# A pluralistic framework for the psychology of norms

Evan Westra[1] & Kristin Andrews[2]

[1]Department of Philosophy, Purdue University

[2]Department of Philosophy, York University

Forthcoming in *Biology & Philosophy*

**Abstract:** Social norms are commonly understood as rules that dictate which behaviors are appropriate, permissible, or obligatory in different situations for members of a given community. Many researchers have sought to explain the ubiquity of social norms in human life in terms of the psychological mechanisms underlying their acquisition, conformity, and enforcement. Existing theories of the psychology of social norms appeal to a variety of constructs, from prediction-error minimization, to reinforcement learning, to shared intentionality, to domain-specific adaptations for norm acquisition. In this paper, we propose a novel methodological and conceptual framework for the cognitive science of social norms that we call *normative pluralism*. We begin with an analysis of the (sometimes mixed) explanatory aims of the cognitive science of social norms. From this analysis, we derive a recommendation for a reformed conception of its explanandum: a minimally psychological construct that we call *normative regularities*. Our central empirical proposal is that the psychological underpinnings of social norms are most likely realized by a heterogeneous set of cognitive, motivational, and ecological mechanisms that vary between norms and between individuals, rather than by a single type of process or distinctive norm system. This pluralistic approach, we suggest, offers a methodologically sound point of departure for a fruitful and rigorous science of social norms.

**Keywords:** social norms, normative cognition, pluralism, punishment, conformity.

## *1. Lessons for the cognitive science of social norms from empirical moral psychology*

By the second decade of the 21st century, a crisis of sorts had emerged in the field of empirical moral psychology. By that time, psychologists, neuroscientists, and empirically inclined philosophers had developed a number of competing theories about the cognitive processes that produce moral judgments (Turiel 1983; Greene et al. 2001; Nichols 2004; Prinz 2007; Dwyer et al. 2010; Mikhail 2011; Gray et al. 2012). Despite a shared sense that any such theory should explain certain paradigmatic kinds of judgment – for example, that it is permissible to flip the switch in the traditional version of the Trolley Problem, or that intentionally harming others is wrong – many of these theorists struggled to characterize the boundaries of their subject matter. Most notably, a growing chorus of voices started to question the distinction between moral and conventional reasoning (Turiel 1983), pointing out that, in some cases, reasoning about paradigmatically moral transgressions like murder looks quite conventional, while in other cases reasoning about putatively conventional matters takes on a moralistic character (Maibom 2005; Kelly et al. 2007). Others argued that Turiel's characterization of the moral domain was artificially narrow due its commitment to moral rationalism and focus on WEIRD populations (Shweder et al. 1987; Haidt et al. 1993; Haidt 2001). If we instead adopted a sentimentalist approach or paid more attention to cross-cultural diversity, these critics argued, the moral domain might also encompass a much wider set of values, including purity, ingroup loyalty, and respect for authority (Nichols 2004; Prinz 2007; Graham et al. 2009). Around the same time, a number of philosophers and neuroscientists began to comment on the apparent heterogeneity in the cognitive processes and neural mechanisms engaged in paradigmatic forms of moral judgment, writing articles with skeptical titles like "Where in the brain is moral cognition? Everywhere and maybe nowhere" (Young and Dungan 2012), "Are moral judgments unified?" (Sinnott-Armstrong and Wheatley 2014), "Is morality an elegant machine or a kludge?" (Stich 2006); Joshua Greene put the point succinctly

when he opined that "the field of moral cognition does not study a distinctive set of cognitive processes" (Greene 2015, p. 40). Despite the fact that the field of moral psychology was flourishing, it seemed that moral psychologists harbored considerable doubts about what – if anything – makes cognition *moral*.[1]

These concerns prompted a number of authors to doubt that "moral cognition" really picked out a natural psychological kind at all (Stich 2019). The problem was twofold: on the one hand, it was unclear how to define the boundaries of moral cognition such that it might be clearly distinguished from its non-moral neighbors. On the other hand, putative cases of moral cognition did not seem to have all that much in common with one another, whether at the level of content or at the level of underlying processes. While this fuzziness has not prevented moral psychologists from making important advances in specific kinds of morally relevant thinking – such as empathy (de Waal 2008), character (Miller 2014), self-control (Sripada 2020), and the concept of the moral self (Strohminger and Nichols 2014) – it is no longer clear what all these instances of moral thinking have in common. The borders of moral cognition seem to be determined more by the pragmatic interests of moral psychologists and philosophers than by any real joints in nature.[2]

Against this backdrop, many researchers have begun to search for different ways to carve up the same empirical territory. One very promising avenue of research has been to shift focus towards the psychology of *social norms* (Cialdini et al. 1991; Bicchieri 2006; Sripada and Stich 2006; Colombo 2014; Bicchieri 2017; Kelly and Davis 2018; Andrews 2020; Birch 2020; Fitzpatrick 2020).[3] This research project shifted away from moral thinking as such and towards the broader tendency to conform to and enforce social rules that dictate which behaviors are required, allowed, or forbidden for members of a given community (Kelly and Setman 2020). Norms, so construed, might govern prototypical moral domains concerning matters of harm and fairness, while also regulating a much wider range of so-called "conventional" activities, from rules about queuing at the grocery store to the proper preparation of food. While this approach leaves open the possibility that moral thinking might constitute an empirically distinct type of normative cognition, it does not start with the *a priori* assumption that this must be the case. Indeed, it would be no embarrassment at all if the idea of a distinctively "moral" psychology turned out to be a largely WEIRD phenomenon, a product of the West's peculiar history of normative institutions (Henrich 2020). Rather, the new cognitive science of norms takes heterogeneity and cross-cultural diversity in the contents of normative thinking as its starting point.

One of the central goals of this new, growing program of research has been to give an account of the mechanisms and processes that undergird our adherence to social norms. Some of these accounts take a "cognitive-evolutionary" approach (Kelly and Setman 2020), situating their theories of socially normative cognition in the broader framework of gene-culture co-evolution defended by authors like Robert Boyd, Peter Richerson, and Joseph Henrich (Richerson and Boyd 2005; Henrich 2017). These accounts usually posit a distinctive "norm system" or "norm psychology" constituted by a set of mechanisms and adaptive heuristics for acquiring the norms of one's local community (for example, a prestige bias to preferentially learn from high status individuals (Chudek et al. 2012)), as well as intrinsic motivational processes that dispose agents to conform to these norms and enforce them on others (Sripada and Stich 2006; Chudek and Henrich 2011; Kelly and Davis 2018). Others have argued

[1] For some recent rejoinders to these doubts, see Kumar (2016) and Curry et al. (2019).

[2] Understood this way, moral psychology might be best understood as a form of applied psychology like the psychology of sport, or consumer behavior, which do not presume that their objects of study carve nature at its joints.

[3] All references to "norms," "normativity," or "normative cognition" in this paper should be understood as referring to social norms, unless we specify otherwise.

that the psychology of social norms develops from a more basic capacity for shared intentionality, which enables children to acquire the conceptual abilities necessary for representing normative rules (Schmidt and Rakoczy 2019; Tomasello 2019). Still others have suggested that the mechanisms of norm psychology might be explained in terms of domain-general processes, such as reinforcement learning (Colombo 2014), prediction-error minimization (Veissière et al. 2019; Theriault et al. 2021), and model-based control (Birch 2020).

We find many of these proposals to be quite plausible, and we suspect that several of them do correctly describe mechanisms that support some social norms. However, we worry that there is something about this approach to the psychology of social norms that risks repeating the mistakes of recent moral psychology. Current models of normative cognition have by and large learned from only one of the two lessons that emerged from the crisis in moral cognition: while they have embraced the critique that there are no clear boundaries between moral cognition and normative thinking more broadly, they have not quite grappled with the possibility that the mechanisms of normative cognition might themselves be quite heterogenous, more "kludge" than "elegant machine" (Stich 2006).

Instead, most of the existing proposals in this area seem to begin from the tacit assumption that there is something cognitively or motivationally homogeneous or unified about the psychology of social norms, a discrete set of mechanisms or processes that underpin all forms of social norm acquisition, conformity, and enforcement. For example, Sripada and Stich begin their seminal paper on the psychology of normative cognition with the starting assumption that norms constitute a *natural kind* (Sripada and Stich 2006). In their articulation of the cognitive-evolutionary approach to normative cognition, Kelly and Davis (2018) sketch out the contours of a functionally specific, adaptive *norm system* for the transmission of cultural information. In his skill-based account of the evolution of norms, Jonathan Birch (2020) proposes the hypothesis that all forms of normative cognition share a common evolutionary history grounded in mechanisms that originally evolved in the context of standardized toolmaking (Birch 2020). Similar assumptions of homogeneity appear to pervade attempts to explain norm-guided cognition by appealing to domain-general processes (Colombo 2014; Bicchieri and McNally 2018; Veissière et al. 2019). The main aim of this paper will be to argue that this way of thinking about the cognitive science of social norms is too narrow. While it might make sense to treat the community-level phenomena that we explain in terms of social norms as a functional category, it is most likely not the product of a unified cognitive system or a homogeneous set of processes. The psychology of social norms, we will argue, is pluralistic.

Notably, researchers in this area have quite explicitly recognized a few forms of pluralism about the psychology of norms. In a recent review of the many different ways that researchers have attempted to carve up the normative domain, Elizabeth O'Neill explicitly advocates for a pluralistic approach to the way we *classify* social norms based on our pragmatic goals (O'Neill 2017) – a suggestion also echoed by Daniel Kelly, who calls the proliferation of normative classification schemes "an embarrassment of riches" (Kelly forthcoming). Norm theorists have also been happy to accept a form of pluralism about the *contents* of social norms, acknowledging that normative cognition might govern an extremely wide array of culturally variable social practices. But this pluralism about the best way to classify norms and norm content is belied by apparent assumptions about homogeneity at the level of cognitive mechanisms. This is where we see the greatest need for a pluralistic perspective.

