms and social change. Perhaps they will also be of use in trying to produce and guide it as well (Bicchieri 2016, c.f. Wilson 2016, Wilson et al 2014). While we hold that evolutionary thought has a foundational role to play in this coming synthesis, we disagree with the claim that "**nothing** about norms and institutions the makes sense except in the light of evolution."[54] On the contrary, Christina Bicchieri and her colleagues working on the social expectations account have provided a clear, useful, and high-resolution understanding of an important subset of social norms and norm-related social phenomena. While we have highlighted the differences and limitations of that account, particularly with regard to motivation and acquisition, we have also argued that it can ultimately be integrated into the evolutionarily based picture we have advocated. Together, they can shed more light on all kinds of norms, and on different aspects of the complicated tapestry of human normative psychology, cultural transmission, and social change.

[1] We received much useful feedback on this material, and would like to thank: an anonymous reviewer, Lacey Davison, Nicolae Morar, and especially Gerald Gaus, Shaun Nichols, and the participants of The University of Arizona's Learning and Changing Norms Conference.

[2] For another prominent approach see Nicholas Southwood and Lina Eriksoon, "Norms and conventions" *Philosophical Explorations*, 14, 2 (2011): 195-217 and Geoffrey Brennan, Lina Eriksson, Robert E. Goodin, and Nicholas Southwood *Explaining Norms* (Oxford, Oxford University Press, 2013), and for some useful overview discussions see Leigh Raymond, S. Laurel Weldon, Daniel Kelly, Ximena Arriaga, and Ann Marie Clark, 'Making Change: Norms and Informal Institutions as Solutions to "Intractable" Global Problems', *Political Research Quarterly*, 67, 1 (2013): 197 – 211 and Cristina Bicchieri and Ryan Muldoon, "Social Norms", The Stanford Encyclopedia of Philosophy (Spring 2014 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2014/entries/social-norms/>.

[3] See especially Cristina Bicchieri, *The Grammar of Society: the Nature and Dynamics of Social Norms* (New York: Cambridge University Press, 2006), Cristina Bicchieri, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms* (New York: Cambridge University Press, 2016), and Cristina Bicchieri and Ryan Muldoon, "Social Norms", The Stanford Encyclopedia of Philosophy (Spring 2014 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2014/entries/social-norms/>

[4] Cristina Bicchieri and Peter McNally, P. "Shrieking Sirens Schemata, Scripts, and Social Norms: How Change Occurs" (Manuscript).

[5] Ibid, chapter 5.

[6] The 'minimal' here is meant to capture that we wish to commit to, and take a stand on, as few of the many interesting and important open questions that are still being debated about human normative psychology as we can, while still putting forward a view that is plausible, that captures many of the key points of agreement in the literatures we draw on. As will become clear, however, 'minimal' should not be interpreted as suggesting, for instance, minimal appeal to innate mental structure or content, or to specialized psychological machinery that goes beyond the domain general, barebones repertoire associated with blank slate models of the mind; for some discussion, see Stephan Linquist and Alex Rosenberg, "The Return of the 'Tabula Rasa': Reviewed of *Thought in a Hostile World: The Evolution of Human Cognition* by Kim Sterelny," *Philosophy and Phenomenological Research*, 74, 2 (2007): 476-497).

[7] For overviews of the recent emergence of the richly interdisciplinary field of empirical moral psychology, see John Doris and Stephen Stich "Moral psychology: Empirical approaches." The Stanford Encyclopedia of Philosophy (Summer 2006 Edition), Edward N. Zalta (Ed.), URL = http://plato.stanford.edu/archives/sum2006/entries/moral-psych-emp/, John Doris and The Moral Psychology Research Group (eds.), *The Moral Psychology Handbook*. New York: Oxford University Press (2010), Joshua Greene, *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them* (New York: Penguin Books, 2014), Valerie Tiberius, *Moral Psychology: A Contemporary Introduction* (Oxford, UK: Routledge, 2014) and Mark Alfano, *Moral Psychology: An Introduction* (Cambridge: Polity Press, 2016). We return to the issue of distinguishing distinctively moral norms and moral cognition from other varieties in section four.

