Journal of Personality and Social Psychology
2013, Vol. 105, No. 4, 621–638

0022-3514/13/$12.00 DOI: 10.1037/a0033495

Merely Opting Out of a Public Good Is Moralized: An Error Management Approach to Cooperation

Andrew W. Delton, Jason Nemirow, Theresa E. Robertson, Aldo Cimino, and Leda Cosmides
University of California, Santa Barbara

People regularly free ride on collective benefits, consuming them without contributing to their creation. In response, free riders are often moralized, becoming targets of negative moral judgments, anger, ostracism, or punishment. Moralization can change free riders' behavior (e.g., encouraging them to contribute or discouraging them from taking future benefits) or it can motivate others, including moralizers, to avoid or exclude free riders; these effects of moralization are critical to sustaining human cooperation. Based on theories of error management and fundamental social domains from evolutionary psychology, we propose that the decision to moralize is a cue-driven process. One cue investigated in past work is observing a person illicitly consume collective benefits. Here, we test whether the mind uses a 2nd cue: merely opting out of contributing. Use of this cue creates a phenomenon of *preventive moralization*: moralization of people who have not yet exploited collective benefits but who might—or might not—in the future. We tested for preventive moralization across 9 studies using implicit and explicit measures of moralization, a behavioral measure of costly punishment, mediation analyses of the underlying processes, and a nationally representative sample of almost 1,000 U.S. adults. Results revealed that merely opting out of contributing to the creation of exploitable collective benefits—despite not actually exploiting collective benefits—elicited moralization. Results further showed that preventive moralization is not due to the moralization of selfishness or deviance but instead follows from the uncertainty inherent in moralization decisions. These results imply that even people who will never exploit collective benefits can nonetheless be targets of moralization. We discuss implications for social and political dynamics.

Keywords: evolutionary psychology, error management, moralization, cooperation, collective action

Humans regularly work in groups to produce a good that is then shared. Many collective benefits, however, are vulnerable to free riders, people who take collective benefits despite not contributing to their creation. In response, free riders are often moralized—they are targets of negative moral judgments, anger, ostracism, or punishment (Fehr & Gächter, 2000; Kiyonari & Barclay, 2008; Masclet, Noussair, Tucker, & Villeval, 2003; Price, Cosmides, & Tooby, 2002; Tybur, Lieberman, & Griskevicius, 2009; Yamagishi, 1986).[1] Moralization can change free riders' behavior, encouraging them to contribute or discouraging them from taking benefits later. Moralization can also motivate others (including moralizers) to avoid free riders or exclude them from future cooperation. These functions of moralization are critical to sustaining human cooperation (e.g., Boyd & Richerson, 1992; Hauert, De Monte, Hofbauer, & Sigmund, 2002; Panchanathan & Boyd, 2004; Sasaki & Uchida, 2013; Tooby, Cosmides, & Price, 2006). Moral and social decisions, however, often involve uncertainty and potential error. How does uncertainty affect decisions about moralization? Does it predispose identifying a person as a cooperator—or as a potential free rider?

We address the relation between uncertainty and moralization by combining two approaches from evolutionary social psychology. First, we use a fundamental social domains approach to

This article was published Online First July 1, 2013.

Andrew W. Delton, Jason Nemirow, and Theresa E. Robertson, Center for Evolutionary Psychology, Department of Psychological and Brain Sciences, University of California, Santa Barbara; Aldo Cimino, Center for Evolutionary Psychology, Department of Anthropology, University of California, Santa Barbara; Leda Cosmides, Center for Evolutionary Psychology, Department of Psychological and Brain Sciences, University of California, Santa Barbara.

This research was supported by a National Institutes of Health (NIH) Director's Pioneer Award to Leda Cosmides; National Science Foundation (NSF) Grant 0951597 to Leda Cosmides and John Tooby; University of California, Santa Barbara, Undergraduate Research and Creative Activities funds to Jason Nemirow; and a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the NIH, the NSF, or the John Templeton Foundation. We thank Mike McCullough, Jon Maner, and Peter DeScioli for their comments.

Correspondence concerning this article should be addressed to Andrew W. Delton, Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93106-9660. E-mail: andy.delton@gmail.com

[1] Our usage of the term *moralization* departs somewhat from past usage. In past usage, moralization has denoted the process by which an action, often previously lacking a moral component, comes to have a moral component—thereafter, engaging in (or failing to engage in) the action leads to moral condemnation and related responses (Rozin, 1999). We are using moralization as an umbrella term for a variety of negative moral and punitive responses to avoid having to use lists, such as "moral judgment, moral condemnation, and punitive sentiment," when referring to the phenomena under investigation. Ultimately, however, our analysis ends up having much in common with moralization traditionally defined.

uncover some of difficulties involved in successfully creating and maintaining group cooperation (Bugental, 2000; Fiske, 1992; Kenrick, Li, & Butner, 2003; Kenrick, Neuberg, Griskevicius, Becker, & Schaller, 2010). Second, we use an error management approach to analyze how adaptive decisions about these problems should be made under uncertainty (Haselton et al., 2009; Haselton & Nettle, 2006). Combining both approaches, we show that people who opt out of some forms of cooperation nonetheless become targets of moralization—despite not having taken group benefits. As a consequence, there may be errors of moralization of the innocent—moralization of people who will never illicitly take collective benefits.

