Penultimate Draft

Forthcoming in *Review of Philosophy and Psychology*, Special Issue on Hostile Emotions

A Framework for the Emotional Psychology of Group Membership

By Taylor Davis and Daniel Kelly[1]

Abstract: The vast literature on negative treatment of outgroups and favoritism toward ingroups provides many local insights but is largely fragmented, lacking an overarching framework that might provide a unified overview and guide conceptual integration. As a result, it remains unclear where different local perspectives conflict, how they may reinforce one another, and where they leave gaps in our knowledge of the phenomena. Our aim is to start constructing a framework to help remedy this situation. We first identify a few key ideas for creating a theoretical roadmap for this complex territory, namely the principles of etiological functionalism and the dual inheritance theory of human evolution. We show how a "molecular" approach to emotions fits into this picture, and use it to illuminate emotions that shape intergroup relations. Finally, we weave the pieces together into the beginnings of a systematic taxonomy of the emotions involved in social interactions, both hostile and friendly. While it is but a start, we have developed the argument in a way that illustrates how the foundational principles of our proposed framework can be extended to accommodate further cases.

I. Introduction

Research on negative treatment of outgroups and favoritism toward ingroups is rich and voluminous, and can be bewildering. Prominent theories in this domain focus on different aspects of this complex subject matter, and often do so using different concepts and methods. We will argue that as a result of this, it is hard to tell when these theories actually conflict, when they simply address different phenomena, and when they describe and explain the same phenomena in different vocabulary. It is thus difficult to discern the scope and limits of our current theoretical coverage of this domain.

In response, we suggest a back-to-basics approach to help systematize what is known and to help point the way forward. Rather than trying to comprehensively review the extant literatures, we begin by identifying certain fundamental principles of psychological explanation, and showing how they inform a certain approach to emotions. More specifically, we adopt a "molecular" view of emotions, and explain how it fits with the principles of etiological functionalism. We then use this perspective to assess several prominent theories of outgroup negativity, showing how they might sit with respect to each other.

In the process, we single out one account for special attention. This account, based on the idea of a *behavioral immune system*, overlaps with ours, but also remains incomplete in an instructive way: it fails to countenance the role of cultural evolution in the relevant etiologies. Accordingly, we introduce *dual inheritance theory* into the discussion (Richerson and Boyd 2005; Henrich 2016), and incorporate the ideas developed in previous sections into the larger framework it offers for the study

[1] Many thanks to Michael Brownstein, Alex Madva, Uwe Peters, and an anonymous reviewer for generous and useful comments on drafts of this paper.

of human evolution. The paper culminates in a first pass at a taxonomy that systematizes central components of the psychology of group membership, and that is pluralistic, conceptually parsimonious, and open ended. It is pluralistic in that it includes a wide range of disparate psychological mechanisms, but parsimonious in that these mechanisms are all identified on the basis of the same few theoretical principles. Completing the taxonomy falls beyond the scope of this paper, so our primary aim is to begin this project, and to illustrate the principles that can be used to extend it.

II. Background: Etiological Functionalism and Emotional Molecularism

The psychology of group membership is a motley patchwork of psychological processes, cobbled together over the course of our species' unique and complex evolutionary history. Humans have quite recently (by evolutionary standards) become *by far* the most cultural species of primate, developing hypertrophied capacities for social learning that enable us to achieve unprecedented heights of cooperation (Boyd and Richerson 2009; Chudek and Henrich 2011; Mathew and Perreault 2015). However, these advanced forms of sociality have a dark side, as sophisticated cooperation within groups gives rise to new and heightened forms of conflict between groups. As we will show, culturally inherited norms, institutions, and ideologies have become a common source of intergroup friction even in the absence of direct competition for material resources. The modern human mind reflects this history, containing cognitive and motivational capacities that evolved at very different times to perform very different functions, and which are inherited via both genetic and cultural pathways.

In order to gain an overview of this motley patchwork, we begin with a familiar functional distinction between cognition and motivation. By *cognition* we have in mind representations, as well as the inferences that take us from one representation to another. For example: representations involved in the classification of individuals as belonging to certain groups, representations of stereotypes that apply to that group, and inferences about how individuals are likely to behave that follow from such classifications and stereotypes. We take *motivation*, by contrast, to capture non-cognitive, often affective psychological processes which yield behavioral tendencies of approach and avoidance. We assume that cognitive capacities and motivational capacities interact constantly in complex and dynamic ways, and that appeals to both are indispensable. However, we argue that motivations have a special role in organizing the emotions of group psychology.

Another basic feature of our account will be a general conception of psychological explanation that we refer to as *etiological functionalism*. According to this doctrine, also discussed as "homuncular functionalism" (Lycan 1981, 1995, following Dennett 1978) and "functional analysis" (Cummins (1975; 1983; 2000), psychological explanations ideally begin by individuating behavioral capacities, or identifying distinct abilities in terms their functions—what they are *for*. Explanation then proceeds by construing the performance of functions in a hierarchical fashion, such that more complex functions are analyzed into component parts which are ascribed their own, simpler functions, each of which is in turn further analyzed into its own component parts with their own

functions, and so on. What makes this kind of hierarachical analysis *etiological*, however, is a further commitment to a specific way of identifying and individuating functions.

As the name suggests, etiological functionalism is backwards looking. More specifically, it holds that what a given psychological capacity is *for*—its function—is whatever it has been *selected for* in the past. Put differently, the functions of psychological traits are determined by their histories of evolutionary selection (Wright 1976; Millikan 1984; Dennett 1988; Griffiths 1993). Importantly, however, we interpret the concept of selection broadly, to include more than just selection acting on genetically inherited traits. Selection acting on culturally inherited traits also plays an enormous role in human evolution, and as we will argue, cultural traits loom large in the psychology of group membership.

This etiologically functionalistic perspective also reinforces the need to distinguish between cognition and motivation. For example, a lamentably common form of outgroup hostility results from interactions between motivations of fear, on one hand, and the cognition of racial stereotyping, on the other. Yet the selection pressures that produced the fear system stretch much deeper into our phylogenetic past than those that produced racial stereotypes. Indeed, selection pressures of the sort that formed the human fear system were already present in the environment of our fish and reptile ancestors, hundreds of millions of years before humans evolved (Panksepp and Biven 2012), and thus long before representations of distinct human races even existed. Appreciating this fact focuses questions on the kinds of selection pressures that have, more recently, forged an evolutionarily novel (and probably klugy, see Markus 2007) *connection* between fear and racial stereotypes (see Machery and Faucher 2005, Kelly et al. 2010).

The distinction between cognition and motivation also underpins the "molecular" approach to emotions we adopt. The literature on emotions is also vast, and while it is not our aim to develop a new theory of emotion, it will help to articulate how a molecular view confortably fits with the principles of etiological functionalism.

We understand emotions as complex functional systems composed of more basic "elements," having both motivational and cognitive elements as component parts.[2] Thus, each distinct combination of motivational and cognitive elements—each distinct molecule—constitutes a distinct emotion. This approach falls squarely within what Scarantino (2016) calls the *motivational tradition* of theories of emotion, which he contrasts with the *feeling tradition* or the *evaluation tradition*. Unlike theories in the feeling tradition, our molecular approach is not primarily concerned with the phenomenological, qualitative experiences that accompany various physiological states. Rather, it identifies emotions by reference to clusters of psychological processes, regardless of what kinds of feelings those processes contribute to first person, subjective experience. In this sense, our molecular view has more in common with theories in the evaluation tradition. However, while our view countenances an important role for the cognitive processes of evaluation, appraisal, and categorization, it gives pride of place to the motivational elements of emotional molecules.[3]

[2] For other recent approaches that are similarly self-consciously "molecular," seeing complex psychological states and processes as being composed of simpler, more elemental ones, see Curry et al. (ms) on morality and Sripada (2020) on self-control.

[3] Within the motivational tradition, the view of emotions most closely aligned with our molecular approach is Scarantino's New Basic Emotions Theory (2015).