The kind of pluralism we have in mind for the psychology of norms is modeled after pluralistic approaches to folk psychology and social cognition (Andrews 2012; Fiebich and Coltheart 2015; Spaulding 2018; Andrews et al. 2020). While early approaches to that field took it for granted that we came to understand the social world by predicting and explaining behavior in terms of propositional attitudes, pluralistic approaches have stressed the importance of alternative strategies for prediction

and explanation, including representations of the situation, traits, stereotypes and (notably) social norms; pluralists also emphasized the importance of regulative or "mindshaping" processes that did not involve prediction or explanation at all (McGeer 2007; Zawidzki 2013). The key insight of pluralistic approaches to social cognition has been that understanding the social world is a messy, complex process, and that there are a variety of ways of *doing* social cognition. Similarly, we suggest there may be many ways of *doing* social norms, and this should be reflected in their psychological study.[4]

In the remainder of this paper, we sketch out the contours of a pluralistic approach to the psychology of social norms. First, we'll motivate and lay a reformed conception of the explanatory goals of the cognitive science of social norms and how they should be studied. Then, using this framework, we'll make the case for our substantive psychological proposal about socially normative cognition. We then conclude by considering a few potential objections to our proposed framework.

### 2.1. The explananda of the cognitive science of social norms

Our case for normative pluralism begins with a very basic question: what are the phenomena that the cognitive science of social norms aims to explain?

One way to answer this question might be to think of social norms as a type of *cultural attractor*, following the epidemiological approach to culture (Sperber 1996; Heintz 2018). Cultural attractors are statistical types that describe the distribution of token cultural phenomena as they occur within the minds of individuals or in their habitats. They are macro-level structures that emerge from the accumulation of many individual-level social interactions mediated by nonrandom psychological and ecological processes, or "factors of attraction." Importantly, cultural attractors have no causal or explanatory power of their own. Rather, they are a way of describing phenomena in need of explanation. The real explanatory work in cultural attraction theory comes from the identification of the psychological and ecological processes that causally contribute to the emergence of cultural attractors (Scott-Phillips et al. 2018).

If social norms are understood as a kind of cultural attractor, then this would mean that the goal of the cognitive science of social norms would be to identify the various psychological and ecological factors of attraction that produce them. However, what distinguishes social norms from other cultural attractors is not immediately obvious. As a rough, first-pass characterization, we shall say that are social norms generally consist in patterns of behavioral conformity maintained to some degree by social pressure or sanctions. They are widely believed to be ubiquitous in everyday life, present in all human populations and manifested across an extremely diverse set of behavioral domains, from children's games (Schmidt et al. 2016a), to standardized tool production (Birch 2020), to littering (Cialdini et al. 1990), all the way to female genital cutting and child marriage practices (Bicchieri and

[4] There are some signs that certain norm theorists could be open to this kind of cognitive-level pluralism. For example, Theriault and colleagues distinguish between four different potential motivations for social conformity: *informational influence*, when a person copies others because they are perceived to be knowledgeable; *reputation-seeking*, when a person conforms in order to seek praise or avoid blame; *social obligation*, when a person feels obligated to conform to others' expectations; and *moral obligation*, when a person conforms because they are motivated by independently held values or convictions (Theriault et al. 2021). Daniel Kelly has also distinguished between at least two different cognitive pathways for adopting a norm, which he calls *internalization* and *avowal*. Norm internalization occurs when a normative rule is adopted via automatic, functionally specific "System 1" processes that also cause us to become intrinsically motivated to conform to and enforce that rule. Norm avowal occurs when a norm is intentionally adopted via slow, effortful "System 2" processes, with the explicit goal of self-regulation (Kelly 2022).

McNally 2018). They are also thought to have played an important role in the evolution of human cooperation (Boyd and Richerson 1992; Henrich 2017), and they are an important factor in the explanation of human cultural variation (Gelfand and Jackson 2016; Henrich 2020). We will refer to this broad class of cultural attractors as *normative regularities*.

Alternatively, the cognitive science of social norms might be engaged in an explanatory project more in line with what Robert Cummins has called *functional analysis* (Cummins 2000). The primary explananda of functional analysis (and, according to Cummins, of psychology more generally) are psychological *capacities*, which are complex dispositional properties specified at Marr's computational level of analysis (Marr 2010); familiar psychological capacities include such as the capacity to represent cardinal numbers, the capacity to represent beliefs, and the capacity for episodic memory. Functional analysis proceeds by taking a well-specified capacity of this sort and then breaking it down into a series of better understood dispositions or processes that enable the overarching capacities to be realized, thus revealing its internal functional structure.

In our view, all of the accounts of the psychology of social norms that we have discussed thus far – including theories that invoke propositional attitudes (Bicchieri 2006), predictive coding (Veissière et al. 2019; Theriault et al. 2021), reinforcement learning (Colombo 2014), and domain-specific cognitive adaptations (Sripada and Stich 2006; Chudek and Henrich 2011; Kelly and Davis 2018) – are best understood as engaged in this latter kind of project. For these theorists, the cognitive science of norms is primarily concerned with specifying and functionally analyzing a species-typical psychological capacity to acquire, enforce, and comply with the norms of one's community. We will call this capacity *ought-thought*.

What is the relationship between normative regularities this posited capacity for ought-thought? For theorists who adopt the functional analysis approach, the answer is simple: the latter explains the former. The species-typical disposition for ought-thought is what accounts for the ubiquity of normative regularities across human communities. The phenomena that we call social norms are thus unified by their common origins in our capacity for ought-thought. Consequently, the activity of ought-thought processes is also criterial: if a pattern of behavioral conformity and sanction is *not* driven by ought-thoughts, then it is not a true social norm, and not part of the explananda for the cognitive science of norms (e.g. Bicchieri, 2017; Theriault et al., 2021).

A major challenge for the functional analysis approach to the cognitive science of social norms is that there is little consensus about what the capacity for ought-thought consists in. Norm theorists specify this capacity in many different ways, each focusing on a slightly different cluster of motivational, representational, and phenomenological features. For example, some authors emphasize the centrality of representations of *rules* in guiding social norm conformity (Schlingloff and Moore 2017; Fitzpatrick 2020). Some stress the importance of *intrinsic motivations* to conform to and enforce social norms (Kelly 2020). Others define social norms as constituted by certain types of *beliefs* and *preferences* (Bicchieri 2017). Still others argue that genuine social norm conformity is accompanied by a distinctive kind of *affective experience* (Theriault et al. 2021). Michael Tomasello's account emphasizes the importance of representations of shared social goals and collective intentionality (Tomasello 2020). These differences in the initial characterizations of ought-thoughts are in turn reflected in the functional analyses offered by different theories, leading to widely varying proposals about underlying cognitive architecture.

One domain where this lack of consensus about the specification of ought-thought has led to significant empirical problems is in the debate about social norms in non-human animals. While some researchers have suggested that human beings are the only "normative animal" (Schmidt and Rakoczy 2019; Tomasello 2020), many animals engage in patterns of behavior that seem to have socially

normative characteristics, from inequity aversion in capuchin monkeys (Brosnan and De Waal 2003), to interventions to prevent infanticide in chimpanzees (von Rohr et al. 2011), to play rituals with rules about aggressiveness in dogs, wolves, coyotes and dolphins (Bekoff and Pierce 2009). However, these claims about animal social norms are subject to shifting standards of evaluation. For example, Laura Schlingoff and Richard Moore – who define norms in terms of explicit representations of rules – have argued that all of the chimpanzee behaviors that theorists have described as social norms can be more parsimoniously explained by lower-level cognitive processes, without appeal to representations of normative rules (Schlingloff and Moore 2017). For proponents of rule-based approaches to social norms, such as Laura Danón (2019) and Simon Fitzpatrick (2020), addressing this challenge would mean developing a research program that could systematically test all the predictions of a particular normative rule, thereby eliminating all the lower-level, non-rule-based explanations of the normative behavior in question. Such a research program would need to proceed in the same manner as research on altruistic explanations of spontaneous helping behavior, which involved the systematic elimination of alternative, egoistic explanations of helping behaviors (Batson et al. 1988). However, for researchers who adopt affective accounts of ought-thoughts (Andrews 2020; Theriault et al. 2021), or who alternatively demand more advanced capacities for meta-representation (Korsgaard 2006) or shared intentionality (Tomasello 2019), such a research program would hold no probative value: for affective approaches, representations of rules are not necessary for the presence of social norms, while for more cognitively demanding accounts, they are not sufficient. Thus, the animal norms debate is left at an impasse: without any consensus about what the core psychological features of social norms are, there can be no agreement about the kind of empirical research agenda that would be suitable to study them.

These problems are not unique to the animal norms debate: the same definitional and empirical issues could easily emerge in other borderline cases where the norm in question is not explicit, as in the study of apparently normative behaviors in young, preverbal children (Warneken and Tomasello 2007; Hamlin and Tan 2020), or when positing (as many norm theorists do) that many social norms are sustained by nonverbal, unconscious, implicit mental representations(Kelly 2022). Only in paradigmatic cases where the relevant social norms are expressed as explicit linguistic rules (e.g. norms of etiquette) can there be anything approaching a consensus about whether a social norm is present at all. However, these clear-cut cases may set the bar too high, and only reflect social norms in their most cognitively advanced, institutionalized forms. They may not provide a reliable guide to understanding how social norms develop, how they have evolved, or how they manifest themselves across different social environments.

Treating ought-thought as necessary for criterial also implies that social norms are psychologically homogeneous. While agents enforcing and complying with social norms might differ from one another in a number of ways and occupy very different environments, according to this view, their behaviors will still arise from the same basic psychological capacities that constitute ought-thought. This will be generally true of agents in the same community participating on the same normative regularity, of individual agents across all the different regularities that they participate in, and of agents in separate communities adhering to different regularities. Of course, such a framework would allow for some degree of variability at the level of ought-thought (e.g. among neuroatypical individuals, or in individuals at different developmental stages), but in general, ought-thought is understood to be a robust species-typical disposition. Considering the sheer variety of normative regularities and the different domains of human life in which they are thought to appear, assumption seems premature.

This reveals the fundamental problem with trying to give a functional analysis of ought-thoughts. As a psychological capacity, they are not well-defined. There is no specific thing that this capacity enables us to do, no agreed upon behavioral or psychological signature of ought-thought. In this respect,

ought-thoughts are quite unlike other well-known targets of functional analysis in cognitive science like belief-attribution (Apperly and Butterfill 2009) or the approximate number system (Clarke and Beck 2021). In these cases, the capacities in question are well-defined, even though there are many different proposals for how they should be functionally analyzed. In contrast, each theory of social norms characterizes its explanandum in its own way, and this is reflected in their functional analyses and in debates about non-paradigmatic cases of social norm conformity and enforcement. And yet, ought-thoughts are also posited as the primary *explanation* of normative regularities – a vast and variable set of phenomena encompassing all human communities and stretches back deep into the evolutionary past of our species. Such an amorphous construct seems distinctly unsuited for this explanatory role.

