This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

Empathy and the Evolutionary Emergence of Guilt

Grant Ramsey
Institute of Philosophy, KU Leuven, Belgium
ORCiD: 0000-0002-8712-5521
grant@theramseylab.org
www.theramseylab.org

Michael J. Deem
Center for Bioethics and Health Law
Department of Human Genetics, Graduate School of Public Health
University of Pittsburgh
Pittsburgh, PA
ORCiD: 0000-0002-4379-8257
mdeem@pitt.edu

Abstract

Guilt poses a unique evolutionary problem. Unlike other dysphoric emotions, it is not immediately clear what its adaptive significance is. One can imagine thriving despite or even because of a lack of guilt. In this paper, we review solutions offered by Scott James, Richard Joyce, and Robert Frank and show that, although their solutions have merit, none adequately solves the puzzle. We offer an alternative solution, one that emphasizes the role of empathy and post-transgression behavior in the evolution of guilt. Our solution, we contend, offers a better account of why guilt evolved to play its distinctive social role.

Acknowledgments

A previous version of this paper was presented at the University of Kansas. We thank the audience members for their feedback. And we thank Cailin O'Connor and the anonymous referees for their valuable feedback.

1. Introduction. Many emotions strike us as biologically adaptive. Fear, for example, fires our body into action, increasing our heart rate, heightening our awareness, and setting us poised for fighting or fleeing. The individual for whom a charging lion or looming bus does not summon fear will not fare well in the game of life. Some unusual fears—of clowns or escalators, say—might be sources of debilitating phobias, but such cases tend not to lead us to question whether fear plays an overall adaptive role in our species. Guilt, in contrast, presents more of an evolutionary puzzle. While it is hard to imagine someone who never experiences fear flourishing in life, it is not difficult to imagine an individual thriving despite—or perhaps even because of—a lack of guilt, particularly if that individual is adept at concealing transgressions. Might not such individuals have had an evolutionary advantage over those who experience guilt? If so, how do we account for guilt's evolution?

Perhaps the most prominent evolutionary accounts of guilt are those developed by Frank (1988), Joyce (2006), and Scott (2011). Their accounts converge on a similar conclusion regarding the evolutionary role of guilt; namely, experiences of guilt were adaptive for individuals because they served as a countervailing force against motivations to defect on cooperative arrangements or transgress communally accepted normative standards. These accounts take guilt to have evolved to reinforce important prosocial behaviors and to help sustain cooperative arrangements among early humans.

While we agree with these broad claims, we argue that they are insufficient for explaining how guilt achieved an evolutionary foothold in humans. Focusing almost exclusively on what some philosophers have dubbed "anticipatory guilt" (e.g., Greenspan 1995), these accounts say relatively little about two other important aspects of guilt and guilt-induced behavior, which also need explanation. First, empirical research in the psychological sciences has linked guilt to a number of maladaptive effects on the individual, including social withdrawal and psychopathologies (Averill et al., 2002; Bybee and Quiles, 1998; Harder, 1995; Luyten, et al., 2002). This is not to say that if an emotion is an adaptation it can have no associated maladaptive effects. But from a retrospective evolutionary account, we should ask how the benefits of the trait might have overweighed its plausible costs. Second, and perhaps more importantly, any plausible individual-level[1] evolutionary account of guilt must explain not just why the effects of guilt were on balance adaptive for the individual, but also why members of a community would have responded *positively* toward individuals who expressed transgression-induced guilt. Why would the community members respond favorably to such individuals through, say, forgiveness or restoration to communal grace, rather than, say, taking the opportunity to exploit or severely punish them? From an evolutionary standpoint, explaining stereotypical behaviors motivated by *posttransgression* guilt seems to be especially important.

In this paper, we argue that these previous accounts fail to explain fully why guilt was adaptive at the individual level, and we provide the outlines of an evolutionary explanation of guilt that aims to account not just for its commonly identified adaptive features, but also for its apparent maladaptive effects and for the tendency of social groups to forgive and reincorporate guilt-prone individuals. Drawing from the philosophical and empirical literature on guilt, we provide in section 2 a characterization of guilt, its action tendencies, and the role it plays in

[1] It is also possible that guilt evolved in part due to group-level selection (Deem and Ramsey 2016). We focus here on individual-level accounts of guilt, not because we hold that group-level selection for guilt did not occur, but because we want to investigate whether a purely individual-level account is plausible—and if so, what it must accomplish.

This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

contemporary social contexts. In section 3, we assess the details of Frank's, Joyce's, and James's evolutionary accounts of guilt. While we acknowledge that each sheds some light on guilt's origin, we argue that each falls short of explaining guilt within an individual-level selectionist framework. Finally, in section 4, we argue that the previous accounts are missing an important component—empathic concern—which can help explain communal responses to guilt, and how guilt-proneness can be individually adaptive. Our shift toward communal responses to guilt does not diminish the significance of these accounts. Rather, our explanation complements them, showing that the evolution of guilt was likely the result of the interplay of the early emergence of posttransgression feelings of self-recrimination in individuals, an established suite of prosocial emotions underwriting cooperative enterprises, and an informationally and normatively rich social environment.

2. Guilt's nature and function. Following recent trends in emotion research and developmental psychology, we consider guilt to be a distinct emotion, differentiated from other emotions according to its relatively late developmental emergence in humans, its observed cross-cultural presence, and its unique behavioral profile (Fowers 2015; Tangney et al. 2013; Teroni and Deonna 2008; Treeby et al. 2016).

Because many emotions evolved to play specific adaptive roles within ecologies significantly different from those of the present, evolutionary accounts of such emotions will be speculative to some degree. This is particularly true of an emotion like guilt, which is tightly linked to complex social behaviors that play an important role in the moral, social, and legal lives of contemporary humans, but might have played other social roles at its evolutionary emergence.

Our account here operates under the presupposition and not proof that guilt is an adaptation.[2] Despite this limitation, one can hedge an evolutionary account of guilt against the charge of being another just-so story by providing a conceptually clear and empirically informed picture of guilt and its behavioral profile. Drawing from contemporary psychology, comparative biology, phylogenetics, and social scientific research programs on guilt and its social role, we can find clues about guilt's original function and thereby avoid an altogether speculative picture of how guilt was initially adaptive. And while guilt plays a somewhat heterogeneous set of roles within and across contemporary cultures, there are also patterns of similarity in the psychology and social roles it plays—patterns that are likely to extend into the distant past. Thus, the effects guilt has on the individual and the roles its expression plays in contemporary social contexts provide some evidence about its original evolutionary function. Further situating guilt within the ecological and social conditions in which it likely emerged in humans, and considering other cognitive and affective traits that we have good reason to think were already present at this emergence, prevents our account from being overly speculative (Griffiths 1997; Sterelny 2012).

[2] Prinz offers a non-evolutionary account of guilt, in which he asserts, but does not defend, the view that guilt is "a product of nurture that builds on other emotions, a desire for affection, and a general capacity for learning" (Prinz 2004, 129). It is not clear, however, how one adjudicates simplicity or conservativism among evolutionary accounts and Prinz's multi-factorial composite account. Further, the neurological, biological, and psychological research largely portrays guilt to be a distinct adaptation that is modulated by culture, defying a simple innate vs. learned dichotomy. Deem and Ramsey (2016) provide a summary of this research.