In ordinary usage, emotion terms such as "fear," "anger," or "disgust" often pick out motivational elements by themselves. Both fear of predators and fear of public speaking qualify as fear. They are both subtypes within the fear "genre," in virtue of sharing a motivational core that originally evolved from selection for avoiding predators. However, they are also importantly different psychological states, in virtue of cognitive elements they do *not* share. These include different representations and inferences serving as inputs, or triggers, for the fear system, as well as different representations and inferences produced as outputs from the fear system, which channel fear motivation into very different forms of behavior. The molecular approach to emotions preserves the intuitive, ordinary-language practice of classifying emotions primarily on the basis of motivational elements—at the genre level—but can also distinguish subtypes within those genres in virtue of their distinct cognitive elements. Thus, fear of bodily harm from a predator and fear of embarrassment from public speaking fall in the same genre because they share a motivational core, but since their associated cognitive elements are different, they are distinct emotions, distinguishable "fear molecules".

The molecular view of emotions is not an analysis of the folk concept of emotion, or a description of how emotion terms are used in the vernacular. Rather, it is theoretically oriented, and free to part ways with folk concepts of emotion when doing so serves the explanatory and predictive aims of psychologists and other empirical theorists. In this case, however, the systematizing goals of science appear to be well served by the folk intuition that motivation is taxonomically primary and cognition secondary. The number of distinct motivational elements is relatively small compared to the number of cognitive elements that may be combined with them. Thus, we will identify the principal genres of emotion (e.g. fear, disgust, anger) by appeal to motivational functions, reserving cognitive functions to draw more fine-grained distinctions within these genres.[4]

Bringing together etiological functionalism with this molecular view of emotions yields a two-step procedure for organizing the psychology of negative treatment of outgroups and favoritism toward ingroups. First, we identify a range of basic motivational capacities involved, in virtue of the selection histories that produced them. Second, we construct a more high-resolution map of this domain by identifying various distinct cognitive elements that work in conjunction with those motivational capacities, again by appeal to the selection histories that forged functional connections between the relevant motivational and cognitive elements. We can even sloganize the main idea of our approach to identifying molecules of emotion: *follow the selection histories*. The rest of the paper spells out what this slogan means in technical terms.

III. Prominent Accounts of Outgroup Negativity

[4] Note that nothing in this picture limits emotions to only involving a *single* motivation. For example, researchers studying "empty nest syndrome" might identify a distinct emotion capturing the combination of affectively positive pride and affectively negative loss (along with the associated cognitive processes) that parents experience when their children leave home. Likewise, researchers studying those who provide care for elderly relatives might identify a distinct emotion capturing the combination of grief, relief, and guilt about feeling relief that often occurs when a person they've been caring for passes away. Other kinds of complex, "mixed" or "bittersweet" emotions, might be similarly construed as containing multiple motivational and cognitive components.

A thoroughgoing review of the existing literature on outgroup negativity would extend beyond the scope of this article, but a brief comparison of some prominent landmarks will be sufficient to illustrate the difficulties of theoretical coverage and conceptual integration we hope to ameliorate.

For example, Dovidio and Gaertner (2000) identify what they call *aversive racism*, in which individuals who are "averse" to racism nevertheless harbor implicit racist attitudes. Alternatively, Plant and Devine (1998) separate "internal and external motivations to respond without prejudice," noting that personally rejecting prejudice is different from merely seeking to avoid punishment and gain approval from others who reject prejudice. Elsewhere, Fiske and colleagues (Fiske et al. 2002, also see Abele et al. 2020) propose the Stereotype Content Model, which posits a two-dimensional space defined by stereotypes of warmth and competence, yielding four emotions toward outgroups: *admiration* toward groups stereotyped as high in both warmth and competence, *envy* toward groups seen as high in competence but low in warmth, *pity* toward groups seen as high in warmth but low in competence, and *contempt* toward groups seen as low in both warmth and competence. Finally, Intergroup Emotions Theory (Devos et al. 2002) attempts to combine appraisal theories of emotion (Smith and Ellsworth 1985) with self-categorization theory (Turner 1985), yielding a single theory that identifies intergroup emotions as appraisals of situations in which outgroups affect the goals and interests of the agent's ingroup.

Each of these four accounts offers a valuable local perspective, capturing certain important aspects of outgroup negativity. But none provides—or, to be fair, aspires to provide—a comprehensive account of the entire domain. Nor do any provide a principled way of identifying the specific subdomain it addresses within the larger, shared domain. Moreover, each account employs its own basic set of concepts and methods, developed in isolation from the other accounts, which in turn leaves it unclear how the different accounts relate to one another. For example, one might wonder to whether the posited "internal motivations to respond without prejudice" are the same as the motivations that make individuals "averse" to racism, or if not, how the two types of motivations compare and contrast. One might also want to know what kind of emotion occurs when one's *stereotypes* of a specific outgroup yield admiration, but one's appraisal of a specific situation is that the same outgroup is interfering with the goals of one's ingroup. The theories themselves offer little guidance. Indeed, it may be the case that some of them are alternative and incompatible accounts of the same explanatory target, while others have different explanatory targets, and so are not in competition with each other at all. But even where the theories do not compete or conflict, it is an open question whether, or how, they could be integrated into a single, logically coherent account of the overarching subject matter of the psychology of group membership. The brute differences between them leave would-be unifiers with a patchy, incomplete understanding of the subject matter, and no conceptual means for identifying where the gaps in our knowledge lie.

A fifth theory, based on the notion of a behavioral immune system, points to a way forward. We will argue that this theory is also incomplete, but it provides a promising basis on which to build something more comprehensive. Its central idea is that alongside a physiological immune system that attacks pathogens after they have already entered the body, many animals are also equipped with a *behavioral* immune system that is oriented toward preemption, employing a suite of psychological

adaptations for avoiding contact with pathogens in the first place (Schaller 2011). Two more ideas establish the connection to outgroup avoidance. First, an important component of the behavioral immune system is disgust. Second, an especially potent vector of disease transmission is other people, especially strangers from other groups (Kelly 2011, Curtis 2013).

This theory predicts, and has found evidence for, reliable connections between disgust sensitivity and attitudes toward outgroups.[5] Humans live in groups, and members of the same group interact much more with each other than with people from other groups. As a result, the physiological immune systems of individuals from the same group develop together, producing antibodies for those diseases to which they are collectively exposed. Ingroup members thus come to have the same immuno-strengths and, more importantly, immuno-weakness. Thus, for any given individual, people from other groups are much more likely to carry diseases to which that individual's physiological immune system has no response (Faulkner et al. 2004; Navarrete and Fessler 2006; Navarrete, Fessler, and Eng 2007). To protect against this kind of threat, the preemptively-oriented behavioral immune system combines the motivational element of disgust with cognitive elements involving representations of members of foreign groups.

Schaller and Neuberg (2012), who defend this theory, also claim that it fits comfortably within Cottrell and Neuberg's (2005) more general sociofunctional theory of prejudice. This account holds that different forms of prejudice result from different adaptive threats, and since infection from outgroup members constitutes a serious adaptive threat, the behavioral immune system can be identified as one important source of prejudice.

We endorse this (and many other) use(s) of evolutionary theory in the psychological and behavioral sciences (e.g., Muthukrishna and Henrich 2019). We also applaud Schaller and Neuberg (2012) for leveraging the evolutionary reasoning that animates their view to help situate it within a broader theoretical landscape. They thus make significant progress on the problems of theoretical coverage and conceptual integration mentioned above. However, this progress is limited in two ways. First, not all selection pressures count as threats, and there is no reason to focus specifically on those selection pressures that do. Second, while they (along with Cottrell and Nueberg (2005)) recognize the relevance of cultural evolution in principle, their account never actually appeals to any cultural selection histories. Accordingly, the next section will use the sociofunctional theory as a stalking horse, a foil against which to compare and contrast our preferred approach.

IV. Etiological Functionalism at Work: Adding Motivations of Approach

Cottrell and Neuberg's (2005) "sociofunctional, threat-based approach" puts etiological functionalism to work in the study of prejudice. In their own terms (2005, p. 771), "…individuals possess psychological mechanisms 'designed' by biological and cultural evolution to take advantage of the opportunities provided by group living and to protect themselves from threats to group living." Schaller and Neuberg (2012) then integrate the behavioral immune system into this framework, identifying

[5] Recent research has also found intriguing relationships between aspects of disgust and political attitudes; see Aarøe et al. 2020 and Ruisch et al. 2020.