### 2.2. Defining normative regularities

Our diagnosis of the problem with taking ought-thoughts as the primary *explananda* of the cognitive science of norms is simple: relying on the capacity for ought-thought to explain normative regularities gets things *inside-out*. In order to understand the cognitive underpinnings of social norms, norm theorists should not start by stipulating an idiosyncratic specification of the capacity normative cognition and then use the presence of this capacity as a criterion for identifying social norms. Instead, the cognitive science of norms needs proceed from the *outside-in*, starting with normative regularities and the patterns of social interaction that constitute them. Only after identifying these patterns can we make progress on uncovering factors of attraction that bring them about.

First, however, the nature of normative regularities needs to be made more precise. This is a delicate task. To avoid importing a tacit conception of ought-thought, any specification of what constitutes a normative regularity must remain neutral with respect to their psychological underpinnings. It should also strive to encompass points of general consensus among different theories of social norms about their distinctive social and behavioral characteristics, so as to better capture the entire spectrum of behavioral phenomena that we call social norms. The combination of these two constraints is likely to yield a very broad construct that almost certainly will not match any a priori conceptual analyses of what we ordinarily *mean* when we say, "social norm." However, satisfying armchair intuitions about social norms is less important than generating a useful, ecumenical notion that will capture all the different things researchers have called social norms within its wide net.

With these constraints in mind, we propose that a normative regularity be defined as follows:

> **Normative regularity:** A socially maintained pattern of behavioral conformity within a community.

*Patterns of behavioral conformity* feature in almost all accounts of normative cognition, even those that adopt very different theories of cognitive architecture. For example, Bicchieri's account of how normative motivation includes a central role for what she calls "empirical expectations," which are an agent's beliefs about the behaviors that are typical in their community (Bicchieri 2006, 2017). Kelly and Setman likewise describe the behavior produced by the norm system as a "robust and multifaceted type of response that is centered on conformity and punishment" that "produce[s] stabilizing group-level effects on patterns of collective social organization" (Kelly and Setman 2020). Matteo Colombo's account stresses the predictive benefits of norm-driven conformity: when most people in the community conform to the same social norms, this reduces their uncertainty and makes it easier for them to navigate the social environment (Colombo 2014). Veissière and colleagues describe these patterns of conformity as a way for agents to "outsource their policy selection to relevant others and

to aspects of their material niche" (Veissière et al. 2019, p. 17). This suggests that norm theorists see patterns of behavioral conformity as a primary feature of social norms.

This emphasis on real patterns of behavioral conformity distinguishes normative regularities from purely prescriptive norms that are not widely adopted by a community, and hence not manifested in concrete behavioral patterns. Obviously, it is possible for agents to represent prescriptive norms of this sort, and to think about how the world falls short of an unachieved ideal. But this variety of ought-thought – which is more commonly associated with ethics than with social norms – is not typically among the explananda for the cognitive science of social norms, even for those who do understand social norms in psychological terms. Social norms in the sense we are interested in thus have a strong descriptive, statistical component that captures "what is done" or "the way we do things around here." In other words, the "oughts" of normative regularities really do imply "is."

Our notion of *social maintenance* is intended to capture a feature that is central to many accounts of the psychology of social norms: normative regularities are not merely patterns of behavioral conformity, but rather patterns that are incentivized by the behaviors of other agents within the community. Whether or not a pattern of behavioral conformity counts as a normative regularity depends on how members of a given community respond to individual cases of conformity and nonconformity. To illustrate, consider a few paradigmatic examples of social norm violations: if an individual violates a queuing norm by cutting in line or started talking loudly on a cellphone during a funeral, we would reliably expect other individuals to respond negatively towards that individual in a way that would incentivize normative conformity. In contrast, consider the patterns of behavioral conformity that emerge when we open doors with our right hands, or when savannah animals repeatedly congregate at the same watering hole during the dry season. These regularities are easily explained by nonsocial factors, like the fact that most humans tend to be right-handed or facts about the scarcity of water in savannah ecosystems. It would be very surprising to learn that violating one of these regularities would trigger enforcement behaviors among onlookers – or indeed, any response at all. These forms of behavioral conformity are not subject to social pressure. The presence of social pressure is what distinguishes mere behavioral regularities from normative regularities: while both types of behavioral pattern are statistically common, only normative regularities are enforced by agents upon one another.

In many accounts of the psychology of social norms, what we are calling social maintenance is often framed in terms of *punishment* or *punitive attitudes* (Bicchieri 2006; Sripada and Stich 2006; Rakoczy et al. 2008; Kelly and Davis 2018). However, "punishment" in this sense is usually understood quite broadly: many of these theorists hold that social maintenance can include a wide range of negative reactions to norm-violating behavior, from physical violence to "correcting, withholding cooperation, communicating disapproval through body language or explicit criticism, [and] ostracizing or gossiping about norm violators" (Kelly and Setman 2020). This diverse set of behaviors is unified by the fact that they all incentivize norm conformity by imposing costs on non-conformers; we will refer to this as "negative social maintenance." Later in the paper, we will also explore the possibility that there exist normative regularities sustained by non-punitive, "positive" forms of social maintenance. Thus, we define social maintenance in broad terms as *any social response to behavioral conformity or nonconformity that incentivizes conformity,* be it through positive or negative incentives.[5]

[5] One concern about tying the concept of normative regularities to patterns of behavioral conformity is that many familiar examples of social norms focus on behavioral prohibitions about what agents *must not do* rather that behavioral prescriptions about what they should do. Intuitively, members of a community *not* taking a certain action does not seem like the sort of thing that one could easily observe. Fortunately, there are ways around this obstacle. One way to infer

We will leave the notion of a *community* unspecified. Which population of individuals counts as the community in which a given norm holds is invariably context-dependent. Consider, for example, Bicchieri's notion of a *reference network*, which refers to "the range of people whom we care about when making particular decisions" (Bicchieri 2017, p. 14). In her account, members of a community need not be physically present to influence norm-conformity, nor do they need to be geographically local: immigrants might conform to the normative regularities of their native countries instead of adopting new norms because they are more strongly influenced by social pressures from that community. By the same token, individuals can be members of many intersecting and hierarchically nested communities, and thus subject to conflicting social pressures. As such, it is incumbent upon researchers to specify the community in which a given normative regularity holds, and to give an independent justification for why that specification is appropriate.

With this characterization of normative regularities in place, we are now in a position to fully articulate our central empirical thesis:

> **Normative pluralism:** normative regularities are the products of a variety of different underlying cognitive, affective, and ecological processes of varying degrees of complexity.

The core argument for normative pluralism is simple: once we stop taking it for granted that "ought-thought" refers to a discrete type of cognitive process or natural psychological kind, and we instead focus on normative regularities, then it becomes quickly clear that the psychology of social norms can be realized by many different processes. What these processes have in common is not a unified cognitive architecture, but rather a shared causal role as factors of attraction that sustain normative regularities. Or so we will argue in the sections that follow, laying out in very broad terms the three main aspects of normative regularities that the cognitive science of social norms might aim to explain, which we model after the framework of Sripada and Stich (2006):

1. Social norm acquisition: how individuals learn about the normative regularities in their communities.
2. Social norm conformity: Why individuals are motivated to *adhere* to normative regularities.
3. Social norm maintenance: How normative regularities are *maintained* by the behaviors of others in the community.

## 3. *Social norm acquisition is pluralistic*

How does a person come to learn the normative regularities in their community, such that they are able to conform to and enforce them? The literature offers many plausible, mutually compatible possibilities, which range from cognitively demanding and representationally complex to extremely minimal. One point of tension in existing accounts of socially normative cognition concerns representations of *rules* – whether these must be explicit or implicit, or whether they are required at all (Railton 2006; Schlingloff and Moore 2017; Danón 2019). Coming from the perspective of normative pluralism, this dispute is avoided. Representations of rules *can but need not* figure in the psychology of social norms. In some cases, representations of rules might turn out to be the best explanation of a

whether a behavior of a certain type is prohibited is to look for signs that it elicits negative social maintenance. If every time a child writes with their left hand, an adult strikes them and forces them to use their right hand instead, that is a sign that writing with the left hand is prohibited. Notably, this approach would have limited use when it comes to studying strong behavioral prohibitions that are rarely or never violated. However, the presence of such normative prohibitions can be inferred by drawing comparisons between otherwise similar communities: if, for example, left-handedness is relatively common in some communities but is rare in others, that is a sign that left-handedness might be prohibited, even if one never observes it being punished. Together, evidence of negative social maintenance and comparisons across communities can enable us to detect normative prohibitions even in cases where these prohibitions are not explicitly avowed.

given normative regularity. In other cases, positing such representations might actually overintellectualize a much simpler form of cognition. In what follows, we will describe five ways in which norms can be acquired, starting with cases that look the most like rule-learning and ending with cases in which representing a rule as such may not be required. This is not meant to be an exhaustive account of social norm acquisition, but we take it to illustrate the pluralism apparent in this aspect of the psychology of social norms.

### 3.1. Direct instruction

At the most cognitively complex end of the spectrum, it seems quite plausible that a lot of norm learning occurs via *direct instruction* or *testimony*, through explicit, linguistic channels, especially in WEIRD cultures. This form of norm acquisition is sometimes passed over too quickly, particularly in accounts that draw analogies between normative and grammatical rules, stressing the tacit, inarticulable nature of normative knowledge. While some norms are certainly acquired without direct instruction, there are many cases where we learn a norm because *someone told us about it*. As children, our parents proactively remind us to mind our manners ("What do you say?" "*Thank you!*"). When we take our pets to the park, strategically placed signs sternly remind us to pick up after our dogs. Sexist norms of chivalry are sometimes explicitly repeated as they are enacted, like when a man announces, "Ladies first!" while opening a door for a woman. We have very explicit beliefs about how much one is supposed to leave as a tip in a restaurant; when we go abroad, we look to travel guides to learn whether the local tipping customs differ from our own. All this has important implications for the cognitive underpinnings of our knowledge of social norms. At least some of the time, norms are acquired via language comprehension processes, stored as explicit linguistic representations in semantic memory, and activated in the form of explicitly articulated or mentally simulated speech.