This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

Much of the psychological research on guilt converges on several core phenomenological and cognitive elements. Research subjects routinely group guilt with other psychologically painful emotions, such as shame (Harder 1995; Tracy and Robins 2006), and typically identify guilt's object as a particular set of past actions that constitute transgressions against accepted normative standards (Tangney et al. 2011). The tight link between guilt experiences and the attending judgment that one is *responsible for* such transgressions underwrites one dominant characterization of guilt by the psychological sciences as a primarily *action-focused* emotion, in contrast to other psychologically painful emotions, such as shame, that focus chiefly on some aspect of the self (Barrett 1995; Drummond et al., 2017; Tangney 1996).[3]

In addition to these core phenomenological aspects of guilt experiences, the anticipation of guilt and posttransgression experiences of guilt powerfully alter an individual's motivational profile and behavior. Anticipatory guilt can serve as a powerful counterweight to motivation for actions that transgress accepted standards, defect on cooperative arrangements, or harm others

[3] We remain neutral on whether guilt arises from the transgression of a moral standard or social norm itself, or from undermining a standard or norm that one personally values. Batson, for example, contends that guilt arises when one anticipates violating a norm that one *values*, rather than from norm violation generally (Batson 2015). According to Batson, this perspective on guilt helps to explain why individuals are sometimes disposed to merely appear to adhere to some social norms while concealing their transgression of these same norms. Our account of guilt's evolution is compatible with Batson's picture, since we do not argue that guilt is experienced and acted upon with every conscious transgression. However, as even Batson notes with respect to norms that the individual values, guilt displays do regularly occur after some transgressions.

This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

(Batson 2015; Svensson et al. 2013). In posttransgression scenarios, guilt typically motivates reparative actions on the part of the transgressor, particularly toward those directly harmed, as well as self-punitive behavior, including acceptance of punishment or self-administered penance (Lindsay-Hartz et al. 1995; Radzik 2009; Silfver 2007).

The sociological and legal literature provide a fuller picture of the emotion's social function. The expression of guilt pulls in two directions within the legal arena. If someone who is accused of a crime exhibits remorseful behavior, this behavior will often be taken as evidence that they are responsible for the crime (Bornstein et al. 2002; Jehle et al. 2009). While displays of remorse make it more likely that one will be convicted of a crime, in Western legal systems remorse generally has a dampening effect on sentencing (Garvey 1998; Gold and Weiner 2000).

This phenomenon might be explained in two ways. One is that the experience of remorse itself could be considered punishment, so the court is not moved to inflict as much external punishment in order to receive parity of punishment with the remorseless. Another explanation is that individuals who exhibit remorse are indicating that they are unlikely to recidivate, that is, commit the same or a similar crime again (Hosser et al. 2008). Either option, or even a combination of the two, explains why expressions of guilt can benefit individuals in certain contexts, despite incurring some cost. And with additional premises (like transgressions should receive parity of punishment, or that the function of punishment is crime reduction), one can account for why judges or juries should view guilt expressions as mitigating factors in sentencing. But both explanations already assume at least a loosely structured penal procedure that already acknowledges that posttransgression expressions of guilt regularly occur. What evolutionary pressures would have resulted in these posttransgression expressions of guilt being common among humans in the first place? We now turn to consider and critique three representative accounts of the evolution of guilt from the behavioral scientific and philosophical literature. While these accounts provide important insights into the effects of guilt on the individual, we contend that each contains serious deficiencies in accounting for the evolutionary emergence and maintenance of guilt.

3. Recent individual-level accounts and their shortcomings. Recent evolutionary accounts of guilt in the philosophical and scientific literature have drawn the conclusion that guilt proneness is a straightforwardly adaptive trait, given that it prompts prosocial and reparative behaviors (e.g., Broom 2003; de Waal 1996; James 2011; Joyce 2006). But such prosocial and reparative behavior cannot be taken for granted. It is not clear why at guilt's emergence guilt-prone individuals were not just taken advantage of when guilt was expressed after transgression, quashing its subsequent evolution. Even if we were to conjecture that an important benefit of guilt expression is that the group views guilt experiences as the individual imposing self-punishment, which helps to explain dampening effects of guilt expression on sentencing in contemporary legal contexts, we would still need to determine why members of early human groups *initially* responded to guilt expressions in this way.

The forgoing considerations of the nature and contemporary effects of guilt show us precisely what an individual-level evolutionary account of guilt requires. Such an account must explain not only how the prosocial and reparative actions that guilt induces would have provided benefits to the individual, but also why the guilt expressions of norm transgressors tended to influence group response in ways that were beneficial to individuals. And, more specifically, they show why focusing on anticipatory guilt alone will occlude what appears to be the more difficult evolutionary story tell—namely, why *posttransgression* psychosocial effects and behaviors that appear to be costly to the individual would have been favored by selection.

3.1 Frank's commitment model of guilt. Frank (1988) includes guilt among a suite of emotions that, he contends, evolved to enable individuals to make credible commitments with one another, yielding long-term payoffs. On Frank's view, these payoffs are more likely to be realized if individuals maintain a firm commitment to cooperative arrangements, even when one or more party stands to benefit more by pursuing a strictly self-interested course of action at the expense of other group members. According to Frank, emotions underwrite cooperation in two ways. First, emotions such as love, envy, and guilt incentivize individuals to follow cooperative terms and provide a counterweight to impulses to cheat or defect. For instance, the anticipation of guilt can diminish the allure of cheating for a larger individual payoff. Second, individuals who are recognized as disposed to experience emotions like guilt and sympathy will be sought out by others for cooperative ventures.

Frank suggests two evolutionary pathways by which emotions might have emerged. First, along the "reputation pathway," individuals who consistently resist the strong urge to cheat acquire a good reputation, the transmission of which leads to further opportunities to benefit through cooperative ventures. Second, along the "sincere-manner pathway," the experience of emotions is associated with involuntary, hard-to-fake facial expressions, which others can use to draw inferences about whether an individual is a reliable cooperator. In both scenarios, the choice of reliable cooperators increases the selective pressure on emotional dispositions.

Because Frank counts guilt among the emotions favored by selection to serve as commitment devices, we should ask whether either of his proposed pathways is a plausible evolutionary scenario for guilt's emergence. Consider first the sincere-manner pathway. Empirical studies on emotions and their associated facial expressions have shown that, in contrast to other social emotions, there is little evidence that guilt is associated with a stereotypical bodily signal by which it can be readily identified (Keltner et al., 1996; Wallbott 1998). Prinz (2004), for one, takes the lack of evidence for a distinctive physiological signal for guilt as reason to suppose that guilt did not emerge along the sincere-manner pathway. However, as we will argue in section 4, a modified version of this account yields considerable explanatory power. Much turns on whether the sincere signal needs to be an involuntary physiological change. But for now, let's consider whether Frank's reputation pathway might fare better.

While reputation might have played some role in the evolution of guilt proneness, Frank's account does not explain why guilt proneness was itself a particular target of selection. Outside parties would be making inferences only about an individual's adherence to cooperative terms or general disposition to experience some set of social emotions. But whether the anticipation of guilt—and not, say, sympathy or fear—induces commitment in a given instance of cooperation would be opaque to observers. By merely positing that guilt was among a host of prosocial emotions that came under selection pressure at some point because they contributed to beneficial cooperative ventures, the reputation pathway fails to explain why selective pressures would have targeted guilt *specifically*, leaving guilt's function undifferentiated from those of other prosocial emotions. On this view of guilt, we are warranted only in concluding that these pressures targeted whichever dispositions happened to be present, guilt or no guilt. This seems correct, as far as it goes, but the reputational pathway account does not provide much by way of explaining why guilt itself evolved, since it dilutes the role guilt plays in prosocial behavior among other social emotions. Reputations would only need to track the resulting cooperation, not

This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

the particular disposition or motivation to cooperate. In our critique of Joyce's and James's evolutionary accounts below, we provide reasons why a predefection role for guilt does not fully explain why it might be an adaptation.