[8] Indeed, on the Minimal Account many of the most important features of human normative psychology are underpinned by what Daniel Kahneman, *Thinking Fast and Slow* (New York: Farrar, Straus and Giroux, 2011) describes as System 1 mechanisms and processes; they are fast, intuitive, automatic, and unconscious.

[9] See Shaun Nichols and Ron Mallon, "Moral dilemmas and moral rules" *Cognition*, 100 3 (2006): 530 – 42 and Ron Mallon and Shaun Nichols, "Rules," *The Moral Psychology Handbook*, Eds. J. Doris et al. (New York: Oxford University Press 2010, pages 297 – 320) on rules. However, what we are delineating as normative psychology is almost certainly not *exhaustive* of the psychology of rules, as people can be well aware of rules that are not intuitively norms (the rules of chmess, for an extreme example provided by Daniel Dennett, "Higher-order truths about chmess," *Topoi* 1 (2006): 39–41), and also aware of rules which have no bearing on their own behavior. For example, a person might have read about the taboos and norms of cultures and societies to which she does not belong, and feels no motivation to enforce or comply with. In such a case, those rules are represented, perhaps more purely cognitively, in her mind, but in some component other than in her norm system.

[10] The quoted passage is from Maciej Chudek, Wanying Zhao, and Joseph Henrich, "Culture-Gene Coevolution, Large-Scale Cooperation, and the Shaping of Human Social Psychology" *Cooperation and Its Evolution*, Eds. Kim Sterelny, Richard Joyce, Brett Calcott, and Ben Fraser. Cambridge, MA: The MIT Press (2013): 425 – 458; also see Peter Richerson and Robert Boyd, *Not By Genes Alone: How Culture Transformed Human Evolution* (Chicago: University of Chicago Press, 2005), Maciej Chudek, and Joseph Henrich, 'Culture–

gene coevolution, norm-psychology and the emergence of human prosociality,' *Trends in Cognitive Sciences*, 15, 5 (2011): 218-226, and Michelle Gelfand and Joshua Jackson, "From one mind to many: the emerging science of cultural norms," *Current Opinion in Psychology*, 8 (2016): 175–181.

[11] Daniel Fessler and Edouard Machery, "Culture and cognition" In E. Margolis, R. Samuels, and S. P. Stich (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science.* Oxford: Oxford University Press (2012), pages 503-527.

[12] In the main text we are making a fairly straightforward point about the similar division of explanatory labor between appeals to innate versus learned traits in explanations of both linguistic and normative capacities. The strengths and weaknesses of a more developed analogy between language, on the one hand, and social rules and morality, on the other, have been well explored in recent years. For a book length development and defense of the analogy from a cognitive scientific and legal point of view, see John Mikhail, *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment* (Cambridge University Press, 2011), and for a shorter version John Mikhail, "Universal Moral Grammar" *Trends in cognitive science,* 11 (2007): 143-152. Also see Erica Roedder and Gil Harman, "Linguistics and Moral Theory," *The Moral Psychology Handbook*, (Eds. J. Doris et al. New York: Oxford University Press, 2010, pages 273 – 296), Susan Dwyer, Bryce Huebner and Marc Hauser, "The linguistic analogy" *Topics in Cognitive Science*, 2, 3 (2009): 486–510, and Marc Hauser, Liane Young and Fiery Cushman, "Reviving Rawls' linguistic analogy," In W. Sinnott-Armstrong (Ed.), *Moral psychology: Vol. 2. The cognitive science of morality*. Cambridge, MA: MIT Press (2008). For a dissenting voice concerning the utility of the linguistic analogy, see Jesse Prinz, "Resisting the Linguistic Analogy: A Commentary on Hauser, Young, and Cushman," In W. Sinnott-Armstrong (Ed.), *Moral psychology: Vol. 2. The Cognitive Science of Morality: Intuition and Diversity*. Cambridge, MA: MIT Press (2008), and for an orthogonal but suggestive inversion of the usual explanatory order between language and morality (broadly construed as norm-governed cooperation), see Peter Richerson and Robert Boyd, "Why Language Possibly Evolved," *Biolinguistics* 4, 2–3 (2010): 289–306.

[13] Chandra Sripada and Stephen Stich, "A framework for the psychology of norms" P. Carruthers, S. Laurence, & S. Stich (eds.), *The innate mind Vol 2: Culture and cognition.* New York: Oxford University Press, 2007, pages 280-301.