### 3.2. Mentalizing

Slightly less explicit but still cognitively complex, some social norms are not directly taught, but rather inferred via mentalizing – that is, inferring what one is supposed to do based on the normative expectations that we attribute to those around us. The best-known theory of norm-learning along these lines is probably that of Bicchieri, whose notion of social norms is explicitly grounded in belief attributions: we infer that a rule is a social norm just in case members of our reference network (1) conform to it and (2) believe that others ought to conform to it as well (though this model is often treated as a high-level rational reconstruction of normative cognition, rather than a literal description of the cognitive processes involved). We also see another mentalistic approach to social norm learning in Theriault and colleagues' account of the *sense of should*, where people infer what they are *supposed* to do by modeling what other agents *expect* they *will* do (Theriault et al. 2021) – though notably, this account is not framed in terms of inferences about rules. And intuitively, we have all had the experience of discovering the existence of a norm by accidentally violating it and then becoming aware of the displeasure of one's companions. We explain the displeasure by inferring the existence of a norm we hadn't yet noticed. Having the ability to infer what members of a community are thinking provides a valuable window into that community's norms.

### 3.3. Social learning heuristics

Another way that norm learning can occur is when norms are inferred directly from certain readily observed behavioral cues. These direct inferences might be supported by a variety of social learning heuristics that do not obviously involve any mentalistic inferences, such as "copy the majority" or "learn from prestigious individuals" (Chudek et al. 2012). There is also some evidence that young children sometimes adopt a strategy of *promiscuous normativity*, inferring that whatever adults do is what is *supposed* to be done, and spontaneously enforcing these inferred rules upon others (Schmidt et al.

2016b). This promiscuous normativity heuristic may also be what drives young children to *overimitate* intentionally performed but unnecessary actions (Kenward et al. 2011; Keupp et al. 2013). In these cases, children seem to be able to detect and correct violations of a norm (albeit one they may not be able to verbally articulate) but are unlikely to be making an explicit inference about agents' normative beliefs, given their age and minimal competence with belief-attribution.

### 3.4. Reinforcement learning

In some cases, norms may be learned tacitly through the operation of simple, domain-general mechanisms such as reinforcement learning. With experience, different action types that appear in recurring social decision problems may come to be associated with specific expectations about social rewards and punishments. Social rewards and punishments can be construed as intrinsically rewarding or aversive responses from a member of the community, which might come in the form of facial expressions, gestures, vocalizations, touch, and so on (Colombo 2014). As an example, consider the familiar choice of whether to move forward or stand still on an escalator. At first, the value of these two courses of action might be roughly equivalent: the amount of time saved by walking is equivalent to the amount of energy saved by standing still, and so you choose your course of action at random. But if others around you are walking while you are standing still, you might get jostled as people walk past, which you experience as more aversive than walking. So, you come to associate a higher value with walking than staying still while on the escalator. This kind of strategy would not require any kind of representation of agents' mental states or their normative beliefs; indeed, it would not require a representation of a normative rule at all, above and beyond a stored map of the reward structure of the social environment.

### 3.5. Biological inheritance

Finally, some elements of social norms may be part of one's biological inheritance. Norms governing the care of immature infants are likely supported by psychological mechanisms that evolved via kin selection (Hrdy 2011). Norms governing cooperation with group members may be facilitated by evolved dispositions for social tolerance that evolved in the context of obligate mutualistic foraging (Tomasello 2012). Innate dispositions towards pair-bonding present across many primate species might provide a foundation for the culturally specific norms that comprise marriage institutions (Henrich 2017). The developmental effects of these evolved psychological foundations are in turn supplemented by distinctly cultural modes of transmission, as new parents in industrialized countries might learn how they should treat their infant by reading books like *The Happiest Baby on the Block*. What is inherited and shared across cultures and across species may be a greater sensitivity to infants, some degree of know-how, or motivation to care for infants.

This non-exhaustive survey already illustrates the sheer variety of cognitive processes that plausibly underlie the acquisition of normative regularities. Some of these processes might be ontogenetically prior to others, or more cross-culturally ubiquitous, or more phylogenetically widespread. It might also be that some of these norm learning mechanisms operate automatically or as a default, while others are deployed reflectively and only in certain contexts. It is also possible that over time, normative regularities that first emerged as the result of a low-level process like reinforcement learning might eventually become more explicit, rule-based representations, as we reflect upon and codify the various patterns in our social environment. All of these are important empirical questions that a complete psychology of norms should set out to answer. Just as important, however, is the observation that all these strategies plausibly contribute to the existence of normative regularities.

## 4. *Social norm conformity is pluralistic*

The conformity component of a psychology of social norms should explain why individuals are motivated to adhere to normative regularities. In several accounts, it has been suggested that social norms have *intrinsic* motivational properties: to have fully internalized a norm from one's community *just is* to find it primitively motivating, independently of any other costs that might come from not adhering to it, or any benefits that result from conformity (Sripada and Stich 2006; Chudek and Henrich 2011; Kelly and Davis 2018; Kelly 2020; Theriault et al. 2021). There are several plausible explanations of how this property of social norms might emerge from more basic computational and biological processes, such as the allostatic maintenance of metabolic resources by minimizing prediction errors (Theriault et al. 2021) or model-free reinforcement learning processes (Colombo 2014). However, intrinsic motivations are not a defining feature of normative regularities. Instead, these intrinsically motivating processes become just a few among the many mechanisms that cause individuals to adopt the relevant behaviors (c.f. Kelly, 2020a). Widening the scope of our inquiry in this way gives us the opportunity to understand not just what different normative regularities have in common, but how and why adherence to them might vary – both from one norm to another (Nichols 2004), and on a broader cultural level (Gelfand et al. 2011). It also puts us in a better position to understand the psychological factors that cause individuals to vary in their propensity to adhere to social norms (DeYoung et al. 2002; Bègue et al. 2015; Kosloff et al. 2017).

### 4.1. Avoiding punishment

One of the most obvious non-intrinsic motivations to conform to a normative regularity is the motivation to avoid punishment, as well as other negative forms of social maintenance. It is well-established that third-party punishment can stabilize all kinds of social behaviors, regardless of their immediate individual cost (Boyd and Richerson 1992; Fehr and Fischbacher 2004). Besides fear of third-party punishment or second-party punishment, norm-conformity might also be sustained via the fear of ostracism, if failure to comply with a norm affects the partner-choice decisions of their conspecifics (Baumard et al. 2013). In general, if the risks of punishment outweigh the benefits of not conforming to the norm, we should expect widespread norm-conformity. These cost-benefit calculations could be explicit, or they could be implicit in the learning process.

### 4.2. Social rewards

Alternatively, individuals might find that social norm conformity confers a variety of rewards or benefits. These might come in the form of material rewards, if abiding by the norm in question is just a solution to a coordination problem that allows us to achieve our end goals (Lewis 1969). For example, in North America we often abide by the norm of walking on the right side of a stairway, because it is so much easier to get to the top of the steps by conforming to the behavior of others around us, rather than trying to fight against the current of people on the left. The convention could have just as easily been "walk on the left," but we conform and walk on the right because following the norm is an effective means to achieve our goal. In these cases, the best explanation for norm conformity is that following the norm is the most straightforward way to get what we want.

Following norms also sometimes yields social rewards by satisfying more basic desires for social interaction and play, like when we abide by the rules of a game. Along these lines, Tomasello has argued that the motivation to conform to social norms stems from a more basic motivation to engage in shared activities, and to take part in a collective "we" (Tomasello 2019). This approach is also consistent with theories that tie social norm conformity to subjective group dynamics, and the need to maintain a sense of sense of social identity by, for example, putting a "In this house we believe…" sign in the yard or wearing a T-shirt advertising a favorite baseball team (Hogg et al. 2017). Thus,

adhering to social norms can yield a range of payoffs that stand to incentivize habits of norm conformity across many domains.

### 4.3. Niche construction and environmental scaffolding

It is also important not to overlook the ways that niche construction contributes to social norm conformity. Consider, for example, a ski hill where there is a norm against going out-of-bounds, and skiers are supposed stick to designated runs. If someone were to ski out-of-bounds, they might find themselves on the receiving end of dirty looks, or even a mild scolding from the ski patrol. However, the ski hill has also been intentionally designed in such a way that the runs are also the best way to get to various desirable locations, such as the chairlift, the chalet, and the parking lot. Going off-course, meanwhile, poses a variety of risks and costs, from a laborious hike back to the car to the possibility of serious injuries or avalanches. As a result, most people stick to the designated runs, and only an adventurous minority ever violates the "don't ski out of bounds" norm.

What is interesting about this case is that widespread adherence to the norm arguably requires very little cognition at all, above and beyond the detection of various environmental affordances and responding appropriately to them. In order to conform to the relevant normative regularity, all you have to do is grasp that the runs are clear and well-groomed, that they lead to various destinations, and that going out-of-bounds poses a variety of practical obstacles. The structure of the built environment serves as a tacit guide towards the normatively prescribed path, whether or not one recognizes it as such. Indeed, in this case a person's normative knowledge might be located entirely in the environment, not in the head of the norm adherent.

Here, it might be argued that this case does not really describe a genuine normative regularity, but rather something more like a *custom*, such as using an umbrella when it is raining (Bicchieri 2017). But anyone who has been skiing will be able to tell you that the prohibition against skiing out of bounds is not *just* a matter of individual-level practical reasoning: it is also enforced by third parties, and is often explicitly encoded in the rules of the ski hill. This suggests that the psychological processes that underpin conformity to the out-of-bounds prohibition and those that underpin its social maintenance can be of a very different character. The norm in question is at once explicit and highly cognitive, and also implicit and grounded in environmental affordances shaped by niche construction. Both types of process contribute to the persistence of this normative regularity. While one does not need to know anything about the rule in order to reliably conform to it, sanctions for its violation might serve as a kind of normative failsafe layered on top of the existing practical incentives for conformity inherent in the designed environment.

## 5. *Social maintenance is pluralistic*

While norm acquisition and conformity are about individuals learning to follow social norms, social maintenance is about how other community members respond to such individuals. Most proposals about the psychology of social norms focus on negative forms of social maintenance, such as punishment and displays of social disapproval. In this section, we identify four types of negative social maintenance – third-party punishment, gossip, second-party punishment, and restorative justice – and show how their psychological underpinnings are also pluralistic. Then, we consider how more positive forms of social maintenance that are not typically discussed in theories of socially normative cognition can contribute to normative regularities.