But consider a more serious problem for this account. Frank's reputational pathway account considers reputation to be an important evolutionary driver for guilt. Now, the cognitive load involved in receiving, retaining, and transmitting information about the shifting reputations of multiple potential cooperators seems to require the sort of complex psychological machinery that would have evolved only in the context of complex human social interaction (Deem and Ramsey 2016; Sterelny 2012). Thus, Frank's view of the reputation pathway appears to presuppose that which it wishes to explain, namely, communal stability and commitment within complex social environments.

3.2 Joyce's and James's self-recrimination models of guilt. Like Frank, Joyce and James conceptualize guilt primarily as a kind of internal check on urges to defect or cheat on cooperative ventures. In contrast to Frank, Joyce and James seek to develop in more detail the specific role guilt plays in cooperation rather than leave its function undifferentiated from that of other social emotions. Joyce characterizes guilt as an "internal self-punishment system" (2006, 70) that "guides action 'from the inside'" (101). The action-guiding element of guilt stems from what Joyce sees as its close association with moral judgments about particular types of actions as deserving of punishment. On his view, this package of moral judgment and guilt was selected for because it "reinforced in a motivation-boosting way" (113) other social emotions, increasing the "likelihood that certain adaptive social behaviors [would] be performed" (114). James (2011) follows Joyce's self-recrimination model, suggesting that guilt involves "feeling our wrongdoing deserves punishment" (56), functioning as a "check" (75) on temptations to transgress norms. On the self-recrimination model, individuals wish to avoid the painful experience of guilt, making them more compliant with group norms and, consequently, better cooperators.

Joyce's and James's accounts face two problems, however. First, they portray guilt as straightforwardly adaptive, underwriting an individual's prosocial behavior. But the question of whether guilt really is adaptive for the individual is more difficult than either lets on. Guilt has the potential to incur significant burdens for the individual, even as it plays the self-recrimination role with which Joyce and James associate it. Clinical studies of guilt proneness in individuals show a significant correlation between guilt experience and individual psychopathology, including depression, self-loathing, and heightened anxiety (Harder 1995; Zahn-Waxler et al. 2012). Any individual-level account of guilt, then, must explain how these ostensibly maladaptive effects on the individual were offset by the benefits accrued by being guilt prone.

Perhaps one might argue that it is enough that guilt disposes individuals to resist temptations to defect on cooperative arrangements, thereby further strengthening cooperative tendencies that are underwritten by other prosocial emotions. This line of argument faces a significant challenge: Why would selection favor a novel, complex emotion with a presumably high maintenance cost just to reinforce these tendencies? Consider that recent neurobiological and primatological research shows guilt to be a cognitively complex emotion that might be unique to humans. For example, neurobiological research takes guilt experiences to be produced by complex subcortical and neocortical processes, which are associated with reduced asymmetry in right and left cortical activity and indicative of a unique simultaneous orientation toward withdrawal and approach behaviors (Amodio et al. 2007; Moll et al. 2008; Panksepp and Biven 2012). This research supports both primatological research and philosophical analyses that posit guilt as a relatively late phylogenetic addition in humans, which emerged only after early human communities had developed sustained cooperative structures and cultural systems of transmission and enforcement of social norms (Boehm 2012; Fessler and Gervais 2010). Moreover, as the contemporary psychological and legal literature we discussed in section 2 shows, an important background condition for guilt experience is the individual's capacity to recognize norms of behavior and to evaluate and take responsibility for one's actions. As several biologists and philosophers have argued, the capacities to accept social norms and make evaluative judgments about one's actions and those of others likely emerged comparatively late in human evolutionary history (Deem 2016; Laland and Brown 2011; Silk and Boyd 2010; Sterelny 2012). As human social arrangements achieved sufficient complexity, phylogenetically older tendencies instilled by, say, kin selection or reciprocal altruism likely became insufficient for generating cooperation on large scales.

If, as this broad set of research suggests, guilt is a neurobiologically complex and cognitively demanding emotion that emerged phylogenetically late within systems of human cooperation and cultural transmission, then it seems unlikely that selection would favor a novel and complex emotion *simply* to serve as a psychological reinforcement of cooperative tendencies that were underwritten by a suite of other, more phylogenetically ancient, emotions, such as sympathy, empathy, or fear. This is not to mention the potential fitness costs that guilt incurs pre- and post-transgression. For this scenario to be plausible, we need to know ways in which the fitness landscape changed, rendering these more phylogenetically ancient emotions and tendencies less reliable for underwriting cooperative arrangements and preventing individual defection. Importantly, Joyce and James ascribe to guilt the primary evolutionary function of reinforcing fixed cooperative tendencies but do not show how this outweighs the costly

This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

behaviors associated with guilt experiences that seem to have the most important social ramifications for the individual, namely, postdefection expressions of guilt. One current role of guilt may well be the regulation of norm transgression, but in order for this to be a plausible candidate for guilt's evolutionary function, we would need to know how it yielded sufficient biological benefit to offset the potential fitness costs of postdefection, guilt-induced behaviors.

This is not to deny that guilt *can* serve as a motivational counterweight to temptations to defect on cooperative schemes. Indeed, as we noted in section 2, the anticipation of guilt often modifies our motivational profile, dampening the allure of violating norms. But it is to cast doubt on the notion that guilt was favored by selection *primarily* to serve this purpose or merely to increase the aggregate strength of the set of more phylogenetically ancient prosocial emotions. When we consider guilt's late evolutionary emergence in humans along with its unique social and behavioral profile *after* transgression, we see that attributing the biological function of guilt solely to its role as a motivational counterweight leaves significant explanatory gaps in the evolutionary account.[4]

This leads to a second and more significant problem for Joyce's and James's accounts: neither considers how the action tendencies of guilt in postdefection scenarios would have been adaptive for individuals. Recall from the discussion in section 2 that guilt frequently induces a number of potentially costly behaviors for the individual, including public confession to wrongdoing, submission to punishment, and self-penance. In contemporary social and legal

[4] Whether the anticipation of guilt and its role in modifying one's motivational profile are incidental effects of the main evolutionary function of guilt, or are themselves among the original evolutionary functions of guilt, is a question on which we remain neutral.

settings, guilt-induced behaviors tend to be met by specific responses by the community that benefit or mitigate harm to the individual who exhibits them. Neither Joyce nor James considers how guilt-induced behaviors render the individual vulnerable to group response to expressions of guilt, or why group response carries potential benefit to the individual. Why would members of early human groups respond positively to guilt-prone individuals? Guilt would hardly have been a boon to the individual if the expression of guilt were routinely discounted, ignored, or exploited by the community.[5]

Alternatively, and we think more plausibly, guilt-proneness might indeed reinforce these tendencies, but was likely favored, at least initial, by selection for its role in *restoring* cooperative arrangements after transgression. If this is right, guilt would still largely fit the descriptive profile that Joyce and James sketch when an individual considers defecting on communal norms or moral commitments, but would also enhance an individual's derived benefits from cooperative enterprises insofar as guilt-induced behaviors play important roles in the restoration of relationships that norm transgressions altered. Joyce's and James's accounts do, however, shed significant light on *why* individuals might have been disposed to reveal

[5] One might object that the presence of prosocial emotions would preclude or dampen drives to exploit the guilt prone. We might respond by noting that while prosocial emotions such as sympathy and empathy dispose individuals to help, assist, refrain from harm, etc., they are variably expressed in behavior and might be counterbalanced by other evolved dispositions. We could think of exploitation in terms of sadistic expressions of, say, anger or even malice, but we can also think of it in terms of severe punishment that is viewed as justified within contexts where punishment norms are well established.

otherwise concealed transgressions or perform actions such as apology and restitution, despite facing punishment from others. On these self-recrimination accounts, individuals often feel they ought to be punished. But even if this self-recrimination prompts such behaviors, we still need an explanation for why revealing one's transgressions or placing oneself before the mercy and judgment of the group appears to be a stable adaptive strategy. In the following section, the account of the evolutionary origins of guilt we introduce uses an ingredient absent from those of Frank, Joyce, James, and others: empathy. Empathy in humans, we argue, likely preceded the evolution of guilt, and this fact is a key to the full understanding how guilt evolved. As we will show, the inclusion of empathy in an evolutionary account of guilt uniquely enables us to make sense of posttransgression responses to expressions of guilt, thereby laying the groundwork for an individual-level adaptive story for guilt's origin and maintenance.