> a set of qualitatively distinct prejudices rooted in distinct sets of psychological processes, each of which can be understood as an adaptive consequence to a distinct kind of threat that imposed evolutionary selection pressures on ancestral populations. Within this broad framework, individual lines of research have focused on two specific kinds of threat—the threat of interpersonal violence and the threat of infectious disease—and their separate implications for different kinds of prejudices pertaining to different categories of people" (p. 4).

Even if these authors never use the terms "etiology" or "functionalism," they employ these principles adeptly, allowing Cottrell and Neuberg to give penetrating analyses of several other theories of prejudice, including two of those mentioned above. Regarding the Stereotype Content Model, they correctly point out (p. 775) that a four-way classification scheme identifying admiration, envy, pity and contempt is both too coarse-grained and too narrow in scope. For example, they object that *contempt* lumps together anger and disgust, each of which evolved in response to a different adaptive threat, and each of which drives different forms of prejudicial behavior. They also object that the Stereotype Content Model fails to address the role of fear in outgroup avoidance, since fear is not a common response to either the warmth or the competence of a foreign group. Cottrell and Neuberg then criticize Intergroup Emotions Theory on similar grounds of scope. Its authors, they argue, "have limited their explorations to the emotions of anger and fear, within the context of having experimental participants imagine interacting with groups designed to differ in the strength of threat they posed to participant in-groups" (p. 775). Thus, they claim the theory is ill-equipped to account for the role of disgust and other components of the behavioral immune system.

These specific critiques highlight the more general problems we raised above. Despite the important insights yielded by more local theories, they ultimately provide a patchwork of narrowly focused and difficult to integrate fragments, leaving possible gaps in our knowledge of the general domain, and no clear way of identifying where the holes lie. That such shortcomings are clarified by evolutionary theory also illustrates the potential of etiological functionalism for making headway on this problem. Yet the sociofunctional account fails to fully realize this potential. As noted above, nothing in evolutionary theory suggests that selection is only about avoiding threats. Some adaptive pressures are indeed rooted in threat, resulting in motivations of *avoidance* such as fear or disgust. But contrasting with these are motivations of *approach* (see Elliot 2006), many of which evolved to ensure that animals take advantage of various adaptive benefits, rather than avoid threats. Ancestors who were better able to acquire food, water, high-quality mates, and other goods had higher fitness, leading to the evolution of approach-based motives like hunger, thirst, and sexual attraction.[6]

More to the point, motivations of approach are likely to play a significant role in the psychology of group membership, and can lead to consequences that are negative for outgroups. An

[6] In some cases, of course, a *lack* of beneficial resources could be called a threat, so hunger could be interpreted as the avoidance of starvation, and thirst as the avoidance of dehydration. But it stretches the meaning of "threat" to say that a person with a kitchen full of food is facing threat of starvation when his hunger drives him downstairs for a snack. Similarly, when a person who already has five children feels sexual attraction, it isn't clear what adaptive threat is being avoided, and yet clearly the motivation to reproduce is still doing exactly what it was selected for. More generally, we are skeptical that the entire category of approach-based motives needs to be, or could usefully be, reinterpreted in terms of adaptive threats, understood as the lack of adaptive benefits or any other form of "threat".

unfair hiring decision, for example, can just as easily result from the members of the hiring committee being favorably disposed towards an ingroup applicant as from their being averse toward outgroup applicants. Indeed, there is reason to think that such positive, approach-based motivations can take multiple forms as well. A recent analysis (Moya and Boyd 2015) identified two distinct pressures selecting for two distinct forms of affiliation, one that is operative in the context of *coordination*, the other in the context of *cooperation*.[7] Neither pressure, however, would select for motivations of *avoidance* toward outgroup members. Rather, in cases of coordination and cooperation alike, humans will be *more* motivated to *approach* ingroup members than outgroup members. This is still compatible with the possibility of individuals also approaching outgroup members, rather than avoiding them. Attempting to work with foreigners or unknown outsiders can be a more promising option than not working with anyone at all, and thus foregoing even the possibility of benefits from social interaction.

Therefore, Moya and Boyd also apply the principles of etiological functionalism when they appeal to distinct selection pressures to identify distinct forms of motivation. However, unlike Cottrell, Neuberg, and Schaller, they focus on selection pressures beyond those that take the form of threats. This focus allows them to recognize approach-based, *affliative* motivations that lead to ingroup favoritism. We will return to this in our penultimate section.

V. Gene-Culture Coevolution, Norms, and Tribal Social Instincts: Adding Normative Motivations

In addition to using etiological functionalism to include approach-based motivations within a more comprehensive and integrated picture of group psychology, Moya and Boyd's discussion also broadens the picture along another dimension, which will be our focus in this section. More specifically, it gives an example of how to incorporate culture, cultural evolution, and *cultural selection* into the picture, and illustrates the benefits of doing so.

In cooperative endeavors, they note, a preference for ingroup partners is adaptive because the interactants "have recourse to group-based punitive institutions were their partner to defect" (2015). This claim is supported by a large body of work arguing that culturally inherited

[7] In their own words:

> "There are a number of reasons why people may be motivated to assort with others from the same social category. First, people may be ethnocentric so they can avoid coordination costs by interacting with others who share their same preferences, expectations, or personality characteristics (McElreath et al. 2003). Alternately, people may be motivated to interact with others from the same group for cooperative endeavors, knowing they will have recourse to group-based punitive institutions were their partner to defect (Bowles and Gintis 2004; Boyd and Richerson 1992). The former interactions, which are pure coordination games, differ from the latter, cooperative ones in that there is no incentive to defect on one's partner. However, behavioral patterns of assortment may reflect motivations for coordination, cooperation, or both, and without direct interventions it is nearly impossible to distinguish between them. Therefore, we discuss in-group preferences that may arise from either selection pressure jointly. (Moya and Boyd 2015, p. 15)

Note that while "coordination costs" might sound like a reference to an adaptive threat, the term captures the fact that interactions with outgroup partners are merely less effective at producing adaptive benefits. Being less beneficial is not the same as being a threat or a harm, however, and neither of the two pressures described by Moya and Boyd would would select for motivations of *avoidance* toward outgroup members.

institutions—taken to include norms, laws, and policies—will be favored by selection when they effectively suppress free riding and defection in cooperative interactions (Boyd & Richerson 2009; Chudek & Henrich 2011; Henrich 2004). This research is based on dual inheritance theory, whose central idea is that individual humans inherit traits both genetically, through reproduction, and culturally, through social learning.

Each of these two streams of inheritance gives rise to its own form of selection and fitness. Thus, to say that cultural traits are selected is not to say that they spread by increasing the *genetic* fitness of individuals. Quite the contrary, in some cases cultural traits may spread even while *reducing* the genetic fitness of individuals who adopt them (consider norms of celibacy, or the recent spread of the anti-natalist movement c.f. Brown and Keefer 2020; Richerson and Boyd 2005, Chapter 5). Rather, the claim is that groups in which rules against cheating and free riding are enforced are more capable of collectively generating nonexcludable public goods, such as military defense and public infrastructure. As a result, more cooperative groups grow faster and outcompete less cooperative groups, enabling their culturally inherited traits to spread more widely through the overall human population—including their norms against cheating and free riding. In technical terms, dual inheritance theorists hold that *cultural group selection* has favored the evolution of cultural traits that promote large-scale, prosocial cooperation (Richerson et al. 2016).

These two streams of inheritance interact with each other as well, making it important to distinguish between cultural and genetic selection pressures. For the sake of clarity, we will call those adaptive pressures that *act on genes* "genetic selection pressures" regardless of the "source" of those selection pressures, and so regardless of whether they are generated by the physical, biological, social, or cultural features of the environment. Correspondingly, we will call those adaptive pressures that *act on culture* and cultural items "cultural selection pressures," regardless of the source of those pressures.

To illustrate, Moya and Boyd claim that preferring ingroup partners to outgroup partners is *genetically* adaptive for individuals, because it leads to more adaptive benefits from cooperation. Individuals from the same group will share a common set of norms, and know they will likely be subject to punishment if they violate those norms. By contrast, potential cooperative partners from different groups will not be bound by each other's norms, making cheating and free riding easier and more likely, thus reducing the likely benefits of cooperating with outgroup members. As a result, the genetic fitness payoffs of cooperation with outgroup members are, on average, lower than the fitness payoffs of cooperation with ingroup members. The "source" of these differential selection pressures acting on genes, however, is a social environment filled with prosocial norms and institutions. And these norms and institutions consist of traits that individuals inherit culturally, through social learning. Thus, the genetic selection pressures Moya and Boyd refer to are generated by culturally transmitted norms, and would only have been operative *after* the relevant cultural traits evolved and spread. This is an instance of what dual inheritance theorists call *culture-driven genetic selection* (Henrich 2016), or *gene-culture coevolution* (Richerson and Boyd, 2005).