### 5.1. Third-party punishment

To get a sense of the different processes driving negative social maintenance, we can first consider the different reasons that we punish norm violations.[6] In traditional analytic philosophy, the two major rationales for punishment are retributivism and deterrence (Bedau and Kelly 2019). There is some evidence that participants in WEIRD populations have been found to reason about punishment in a retributive manner, with a specific focus on psychological factors like the perpetrator's intentions and foresight (Carlsmith 2006; Kneer and Machery 2019). However, this pattern does not hold cross-culturally; in many non-WEIRD populations, judgments about punishment are much less sensitive to intentions than to outcomes (Barrett et al. 2016; McNamara et al. 2019; Curtin et al. 2020). Even within WEIRD participant populations, we see considerable individual differences in deterrence versus retributivist intuitions about punishment (Crockett et al. 2014; Clark et al. 2017), and in the motivation to engage in punishment more generally (Hofmann et al. 2018). This suggests that both the disposition to punish norm violations and the psychological motivations for doing so are quite varied.

Besides retribution and deterrence, third party punishment is often motivated by communicative, pedagogical intentions. When an individual transgresses a norm, an observer can use punishment to let them know that their behavior is not acceptable, in the hopes of shaping their future behavior. In these cases, punishers do not treat the act of punishment as an end unto itself; rather, they view the punishment as effective only insofar as the right message has been received by the norm violator (Crockett et al. 2014; Funk et al. 2014; Sarin et al. 2021). Cognitively speaking, this form of punishment is relatively complex, since it requires both punishers and violators to treat an act of punishment as a kind of Gricean communication involving a series of mentalistic inferences about communicative intentions. One therefore wouldn't expect this pedagogical sort of punishment to play a significant role in creatures without these abilities, though it may be quite pervasive in mature human populations.

### 5.2. Gossip and reputation management

Sometimes, we engage in acts of punishment to enhance our reputations in the eyes of onlookers. Evolutionary game theorists have long suggested that punishing those who violate group-beneficial norms can serve as a costly signal of one's trustworthiness as a potential cooperative partner, and thereby stabilize cooperation in the group as a whole (Gintis et al. 2001). In line with this proposal, there is behavioral evidence that people tend to punish more readily when they are being observed by third parties, and that they abstain from costly punishment when they have already demonstrated their trustworthiness to onlookers in some other way (Jordan et al. 2016; Jordan and Rand 2020). These behaviors do not appear to be supported by conscious, reflective processes; rather, they are the product of an unconscious heuristic to behave as though reputation is at stake, even if one has not explicitly reasoned about who is watching.

While we are sometimes motivated to maintain social norms out of a desire to protect our reputations, we also maintain social norms by sharing information *about the reputations of others*. Gossip and the establishment of reputation systems can serve as a powerful form of social maintenance (Wu et al. 2016).[7] As many authors have noted, the desire to maintain one's reputation as good social partner functions as a powerful motivator to conform to prosocial norms (Feinberg et al. 2014; Számadó et

[6] A complicating factor here is that how and when we punish others for norm violations is often shaped by other norms, values, and institutions. In practice, the negative social maintenance practice for one normative regularity may depend upon a whole network of other norms regulating whether or not punishment is permissible and who is permitted to carry it out.

[7] Gossip can also function as a vector for the acquisition of social norms: negative gossip about a behavior can often serve as evidence that it is normatively prohibited (Westra 2021).

al. 2021). The prevalence of gossip about about norm violating behaviors, on the other hand, likely stems from a variety of motives, such as the prosocial desire to protect vulnerable individuals from exploitation (Feinberg et al. 2012), as well as the affiliative desire to establish and maintain social bonds (Dunbar 2004).

### 5.3. Second-party retaliation

While third-party forms of negative social maintenance have played an important role in theories of the evolution of cooperative norms, it is important not to overlook the role of second-personal, retaliatory punishment in the psychology of norms – i.e. revenge (McCullough et al. 2013). Many norms regulate interpersonal interactions in a way that allows us to avoid conflict: queuing norms, norms for fairly dividing resources, property norms, norms of personal space and appropriate physical contact, and more. Violating any of these norms can inflict direct harm or material losses on other agents, who in turn have the option of retaliating. Retaliation could be triggered directly by the experience of being harmed, which in turn triggers a reflexive desire for revenge (McCullough et al. 2013), and an experience of pleasure after the retaliatory act has been carried out (Chester and DeWall 2016). Alternatively, the aggrieved individual might also view the sheer fact of the norm violation as intrinsically motivating, over and above the harm it may have caused them. Or they might experience both forms of motivation simultaneously. From the perspective of a prospective norm violator, retaliation motivated in either way would have the same practical impact – namely, deterrence against violating this norm in the future. The same thing is true at the population level: the same normative regularity could be sustained by either kind of retaliatory motivation, or even a mixture of the two.

This point about mixed motivation generalizes to all the punitive motivations discussed in this section. Individuals engaged in social enforcement behavior can be driven by a variety of different goals, sometimes at the same time. The same act of normative punishment might be construed by the punisher as just deserts, deterrence, a "teachable moment" for the norm violator, a signal to onlookers of the punisher's reputation, and a hedonistic relief of a retaliatory urge. Norm enforcement stemming from such mixed motivations would be just as effective in supporting a normative regularity in the broader community, as would punishments stemming solely from one of these motivation types.

### 5.4. Restorative justice

It is also worth noting that not all forms of negative social maintenance responses are punitive even in a broad sense. In some cultures, community-members respond to norm violators by attempting to repair relationships through restorative justice (Braithwaite 2002; Wiessner and Pupu 2012). To avoid the negative costs of punishment, group members can engage with a norm violator to encourage mutual acts of reconciliation that take on a quasi-pedagogical flavor. In the Navajo peacemaking tradition, for example, norm violations are treated as an opportunity to educate the norm violator on the consequences of their actions while also providing them with a path to reintegrate into the community in a position of respect (Zion 1998). This allows the norm violator to maintain their status as a person of worth, someone who could be a productive future member of society (Petersen et al. 2012).

Finally, some forms of negative social maintenance might arise from the withdrawal of rewards for norm conformity, or positive social maintenance. If an agent comes to expect and rely upon the social benefits that come with norm conformity, then simply taking those benefits away can be a powerful form of social maintenance unto itself. We take up the topic of positive social maintenance in the next sub-section.

### 5.5. Praise

One reason that individuals respond positively to social norm conformity is as a means of providing evaluative feedback in pedagogical contexts. This can be thought of as the social counterpart to explicit, verbal forms of norm acquisition mentioned in section 3.1: in order for children to learn norms via teaching, there must also be individuals who are motivated to teach norm compliance. And one of our key pedagogical strategies is, of course, giving praise for normatively correct behavior: "What do you say?" "Thank you!" "*Very good!*" Like pedagogical uses of punishment, pedagogical praise can be thought of as a way of communicating information about the value of certain actions, either as an instrumental means to a desired end or as an end in of itself (Ho et al. 2017). In these contexts, it is quite likely that teachers' actions are guided by explicit representations of rules or norms, and thus would constitute a more sophisticated form of normative cognition.

### 5.6. Identifying ingroup members

Another reason that an individual might respond positively towards norm-conforming behavior is because it helps them identify ingroup members – for example, when sports fans wearing the same team jersey greet one another on the street (Abrams et al. 2000). Different social groups engage in distinctive normative regularities, and so norm-conforming behavior can be a useful way of distinguishing between groups, a process sometimes referred to as *normative differentiation* (Marques et al. 1998). One psychological explanation for positive reactions to norm-conforming ingroup members is that norm conformity promotes between-group distinctiveness, which in turn helps to maintain their social identity (Tajfel 1981). A distinct but not incompatible explanation is that norm conformity might function as "ethnic markers" that help us to identify reliable partners for the purposes of coordination and the pursuit of mutual goals (McElreath et al. 2003; Jensen et al. 2015).

### 5.7. Social fluency

A much more basic explanation of positive responses towards norm-conformers is that we find them predictable. By hypothesis, abiding by normative regularities means acting in a way that is statistically common and hence expected in a given context. On prediction-error minimization models of cognition, unexpected behaviors can generate metabolic costs, which are aversive. One strategy for agents to minimize prediction errors is to actively seek out more predictable social environments where other agents largely conform to their expectations – a kind of active inference (Colombo 2014; Theriault et al. 2021). When agents conform to the social norms, this can be a sign that they won't be a source of many prediction errors. This will in turn lead us to associate norm-conformers with positive affect and higher levels of social fluency (Reber and Norenzayan 2012). Less technically, norm-conformers just feel nice and familiar, and so we respond positively towards them.

All of these positive reactions towards norm-conformity will serve to reinforce norm adherence in the broader population, over and above any negative social maintenance behaviors like punishment. In some cases where positive social maintenance is strong enough, norm conformity might become so uniform that acts of negative social maintenance never actually occur in practice. Under these circumstances, naturalistic observations of negative social maintenance behaviors would be impossible, which poses a significant empirical challenge for norms researchers when trying to establish the presence of a normative regularities. In these cases, we suggest, social maintenance should be understood in counterfactual terms: *if* the behavioral regularity *were* violated, we should expect some sort of negative social maintenance behaviors in response. Such counterfactual predictions would need to be operationalized experimentally.

## 6. *Objections to normative pluralism*

In this section, we consider three objections to normative pluralism: (1) whether our view over-extends the notion of social norms; (2) whether our view is compatible with the existence of an evolved norm system; and (3) whether our view entails a form of eliminativism about social norms.

### 6.1. Is the normative regularities construct too permissive?

One potential criticism of our approach to social norms is that the notion of normative regularities is too permissive, making normative pluralism trivially true. One way of developing this criticism would be to frame it as a *reductio*: given the set of norm acquisition, norm conformity, and norm enforcement processes we have outlined, normative regularities could conceivably arise even in creatures that are completely unaware of anything about the norms they follow – species with extremely minimal cognitive capacities, with no normative concepts, no representations of rules, and no corresponding affective discomfort stemming from norm violations. Such "social norms" could simply arise as emergent properties of the way such creatures interact with one another and with their social environment.

One possible example of this kind of normative regularity might be the phenomenon of "marching bands" in Mormon crickets, when millions of insects gang together to form swarms of up to 10 km long and traveling up to 2 km in a day (Simpson and Raubenheimer 2012). Importantly for our account, this pattern of social conformity is the result of a distinctive form of social maintenance: *cannibalism*. Marching bands occur when a population of crickets has depleted the local environment of protein and salt, and the only remaining source of these key nutrients is other crickets. The coordinated movement of millions of crickets that emerges is the indirect effect of each individual simultaneously chasing the insect in front of it while fleeing the one behind. This, according to our framework, would count as a normative regularity.