[14] Marco Schmidt, Lucas Butler, Julia Heinz and Michael Tomasello, "Young Children See a Single Action and Infer a Social Norm: Promiscuous Normativity in 3-Year-Olds," *Psychological Science*, (2016): 1-11.

[15] For different kinds of arguments in favor of the claim that not all intrinsic motivation need be innate or have innately specified aims, and that ultimate ends can be acquired and changed in the ways implied by this claim about human normative psychology, see e.g. Chandra Sripada, 'Adaptationism, Culture, and the Malleability of Human Nature,' *The Innate Mind Vol 3.: Foundations and Future Horizons*, Eds. Peter Carruthers, Stephen Laurence and Stephen Stich. New York: Oxford University Press, 2007, 311 - 329), Stephen Stich, "Why there might not be an evolutionary explanation for psychological altruism," *Studies in History and Philosophy of Biological and Biomedical Sciences* 56 (2016): 3-6, and, from a different angle, Elijah Millgram, *The Great Endarkenment: Philosophy for an Age of Hyperspecialization* (New York: Oxford University Press, 2015) (especially chapters 3 and 10).

[16] See, e.g. Scott Atran, "The moral logic and growth of suicide terrorism," *The Washington Quarterly*, 29, 2 (2006): 127-147.

[17] See Richard Shweder, Nancy Much, Manamohan Mahapatra and Lawrence Park, The "big three" of morality (autonomy, community, and divinity), and the "big three" explanations of suffering. In A. Brandt & P. Rozin (eds.), *Morality and Health*. Routledge. (1997), Paul Rozin, Laura Lowery, Sumio Imada and Jonathan Haidt, "The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity)" *Journal of Personality and Social Psychology*, 76, 4 (1999): 574-586; c.f. Jonathan Haidt and Craig Joseph, "The moral mind: How 5 sets of innate moral intuitions guide the development of many culture-specific virtues, and perhaps even modules," In *The Innate Mind*, vol. 3, ed. P. Carruthers, S. Laurence, and S. Stich (2007), 367–391. New York: Oxford University Press.

[18] In confining our characterization of some of the key mechanisms of the human norm system, we are sketching part of a *proximate* psychological explanation of human normativity. There are also complimentary *ultimate* explanations of how those psychological mechanisms evolved, the most promising of which appeal to culture-driven genetic evolution. Most of the details of these explanations fall beyond the scope of this paper, though see Maciej Chudek, and Joseph Henrich, 'Culture–gene coevolution, norm-psychology and the emergence of human prosociality,' *Trends in Cognitive Sciences*, 15, 5 (2011): 218-226, Maciej Chudek, Wanying Zhao, and Joseph Henrich, "Culture-Gene Coevolution, Large-Scale Cooperation, and the Shaping of Human Social Psychology" *Cooperation and Its Evolution*, Eds. Kim Sterelny, Richard Joyce, Brett Calcott, and Ben Fraser.

Cambridge, MA: The MIT Press (2013): 425 – 458, and Kim Sterelny, "Cooperation, Culture, and Conflict" *The British Journal for the Philosophy of Science*, 67, 1 (2014): 1 – 31 for recent paper length discussions, Joseph Henrich, *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter* (Princeton, NJ: Princeton University Press, 2015) for a book length discussion of the distinct processes of cultural evolution and culture-driven genetic evolution, and Daniel Kelly and Patrick Hoburg, 'A Tale of Two Processes: On Joseph Henrich's *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*,' *Philosophical Psychology* (2017) for an overview for a philosophical audience.

[19] It is entirely possible that not *all* behavior-guiding rules of social interactions are socially learned. Whether any particular such rules, perhaps some putatively moral rules that prohibit incest or battery, are part of the innately specified human psychological endowment is one of those issues on which we take no stand, though see Daniel Kelly, *Yuck! The Nature and Moral Significance of Disgust* (Cambridge, MA: The MIT Press, 2011, page 98, and John Mikhail, *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment* (Cambridge University Press, 2011), for discussion and references for each case.