As a result, the affiliative motivations Moya and Boyd identify are hypothesized to contribute to a set of distinctively human "tribal social instincts" (Richerson and Boyd 2001; Richerson and Henrich 2012; also see Kelly 2013). These genetic adaptations ("instincts") evolved in response to

selection pressures generated by a culturally evolved social environment characterized by large-scale ("tribal") cooperation, in groups of a few hundred to a few thousand people.[8] This coevolutionary dynamic results in a positive feedback loop favoring cooperation. The more cultural selection pressures favor reliance on culturally local norms, the stronger the genetic selection pressures become favoring tribal instincts for internalizing norms and for motivating individuals to comply with and enforce them. But at the same time, genetic adaptations for norm internalization also render cooperative *norms* more effective at producing cooperative *behavior*, thereby giving cultures in which such norms are common further selective advantages over competing cultures.

This coevolutionary dynamic has wide-ranging implications for human social psychology, but for present purposes, the most important ones concerns the capacity to *internalize* norms. Individuals who internalize the norms of their culture become intrinsically motivated to follow them (Sripada and Stich 2007; Chudek and Henrich 2011; Kelly and Davis 2018). To follow a norm out of intrinsic motivation is to "do the right thing" (as specified by the norm) for its own sake, regardless of the consequences of the action, or the instrumental value of obtaining approval or avoiding punishment. Moreover, there is reason to think that when a norm is internalized, individuals thereby acquire intrinsic motivations to enforce the norm as well, sanctioning others who fail to comply.[9] This is striking, since from a functional point of view obeying a norm oneself is quite distinct from punishing those whose break it (see Boyd 2017 for discussion).

Gene-culture coevolution appears indispensable in accounting for internalization and intrinsic normative motivations (Gavrilets and Richerson 2017), a fact which lends credence to the idea that these features of human norm psychology are distinct from the psychological underpinnings of other types of social and rule-governed behavior (also see Kelly 2020, forthcoming). For example, a person can act in accordance with a rule simply out of fear of being punished. From a psychological point of view, this is not the intrinsic motivation associated with internalized norms, but rather a merely instrumental form of motivation. Indeed, the relevant motivational element here would be ordinary fear, rather than concerns about doing the right thing or being a good person. This kind of fear of punishment is just an instance of fear of aggression from conspecifics, an avoidance-based motivational capacity that, as noted above, has a selection history that is shared with other species and that originates much further back in our phylogeny than the emergence of culture and cultural norms. The same can be said of approach-based instrumental motives to follow norms in order to obtain social approval and increased status. Genetic selection

[8] In their own words, "Cultural evolution created cooperative groups. Such environments favoured the evolution of a suite of new social instincts suited to life in such groups, including a psychology which 'expects' life to be structured by moral norms, and that is designed to learn and internalize such norms. New emotions evolved, like shame and guilt, which increase the chance the norms are followed. Individuals lacking the new social instincts more often violated prevailing norms and experienced adverse selection. They might have suffered ostracism, been denied the benefits of public goods, or lost points in the mating game." (Boyd and Richerson 2009, p. 3286)

[9] As Sripada and Stich put it (2007, p. 289), "children who learn that hitting babies is wrong do not need to be taught that one should exhibit anger, hostility, and other punitive attitudes toward those who hit babies." Also see Kelly and Setman 2020 for discussion and review of recent evidence, especially from developmental psychology.

pressures favoring motivations for seeking status were already in place in our primate lineage well before emergence of human-like levels of culture.[10]

By contrast, only after the emergence and proliferation of culturally inherited norms would a distinct "instinct" to follow and enforce rules have enhanced genetic fitness, and only then would genetic selection favor *intrinsic* normative motivations to do the right thing for its own sake. In a social environment filled with norms, it is simply too risky, and too cognitively demanding, to try to make instrumental calculations about all of the rules one needs to be sensitive to at a given time (Gintis 2003; Sperber and Baumard 2012). But reliable enforcement of norms only became a common feature of the human social environment within roughly the last one million years, emerging along with the hypertrophied social learning capacities that make us cultural creatures in general (Boyd and Richerson 2009). Thus, on this picture, while the *genetic* evolution of cultural learning capacities is a necessary condition for the *cultural* evolution of norms, the *cultural* evolution of norms is likewise a necessary condition for the *genetic* evolution of capacities dedicated to norm internalization. In the early days of norm evolution, *all* motivations for following and enforcing norms were merely instrumental. Only after thousands of years of the cycle of gene-culture coevolution would more fully evolved tribal social instincts have emerged and spread, complete with intrinsic normative motivations.

VI. Etiological Functionalism at Work: Righteous Anger, Righteous Disgust, and Other Culturally Inflected Emotional Molecules

These details about cultural evolution are relevant to the psychology of intergroup interactions in a number of ways. For just as individuals may exhibit more or less negativity toward outgroups, so, too, may whole cultures.[11] A cultural group's level of xenophobia, ethnocentrism, or other form of outgroup negativity will be a function of their shared cultural values, expressed in, for instance, norms that license withholding fair and equal treatment to members of other races, and shared belief-like states such as negative racial stereotypes, oppressive scripts, and prejudicial schemas. As noted above, the precultural, genetic selection pressures that Neuberg et al. identify explain why outgroup members are easy targets for ancient motives like fear, disgust, and anger. Against this psychological background, it is all too easy for hostile and avoidant norms to spread through cultural transmission, and the more common such norms become, the more likely it is that they will be internalized by more members of the group.[12] As a result, many members of such a

[10] We should note that in humans, status appears to take two functionally distinct forms: dominance, which has a deep evolutionary history and is found in other species, and prestige, which is culture based and unique to us (Henrich and Gil-White 2001, Cheng et al 2012). We acknowledge this complication, and its implication that there may be at least two distinct types of emotional molecules related to hierarchy in the human psychology of group membership, but set it aside for development in later work.

[11] Also see Davidson (2019) for a pluralist account of racism that appears extendable to other notions associated with outgroup negativity, like xenophobia, prejudice, bigotry, etc. According to Davidson, all kinds of different entities can properly be called racist, including individual people, beliefs, motivations, actions, norms, laws, cultural groups, institutions, etc. Furthermore, on her account no one of those types of entities is more basically or primarily racist than any of the others.

[12] See Buskell (2017) for a discussion of how such precultural cognitive mechanisms can serve as "cultural attractors" that boost the fitness of cultural variants, Nichols (2002) for evidence of a specific case involving culturally transmitted

group will not just experience fear, anger, or disgust toward outgroups. They will also view those feelings as *justified*, such that outgroup negativity is not just common, it is also seen as *right*.

As a result, culturally inherited norms that prescribe outgroup negativity give rise to at least three distinct sources of motivation. First, individuals might be *instrumentally* motivated to follow such norms, in order to avoid reprimand from and to gain the approval of their ingroup fellows. Here, norms that are culturally inherited can leverage more general motivational capacities that predate culture itself, such as desire for status and fear of punishment by ingroup members. Second, to the extent that individuals internalize such norms, they will also become *intrinsically* motivated to obey them. They will thus be be disposed to treat outgroup members poorly because they see it as the right thing do, and will be so motivated independently of any desire for social rewards or fear of potential punishment. Third, on the hypothesis that the process of norm internalization bundles *intrinsic motivations of enforcement* together with intrinsic motivations of compliance, individuals who internalize negative outgroup norms will also be intrinsically motivated to punish other members of their own group who violate such norms, and to reward those who follow them.