This might seem like a counterintuitive consequence of normative pluralism: in the absence of any psychological criteria for normative regularities, "social norms" start to turn up all throughout the natural world, even in crickets. We suspect that this intuition stems in part from a broader tendency to underrate the social and cognitive complexity of invertebrates (Mikhalevich and Powell 2020). It may also hinge upon a folk psychological conception of ought-thought: if one's intuitive criteria for social norms hinge upon something like the representation of intrinsically motivating rules, then *of course* it would seem absurd to claim that social norms in insects. But this intuitive way of thinking about social norms is precisely what we have argued should have no place in our understanding of the explanandum of the cognitive science of social norms. In our view, if it turns out that there are normative regularities even in phylogenetically distant species that are underpinned by a radically different set of psychological processes, this only serves to broaden our understanding of how such community-level patterns of behavior might come about. Indeed, some normative regularities in humans that we quite comfortably refer to as social norms might actually resemble those at work in the crickets (for example, the case of the escalator norms). More generally, observing how normative regularities emerge in other species can help us to recognize how certain types of multiply realizable social structure – such as social hierarchies – can create conditions under which social norms reliably tend to emerge (Jebari 2018).

In short, we see great value in an open-minded approach to studying social norms. However, researchers who are interested in studying a more limited range of phenomena need not retreat to the ought-thought conception of the psychology of norms criticized in 2.1. Instead of imposing additional psychological criteria on which normative regularities "count" as social norms, norm theorists have the option of drawing additional non-psychological distinctions between different types of normative regularities in order to narrow their scope of inquiry. For example, some norms researchers (e.g. Kelly

2020) have an explicit focus on *internalized* norms that guide behavior even when nobody is there to enforce it (think of the times you have waited for a walk signal before crossing the street when there are no cars about). One way to capture this notion while still adhering to an outside-in, pluralistic approach would be to distinguish between normative regularities that where behavioral conformity persists only as long as there is active social maintenance from members of the community, and normative regularities that persist even after social maintenance behaviors have ceased. Some theorists might also be especially interested in normative regularities where the relevant social maintenance behaviors are *themselves* socially maintained (e.g. via higher-order punishment). Articulating distinctions like these offers norm theorists a way to restrict their domain of inquiry without taking a conception of ought-thought for granted.

### 6.2. Is normative pluralism consistent with a dedicated architecture for social norms?

So far, we have argued that normative regularities can be realized by a wide range of cognitive and affective processes, and that no single psychological mechanism or set of mechanisms stands out as *distinctively* normative. However, proponents of the cognitive-evolutionary approach to the psychology of norms might well argue that the pluralistic collection of processes that we have identified *just is* the norm system. On this view, the fact that normative cognition is a kludge and that its component parts have evolved piecemeal does not preclude it from being understood as a complex adaptation with a unified function.[8] Because normative regularities have been so relevant to human fitness, the objection goes, it could be that natural selection – operating via genetic or cultural evolution – has come to treat the many mechanisms and processes that sustain them as a single, cohesive unit.

Understood this way, some versions of cognitive evolutionary approach (e.g. Chudek & Henrich, 2011; Kelly & Davis, 2018) could be construed as *weakly* pluralistic: the mechanisms and processes that bring about normative regularities are variable but still share a common adaptive function and an integrated information-processing architecture. Our own view is *strongly* pluralistic: the wide range of mechanisms and processes that bring about normative regularities do not share a common adaptive function, and there is little architectural integration among them. Many of the mechanisms we discussed in the preceding sections are domain general. Other factors contributing to normative regularities are entirely non-psychological, emerging instead from the properties of the local environment and the constraints it imposes. Still others have other evolutionary functions that do not depend upon their role in normative regularities. This collection of mechanisms is functionally unified only in the sense that they all play some causal role in bringing about complex social effects, not in virtue of their evolutionary history.

The distinction between weak and strong pluralism in the architecture of normative cognition could have several empirical implications. A weakly pluralistic architecture assumes that acquiring and enforcing normative regularities was selected for at some point in our ancestral history. Depending on when in history this selection process is hypothesized to have occurred, this could have implications for whether the proposed norm system is uniquely human (c.f. Birch 2020). It might also generate predictions about the paleo-archeological record (Sterelny 2021). Claims of functional integration within the norm system should also yield specific predictions about how norm-relevant information is processed. A strong pluralistic approach, in contrast, makes very few predictions of

[8] This view is analogous to Pinker and Jackendoff's (2005) claim that the human language faculty is a complex system that evolved piecemeal but nevertheless has a single adaptive function (communication).

this sort, since it rejects claims about broad generalizations across all different forms of normative cognition, and instead advocates for explanations tied to specific normative regularities.

### 6.3. Norm eliminativism?

A final worry about this approach is that the truth of a strong form of normative pluralism would a kind of norm eliminativism: if the social norms are so heterogeneous that it cannot support useful psychological generalizations, then perhaps we should not employ the concept of social norms in cognitive science at all.

While our proposal is in many ways a deflationary one, we are not advocating for wholesale eliminativism about social norms. Instead, we think that social norms are best understood as an instance of what Daniel Dennett calls *real patterns* (Dennett 1991). Real patterns are pragmatic, instrumental constructs that are readily discernable and afford robust predictive powers, but whose underlying mechanisms are unknown and whose ontological status may be uncertain. In his theory of the intentional stance, Dennett suggests that propositional attitudes like belief and desire might be real patterns in this sense: attributing them enables us to reliably predict behavior and facilitate social interaction, but we are not certain about the mechanisms that generate them. Once a real folk psychological pattern has been identified, there is no further question about whether the attitudes in question really exist. The patterns *are* the attitudes.

Something similar is true of social norms. Identifying a pattern of social behavior in a community as a social norm offers us a compact way of describing of a large set of behaviors. Once identified, social norms can be robustly predictive and facilitate fluent social interaction. Social norms can be clearly recognized even when we do not know the psychological mechanisms that cause members of a community to adhere to them. And once a real pattern has been identified, there need be no deeper question about whether the norm really exists; the pattern *is* the social norm.[9]

Moreover, denying that social norms constitute a natural psychological kind does not make them any less interesting, or any less amenable to psychological study. Just as relaxing assumptions about the unity of moral cognition has not prevented researchers from learning about a variety of morally relevant processes, adopting a pluralistic approach to social norms does not preclude us from learning about the mechanisms underpinning normative regularities that matter a great deal in our day-to-day lives. Instead, normative pluralism should simply lead cognitive scientists studying social norms to aim for narrower, more modest generalizations about specific normative regularities and particular mechanisms for norm acquisition, conformity, and maintenance.

Appreciating the psychological variability underlying different types of social norm can also give us some insight into the patterns of normative variability we see across cultures – for example, the differences between human societies with "tighter" or "looser" social norms (Gelfand et al. 2011). Pluralism at the psychological level might also help to explain why some norms prove especially robust. If a normative regularity depends on the population sharing a single, highly specific set of motivations, then any motivational heterogeneity in that population might destabilize it. If, however, the norm can be sustained by a wider range of motivations, then it is potentially more robust to change. Alternatively, if agents have multiple, redundant motivations to conform to a given norm, that might make them more likely to conform than if they had only one.

In short, normative pluralism does not lead to eliminativism about social norms. Instead, it offers us a more disciplined and consistent way of studying social norms, free from strong theoretical

[9] Of course, one need not accept Dennett's position on the nature of the propositional attitudes in order to think of social norms as real patterns in this sense.

assumptions about their core psychological characteristics. It puts us in a position to learn *more* about the cognitive science of social norms, not less.

## 7. *Conclusion*

The central thesis of this paper – what we've called *normative* pluralism – is that we should not take the psychological unity of social norms for granted. In our methodological proposal, we argued that the best candidate explananda for the cognitive science of social norms should not be construed as a complex psychological disposition – i.e. as ought thoughts – but rather as real patterns of socially maintained community-level behaviors, which we have called *normative regularities*. In our empirical proposal, we argued that learning about, conforming to, and maintaining normative regularities are most likely not the product of a single kind of process or system, but rather the products of a heterogeneous set of cognitive, affective, and ecological mechanisms.

We suggest that this pluralistic perspective is useful in a number of respects. First, it avoids pinning the explanatory targets of the cognitive science of norms to controversial, theory-driven conceptions of norms, and instead focuses on their more readily observed and measurable attributes. Second, once we come to expect that different social norms might be driven by different psychological processes, and that individuals can differ in their motivations for adhering to or enforcing any given norm, we will be in a better position to explain patterns of variability in norm-relevant behavior.

Thinking about social norms in this way will undoubtedly make the cognitive science of norms more complex and messy. If we are correct, however, then this will simply be a reflection of the complexity and messiness of social norms themselves. Taking a pluralistic approach to social norms allows us to explore the potential variability inherent to norm-governed behavior, which can help us to better understand how social norms shape our lives, and how they manifest themselves throughout the natural world.[10]

[10] We are grateful to two anonymous reviewers from this journal, Jonathan Birch, Laura Danón, Simon Fitzpatrick, Cecilia Heyes, Joseph Jebari, William O'Shea, Stephen Stich, Jordan Theriault, and especially Daniel Kelly for their detailed comments on earlier drafts of this paper. We are also grateful to audiences at the Normative Animals conference, the 2022 meeting of the Southern Society for Philosophy and Psychology, the ASENT Research Group at the London School of Economics, the Department of History and Philosophy of Science at the University of Cambridge, the Situated Cognition Research Group at Bochum University, the Department of Philosophy and Religious Studies at Peking University, and to the participants of the Evolution of Normativity Workshop.