[20] See Peter Richerson and Robert Boyd, *Not By Genes Alone: How Culture Transformed Human Evolution* (Chicago: University of Chicago Press, 2005) for an overview; c.f. Samuel Bowles, Jung-Kyoo Choi and Astrid Hopfensitz, "The coevolution of individual behaviors and group level institutions," *Journal of Theoretical Biology*, 223 (2003): 135–-147.

[21] Gerald Gaus and Shaun Nichols, "Moral Learning in the Open Society: The Theory and Practice of Natural Liberty," *Social Philosophy and Policy* (forthcoming).

[22] See Robert Boyd and Peter Richerson, "Punishment allows the evolution of cooperation (or anything else) in sizable groups," *Ethnology and Sociobiology* 13(1992): 171–95, Ernest Fehr and Simon Gachter, "Altruistic punishment in humans," *Nature,* 415 (2002): 137-140, and Ernest Fehr and Urs Fischbacher, "Third party punishment and social norms," *Evolution and Human Behavior*, 25, 2 (2004): 63-87.

[23] Such feedback loops and endogenously maintained equilibria are common properties of complex self-organizing systems. What is interesting and perhaps unique to groups of human beings, however, are some of the central mechanisms by which stable equilibria are achieved and sustained, namely culturally transmitted norms, punishment, and human normative psychology. But also, this is why social arrangements can be durable without any Leviathan-like entity to serve as a foundational stabilizer or ultimate norm enforcer; stable social arrangements stabilize themselves.

[24] See especially Robert Boyd and Peter Richerson, "Gene-Culture Coevolution and the Evolution of Social Institutions", In: *Better than Conscious? Decision Making, the Human Mind, and Implications for Institutions.* (C. Engel and W. Singer eds, MIT Press, Cambridge, 2008), 305-324, Peter Richerson and Joseph Henrich, "Tribal Social Instincts and the Cultural Evolution of Institutions to Solve Collective Action Problems," *Cliodynamics*: *The Journal of Theoretical and Mathematical History*, 3, 1 (2012): 38-80, and Robert Boyd, "A Different Kind of Animal: How Culture Made Humans Exceptionally Adaptable and Cooperative" The Tanner Lectures on Human Values, (manuscript).

[25] See Joseph Henrich, Robert Boyd and Peter Richerson, 'Five Misunderstandings about Cultural Evolution,' *Human Nature,* 19 (2008): 119-137 for diagnosis and correction of common types of confusion, though.

[26] No attempt to provide a strict and univocal definition of a culture has won widespread acceptance. We have been emphasizing the role of social learning a bit more, but Grant Ramsey, "Culture in humans and other animals," *Biology and Philosophy* 27 (2013): 457-479 defends an explication that is useful and largely aligns with the way we are understanding the term: "Culture is information transmitted between individuals or groups, where this information flows through and brings about the reproduction of, and a lasting change in, the behavioral trait" (ibid, 466).

[27] See Grant Ramsey and Andreas De Block, "Is Cultural Fitness Hopelessly Confused?" *The British Journal of the Philosophy Science*, (2015): 1-24 for discussion of the idea of cultural fitness, and Robert Boyd and Peter Richerson, "Memes: Universal acid or better mousetrap?" In *Darwinizing Culture*, ed. R. Aunger. Cambridge: Cambridge University Press, 2000 for a convincing case that there are better tools for theorizing about cultural evolution than the most well known one, the "meme."

[28] See Peter Richerson and Robert Boyd, *Not By Genes Alone: How Culture Transformed Human Evolution* (Chicago: University of Chicago Press, 2005, chapter 3 for a nice presentation of a simple model that illustrates the idea, for instance how "Conformity bias at the level of the individual leads to reasonably accurate replication at the population level" (ibid, 86).

[29] (Joseph Henrich and Francesco Gil-White, 'The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission,' *Evolution and Human Behavior*, 22 (2001): 165-196.