Thus, one way in which norm psychology interactions with emotion is by providing the cognitive elements for certain emotional molecules. For example, the intrinsic motivation to punish someone for violating a norm may combine representations of wrongness, or transgression, with the emotional core of anger. We will call this emotional molecule *righteous anger* (c.f. Rozin et al 1999, Cherry and Flanagan 2018), to distinguish it from other emotional molecules in the anger genre, which fall within the same genre in virtue of sharing the core motivational element of an aggressive, approach-based impulse to attack. For example, another emotional molecule in this genre produces the "fight" part of the "flight-or-fight" response. Evolution has selected for what can be called *defensive anger*: an hostile, aggressive response to situations in which an animal believes it is cornered and flight is not an option. Similar aggressive behavior is adaptive in other contexts as well, including those that involve competition for resources. The *competitive anger* that motivates individuals to fight over food scraps is thus yet another distinct molecule in this emotional genre. Righteous anger, then, is a norm-specific subtype of anger. The selection pressures that gave rise to it evolved long after those that gave rise to defensive or competitive anger, because they arose only after the relatively recent evolution of human culture in general.

Consider, as a simplified example, the violent aggression of white segregationists toward civil rights activists participating in lunch-counter sit-ins during the 1960s in the United States. The attackers obviously were not defending themselves, since the activists rigorously adhered to a strategy of passivism. Competition for proximate resources was not at issue, either. White patrons were not concerned that if lunch counters began serving African-Americans, they would run out of food or seating for white customers. Instead, the conflict centered around the *culture* of segregation.[13] The attackers had grown up in the Jim Crow South, internalizing norms according to which segregation is *right*, and integration *wrong*. The civil rights activists were deliberately and flagrantly violating these norms, so righteous anger was a lamentable but preditable response.

---

norms and the emotion of disgust, and Buchanan and Powell (2018) for a compatible but much broader picture concerning disgust, "threat cues", and the spread of "exclusivist" norms and values.

[13] Or, in Wilkerson's (2020) provocative terminology, the conflict was about the American *caste* system.

Of course, cultural evolution has also produced norms *against* outgroup negativity, ("inclusivist" norms in Buchanan and Powell's (2018) terminology, "impartial" norms in Henrich's (2020)). Indeed, the widespread internalization of such norms is a key element in explaining many instances of moral progress, such as the successes of the civil rights movement. Those who fought for it were quite aware of the violent anger they would face, even given their peaceful method of passive resistance. This naturally triggered fear of punishment, an instrumental motivation to *comply* with segregationist norms, rather than flout them. Nevertheless, the protesters persisted. As strong as their fears must have been, other motives were stronger still. Fear was suppressed and overridden by powerful and countervailing motivations derived from internalized norms of justice and equality. Accordingly, both the segregationists and the activists were likely acting from intrinsic normative motivations associated with the norms they acquired from their respective cultures. The difference was that they came from very different cultures, with clashing sets of norms.

A similar account can be given for *righteous disgust*.[14] Just as representations of norm violations can activate the motivational core of anger, they can also activate that of disgust (Haidt et al. 1993; Rozin et al. 1999; Nichols 2004; Kelly 2011; Graham et al. 2013). Righteous disgust should thus be distinguished from the *direct pathogen disgust* that is trigged by outgroup members, as the shared motivational element of disgust is paired with different cognitive elements in each case. While direct pathogen disgust is activated by the classification of individuals as members of foreign groups who may carry dangerous pathogens, righteous disgust is activated by representations of norm-violating behaviors. As noted above, Schaller and Neuberg's (2012) discussion of the behavioral immune system provides an excellent account of direct pathogen disgust, but it fails to address righteous disgust. Both emotional molecules can contribute to poor treatment of outgroup members, but will do so in different ways.

VII. (The Beginnings of) A Taxonomy of the Psychology of Group Membership

In this section we compile and integrate many of the points we have made along the way, recasting them in our own terms. We are hopeful that the taxonomic structure we offer will be able to accommodate and situate more local accounts of different aspects of the psychology of group membership. The recipe for fitting in other emotional molecules will remain the same as the one used here: first identify the relevant motivational capacities, by appeal to selection history, then identify the various cognitive elements that are paired with that motivational core, again by appeal to selection history. Following this recipe, we enumerate eleven distinct molecules that contribute to outgroup negativity and ingroup favoritism, which are simply the ones that we have had the opportunity to consider in developing our argument. This list is, we realize, far from exhaustive.

Anger

[14] Something similar often gets called "moral disgust" (Chapman and Anderson 2013, Kelly 2013, Plakias 2018). We avoid that label here and in our above discussion of anger (i.e see Russell and Giner-Sorolla 2011) to signal that not all internalized norms need be moral norms, and disgust (or anger, or any other emotion) can be activated by a norm transgression whether or not that norm counts as moral (Davis and Kelly 2018, Stich 2018).

The phrase "hostile emotion" may most naturally be associated with anger. As noted above, there are many emotional molecules in this genre, many different cognitive elements that can combine with anger's powerful motivational core. Morever, it seems to us that all of the anger-based emotions identified above may be directed toward outgroups under certain conditions.

Chimpanzees and hunter-gatherer humans alike tend to be territorial and aggressive toward trespassers (Schaller and Neuberg 2012, p. 17), and recorded human history is full of instances of raiding and conquest. Thus, both (1) *competitive anger* and (2) *defensive anger* toward outgroups would have been adaptive throughout much of our evolutionary past. Competitive anger toward outgroup members helps obtain important resources, often by securing territory, while defensive anger helps when one's group is under attack from outsiders trying to seize its resources. The existence of group-level conflict among chimpanzees suggests that the associated selection pressures were already acting upon the genetically inherited traits of individuals before the evolution of human culture. However, raids and conquests are the kinds of collective activities made more effective by cultural norms, so an environment regulated by norms may have further augmented the genetic selection pressures already in place.

By contrast, (3) *righteous anger* would only have come into existence after evolution equipped humans with tribal social instincts and the capacity to internalize norms, since these provide the cognitive elements of this emotional molecule. Righteous anger directed at outgroup members will not be triggered in virtue of their being recognized as outgroup members, but in virtue of their norm-violating behaviors. However, since outsiders will typically not have internalized the local set of norms, they are are also more likely than ingroup members to violate those local norms.

Fear

Unlike anger, fear leads to avoidance, but the two share similarly long and complex etiologies. Like anger, fear was selected to help deal with adaptive threats, some of which took the form of aggression from outgroup members. The presence of intergroup conflict and aggression among chimpanzees suggests that selection pressures favoring (4) *direct fear* of outgroup members existed well before culture became a major factor in human evolution. But as with competitive anger and defensive anger, the effects of culture would likely have further augmented these selection pressures.

Direct fear can then be distinguished from (5) *fear of sanction* along a number of dimensions. Most obviously, direct fear is triggered by outgroup members, while fear of sanction is triggered by the possibility of disapproval *from members of one's own group*, and the punishment they are likely to administer if they catch you violating a local norm. In this latter case, the motivational core of fear is channeled into instrumental behaviors of norm compliance, rather than intrinsic normative motivations. For example, in the Jim Crow South, a white woman who was too friendly with black men could have suffered significant damage to her reputation, affecting her status, marriage opportunities, etc. The psychological state of someone motivated by fear of sanction from her white counterparts is very different from the psychological state of someone intrinsically motivated by

norms of segregation. It is also very different from the psychological state of someone motivated by fear of black men. But in a culture containing racist norms, both subtypes of fear can be prevalent, and both can be just as grimly effective in shaping behavior.

Disgust

We are in broad agreement with the idea that genetic selection would have favored (6) *direct pathogen disgust* toward outgroup members. This emotion is a key component of the behavioral immune system, some ancestral version of which probably evolved long before gene-culture coevolution became a dominant force in human evolution. We also agree with Schaller and Neuberg's (2012, 36) claim that "outsiders are often ignorant of local behavioral norms that serve as barriers to pathogen transmission (e.g., norms pertaining to hygiene, food- preparation); as a consequence, they may be more likely to violate these norms, thereby increasing the risk of pathogen transmission within the local population". Thus, even before culture or norms were on the scene, disgust was already driving negative behavior and attitudes toward outgroups, and there is reason to think that cultural evolution's influence on inter- and intra-group dynamics would have reinforced the selection pressures that had already forged the function of this emotion.