4. Guilt, empathic distress, and the restoration of cooperation

4.1 Explaining group response to guilt. If guilt is potentially psychologically and socially maladaptive for the individual, as the empirical literature suggests, how might we explain the evolution of guilt without recourse to a group selection model? We are not assuming that group-level selection scenarios are outright untenable. Indeed, it may be that being composed of guilt-prone individuals provided groups with a competitive advantage. There are ongoing debates about the tenability of group selection models and whether group selection can in some cases swamp the effects of individual-level selection. Some group selection models have been strengthened by the addition of culture, since culture can have the effect of increasing intragroup homogeneity and intergroup heterogeneity, increasing group-level selection pressures. Such models are increasingly used to account for the evolution of human cooperation (Henrich 2004;

This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

Richerson et al., 2016). However, it is not yet clear that cultural group selection can provide adequate explanations of the evolution of guilt. As Nesse (2016) argues, cultural group selection "has a hard time explaining the pervasiveness and intensity of guilt, motivations for reparations, extreme sensitivity to what others think, concern for others' welfare, pity, commitment, empathy, philanthropy, and pride in generosity" (35). Furthermore, group-level explanations suffer from the fact that it is unclear whether group-level selection alone was strong enough for guilt to evolve (Deem and Ramsey 2016). Such explanations appear better at accounting for the *spread* of guilt proneness through the species than the *origins* of guilt proneness in individuals.

Even if one thinks that one can produce tenable group selection models of the evolution of guilt, it is important to ask whether such models are necessary—whether, that is, there are viable evolutionary accounts of guilt that do not require the resources of a group-level selection framework. At any rate, our primary aim in this paper is to identify the minimal components required for an individual-level selectionist account of guilt, remaining neutral on whether and to what degree guilt proneness as a trait gained an evolutionary foothold through genetic or cultural (e.g., individual-to-individual, across generations) transmission.

Perhaps the way in which an individual tends to alleviate guilt can provide some guidance here. There is evidence that the psychologically maladaptive effects of guilt are strongly mitigated by opportunities for the guilt-experiencing individual to make amends with those parties who were harmed by a particular transgression (Estrada-Hollenbeck 1998). But the alleviation of guilt along this route presupposes that the group members are willing to restore the guilt-prone individual to some positive degree of social standing and reincorporate the individual in cooperative enterprises. This is precisely what the aforementioned individual-level accounts overlook: guilt-prone individuals would likely be at a significant disadvantage if guilt induced behaviors were *not* met with positive responses from group members. In this scenario, it might indeed pay to be successfully deceptive or demure about one's own transgressions, while signaling falsely one's acceptance of communal norms. Thus, merely pointing to the way guilt checks motivations to defect, or even noting the reparative behaviors it induces, is not sufficient to show why guilt was adaptive for the individual. Moreover, merely positing that guilt made individuals better facilitators of community benefits is too vague; it does little to differentiate the evolutionary and behavioral profile of guilt from those of other social emotions, and altogether ignores what is most puzzling about guilt's evolution. The communal attitudes and responses to postdefection displays of guilt must be given an important place within any plausible individual-level explanation of guilt.

The foregoing discussion suggests that constructing an adequate individual-level account of guilt demands explanation both of why guilt proneness is adaptive for the individual, despite its connection to maladaptive behaviors and psychopathology, and why others tend to forgive and reincorporate, rather than exploit or banish, individuals who perform costly reparative behaviors.

4.2 Behavioral Regulation and Post-transgression Risks. Guilt's role, then, as a regulator of norm transgression and indicator of such regulation, while important, is a poor candidate for the emotion's main evolutionary function. If an account focuses only on how guilt affects the motivational and behavioral profiles of individuals, particularly as it mitigates temptation to defect, then it neglects what is perhaps the more evolutionarily significant dimension of guilt; namely, how it elicits responses from conspecifics that benefit the guilt-prone individual. This is because experiencing and signaling guilt will not be favored by selection if it is met by negative responses from conspecifics that level high costs to the individual. Any plausible evolutionary perspective on guilt, then, must explain why individuals who experienced and displayed guilt altered the motivational and behavioral profiles of *their conspecifics*, effecting the individual's posttransgression reintegration into communal life.

As we discussed in section 2, an individual's expressions of guilt in contemporary social and legal contexts are often met with responses by others that produce some benefit for the individual (e.g., reincorporation into cooperative arrangements), or mitigate costs imposed on the individual due to others' perception of the individual's responsibility for a transgression (e.g., leniency in legal sentencing). An evolutionary account of guilt must consider these posttransgression benefits that expressions of guilt provide for the individual. But what accounts for the broadly positive attitudes of conspecifics toward an individual who, motivated by guilt, indicates that s/he is suffering the pangs of guilt or that s/he wishes to repair damage to relationships caused by norm transgression?

To answer to these questions, we must consider what dispositions were already present in individuals prior to the evolutionary emergence of guilt such that guilt-induced behaviors would have been regarded positively, leading to benefits for the guilt-prone individual. The earlier evolution of empathy in humans, we contend, provides a crucial piece of the explanation for how guilt might have been individually adaptive at its emergence.

4.3 Empathic response as a key evolutionary driver. There is considerable variability within the empirical philosophical literature with respect to the precise nature of empathy, and

This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

we do not attempt here to provide a complete descriptive account.[6] However, there is some convergence within the contemporary literature on at least three key features of empathy, which we take to be important to explaining how guilt gained an evolutionary foothold. First, empathy has an affective aspect: an empathic state involves an experiencing of the positive or negative valence of another individual's affective state (Coplan 2011; de Waal 2006; Hatfield et al. 2009). Whether this includes an additional epistemic state of being aware of how that individual feels or clear self-other differentiation is a matter of considerable philosophical debate into which we do not enter here (Batson and Weeks 1996; Coplan 2011; Smith 2017). Second, empathy tends to be *self-focused*: the experience of empathy primarily involves focus on one's own experience of this negative or positive valence, as opposed to taking on another's perspective or imagining oneself as if being in another's position (Batson et al. 1997; Miller 2011; Snow 2000). Third, empathy is associated with a behavioral response to one's negative or positive affective experience. Some researchers claim that empathy typically motivates behaviors aimed at enhancing the welfare of another individual and to produce a positive affect *in one's self* (de Waal 2008; Eisenberg et al. 2006).

[6] Batson (2009), Coplan (2011), and Smith (2017) note that the term 'empathy' is used by philosophers and scientists to pick out a number of different neurological, psychological, and behavior phenomena, including mirroring or catching others' emotional states, imagining other's affective states, picturing ourselves as experiencing others' affective states, and feeling the others' emotions. The very broad description of empathy on which we rely here is consistent with most of these characterizations, and we do not attempt to provide a fine-grained analysis of the emotion.