**References**


Abrams D, Bown N, Marques JM, Henson M (2000) Pro-norm and anti-norm deviance within and between groups. J Pers Soc Psychol 78:906–912. https://doi.org/10.1037/0022-3514.78.5.906

Andrews K (2020) Naïve Normativity: The Social Foundation of Moral Cognition. J Am Philos Assoc 6:36–56. https://doi.org/10.1017/apa.2019.30

Andrews K (2012) Do apes read minds?: Toward a new folk psychology. MIT Press, Cambridge, MA

Andrews K, Spaulding S, Westra E (2020) Introduction to Folk Psychology: Pluralistic Approaches. Synthese. https://doi.org/10.1007/s11229-020-02837-3

Apperly I, Butterfill SA (2009) Do humans have two systems to track beliefs and belief-like states? Psychol Rev 116:953–970. https://doi.org/http://dx.doi.org/10.1037/a0016923

Barrett HC, Bolyanatz A, Crittenden AN, et al (2016) Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. Proc Natl Acad Sci U S A 113:4688–4693. https://doi.org/10.1073/pnas.1522070113

Batson CD, Dyck J, Brandt J, et al (1988) Five studies testing two new egoistic alternatives to the empathy-altruism hypothesis. J Personal Soc Psychol 55:52–77

Baumard N, André JB, Sperber D (2013) A mutualistic approach to morality: The evolution of fairness by partner choice. Behav Brain Sci 36:59–78. https://doi.org/10.1017/S0140525X11002202

Bedau HA, Kelly E (2019) Punishment. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy, Winter 201

Bègue L, Beauvois J-L, Courbet D, et al (2015) Personality Predicts Obedience in a Milgram Paradigm. J Pers 83:299–306. https://doi.org/10.1111/JOPY.12104

Bekoff M, Pierce J (2009) Wild justice: The moral lives of animals. University of Chicago Press

Bicchieri C (2006) The Grammar of Society. Cambridge University Press, New York

Bicchieri C (2017) Norms in the Wild: How to Diagnose, Measure, and Change Social Norms. Oxford University Press, New York

Bicchieri C, McNally P (2018) Shrieking Sirens: Schemata, scripts, and social norms. How change occurs

Birch J (2020) Toolmaking and the Origin of Normative Cognition. Biol Philos 1–47. https://doi.org/10.1007/s10539-020-09777-9

Boyd R, Richerson PJ (1992) Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethol Sociobiol 13:171–195. https://doi.org/10.1016/0162-3095(92)90032-Y

Braithwaite J (2002) Restorative justice & responsive regulation. Oxford University press on demand

Brosnan SF, De Waal FBM (2003) Monkeys reject unequal pay. Nature 425:297–299. https://doi.org/10.1038/nature01963

Carlsmith KM (2006) The roles of retribution and utility in determining punishment. J Exp Soc Psychol 42:437–451. https://doi.org/10.1016/j.jesp.2005.06.007

Chester DS, DeWall CN (2016) The pleasure of revenge: Retaliatory aggression arises from a neural

imbalance toward reward. Soc Cogn Affect Neurosci 11:1173–1182. https://doi.org/10.1093/scan/nsv082

Chudek M, Heller S, Birch S, Henrich J (2012) Prestige-biased cultural learning: bystander's differential attention to potential models influences children's learning. Evol Hum Behav 33:46–56. https://doi.org/10.1016/j.evolhumbehav.2011.05.005

Chudek M, Henrich J (2011) Culture-gene coevolution, norm-psychology and the emergence of human prosociality. Trends Cogn Sci 15:218–26. https://doi.org/10.1016/j.tics.2011.03.003

Cialdini RB, Kallgren CA, Reno RR (1991) A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In: Advances in experimental social psychology. Elsevier, pp 201–234

Cialdini RB, Reno RR, Kallgren CA (1990) A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. J Pers Soc Psychol 58:1015

Clark CJ, Baumeister RF, Ditto PH (2017) Making punishment palatable: Belief in free will alleviates punitive distress. Conscious Cogn 51:193–211

Clarke S, Beck J (2021) The Number Sense Represents (Rational) Numbers. Behav Brain Sci 1–57

Colombo M (2014) Two neurocomputational building blocks of social norm compliance. Biol Philos 29:71–88. https://doi.org/10.1007/s10539-013-9385-z

Crockett MJ, Özdemir Y, Fehr E (2014) The value of vengeance and the demand for deterrence. J Exp Psychol Gen 143:2279–2286. https://doi.org/10.1037/xge0000018

Cummins R (2000) How does it work?" versus" what are the laws?": Two conceptions of psychological explanation. In: Keil FC, Wilson RA (eds) Explanation and cognition. MIT press, Cambridge, MA, pp 117–144

Curry OS, Jones Chesters M, Van Lissa CJ (2019) Mapping morality with a compass: Testing the theory of 'morality-as-cooperation' with a new questionnaire. J Res Pers 78:106–124. https://doi.org/10.1016/j.jrp.2018.10.008

Curtin CM, Barrett HC, Bolyanatz A, et al (2020) Kinship intensity and the use of mental states in moral judgment across societies. Evol Hum Behav 1–15. https://doi.org/10.1016/j.evolhumbehav.2020.07.002

Danón L (2019) Animal normativity. Phenomenol Mind 17:176–187. https://doi.org/10.13128/pam-8035

de Waal FBM (2008) Putting the Altruism Back into Altruism: The Evolution of Empathy. Annu Rev Psychol 59:279–300. https://doi.org/10.1146/annurev.psych.59.103006.093625

Dennett DC (1991) Real patterns. J Philos Philos 88:27–51

DeYoung CG, Peterson JB, Higgins DM (2002) Higher-order factors of the Big Five predict conformity: Are there neuroses of health? Pers Individ Dif 33:533–552. https://doi.org/10.1016/S0191-8869(01)00171-4

Dunbar RIM (2004) Gossip in evolutionary perspective. Rev Gen Psychol 8:100–110. https://doi.org/10.1037/1089-2680.8.2.100

Dwyer S, Huebner B, Hauser MD (2010) The linguistic analogy: Motivations, results, and speculations. Top Cogn Sci 2:486–510. https://doi.org/10.1111/j.1756-8765.2009.01064.x

Fehr E, Fischbacher U (2004) Third-party punishment and social norms. Evol Hum Behav 25:63–87. https://doi.org/10.1016/S1090-5138(04)00005-4

Feinberg M, Willer R, Schultz M (2014) Gossip and Ostracism Promote Cooperation in Groups. Psychol Sci 25:656–664. https://doi.org/10.1177/0956797613510184

Feinberg M, Willer R, Stellar J, Keltner D (2012) The virtues of gossip: Reputational information sharing as prosocial behavior. J Pers Soc Psychol 102:1015–1030. https://doi.org/10.1037/a0026650

Fiebich A, Coltheart M (2015) Various Ways to Understand Other Minds: Towards a Pluralistic Approach to the Explanation of Social Understanding. Mind Lang 30:235–258. https://doi.org/10.1111/mila.12079

Fitzpatrick S (2020) Chimpanzee normativity: evidence and objections. Biol Philos 35:1–28. https://doi.org/10.1007/s10539-020-09763-1

Funk F, McGeer V, Gollwitzer M (2014) Get the Message: Punishment Is Satisfying If the Transgressor Responds to Its Communicative Intent. Personal Soc Psychol Bull 40:986–997. https://doi.org/10.1177/0146167214533130

Gelfand MJ, Jackson JC (2016) From one mind to many: The emerging science of cultural norms. Curr Opin Psychol 8:175–181. https://doi.org/10.1016/j.copsyc.2015.11.002

Gelfand MJ, Raver JL, Nishii L, et al (2011) Differences between tight and loose cultures: A 33-nation study. Science (80- ) 332:1100–1104. https://doi.org/10.1126/science.1197754

Gintis H, Smith E, Bowles S (2001) Costly Signaling and Cooperation. J Theor Biol 213:1–29

Graham J, Haidt J, Nosek BA (2009) Liberals and Conservatives Rely on Different Sets of Moral Foundations. J Pers Soc Psychol 96:1029–1046. https://doi.org/10.1037/a0015141

Gray K, Waytz A, Young L (2012) The moral dyad: A fundamental template unifying moral judgment. Psychol Inq 23:206–215

Greene JD (2015) The rise of moral cognition. Cognition 135:39–42. https://doi.org/10.1016/j.cognition.2014.11.018

Greene JD, Sommerville RB, Nystrom LE, et al (2001) An fMRI investigation of emotional engagement in moral judgment. Science (80- ) 293:2105–2108

Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol Rev 108:814

Haidt J, Koller SH, Dias MG (1993) Affect, culture, and morality, or is it wrong to eat your dog? J Pers Soc Psychol 65:613

Hamlin JK, Tan E (2020) The emergence of moral responses and sensitivity. In: Jensen LA (ed) The Oxford Handbook of Moral Development: An Interdisciplinary Perspective. Oxford University Press Oxford, UK, Oxford, p 267

Heintz C (2018) Cultural Attraction Theory. Int Encycl Anthropol 1–10. https://doi.org/10.1002/9781118924396.wbiea2311

Henrich J (2020) The weirdest people in the world: How the west became psychologically peculiar and particularly prosperous. Farrar, Straus and Giroux

Henrich J (2017) The secret of our success: how culture is driving human evolution, domesticating

our species, and making us smarter. Princeton University Press

Ho MK, MacGlashan J, Littman ML, Cushman F (2017) Social is special: A normative framework for teaching with and learning from evaluative feedback. Cognition 167:91–106. https://doi.org/10.1016/j.cognition.2017.03.006

Hofmann W, Brandt MJ, Wisneski DC, et al (2018) Moral Punishment in Everyday Life. Personal Soc Psychol Bull 44:1697–1711. https://doi.org/10.1177/0146167218775075

Hogg MA, Abrams D, Brewer MB (2017) Social identity: The role of self in group processes and intergroup relations. Gr Process Intergr Relations 20:570–581. https://doi.org/10.1177/1368430217690909

Hrdy SB (2011) Mothers and others. Harvard University Press

Jebari J (2018) Empirical moral rationalism and the social constitution of normativity. Philos Stud. https://doi.org/10.1007/s11098-018-1134-3

Jensen NH, Petersen MB, Høgh-Olesen H, Ejstrup M (2015) Testing Theories about Ethnic Markers: Ingroup Accent Facilitates Coordination, Not Cooperation. Hum Nat 26:210–234. https://doi.org/10.1007/s12110-015-9229-4

Jordan JJ, Hoffman M, Bloom P, Rand DG (2016) Third-party punishment as a costly signal of trustworthiness. Nature 530:473–476. https://doi.org/10.1038/nature16981

Jordan JJ, Rand DG (2020) Signaling When No One Is Watching: A Reputation Heuristics Account of Outrage and Punishment In One-Shot Anonymous Interactions. J Pers Soc Psychol 118:57–88. https://doi.org/10.1037/pspi0000186.supp

Kelly D Two Ways to Adopt a Norm: The (Moral?) Psychology of Internalization and Avowal. In: Vargas M, Doris JM (eds) The Oxford Handbook of Moral Psychology. Oxford University Press, New York