[30] Michael Muthukrishna, Thomas Morgan and Joseph Henrich, "The when and who of social learning and conformist transmission" *Evolution and Human Behavior* 37 (2016): 10–20.
[31] See Dan Sperber, *Explaining culture: A naturalistic approach* (New York: Blackwell Publishers, 1996), Scott Atran, "Folk Biology and the Anthropology of Science: Cognitive Universals and Cultural Particulars," *Behavioral and Brain Sciences* 21 (1998): 547-609, and Pascal Boyer, "Cognitive Tracks of Cultural Inheritance: How Evolved Intuitive Ontology Governs Cultural Transmission", *American Anthropologist* 100 (1999): 876-889.
[32] Rick O'Gorman, David Sloan Wilson and Ralph Miller, "An evolved cognitive bias for social norms" *Evolution and Human Behavior*, 29 (2008): 71–78.
[33] Shaun Nichols, "On the genealogy of norms: a case for the role of emotion in cultural evolution" *Philosophy of Science*, 69 (2002): 234-255, c.f. Chip Heath, Chris Bell and Emily Sternberg, Emotional selection in memes: The case of urban legends. *Journal of Personality and Social Psychology* 81 (2001): 1028–1041.
[34] For a range of perspectives on this idea, see Jonathan Gottschall, *The Storytelling Animal: How Stories Made Us Human* (New York, Houghton Mifflin Harcourt, 2012), Merlin Donald, "The slow process: A hypothetical cognitive adaptation for distributed cognitive networks," *Journal of Physiology – Paris*, 101 (2006): 214 – 222 and Kristien Tylén et al, "Brains striving for coherence: Long-term cumulative plot formation in the default mode network" *NeuroImage* 121 (2015) 106–114, Cristina Bicchieri and Peter McNally, P. "Shrieking Sirens Schemata, Scripts, and Social Norms: How Change Occurs" (Manuscript) Manjana Milkoreit, "The promise of climate fiction: Imagination, storytelling, and the politics of the future," (in *Reimagining Climate Change*, ed. P. Wapner and H. Elvner, New York: Routledge, 2016.
[35] Deirdre Barrett, *Supernormal Stimuli: How Primal Urges Overran Their Evolutionary Purpose* (New York: W. W. Norton and Co., 2010, Andreas De Block and Bart Du Laing, "Amusing ourselves to death? Superstimuli and the evolutionary social sciences," *Philosophical Psychology*, 23, 6 (2010): 821 – 843, c.f. Daniel Kelly, 'Moral Cheesecake, Evolved Psychology, and the Debunking Impulse', to appear in the *Routledge Handbook of Evolution and Philosophy*, (Ed. R. Joyce, New York: Routledge Press, forthcoming).
[36] Christophe Heintz, "Institutions as Mechanisms of Cultural Evolution: Prospects of the Epidemiological Approach," *Biological Theory,* 2, 3 (2007): 244–249.
[37] We will continue to follow Richerson, Boyd and Henrich in continuing to call these "instincts" but acknowledge that this terminology might be misleading. These psychological capacities and the mechanisms that underlie them are much more sophisticated, sensitive to subtle social cues, and productive of flexible inferences and behavior than the connotations of the term "instinct" suggest. We are thankful to Peter Railton for pressing us on this point. "Instinct" functions primarily to emphasize the fact that the traits in question are inherited genetically, rather than culturally.
[38] Cristina Moya and Joseph Henrich, Culture–gene coevolutionary psychology: cultural learning, language, and ethnic psychology," *Current Opinion in Psychology*, 8 (2016): 112–118.
[39] On norm enforcement see Marco Schmidt, Hannes Rakoczy and Michael Tomasello, "Young children enforce social norms selectively depending on the violator's group affiliation," *Cognition*, 124 (2012): 325–333 and on reputation management see Jan Engelmann, Harriet Over, Esther Herrmann and Michael Tomasello, Young children care more about their reputation with ingroup members and potential reciprocators," *Developmental Science*, 16, 6 (2013): 952–958.
[40] See especially Carsten De Dreu, "Oxytocin modulates cooperation within and competition between groups: an integrative review and research agenda," *Hormones and Behavior*, 61(2012): 419–28 and Carsten De Dreu, and Mariska Kret,"Oxytocin conditions intergroup relations through upregulated in-group empathy, cooperation, conformity, and defense," *Biological Psychiatry,* 79 (2016): 165–73. Daniel Balliet, Junhui Wu, and Carsten De Dreu, "Ingroup Favoritism in Cooperation: A Meta-Analysis," *Psychological Bulletin*, 140, 6 (2014): 1556–1581 provide a largely vindicating meta-analysis of work on ingroup favoritism and Stephanie Hechler, Franz Neyer and Thomas Kessler, "The infamous among us: Enhanced reputational memory for uncooperative ingroup members" *Cognition*, 157 (2016): 1–13 explore differences in how people remember ingroup versus outgroup members. Also see Charles Efferson, Rafeal Lalive and Ernest Fehr, "The Coevolution of Cultural Groups and Ingroup Favoritism" *Science*, 321 (2008): 1844 – 1849 for an account of the co-evolutionary back and forth that selected for cultural groups and ingroup favoritism, Karla Hoff, Mayuresh Kshetramade and Ernest Fehr, "Caste and Punishment: The Legacy of Caste Culture in Norm Enforcement," *The Economic Journal*, 121 (2011): 449–475 for a closer look at a particular case, and how caste membership influences norm enforcement, and Cristina Moya and Robert Boyd, "Different Selection Pressures Give Rise to Distinct Ethnic Phenomena: A Functionalist Framework with Illustrations from the Peruvian Altiplano," *Human Nature*, 26, 1 (2015): 1 – 27 for useful distinctions between putatively different kinds of 'groups'.