This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

At the very least, it seems that empathic response is motivated in large part to enhance one's *own* positive affect or diminish one's own negative affect, and this often involves actions directed toward another insofar as perception of the latter's affective experience plays a determining role in one's empathic experience (Batson et al. 2016). This minimal conception of empathy, which some social psychologists have called "empathic" or "personal" distress (Batson 2009; Hoffman 1981), stands in contrast to more robust conceptions of empathic concern that include eliciting behaviors also aimed at relieving the distress of others.[7] For our purposes here, we assume what seems to be a baseline consensus that empathy involves at least the experiencing of the positive or negative *valence* of another's emotional state and motivates behaviors that are associated with preserving or alleviating this euphoric or dysphoric experience. These two features of empathy, we contend, help to explain how guilt evolved in early human social contexts and rendered guilt adaptive for individuals.

For our claim to be plausible, we first need evidence that empathy preceded guilt on the evolutionary timescale and that empathy plays a significant role in the social restoration of individuals who transgress norms and subsequently express their guilt experiences. There is empirical evidence that the evolutionary emergence of empathy preceded that of guilt. Animal researchers claim that rudimentary forms of empathy are phylogenetically widespread, being found in a range of taxa (Langford et al. 2006; Povinelli et al. 1992; de Waal 2008). De Waal for example, claims that nonhuman primates exhibit susceptibility to other troop members' negative affective states. This phenomenon, frequently dubbed *emotional contagion*, involves negative

[7] For a helpful disentangling of the many senses of 'empathy' in the philosophical and psychological literature, see Batson (2009).

affective states in individuals inducing “a matching or closely related state” in others, and motivating response behaviors aimed at relieving the distress caused by the shared states (2006, 26). In contrast to empathy and its rudimentary forms across taxa, the cognitive machinery underlying guilt experiences is highly complex and perhaps unique to humans, suggesting that guilt might not be phylogenetically widespread (Amodio et al. 2007; Boehm 2012; Deem and Ramsey 2016). This provides evidence that guilt had a later evolutionary emergence than even the more robust form of empathy described by De Waal, and we can plausibly maintain that guilt in humans evolved within a social context in which, minimally, susceptibility to empathic distress was already established.

The claim that empathy provided a pathway for guilt to evolve can be strengthened by considering current psychological research on the experience and effects of empathy. Empirical studies of empathy-related responses of children and adults show that among empathy’s primary functions are to render subjects sensitive to the emotional distress of others, to vicariously participate in this distress, and to prompt behaviors aimed directly at its alleviation in the agent or both parties, which frequently is achieved via the enhancement of the welfare of the party in whom emotional distress was initially detected (Eisenberg et al. 2006; Zahn-Waxler et al. 1995). Successful alleviation of this distress is associated with experiences of positive affect, suggesting empathic response is associated with benefit to the empathic party (Batson and Weeks 1996). Experiences of empathy are also strongly associated with diminished anger and aggression toward others (Harmon-Jones 2004; Jagers et al. 2007; Strayer and Roberts 2004), and there is evidence of a proportional relation between empathic capacity and guilt proneness (Treeby et al. 2016).

This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

From this psychological research on empathy, we can draw two plausible (but non-demonstrative) conclusions about the evolutionary interplay between empathy and guilt. First, as we observed in section 2, guilt experiences are negative affective states. From an evolutionary perspective, early experiences of guilt, if detected by others (more on this below), would likely affect others' emotional states to some degree. Sensitivity to the emotional distress of guilt in others, then, would likely have influenced others' motivational profiles via empathic distress at the very least, potentially prompting bystanders to behave in ways to alleviate or eliminate their own distress, perhaps along with the distress observed in the other. Second, this empathic experience in view of another's distressing guilt experience would likely have dampened anger and aggression toward that individual. While aggression and anger would have been responses to an individual's perceived norm transgression—and, again, we take no stand on whether what is represented is the transgression itself or a relation-dependent property of representing harm—the experience of empathy would potentially reduce urges to severely punish, return harm, or expel from cooperative arrangements.

One effect of empathy is that the perceived suffering of another individual causes in one distress and compels one to relieve one's own negative affect, often via attempts to relieve the other's pain.[8] Now, this need not be motivated by an express concern for the other—indeed, we can imagine that relieving another's distress could be taken merely instrumentally as a way to alleviate one's own empathic distress. Our ancestors, then, would also have been compelled to

[8] Again, we leave aside the question about the actual *motivation* one has in such behaviors—whether it is exclusively to relieve one's own distress or also admits the drive to relieve the perceived distress in others.

reduce the suffering—physical and psychological—of others, either directly or indirectly. The emotional suffering of guilt, of course, would be something transgressors would try to avoid and to ameliorate. But as long as others in the group were susceptible either to empathic distress or empathic concern, many would be inclined to aid in mitigating and eliminating this suffering. The group could do this in two ways: the anticipated pain of guilt would lead to group members encouraging others to avoid transgressions (e.g., "Think of how you'll feel"; "How could you live with yourself?"), and detecting guilt experiences in individuals would reduce aggression toward transgressors, encourage group members to forgive transgressors, or relax the implementation of punishment norms on transgressors. Thus, according to our evolutionary perspective on guilt, the key affective and behavioral aspects of empathy noted in the empirical literature would have contributed to the individual-level benefit of guilt experience and expression at their evolutionary emergence. Furthermore, an individual's expression of guilt in posttransgression scenarios could then be an effective adaptive strategy for alleviation of affective distress and reincorporation in cooperative ventures.[9]

4.4 Reliable guilt signaling as protection against exploitation. For guilt-induced behavior to be a fairly reliable indicator of whether an individual is experiencing emotional distress over transgressing norms and whether they are likely to recidivate, individuals' empathic capacities

[9] Some philosophers and psychologists have argued that empathy does not necessarily lead to norm acceptance, or that actions motivated by empathy are always morally praiseworthy (Bloom 2016; Prinz, 2011). We do not here take a stand on these questions in contemporary moral psychology.

and the group's punitive system cannot be open to easy exploitation. After all, there may be large payoffs to individuals who can successfully feign experiencing guilt and gain forgiveness without the intent of changing their behavior—for example, by exhibiting a "hangdog" look without actually feeling any guilt. Psychopaths, to cite a contemporary example, have high recidivation rates, and yet are the most successful at gaining conditional release when they go up for parole (Porter et al., 2009). Intuitively, one might find it plausible that individuals who were particularly adept at concealing their transgressions or feigning guilt behaviors could reap the benefits of group forgiveness without suffering from feelings of self-recrimination or other maladaptive effects. Wouldn't such individuals be better off than individuals who experience guilt and display guilt-induced behaviors? If all things were equal, this might indeed be persuasive. But all things are not equal. If guilt has a relatively late evolutionary emergence in hominins, as we have supposed, then it seems a relatively stable suite of prosocial emotions—including empathy—would already have been established, driven in part by pressures favoring traits that secured and preserved cooperation. Presumably, this mitigated to some degree the threat of deception within cooperative arrangements for individual gain.

Moreover, as we and others have argued, for guilt to be reliably signaled, significant costs would attach to such signaling. It is reasonable to suppose that the motivation to take on such costs is more reliably produced by the feeling of self-recrimination (à la Joyce and James) than by coolly calculated deception. Further, as Sterelny has argued in response to concerns that deceivers would overrun cooperative systems that rely to some degree on group signaling, even if deception prevailed in a least some individual cases, its threat would have been relatively low within complex social systems in which information flow runs multidirectionally within groups and informational pooling precedes using information for planning and acting (Sterelny 2012). A deceiver likely succeeds, in other words, when no one else checks on the lie.