However, there is again a distinction to be made between direct pathogen disgust and the (7) *righteous disgust* that is triggered by norm violations. The latter is likely to have a shallower selection history than the former, since righteous disgust would only have been adaptive after gene-culture coevolution gave rise to tribal social instincts and norm internalization. Of course, righteous disgust only takes on an ethnocentric or xenophobic flavor when the transgressor of the norm is an outgroup member. But norm violations committed by foreigners and outsiders are not likely to be rare. Schaller and Neuberg's point about outsiders being unfamiliar with local hygiene norms actually applies to *all* local norms. That unfamiliarity will lead to a wide range of transgressions, many of them inadvertent. Nevertheless, the righteous disgust triggered by an outsider's *norm violation* is a different emotional molecule from the direct pathogen disgust triggered by indications of membership in a foreign group, observable as "phenotypic abnormalities" in clothing styles, languages or accents, smells or fragrances, skin color, facial features, etc.

Affiliation

Cooperation and coordination are fundamental features of human life. They generate crucial benefits, and both genetic and cultural selection pressures favor capacities to successfully engage in both kinds of social interaction (Henrich and Muthukrishna 2021). At the core of these capacities are approach-based motivations of affiliation, which cause individuals to seek out, team up with, and bond with others. Individuals need to be discerning about whom they interact with, however, so selection also favors cognitive capacities for assortment, which function to evaluate potential partners, and affiliative motivations toward those partners most likely to secure the adaptive benefits of each form of interaction. In coordination, a poor choice of partner renders the enterprise less effective for both partners; the benefits of working together are not worth the costs from either

party's perspective. In cooperation, a poor choice of partner is comparatively worse. An indiscriminate cooperator can be exploited by partners who defect and take a disproportionate share of the benefits. In each case, both genetic and cultural selective pressures have favored cognitive capacities for identifying and keeping track of potential partners, as well as motivations of affiliation toward the more promising ones.

Here again we see a motivational core, which we'll just call affiliation, that can be paired with distinct cognitive elements to form distinct molecules of the same emotional genre.[15] The distinct social dynamics associated with cooperation and coordination, respectively, would have generated distinct adaptive challenges, resulting in cognitive elements with distinct functions that could shape the inputs and outputs to affiliation in different ways. Thus, we add (8) *coordination affiliation* and (9) *cooperative affiliation* to our taxonomy.

Coordination affiliation will be sensitive to a potential partner's language, internalized norms, customs, and practical goals, since social interactions with a partner who shares these characteristics are likely to be more effectively coordinated; styles will align and things are apt to go smoothly. Ethnic boundary markers and other easily perceivable features of appearance, like skin color, will also be salient, since they are relatively reliable correlates of the basic similarities just listed. Even within the same group, coordination affiliation is likely to be directed toward friends, acquaintances, business partners, and employees who are perceived to be more similar to oneself in these ways. But between groups this effect is amplified considerably, since these are exactly the sorts of traits that people from different groups tend *not* to share. Language, values, customs, practical goals and clothing styles are all culturally inherited traits that tend to diverge, through a combination of selection and drift, as cultural evolution takes its own course in separate societies. Thus, coordination affiliation tends to result in favoritism and preferential treatment toward ingroup members over outgroup members (also see Efferson et al. 2008).

Cooperation affiliation, by contrast, will be sensitive to information that indicates a potential partner is likely to resist temptations to cheat, free ride upon, or otherwise exploit you. This includes memories of one's track record of past interactions with a partner, as well as knowledge about the partner's reputation as a cooperator in general. It also includes information about what norms and values they are subject to, and may have internalized. In addition to reciprocity and reputation, another primary reason why people refrain from exploiting other ingroup members in the pursuit of one's own self-interest is that in most cultures doing so is considered *wrong*.[16] Moya and Boyd's (2015) reasoning illuminates this: since norm enforcement is more effective at maintaining prosocial behavior within a given group than across group boundaries, cooperating with ingroup members is the safer choice. Of course, unknown outsiders may have internalized the specific cooperative norms of their own culture, and evidence that they are motivated to behave cooperatively could

[15] There may not be any clear single term for affiliation in the vernacular ("attachment"? "love"? "admiration"? "fondness"? "concern"? "loyalty"? "team spirit"? "patriotism"?) or at least not as clear of a single correlate as in the cases of disgust, fear, and anger. As noted in Section II, however, our etiological functionalism and molecular view of emotions is free to depart from folk conceptions of emotions. We see this as feature of our approach rather than a bug. (Also see Mallon and Stich 2000 on the semantics of thin and thick ways of slicing emotions.)

[16] From a broad evolutionary and historical view this is likely true, even if it may be false of many cultures today. As Marx notoriously pointed out, exploiting others in the pursuit of self-interest may be the very heart of capitalist culture; thanks to Uwe Peters for reminding us of this.

increase their desirability as cooperative partners. Nevertheless, it is typically more difficult to glean such evidence about outgroup members than about one's fellow ingroup members. This makes ingroup members not just the safer choice but the easier one, and it certainly makes them more likely to activate the affiliative motivation that pushes one cooperate. Thus, both the cooperative and the coordinative molecules in the genre of affiliative emotion will tend to lead to ingroup favoritism.[17]

Normative Motivations

Our discussions of righteous anger (3) and righteous disgust (7) above distinguished these emotions from other forms of anger and disgust. What makes these forms of anger and disgust "righteous" is their functional connection to the cognition of right and wrong; in each case, the motivational core characteristic of the emotional genre is paired with tribal social instincts and an internalized norm. However, both emotions occur in the context of norm *enforcement*, in response to the norm-violating behavior *of others*. By contrast, (10) *intrinsic normative motivations* can drive one's own compliance with internalized norms. Those who have internalized norms that prescribe various forms of outgroup negativity or ingroup favoristism will be directly, non-instrumentally, motivated to engage in such behaviors (e.g., using water fountains reserved for white people), for no other reason than they understand this behavior as *right*.[18]

As noted in Section VII, intrinsic normative motivations are distinct from (11) *instrumental motivations* of all sorts, including instrumental motivations to follow or enforce norms. Yet again, as long as the norms themselves prescribe outgroup negativity or ingroup favoritism, instrumental motivations to follow these norms can produce such behavior. The class of motivations that drive instrumental norm compliance is likely to be heterogenous, since incentives take the form of both sticks and carrots; individuals can instrumentally comply with a norm in order to avoid a sanction, but also to gain a reward. Indeed, we have already named one type of avoidance-based motivation that can drive instrumentally normative behavior, namely, fear of sanction. But other instances of instrumentally motivated norm compliance can be driven by approach-based motivations (status-seeking, affiliation, etc.), rather than fear.[19]

---

[17] In the same spirit as the points made in footnotes 10 and 15, we again acknowledge a complication only to set it aside for future work. For there are likely *many* further emotional molecules in this genre, including those that pair affiliative motivation with the distinct assortative capacities associated with different forms of sociality and cognitive wherewithal, including but not limited to: familial love, genetic relatedness, and kin-related selection pressures; romantic love, mate choice, and child rearing-related selection pressures; friendship, camaraderie, and reciprocity-related selection pressures; the positive associative feelings that accompany the kind of interdependence found in social networks large enough that not everyone interacts on a regular basis, but small enough that members need to be able to keep track of everyone's reputations and interconnections; and team spirit, group pride, patriotism, and forms of positive emotional investment associated with differently structured cultural groups and and larger communities, both real and imagined.

[18] While we follow a trend (e.g. Henrich and Ensminger 2016) in describing some psychological motivations as "intrinsic", it is not trivial to spell out what the term means, perhaps other than serving as a contrast class for "instrumental"; see Kelly 2020 for discussion on this and the connection between normative motivations, emotions, and other psychological sources of motivation.

[19] In many situations, of course, the same norms will elicit both intrinsic and instrumental motivations at the same time. But in other cases, one might follow a xenophobic norm specifically in order to gain approval or seek status. Governor Wallace's famous stand in the schoolhouse door in 1963 was a symbolic attempt to resist forced integration at the

VIII. Conclusion

In sum, then, our initial taxonomy of emotions that contribute to outgroup negativity and ingroup favoritism contains three forms of anger (competitive, defensive, and righteous), two forms of fear (direct fear of outgroup members, fear of sanction), two forms of disgust (direct pathogen disgust toward outgroup members, righteous disgust), two forms of affiliation (coordination, cooperation), and two ways of being motivated by cultural norms (intrinsic, instrumental). We have been careful to note that this is neither the end—there are many more emotional molecules to distinguish, study, and incorporate into the framework—nor the beginning, as we take ourselves to be building on and synthesizing important work previously done by others.