[41] Joey Cheng, Jessica Tracy, Tom Foulsham, Alan Kingston, and Joseph Henrich, "Two Ways to the Top: Evidence That Dominance and Prestige Are Distinct Yet Viable Avenues to Social Rank and Influence," *Journal of Personality and Social Psychology*, 104, 1 (2012): 103–125.
[42] Also see Maciej Chudek, Sarah Heller, Susan Birch, and Joseph Henrich, "Prestige-biased cultural learning: bystander's differential attention to potential models influences children's learning," *Evolution and Human Behavior* 33 (2012): 46–56, c.f. Cristina Bicchieri, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms* (New York: Cambridge University Press, 2016, chapter 5.
[43] Daniel Kelly, *Yuck! The Nature and Moral Significance of Disgust* (Cambridge, MA: The MIT Press, 2011) chapter 4 and Daniel Kelly, 'Moral Disgust and The Tribal Instincts Hypothesis,' *Cooperation and Its Evolution* (Eds. Kim Sterelny, Richard Joyce, Brett Calcott, and Ben Fraser. Cambridge, MA: The MIT Press 2013: pages 503 - 524
[44] Peter Richerson, "Human Cooperation is a Complex Problem with Many Possible Solutions: Perhaps All of Them Are True!" *Cliodynamics: The Journal of Theoretical and Mathematical History*, 4, 1 (2013): 139–152.
[45] For more discussion here see Kim Sterelny, "The Evolution and Evolvability of Culture" *Mind & Language*, 21 (2006): 137 – 165 (cited by Henry Richardson, "Revising Moral Norms: Pragmatism and the Problem of Perspicuous Description," in C. Bagnoli (ed). *Constructivism in Ethics*, Cambridge University Press, (2013)), Peter Richerson, Dwight Collins, and Russell Genet, "Why managers need an evolutionary theory of organizations," *Strategic Organization* 4, 2 (2006): 201-211, and Peter Richerson and Joseph Henrich, "Tribal Social Instincts and the Cultural Evolution of Institutions to Solve Collective Action Problems," *Cliodynamics*: *The Journal of Theoretical and Mathematical History*, 3, 1 (2012): 38-80.
[46] Page 319, Robert Boyd and Peter Richerson, "Gene-Culture Coevolution and the Evolution of Social Institutions", In: *Better than Conscious? Decision Making, the Human Mind, and Implications for Institutions*. (C. Engel and W. Singer eds, MIT Press, Cambridge, 2008), 305-324
[47] Peter Richerson et al., "Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence *Behavioural and Brain Sciences,* 39 (2016): 1-68.
[48] Charles Darwin, *The Descent of Man and Selection in Relation to Sex* (D. Appleton and Co., 1871), page 160.
[49] We will not discuss this idea in detail, but will note that it is easy to misunderstand as a hypothesis about "groups" of people defined as such in virtue of genetic relatedness, and selection favoring genetic adaptations that produce behaviors favoring the group at the expense of those individuals. It would thus be subject to the objections to group selection that dominated evolutionary biology in the later half of the twentieth century. Hopefully our discussion in the main text is enough to head off this common misunderstanding, as the hypothesis is in fact about competition of *culture* at a macro level, or selection between clusters of cultural variants and packages of norms, rather than between genetic adaptations. See Peter Richerson et al., "Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence *Behavioural and Brain Sciences,* 39 (2016): 1-68 for a sophisticated modern formulation of the hypothesis, presentation of a wide array of evidence in support of it, and critical discussion. See Gerald Gaus, "The Egalitarian Species," *Social Philosophy and Policy*, 31, 2 (2015): 1-27 for an exploration of implications of the updated version of the hypothesis for Hayek's political philosophy.
[50] Compare "standards of appropriate behavior for actors with a given identity" in Martha Finnemore and Kathryn Sikkink, "International Norm Dynamics and Political Change" *International Organization* 52 (1998): 887-917, page 891, as cited in Leigh Raymond, S. Laurel Weldon, Daniel Kelly, Ximena Arriaga, and Ann Marie Clark, 'Making Change: Norms and Informal Institutions as Solutions to "Intractable" Global Problems', *Political Research Quarterly*, 67, 1 (2013): 197 - 211) and "learned behavioral standards shared and enforced by a community" in Maciej Chudek, and Joseph Henrich, 'Culture–gene coevolution, norm-psychology and the emergence of human prosociality,' *Trends in Cognitive Sciences*, 15, 5 (2011): 218-226, page 218.
[51] In other words, if there is any stable, interesting, and important set of features that distinguish moral norms from the rest (or moral cognition from the rest, for that matter), finding and specifying it remains a deeply vexed enterprise. See Daniel Kelly, Stephen Stich, Kevin Haley, Serena Eng and Daniel Fessler, "Harm, affect, and the moral/ conventional distinction" *Mind and Language* 22, 2 (2007): 117–131, Daniel Kelly and Stephen Stich, 'Two Theories of the Cognitive Architecture Underlying Morality,' *The Innate Mind Vol 3.: Foundations and Future Horizons*, Eds. Peter Carruthers, Stephen Laurence and Stephen Stich. (New York: Oxford University Press. 2007, page 348-366, and Walter Sinnott-Armstrong and Thalia Wheatley, 'The Disunity of Morality and Why it Matters to Philosophy,' *The Monist* 95, 3 (2012): 355 – 377 for discussions of the problems that arise for attempts to draw such a distinction based on currently available empirical theories, and skepticism about the project itself. See Jonathan Haidt and Jesse Graham, "When morality opposes justice: Conservatives have moral intuitions that Liberals may not recognize," *Social Justice Research,* 20 (2007): 98–116 and Jesse Graham, Jonathan Haidt and Brian Nosek. B. (2009). "Liberals and Conservatives Rely on Different Sets of Moral