But there are two ways by which fake signals of guilt may be rendered less effective within social contexts. First, the capacities for memory and the communication of fine-grained information about individuals' reputation for cooperation render such exploitation much more difficult. While someone may be able to get away with this deception within the relative anonymity of a large prison system, this would not be so easily accomplished in a smaller community of early humans with normalized social relations. If individuals were able to remember and communicate detailed information about the actions of others with whom they have interacted, they would be able to better distinguish earnest expressions of guilt from fraudulent ones, as well as move beyond firsthand experience in judging the relative degree of earnestness in attempts to seek forgiveness. However, there would be a significant cost in terms of time and biological maintenance of these memorial and communicative capacities, and it is reasonable to assume that competing adaptive trade-offs would result in a non-optimal leveling of their power.

However, a second and perhaps more effective way to prevent both the exploitation by fakers of guilt proneness, and the dissemination of inaccurate reputations about the guilt proneness of individuals, would be through reliable signaling of guilt. Recall Frank's sincere-manner pathway, along which hard-to-fake facial expressions serve as signals by which the group can accurately detect in individuals the presence of important social emotions. Guilt, we have seen, has no such telling facial or bodily expression, so Frank's account does not straightforwardly explain the reliable signaling of guilt. But there is no reason to suppose that the primary source of evidence about guilt must come from bodily posture or facial expression. The risk and potentially enormous costs individuals incur by confessing their transgressions, submitting to potential punishment, and performing reparative actions, can serve as credible signals of guilt experience. Mimicry of guilt-prone individuals, then, would not only come at a high price to the individual who attempts to exploit the forgiveness of others and seek the dampening of their punishment after transgression, but would also involve having to perform these costly actions presumably without the guilt experience that typically motivates and sustains them. While this is no guarantee that guilt could never be successfully feigned, the performance of these potentially costly and maladaptive behaviors would have enabled members of groups to infer the high probability that an individual is actually experiencing guilt due to a transgression (Deem and Ramsey 2016).

Other evolutionary accounts of guilt have also converged on the hypothesis that a costly signal might be required for expressions of guilt to yield individual-level benefits. Martinez-Vaquero et al. (2015), O'Connor (2016), and Pereiro et al. (2017a) provide a different avenue to this conclusion through the use of evolutionary game theory modeling. Their models of guilt show that apology after broken commitment and the subsequent restoration of cooperative arrangements between transgressor and transgressed can yield fitness benefits to each under conditions of revenge, apology, and forgiveness. However, in order for the apology to function as a reliable signal of a willingness to recommitment to cooperative arrangements, Martinez-Vaquero et al. (2015) conclude that the apology must cross a "sincerity threshold…where the cost of apologising should exceed that of cooperation" (8), where the cost of cooperation is the risk of defection. In mixed state games, where a population consists of both apology proposers and acceptors, Martinez-Vaquero et al. (2017) conclude that if the apology cost is too low, then apology defectors take over the population in repeated interactions. Similarly, O'Connor (2016) and Pereira et al. (2017b) conclude that a significant cost to apologize is needed to lower the probability that fakers can exploit cooperative arrangements by feigning guilt and that dishonest apologizers will evolve. In addition to concluding that the costs of guilt expression to the individual must cross a high enough threshold in order to be reliable signals of cooperative intent, the Martinez-Varquero et al., O'Connor, and Pereira et al. models suggest that these costs must not be so high that they cannot be absorbed by the individual. In their discussion of their respective models, Martinez-Varquero et al. (2015) and O'Connor (2016) conclude that the costs to the individual must be capped in some way. Otherwise, guilt expressions cannot be an evolutionary stable strategy. Perreira et al. (2017b) conclude that if this cost is too high, then revenge will dominate apology as a cooperative strategy.

None of these models, however, specifies what would create this cap on the costs to guilt-prone individuals. Extensions of these evolutionary game theoretic models in Rosenstock and O'Connor (2018) and Pereira et al. (2017a) provide some basis for explaining this posited cap on costs. On Rosenstock and O'Connor's model, guilt-prone individuals are willing to pay a cost in order to apologize for defection and show a willingness to cooperate in future interactions. Using a model in which one player's guilt expression occurs without detecting guilt proneness in other players, Pereira et al. (2017b) conclude that guilt-prone individuals will be exploited by non-guilt-prone conspecifics. Under such conditions, guilt proneness and its expression appears to carry a very high cost. On a second model that stipulates that an individual will experience postdefection guilt when a co-player acts prosocially toward the individual or has also displayed guilt, Pereira et al. (2017b) conclude that guilt proneness in a population will enhance cooperation and come to dominate.

This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

Both the Rosenstock and O'Connor (2018) and Pereira et al. (2017b) models suggest that the cost of guilt expression is curbed by prosocial behavior toward apology on the part of conspecifics, but neither model provides any specification for what psychological trait might undergird that response. Empathic concern for the guilt-prone individual's distress, we contend, is a good candidate for serving as this effective limit to the costs of guilty apology, since it explains both why players accept apology and why guilt-prone individuals do not exact the kinds of high cost we outline above, such as exploitation, severe punishment, or social exile. Reliable signaling of guilt through costly apology induces empathic concern in the transgressed, motivating forgiveness and the recommencement of cooperation. Our account of the relation between guilt and empathy, then, has the virtue not only of being consistent with these game-theoretic models of the evolution of guilt, but also of providing additional support to them by supplying a key condition for explaining how guilt-induced behavior could yield benefits to the individual.

Let's take stock of the foregoing empirical and conceptual considerations. Our ancestors likely were empathic before they were guilt prone. While the individual's anticipation of guilt could decrease the motivation to violate normative standards, it was the empathic context in which guilt emerged that was the decisive factor in the evolution of guilt. Empathic concern for the emotional distress of guilt likely reduced aggression toward guilt-prone norm breakers, and enabled others to vicariously participate in their emotional distress. Empathy, then, would have prompted behaviors aimed at mitigating guilt distress, including forgiving guilt-laden individuals and reincorporating them into cooperative arrangements. The high cost of expressing guilt, and the capacities to retain and communicate reputations of guilt proneness, would make guilt difficult to fake, and the benefits of genuine experiences of the emotion would amplify. The subsequent benefits of being forgiven and reincorporated into cooperative arrangements, then, would have made expressing guilt a stable adaptive strategy for the individual. Empathy, curiously absent from current evolutionary explanations of guilt, thus is likely a central component in the explanation of the evolution of guilt proneness.

5. Conclusions. It is widely agreed that guilt evolved to play an important role in human cooperation. But, as we have seen, guilt poses a unique evolutionary puzzle, given that its expression occurs postdefection, leaving the guilt-prone individual at the mercy of the group's judgment. Reviewing the solutions offered by James, Joyce, and Frank, we found that although their accounts have merit, none completely solves this puzzle. Our alternative solution, which draws on the contemporary empirical and game-theoretic literature on guilt, emphasizes the role of posttransgression behavior and the centrality of empathy in the evolution of guilt. Our account both leverages the maladaptive features of guilt that the other accounts have a difficult time accounting for, and provides a solution as to why groups would respond positively to individuals whose guilt displays were costly enough to signal reliably an intent to restore cooperative relations. Our account thus offers a better solution to the puzzle of why guilt evolved to play its distinctive social role, and sheds light on the complex relation between guilt expressions and the corresponding group response to their expression.

This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

References

Amodio, David M., Patricia G. Devine, and Eddie Harmon-Jones. 2007. "A dynamic model of guilt implications for motivation and self-regulation in the context of prejudice." *Psychological Science* 18, 524–30.