In addition to the strengths we have explicitly touted, we will end by noting two more. First, a theoretical one. We hope to have illustrated the power of an evolutionary framework, and particularly dual inheritance theory, to organize and illuminate work across the human and behavioral sciences (also see Richerson and Boyd 2005, especially chapter 7; Muthukrishna and Henrich 2019). Second, a more practical one. Different instances of outgroup negativity and ingroup favoritism—nepotism, cronyism, ethnocentrism, partiality, discrimination, prejudice, intolerance, bigotry, racism, xenophobia, dehumanization—are almost certainly underpinned by different psychological mechanisms, with those differences giving rise to different social dynamics. It is also unlikely there will be a single, one-size-fits-all strategy that will effectively ameliorate every form. Getting clear on the functional character of the various mechanisms and dynamics, and the similarities and differences they bear to each other, is a crucial step in designing interventions and policies more finely tuned to better address each particular type. We hope to have contributed to this project as well, even if less directly.

University of Alabama, which clearly served to garner approval from segregationist voters. He very well might have thought he was doing the right thing as well, in which case it would be an example where intrinsic and instrumental motivations were both in play.

References

Aarøe, L., Petersen, M., and Arceneaux, K. (2020). "The Behavioral Immune System Shapes Partisan Preferences in Modern Democracies: Disgust Sensitivity Predicts Voting for Socially Conservative Parties," *Political Psychology*. https://doi.org/10.1111/pops.12665

Abele-Brehm, A., Ellemers, N., Fiske, S., Koch, A., & Yzerbyt, V. (2020, February 28). Navigating the social world. July 24 2020. https://doi.org/10.31234/osf.io/b5nq6

Bowles, S., & Gintis, H. (2004). Persistent parochialism: trust and exclusion in ethnic networks. Journal of Economic Behavior and Organization, 55, 1–23.

Boyd, R. (2017) *A different kind of animal: how culture transformed our species.* Princeton University Press, Princeton

Boyd, R., & Richerson, P. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology, 13*, 171–195.

Boyd, R., & Richerson, P. J. (2009). Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1533), 3281–3288. http://doi.org/10.1098/rstb.2009.0134

Boyer, P. (2001). *Religion explained: The evolutionary origins of religious thought*. New York: Basic Books.

Brown, F. L. and Keefer, L.A. 2020. Anti-Natalism from an Evolutionary Psychological Perspective," *Evolutionary Psychological Science* 6, 283–291. https://doi.org/10.1007/s40806-019-00226-9.

Brownstein, Michael, "Implicit Bias", *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/implicit-bias/>.

Buchanan, A. and Powell, R. (2018). *The Evolution of Moral Progress: A Biocultural Account*. Oxford: Oxford University Press.

Buskell, A. (2017). What are cultural attractors? *Biology and Philosophy*, 32(3), 377–394.

Chapman, H. and Anderson, A. (2013). "Things Rank and Gross in Nature: A Review and Synthesis of Moral Disgust," Psychological Bulletin, 139(2): 300-327. DOI: 10.1037/a0030964

Cheng, J., Tracy, J., Foulsham, T., Kingston, A. and Henrich, J. (2012). "Two Ways to the Top: Evidence That Dominance and Prestige Are Distinct Yet Viable Avenues to Social Rank and Influence," *Journal of Personality and Social Psychology*, 104(1): 103–125.

Cherry, M., & Flanagan, O. (Eds.). 2018. *The Moral Psychology of Anger*. Rowman & Littlefield.

Chomsky, N. 1959. "Verbal Behavior." *Language 35*(1): 26–58.

Chomsky, N. 1976. *Reflections on Language*. Temple Smith.

Chudek, M., & Henrich, J. (2011). Culture-gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences 15*(5), 218–226. http://doi.org/10.1016/j.tics.2011.03.003

Chudek, M., Zhou, W., and Henrich, J. 2013. "Culture-Gene Coevolution, Large-Scale Cooperation and the Shaping of Human Social Psychology." In K. Sterelny, R. Joyce, B. Calcott, and B. Fraser (Eds.) *Cooperation and Its Evolution*. 425-59. MIT Press.

Cottrell, C. A. and Neuberg, S. L. 2005. Different emotional reactions to different groups: a sociofunctional threat-based approach to" prejudice". *Journal of personality and social psychology, 88*(5), 770.

Cummins, R. 1975. "Functional Analysis." *Journal of Philosophy 72*, 741-764.

Cummins, R. 1983. *The Nature of Psychological Explanation*. MIT Press.

Cummins, R. 2000. "'How Does it Work?' vs. 'What are the Laws?': Two Conceptions of Psychological Explanation." In F. Keil and R. Wilson (Eds.), *Explanation and Cognition* (pp. 117-145). MIT Press.

Curry, O., Alfano, M., Brandt, M., and Pelican, C. (manuscript). "Moral Molecules: Morality as a combinatorial system."

Curtis, V. 2013. *Don't Look, Don't Touch, Don't Eat: The Science Behind Revulsion*. Chicago, University of Chicago Press.

Davidson, L. 2019. *That's (Also) Racist! Entity Type Pluralism, Responsibility, and Liberatory Norms*. Purdue University Graduate School. Thesis. https://doi.org/10.25394/PGS.8985791.v1

Davis, T. and Kelly, D. 2018. 'Norms, Not Moral Norms: The Boundaries of Morality Don't Matter,' *Behavioral and Brain Sciences*, 18-19.

Dennett, D. C. (1978). *Brainstorms*: *Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press.

Dennett, D. C. (1988). "Evolution, Error and Intentionality." In Y. Wilks and D. Partridge (eds.), *Sourcebook on the Foundations of Artificial Intelligence*. New Mexico University Press.

Devos, T., Silver, L. A., Mackie, D. M., & Smith, E. R. (2002). Experiencing intergroup emotions. In D. M. Mackie & E. R. Smith (Eds.), *From prejudice to intergroup emotions: Differentiated reactions to social groups* (pp. 111–134). New York: Psychology Press.

Dovidio, J. F. and Gaertner, S. L. 2000. Aversive racism and selection decisions: 1989 and 1999. *Psychological Science 11*(4), 315-319.

Efferson, C., Lalive, R. & Fehr, E. (2008). The coevolution of cultural groups and ingroup favoritism. *Science* 321, 1844–1849.

Elliot, A.J. 2006. The hierarchical model of approach-avoidance motivation. *Motivation and emotion, 30*(2), 111-116.

Faulkner, J., Schaller, M., Park, J. H., & Duncan, L. A. (2004). Evolved disease-avoidance mechanisms and contemporary xenophobic attitudes. *Group Processes & Intergroup Relations, 7* (4), 333–353.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82,* 878–902.

Gavrilets, S. and Richerson, P. (2017). Collective action and the evolution of social norm internalization, *Proceedings of the National Academy of Sciences*, 114 (23) 6068-6073.

Gelfand, M. (2018). *Rule Makers, Rule Breakers: How Tight and Loose Cultures Wire Our World*. New York: Scribner.

Gil-White, F. (2001). Are ethnic groups biological 'species' to the human brain? *Current Anthropology*, 42 (4), 515-554.

Gintis, H. 2003. The hitchhiker's guide to altruism: Gene-culture coevolution, and the internalization of norms. *Journal of theoretical biology, 220*(4), 407-418.

Glaser, J., and Knowles, E. (2008). Implicit motivation to control prejudice. *Journal of Experimental Social Psychology*, 44(1): 164–172.

Godfrey Smith, P. (2011). *Darwinian Populations and Natural Selection*. Oxford: Oxford University Press.

Gould, S. J. & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings Of The Royal Society of London, Series B,* 205 (1161), 581-598.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., and Ditto, P. H. (2013). Moral foundations theory: The Pragmatic Validity of Moral Pluralism", In *Advances in Experimental Social Psychology* (Vol. 47).

Griffiths, P. E. (1993). Functional analysis and proper functions. *The British Journal for the Philosophy of Science*, *44*(3), 409-422.

Haidt, J., S. Koller, and M. Dias. 1993. Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology* 65 (4): 613–628.

Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization*, *53*(1), 3–35. http://doi.org/10.1016/S0167-2681(03)00094-5

Henrich, J. (2016). *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton, NJ: Princeton University Press.

Henrich J. (2020). *The WEIRDest People in the World: How the West Became Psychologically Peculiar and Particularly Prosperous*. New York: Farrar, Straus and Giroux.