Foundations," *Journal of Personality and Social Psychology*, 96(5): 1029–1046 for evidence that the folk seem to conceive of the scope of "morality" differently depending on political orientation and culture; also see Renatas Berniūnas, Vilius Dranseika, and Paulo Sousa, "Are there different moral domains? Evidence from Mongolia,"*Asian Journal of Social Psychology*, 19 (2016): 275–282 for recent cross cultural evidence from Mongolia. See Joseph Henrich, Steven Heine and Ara Norenzayan, "The weirdest people in the world." *Behavioral and Brain Sciences* 33 (2010): 61-135 for a general albeit indirect account that can suggest why, although it might seem obvious that there is such a distinction, the tendency of us WEIRDos to confer special status on some putatively distinctive subset of norms we designate as "moral" is likely a culturally parochial trait rather than a universal one.

[52] See for instance the chart on page 41 of Cristina Bicchieri, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms* (New York: Cambridge University Press, 2016).

[53] For instance, "To uncover the reasons why a collective behavior survives, we have to look beyond attitudes to the beliefs and conditional preferences of those who engage in it. This is why I like to use almost exclusively preferences and expectations in my analysis of norms. They are easy to measure, and measuring them lets us meaningfully classify collective behaviors" (ibid, 10).

[54] Peter Richerson and Joseph Henrich, "Tribal Social Instincts and the Cultural Evolution of Institutions to Solve Collective Action Problems," *Cliodynamics*: *The Journal of Theoretical and Mathematical History*, 3, 1 (2012): 38-80, page 67, our emphasis.