Averill, Patricia M., Gretchen J. Diefenbach, Melinda A. Stanley, Joy K. Breckenridge, and Beth Lusby. 2002. "Assessment of shame and guilt in a psychiatric sample: A comparison of two measures." *Personality and Individual Differences* 32:1365–76.

Barrett, Karen Caplowitz. 1995. "A Functionalist Approach to Shame and Guilt." In *Self-Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride*, ed. June Price Tangney and Kurt W. Fischer, 25–63. New York: The Guilford Press.

Bornstein, Brian H., Lahna M. Rung, and Monica K. Miller. 2002. "The effects of defendant remorse on mock juror decisions in a malpractice case." *Behavioral Sciences and the Law*, 20:393-409.

Baston, C. Daniel. 2009. "These things called empathy: Eight related but distinct phenomena." In *The Social Neuroscience of Empathy*, ed. Jean Decety and William Ickes, 1-16. Cambridge, MA: MIT Press.

Batson, C. Daniel. 2016. *What's Wrong with Morality?: A Social-Psychological Perspective*. Oxford: Oxford University Press.

Batson, C. Daniel, Nadia Y. Ahmad, and Eric L. Stocks. 2016. "Benefits and Liabilities of Empathy-induced Altruism: A Contemporary Review." In *The Social Psychology of Good and Evil*, Second Edition, ed. Arthur G. Miller, 443-66. New York: Guilford Press.

This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

Batson, C. Daniel, and Joy L. Weeks .1996. “Mood Effects of Unsuccessful Helping: Another Test of the Empathy-Altruism Hypothesis.” *Personality and Social Psychology Bulletin* 22: 148-57.

Batson, C. Daniel, Shannon Early, and Giovanni Salvarani. 1997. “Perspective taking: Imagining how another feels versus imagining how you would feel.” *Personality and Social Psychology Bulletin* 23(7): 751-58.

Bloom, Paul. 2016. *Against Empathy: The Case for Rational Compassion*. New York: Ecco.

Boehm, Christopher. 2012. *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. New York: Basic Books.

Broom, Donald M. 2003. *The Evolution of Morality and Religion*. Cambridge: Cambridge University Press.

Bybee, Jane, and Zandra N. Quiles. 1998. “Guilt and mental health.” In *Guilt and children, ed.* Jane Bybee, 269–91. San Diego, CA: Academic Press.

Coplan, Amy. 2011. “Will the Real Empathy Please Stand Up?: A Case for a Narrower Conceptualization.” *Southern Journal of Philosophy* 49(Spindel Supplement): 40-65.

de Waal, Franz B.M. 1996. *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Cambridge, MA: Harvard University Press.

de Waal, Franz B. M. 2006. *Primates and Philosophers: How Morality Evolved*. Princeton: Princeton University Press.

de Waal, Franz B.M. 2008. “Putting the Altruism back into Altruism: The Evolution of Empathy.” *Annual Review of Psychology* 59:279–300.

This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

Deem, Michael J. (2016). "Dehorning the Darwinian Dilemma for Normative Realism." *Biology and Philosophy* 31(5): 727-746.

Deem, Michael J. and Grant Ramsey. 2016. "Guilt by Association?" *Philosophical Psychology* 29(4): 570-85.

Drummond, Jesse D. K., Stuart I. Hammond, Emma Satlof-Bedrick, Whitney E. Waugh, and Celia A. Brownell. 2017. "Helping the One You Hurt: Toddlers' Rudimentary Guilt, Shame, and Prosocial Behavior after Harming Another." *Child Development* 88(4): 1382-97.

Eisenberg, Nancy, Natalie D. Eggum, and Laura Di Giunta. 2010. "Empathy-related Responding: Associations with Prosocial Behavior, Aggression, and Intergroup Relations." *Social Issues Policy Review* 4(1): 143–80.

Estrada-Hollenbeck, Mica, and Todd F. Heatherton. 1998. "Avoiding and Alleviating Guilt through Prosocial Behavior." In *Guilt and Children, ed.* Jane Bybee, 215–31. New York: Academic Press.

Fessler, Daniel M. T., and Matthew Gervais. 2010. "From whence the captains of our lives: Ultimate and phylogenetic perspectives on emotions in humans and other primates." In *Mind the Gap: The Origins of Human Universals*, ed. Peter Kappeler and Joan B. Silk, 261-80. Heidelberg: Spring.

Fowers, Blaine H. 2015. *The Evolution of Ethics: Human Sociality and the Emergence of Ethical Mindedness*. New York: Palgrave MacMillan.

Frank, Robert H. 1988. *Passions within Reason: The Strategic Role of the Emotions*. New York: W. W. Norton and Company.

This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

Garvey, Stephen P. 1998. “Aggravation and mitigation in capital cases: What do jurors think?” *Columbia Law Review* 98:1538–76.

Gold, Gregg J., and Bernard Weiner. 2000. “Remorse, confession, group identity, and expectancies about repeating a transgression.” *Basic and Applied Social Psychology* 22:291–300.

Greenspan, Patricia S. 1995. *Practical Guilt: Moral Dilemmas, Emotions, and Social Norms*. New York: Oxford University Press.

Griffiths, Paul. 1997. *What Emotions Really Are: The Problem of Psychological Categories*. Chicago: University of Chicago Press.

Harder, David W. 1995. “Shame and Guilt Assessment, and Relationships of Shame- and Guilt-Proneness to Psychopathology.” In *Self-Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride, ed.* June Price Tangney and Kurt W. Fischer, 368–92. New York: Guilford Press.

Harmon-Jones, Eddie, Kate Vaughn-Scott, Sheri Mohr, Jonathan Sigelman, and Cindy Harmon-Jones. 2004. “The effect of manipulated sympathy and anger on left and right frontal cortical activity.” *Emotion* 4: 95-101.

Hatfield, Elaine, Richard L. Rapson, and Yen-Chi L. Le. 2009. “Emotional Contagion and Empathy.” In *The Social Neuroscience of Empathy*, ed. Jean Decety and William Ickes, 19-30. Cambridge: Cambridge University Press.

Hosser, Daniela, Michael Windzio, and Werner Greve. 2008. “Guilt and shame as predictors of recidivism: A longitudinal study with young prisoners.” *Criminal Justice and Behavior* 35:138–52.

This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

Jagers, Robert J., Antonio A. Morgan-Lopez, Terry-Lee Howard, Dorothy C. Browne, Brian R. Flay, and Aban Aya Coinvestigators. 2007. “Mediators of the development and prevention of violent behaviors.” *Prevention Science* 8:171–79.

James, Scott M. 2011. *An Introduction to Evolutionary Ethics*. West Sussex: Wiley-Blackwell.

Jehle, Alayna, Monica K. Miller, and Markus Kemmelmeier. 2009. “The influence of accounts and remorse on mock jurors’ judgments of offenders.” *Law and Human Behavior* 33:393-404.

Joyce, Richard. 2006. *The Evolution of Morality*. Cambridge, MA: MIT Press.

Keltner, Dacher, and Brenda N. Buswell. 1996. “Evidence for the distinctness of embarrassment, shame, and guilt: A study of recalled antecedents and facial expressions of emotions.” *Cognition and Emotion* 10:155–71.

Laland, Kevin N., and Gillian R. Brown. 2011. *Sense and nonsense: evolutionary perspectives on human behavior*, 2nd ed. Oxford: Oxford University Press.

Langford, Dale J., Sara E. Crager, Zarrar Shehzad, Shad B. Smith, Susana G. Sotocinal, Jeremy S. Levenstadt, Mona Lisa Chanda, Daniel J. Levitin, and Jeffrey S. Mogil. 2006. “Social modulation of pain as evidence for empathy in mice.” *Science* 312:1967–70.