Kelly D (2020) Internalized norms and intrinsic motivations: Are normative motivations psychologically primitive? Emot Res 36–45

Kelly D (2022) Two Ways To Adopt a Norm: The (Moral?) Psychology of Internalization and Avowal. In: Vargas M, Doris J (eds) Oxford Handbook of Moral Psychology. Oxford University Press, New York, pp 285–309

Kelly D, Davis T (2018) Social norms and human normative psychology. Soc Philos Policy 35:54–76. https://doi.org/10.1017/S0265052518000122

Kelly D, Setman S (2020) The Psychology of Normative Cognition. In: Zalta EN (ed) Stanford Encyclopedia of Philosophy, Spring 202. Metaphysics Research Lab, Stanford University, pp 1–28

Kelly D, Stich S, Haley KJ, et al (2007) Harm, affect, and the moral/conventional distinction. Mind Lang 22:117–131. https://doi.org/10.1111/j.1468-0017.2007.00302.x

Kenward B, Karlsson M, Persson J (2011) Over-imitation is better explained by norm learning than by distorted causal learning. Proceedings Biol Sci 278:1239–46. https://doi.org/10.1098/rspb.2010.1399

Keupp S, Behne T, Rakoczy H (2013) Why do children overimitate? Normativity is crucial. J Exp Child Psychol 116:392–406. https://doi.org/10.1016/j.jecp.2013.07.002

Kneer M, Machery E (2019) No luck for moral luck. Cognition 182:331–348. https://doi.org/10.1016/j.cognition.2018.09.003

Korsgaard C (2006) Morality and the distinctiveness of human action. In: Ober J, Macedo S, de Waal FBM (eds) Primates and philosophers: How morality evolved. Princeton University Press, Princeton, NJ, pp 98–119

Kosloff S, Irish S, Perreault L, et al (2017) Assessing relationships between conformity and meta-traits in an Asch-like paradigm. https://doi.org/101080/1553451020171371639 12:90–100. https://doi.org/10.1080/15534510.2017.1371639

Kumar V (2016) The Empirical Identity of Moral Judgment. Philos Q 66:783–804. https://doi.org/10.1093/pq/pqw019

Lewis D (1969) Convention: A philosophical study. John Wiley & Sons, Oxford

Maibom HL (2005) Moral unreason: The case of psychopathy. Mind Lang 20:237–257. https://doi.org/10.1111/j.0268-1064.2005.00284.x

Marques J, Abrams D, Paez D, Martinez-Taboada C (1998) The role of categorization and in-group norms in judgments of groups and their members. J Pers Soc Psychol 75:976–988. https://doi.org/10.1037/0022-3514.75.4.976

Marr D (2010) Vision: A computational investigation into the human representation and processing of visual information. MIT press, cambride

McCullough ME, Kurzban R, Tabak BA (2013) Cognitive systems for revenge and forgiveness. Behav Brain Sci 36:1–15. https://doi.org/10.1017/S0140525X11002160

McElreath R, Boyd R, Richerson PJ (2003) Shared norms and the evolution of ethnic markers. Curr Anthropol 44:122–129. https://doi.org/10.1086/345689

McGeer V (2007) The regulative dimension of folk psychology. In: Hutto DD, Ratcliffe M (eds) Folk psychology re-assessed. Springer, pp 137–156

McNamara RA, Willard AK, Norenzayan A, Henrich J (2019) Weighing outcome vs. intent across societies: How cultural models of mind shape moral reasoning. Cognition 182:95–108. https://doi.org/10.1016/j.cognition.2018.09.008

Mikhail J (2011) Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment. Cambridge University Press

Mikhalevich I, Powell R (2020) Minds without spines: Evolutionarily inclusive animal ethics. Anim Sentience 5:1

Miller CB (2014) Character and moral psychology. Oxford University Press, Oxford

Nichols S (2004) Sentimental rules: On the natural foundations of moral judgment. Oxford University Press

O'Neill E (2017) Kinds of norms. Philos Compass 12:1–15. https://doi.org/10.1111/phc3.12416

Petersen MB, Sell A, Tooby J, Cosmides L (2012) To punish or repair? Evolutionary psychology and lay intuitions about modern criminal justice. Evol Hum Behav 33:682–695. https://doi.org/10.1016/j.evolhumbehav.2012.05.003

Pinker S, Jackendoff R (2005) The faculty of language: What's special about it? Cognition 95:201–236. https://doi.org/10.1016/j.cognition.2004.08.004

Prinz J (2007) The emotional construction of morals. Oxford University Press

Railton P (2006) Normative Guidance. Oxford Stud Metaethics 3–33

Reber R, Norenzayan A (2012) The Shared Fluency Theory of Social Cohesiveness. In: Proust J, Fortier M (eds) Metacognitive Diversity. Oxford University Press, Oxford

Richerson PJ, Boyd R (2005) Not by genes alone: How culture transformed human evolution. University of Chicago press

Sarin A, Ho MK, Martin JW, Cushman FA (2021) Punishment is Organized around Principles of Communicative Inference. Cognition 208:104544. https://doi.org/10.1016/j.cognition.2020.104544

Schlingloff L, Moore R (2017) Do chimpanzees conform to social norms? In: Andrews K, Beck J (eds) Routledge handbook of philosophy of animals minds. Routledge, New York

Schmidt M, Butler L, Heinz J, Tomasello M (2016a) Young Children See a Single Action and Infer a Social Norm: Promiscuous Normativity in 3-Year-Olds. Psychol Sci 27:1360–1370. https://doi.org/10.1177/0956797616661182

Schmidt MFH, Butler LP, Heinz J, Tomasello M (2016b) Young Children See a Single Action and Infer a Social Norm : Promiscuous Normativity in 3-Year-Olds. https://doi.org/10.1177/0956797616661182

Schmidt MFH, Rakoczy H (2019) On the Uniqueness of Human Normative Attitudes. In: Roughley N, Bayertz K (eds) The Normative Animal? On the Anthropological Significance of Social, Moral, and Linguistic Norms. Oxford University Press, Oxford, pp 121–136

Scott-Phillips T, Blancke S, Heintz C (2018) Four misunderstandings about cultural attraction. Evol Anthropol 27:162–173. https://doi.org/10.1002/evan.21716

Shweder RA, Mahapatra M, Miller JG (1987) Culture and moral development. In: Kagan J, Lamb S (eds) The emergence of morality in young children. University of Chicago press, Chicago, pp 1–83

Simpson SJ, Raubenheimer D (2012) The Nature of Nutrition: A Unifying Framework from Animal Adaptation to Human Obesity. Princeton University Press, Princeton, NJ

Sinnott-Armstrong W, Wheatley T (2014) Are moral judgments unified? Philos Psychol 27:451–474. https://doi.org/10.1080/09515089.2012.736075

Spaulding S (2018) How We Understand Others: Philosophy and Social Cognition. Routledge, London

Sperber D (1996) Explaining culture: A naturalistic approach. Wiley-Blackwell, Oxford

Sripada C (2020) The atoms of self-control. Noûs

Sripada C, Stich SP (2006) A Framework for the Psychology of Norms. In: Carruthers P, Laurence S, Stich S (eds) The Innate Mind Volume 3: Culture and Cognition. Oxford University Press, Oxford, pp 280–301

Sterelny K (2021) The Pleistocene Social Contract. Oxford University Press, Oxford

Stich SP (2006) Is morality an elegant machine or a kludge? J Cogn Cult 6:181–189. https://doi.org/10.1163/156853706776931349

Stich SP (2019) The quest for the boundaries of morality. In: Zimmerman A, Jones K, Timmons M (eds) The Routledge Handbook of Moral Epistemology. Routledge, New York, pp 15–37

Strohminger N, Nichols S (2014) The essential moral self. Cognition 131:159–171. https://doi.org/10.1016/j.cognition.2013.12.005

Számadó S, Balliet D, Giardini F, et al (2021) The language of cooperation: reputation and honest signalling. Philos Trans R Soc B Biol Sci 376:. https://doi.org/10.1098/RSTB.2020.0286

Tajfel H (1981) Human groups and social categories: Studies in social psychology. Cambridge University Press, Cambridge, UK

Theriault JE, Young L, Barrett LF (2021) The sense of should: A biologically-based framework for modeling social pressure. Phys Life Rev 36:100–136. https://doi.org/10.1016/j.plrev.2020.01.004

Tomasello M (2019) The Moral Psychology of Obligation. Behav Brain Sci 43:1–58. https://doi.org/10.1017/S0140525X19001742

Tomasello M (2020) The role of roles in uniquely human cognition and sociality. J Theory Soc Behav 50:2–19. https://doi.org/10.1111/jtsb.12223

Tomasello M (2012) Two Key Steps in the Evolution of Human Cooperation: The Interdependence Hypothesis. Curr. Anthropol.

Turiel E (1983) The development of social knowledge: Morality and convention. Cambridge University Press

Veissière SPL, Constant A, Ramstead MJD, et al (2019) Thinking Through Other Minds: A Variational Approach to Cognition and Culture. Behav Brain Sci. https://doi.org/10.1017/S0140525X19001213

von Rohr CR, Burkart JM, van Schaik CP (2011) Evolutionary precursors of social norms in chimpanzees: A new approach. Biol Philos 26:1–30. https://doi.org/10.1007/s10539-010-9240-4

Warneken F, Tomasello M (2007) Helping and Cooperation at 14 Months of Age. Infancy 11:271–294

Westra E (2021) Folk personality psychology: mindreading and mindshaping in trait attribution. Synthese 198:8213–8232. https://doi.org/10.1007/s11229-020-02566-7

Wiessner P, Pupu N (2012) Toward peace: Foreign arms and indigenous institutions in a Papua New Guinea society. Science (80- ) 337:1651–1654

Wu J, Balliet D, Van Lange PAM (2016) Reputation, Gossip, and Human Cooperation. Soc Personal Psychol Compass 10:350–364. https://doi.org/10.1111/spc3.12255

Young L, Dungan J (2012) Where in the brain is morality? Everywhere and maybe nowhere. Soc Neurosci 7:1–10. https://doi.org/10.1080/17470919.2011.569146

Zawidzki TW (2013) Mindshaping: A New Framework for Understanding Human Social Cognition. MIT Press, Cambridge, MA

Zion JW (1998) The Dynamics of Navajo Peacemaking. J Contemp Crim Justice 14:58–74