Henrich, J. and Ensminger, J. (2014). Theoretical foundations: The coevolution of social norms, intrinsic motivation, markets, and the institutions of complex societies. In J. Ensminger & J. Henrich (Eds.), *Experimenting with social norms: Fairness and punishment in cross-cultural perspective* (p. 19–44). Russell Sage Foundation.

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

Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology*, 12, 123-135.

Henrich, J., & Muthukrishna, M. (2021). The Origins and Psychology of Human Cooperation. *Annual Review of Psychology*. 72: 207–40.

Kelly, D. (2011). *Yuck! The Nature and Moral Significance of Disgust*. Cambridge, MA: The MIT Press.

Kelly, D. (2013). 'Moral Disgust and Tribal Instincts: A Byproduct Hypothesis,' *Cooperation and Its Evolution*, Eds. K. Sterelny, R. Joyce, Calcott, B, & B. Fraser. Cambridge, MA: The MIT Press. Pages 503-524.

Kelly, D. (2020). "Internalized Norms and Intrinsic Motivation: Are Normative Motivations Psychologically Primitive," *Emotion Review*. June 36-45.

Kelly, D. (forthcoming). 'Two Ways to Adopt a Norm: The (Moral?) Psychology of Avowal and Internalization' *The Oxford Handbook of Moral Psychology*, ed. by Manuel Vargas and John Doris.

Kelly, D. and Davis, T. (2018). 'Social Norms and Human Normative Psychology,' *Social Philosophy & Policy*. 35(1): 54 – 76.

Kelly, D., Machery, E. and Mallon, R. (2010).'Race and Racial Cognition,' *The Moral Psychology Handbook*, Eds. J. Doris et al. New York: Oxford University Press, pages 433 - 472.

Kelly, D. and Setman, S. (2020). 'The Psychology of Normative Cognition,' *The Stanford Encyclopedia of Philosophy*. (Fall 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2020/entries/psychology-normative-cognition/>.

Lycan, W. G. (1981). "Form, Function and Feel." *The Journal of Philosophy* 78 (1), 24-50.

Lycan, W. G. (1995). *Consciousness*. Cambridge, MA: MIT Press.

Mallon, R., and S. Stich. 2000. The odd couple: The compatibility of social construction and evolutionary psychology. *Philosophy of Science* 67: 133–154.

Machery, E., & Faucher, L. (2005). Social Construction and the Concept of Race. *Philosophy of Science, 72*, 1208 - 1219.

Marcus, G. 2007. *Kluge: The Haphazard Construction of the Human Mind*. Boston: Houghton Mifflin.

Mathew S, and Perreault C. (2015) Behavioural variation in 172 small-scale societies indicates that social learning is the main mode of human adaptation. *Proceedings of the Royal Society B: Biological Sciences* 282: 20150061.
McElreath, R., Boyd, R. & Richerson, P. (2003). Shared norms can lead to the evolution of ethnic markers. *Current Anthropology*, 44(1), 123–29.
Millikan, R. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press.
Moya, C., & Boyd, R. (2015). Different selection pressures give rise to distinct ethnic phenomena. *Human Nature*, 26(1), 1-27.
Muthukrishna, M., and Henrich, J. (2019). "A problem in theory," *Nature Human Behavior* 3, 221–229.
Navarrete, C., and D. Fessler. 2006. Disease avoidance and ethnocentrism: The effects of disease vulnerability and disgust sensitivity on intergroup attitudes. *Evolution and Human Behavior* 27 (4): 270–282. DOI: 10.1016/j.evolhumbehav.2005.12.001
Navarrete, C., Fessler, D., and Eng, S. 2007. Elevated ethnocentrism in the first trimester of pregnancy," *Evolution and Human Behavior*. 28: 60-65.
Nichols, S. (2002). On the genealogy of norms: a case for the role of emotion in cultural evolution. *Philosophy of Science*, 69, 234-255.
Nichols, S. 2004. *Sentimental rules: On the natural foundations of moral judgment*. New York: Oxford University Press.
Panksepp, J., & Biven, L. (2012). *The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions* WW Norton & Company.
Plakias, A. (2018). "The Response Model of Moral Disgust," *Synthese*, 195(12): 5453–5472.
Plant, E. A. and Devine, P. G. 1998. Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology 75*(3), 811-832.
Richerson, P., Baldini, R., Bell, A. V., Demps, K., Frost, K., Hillis, V., et al. (2016). Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. *Behavioral and Brain Sciences*, *39*, 55–68. http://doi.org/10.1017/S0140525X1400106X
Richerson, P. and Boyd, R. (2001). "The Evolution of Subjective Commitment to Groups: A Tribal Instincts Hypothesis." *The Evolution and the Capacity for Commitment*. (ed.) R. Nesse, 186-220. New York, NY: Russell Sage.
Richerson, P. and Boyd, R. (2005). *Not By Genes Alone: How Culture Transformed Human Evolution*. Chicago: University of Chicago Press.
Richerson, P. and Henrich, J. (2012). "Tribal Social Instincts and the Cultural Evolution of Institutions to Solve Collective Action Problems," *Cliodynamics*: *The Journal of Theoretical and Mathematical History*, 3(1): 38-80.
Rogers, E.M. (2003) *Diffusion of Innovations, 5th Edition*. Free Press, New York.
Rozin, P., Lowery, L., Imada, S., and Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, 76(4), 574-586.
Rozin, P. and Royzman, E. (2001). Negativity Bias, Negativity Dominance, and Contagion, *Personality and Social Psychology Review*, 5(4): 296–320.
Ruisch, B. C., Anderson, R. A., Inbar, Y., & Pizarro, D. A. (2020). A matter of taste: Gustatory sensitivity predicts political ideology. *Journal of Personality and Social Psychology.* Advance online publication. https://doi.org/10.1037/pspp0000365
Russell, P. S., & Giner-Sorolla, R. (2011). Moral anger, but not moral disgust, responds to intentionality. *Emotion, 11*(2), 233–240. https://doi.org/10.1037/a0022598

Scarantino, A. 2016. "The Philosophy of Emotions and Its Impact on Affective Science." In Barrett, Lewis, & Haviland-Jones (Eds.) *Handbook of Emotions* (pp. 3–48). New York, NY: Guilford Press.

Scarantino, A. 2015. "Basic Emotions, Psychological Construction and the Problem of Variability." In Barrett and Russell (Eds.) *The Psychological Construction of Emotion* (pp. 334–376). New York, NY: Guilford Press.

Schaller, M., 2011. The behavioural immune system and the psychology of human sociality. *Philosophical Transactions of the Royal Society B: Biological Sciences 366*(1583), 3418-3426.

Schaller, M. and Neuberg, S. L. 2012. Danger, disease, and the nature of prejudice (s). In *Advances in experimental social psychology* (Vol. 46, pp. 1-54). Academic Press.

Sperber, D., & Baumard, N. (2012). Moral reputation: An evolutionary and cognitive perspective. *Mind & Language*, *27*(5), 495-518.

Smith, C. A., & Ellsworth, P. C. 1985. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology 48*, 813-838.

Sripada, C. and S. Stich. (2007). A framework for the psychology of norms. P. Carruthers, S. Laurence, and S. Stich (eds.), *The innate mind: Culture and cognition*. New York: Oxford University Press, pages 280-301.

Sripada, C (2020). The Atoms of Self-Control. *Noûs*. 1-25. DOI: 10.1111/nous.12332

Sterelny K. (2016). Adaptable individuals and innovative lineages. *Phil. Trans. R. Soc.* B 371: 20150196. http://dx.doi.org/10.1098/rstb.2015.0196

Stich, S. (2018). "The Quest for the Boundaries of Morality," in Karen Jones, Mark Timmons and Aaron Zimmerman, eds., *The Routledge Handbook of Moral Epistemology*, (New York: Routledge).

Turner, J. C. (1985). Social categorization and the self-concept: A social cognitive theory of group behaviour. In E. J. Lawler (Ed.), Advances in group processes: Theory and research (Vol. 2, pp. 77-122). Greenwich, CT: JAI Press.

Wilkerson, I. (2020). *Caste: The Origins of Our Discontents*. New York: Random House.

Wright, L. (1976). *Teleological explanations: An etiological analysis of goals and functions*. Univ of California Press.