Lindsay-Hartz, Janice, Joseph de Rivera, and Michael F. Mascolo. 1995. “Differentiating Guilt and Shame and their Effects on Motivation. In *Self-Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride* ed. June Price Tangney and Kurt W. Fischer, 274–300. New York: Guilford Press.

Luyten, Patrick, Johnny R. J. Fontaine, and Jozef Corveleyn. 2002. "Does the Test of Self-Conscious Affect (TOSCA) measure maladaptive aspects of guilt and adaptive aspects of shame?: An empirical investigation." *Personality and Individual Differences* 33:1373–87

Martinez-Vaquero, Luis A., The Anh Han, Luis Moniz Pereira, and Tom Lenaerts. 2015. "Apology and Forgiveness Evolve to Resolve Failures in Cooperative Agreements." *Scientific Reports* 5:10639.

Martinez-Vaquero, Luis A., The Ahh Han, Luis Moniz Pereira, and Tom Lenaerts. 2017. "When agreement-accepting free-riders are a necessary evil for the evolution of cooperation." *Scientific Reports* 7:2478.

Miller, Christian. 2011. "Defining empathy: Thoughts on Coplan's approach." *Southern Journal of Philosophy* 49(s1): 66-72.

Moll, Jorge, Ricardo de Oliveira-Souza, Roland Zahn, and Jordan Grafman. 2008. "The cognitive neuroscience of moral emotions." In *Moral psychology*, ed. Walter Sinnott-Armstrong, 3–17. Cambridge, MA: MIT Press.

Nesse, Randolph M. 2016. "Social selection is a powerful explanation for prosociality." *Behavioral and Brain Sciences*, 39:35-6.

O'Connor, Cailin. 2016. "The Evolution of Guilt: A Model-Based Approach." *Philosophy of Science* 83(5): 897-908.

Panksepp, Jaak, and Lucy Biven. 2012. *The archaeology of mind: Neuroevolutionary origins of human emotions*. New York, NY: Norton.

Pereira, Luís Moniz, Tom Lenaerts, Luis A. Martinez-Vaquero, and The Anh Han. 2017a. "Evolutionary game theory modeling of guilt." Paper presented at Symposium on

Computational Modelling of Emotion: Theory and Applications 2017, Bath, United Kingdom.

Pereira, Luís Moniz, Tom Lenaerts, Luis A. Martinez-Vaquero, and The Anh Han. 2017b. “Social manifestation of guilt leads to stable cooperation in multi-agent systems.” In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems*, ed. Kate Larson K and Michael Winikoff, 1422–30. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Porter, Stephen, Leanne ten Brinke, and Kevin Wilson. 2009. “Crime profiles and conditional release performance of psychopathic and non-psychopathic sexual offenders.” *Legal and Criminological Psychology* 14:109–118.

Povinelli, Daniel J., Kurt E. Nelson, and Sarah T. Boysen. 1992. “Comprehension of role reversal in chimpanzees: Evidence of empathy?” *Animal Behavior* 43:633–40.

Prinz, Jesse J. 2004. *Gut Reactions: A Perceptual Theory of Emotions*. Oxford: Oxford University Press.

Prinz, Jesse J. 2011. “Against empathy.” *Southern Journal of Philosophy* 49(s1): 214-33.

Radzik, Linda. 2009. *Making Amends: Atonement in Morality, Law, and Politics*. Oxford: Oxford University Press.

Richerson, Peter, Ryan Baldini, Adrian V. Bell, Kathryn Demps, Karl Frost, Vicken Hillis, Sarah Mathew, Emily K. Newton, Nicole Naar, Lesley Newson, Cody Ross, Paul E. Smaldino, Timothy M. Waring, and Matthew Zefferman. 2016. “Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence.” *Behavioral and Brain Sciences* 39:1-68.

Rosenstock, Sarita, and Cailin O'Connor. 2018. "When it's good to feel bad: An evolutionary model of guilt and apology." *Frontiers in Robots and AI*, 5:9.

Scott, James M. 2011. *An Introduction to Evolutionary Ethics*. Oxford: Wiley-Blackwell.

Silfver, Mia. 2007. "Coping with Guilt and Shame: A Narrative Approach." *Journal of Moral Education* 36(2): 169–83.

Silk Joan B., Robert Boyd. 2010. "From grooming to giving blood: the origins of human altruism." In: *Mind the gap: tracing the origins of human universals*, ed. Peter Kappeler and Joan B. Silk, 223–44. Heidelberg: Springer.

Smith, Joel. 2017. "What is Empathy For?" *Synthese* 194:709-22.

Snow, Nancy E. 2000. "Empathy." *American Philosophical Quarterly* 37:65-78.

Sterelny, Kim. 2012. *The Evolved Apprentice: How Evolution Made Humans Unique*. Cambridge, MA: MIT Press.

Strayer, Janet, and William Roberts. 2004. "Empathy and observed anger and aggression in five-year-olds." *Social Development* 13(1): 1–13.

Svensson, Robert, Frank M. Weerman, Lieven J.R. Pauwels, Gerben J.N. Bruinsma, and Wim Bernasco. 2013. "Moral emotions and offending: Do feelings of anticipated shame and guilt mediate the effect of socialization on offending?" *European Journal of Criminology,* 10(1): 22–39.

Tangney, June Price. 1996. "Conceptual and Methodological Issues in the Assessment of Shame and Guilt." *Behavior Research and Theory* 34:741–54.

Tangney, June Price, Jeff Stuewig, and Logaina Hafez. 2011. "Shame, Guilt, and Remorse: Implications for Offender Populations." *Journal of Forensic Psychiatry and Psychology*

This paper is forthcoming in the journal *Philosophy of Science*.
This version is not copyedited and should not be quoted without the permission of the authors.

22(5): 706-23.

Tangney, June Price, Jeffrey Stuewig, Elizabeth T. Malouf, and Kerstin Youman. 2013. "Communicative functions of shame and guilt." In *Cooperation and its Evolution*, ed. Kim Sterelny, Richard Joyce, Brett Calcott, and Ben Fraser, 485-502. New York: MIT Press.

Teroni, Fabrice, and Julien A. Deonna. 2008. "Differentiating Shame from Guilt." *Consciousness and Cognition* 17(3): 725-740.

Tracy, Jessica L. and Richard W Robins. 2006. "Appraisal antecedents of shame and guilt: Support for a theoretical model." *Personality and Social Psychology Bulletin* 32(10): 1339–51.

Treeby, Matthew, Catherine Prado, Simon M. Rice, and Simon F. Crowe. 2016. "Shame, guilt, and facial recognition: Initial evidence for a positive relationship between guilt-proneness and facial emotion recognition." *Cognition and Emotion* 30(8): 1504-11.

Wallbott, Harald G. 1998. "Bodily expression of emotion." *European Journal of Social Psychology* 28:879–896.

Zahn-Waxler, Carolyn, Pamela M. Cole, Jean Darby Welsh, and Nathan A. Fox. 1995. "Psychophysiological correlates of empathy and prosocial behaviors in preschool children with problem behaviors." *Development and Psychopathology* 7: 27–48.

Zahn-Waxler, Carolyn, and Carol Van Hulle. 2012. "Empathy, Guilt, and Depression: When Caring for Others Becomes Costly to Children." In *Pathological Altruism*, ed. Barbara Oakley, Ariel Knafo, Guruprasad Madhavan, and David Sloan Wilson, 321-44. Oxford: Oxford University Press.