Collective Action and Error Management

Successfully navigating social life brings huge benefits: friends and allies, safety and security, mates and children. But a single tool—a general purpose mechanism of sociality—cannot create all of these benefits (Tooby & Cosmides, 1992). Instead, the social mind appears to divide into multiple domains, with each domain having its own set of rules for generating adaptive behavior (Bugental, 2000; Kenrick et al., 2003, 2010). For instance, the decision rules sustaining long-term romantic relationships (Maner, Gailliot, & Miller, 2009; Maner, Miller, Rouby, & Gailliot, 2009) are different from the decision rules guiding intergroup interactions (Ackerman et al., 2006; McDonald, Navarrete, & Van Vugt, 2012).

Here, we focus on the decision rules guiding moralization within the domain of collective action. Collective action involves multiple people collaborating to produce a shared benefit. This type of cooperation appears in all known human societies (Bradfield, 1973; Olson, 1965; Ostrom, 1990). Many collective actions, such as national defense or environmental conservation, produce *public goods*. In public goods, non-contributors can free ride, benefiting from others' contributions (Dawes, 1980; Komorita & Parks, 1995; Van Lange, Liebrand, Messick, & Wilke, 1992). This illicit benefit consumption gives free riders higher payoffs than contributors, which threatens the evolutionary stability of public goods (e.g., Boyd & Richerson, 1992; Hauert et al., 2002; Panchanathan & Boyd, 2004; Tooby et al., 2006). To prevent free riders from proliferating or benefiting at the expense of cooperators, they are often targets of moralization, such as sanctioning or punishment (Fehr & Gächter, 2000; Kiyonari & Barclay, 2008; Masclet et al., 2003; Price et al., 2002; Yamagishi, 1986). Moralization encourages free riders to contribute, prevents them from accessing collective benefits, or otherwise reduces payoffs relative to contributors. But identifying free riders is a difficult problem (Delton, Cosmides, Guemo, Robertson, & Tooby, 2012), and not all potential free riders pose the same challenges (Cimino & Delton, 2010; Delton & Cimino, 2010). What are the decision rules that generate moralization and what are the cues used by these decision rules?

To examine potential cues leading to moralization, we adopt error management theory (Haselton et al., 2009; Haselton & Nettle, 2006), which has successfully been applied to a number of questions involving groups and cooperation (Delton, Krasnow, Cosmides, & Tooby, 2011; Kiyonari, Tanida, & Yamagishi, 2000; Krasnow, Delton, Tooby, & Cosmides, 2013; Maner et al., 2005; Yamagishi, Terai, Kiyonari, Mifune, & Kanazawa, 2007). Error management analyses are based on two assumptions. First, the true state of the world is often ambiguous. Imagine your nighttime route home takes you by dense foliage in an area inhabited by jaguars. Does the foliage hide a hungry jaguar? Second, different types of errors often have different costs. If you wrongly assume there is a jaguar, you pay the cost of a longer route to avoid the area. But if you wrongly assume no jaguar is present, you might face injury or death. When a particular decision problem has been evolutionarily recurrent and has asymmetric error costs, natural selection should shape decision rules to avoid making the costlier error—even at the expense of making a larger number of low cost errors. Thus, if dark, dense foliage predicted the presence of predators over human evolution, then it should generate fear and avoidance—even if predators are rarely present and thus on most occasions avoidance is a waste of time and energy.

Public goods create a similar decision problem: Free riding is a strategic decision but often not an overt, observable behavior; thus, the mind must use fallible cues to identify free riders (see Delton et al., 2012, for further discussion). One confluence of cues that reliably discriminates free riders from others is the taking of collective benefits with the intention of withholding contributions (Delton et al., 2012; Fehr & Gächter, 2000; Masclet et al., 2003; Yamagishi, 1986).

But what predictions follow when considering the component cues in isolation? First, is the overt action of taking collective benefits enough to elicit moralization? Recent research shows that it is not, in part because human cooperation evolved in uncertain, variable environments. In such environments, moralizing people who take collective benefits must be partially decoupled from whether a person has contributed because accidents or bad luck can cause even well-meaning cooperators to fail to contribute (Delton et al., 2012; Delton & Robertson, 2012). What about opting out of contributing? Can non-contribution elicit moralization even if no collective benefits have actually been consumed? This situation creates uncertainty because a person who opts out may—or may not—later take collective benefits; their future actions cannot be perfectly known.

Given the possibility that a person who opts out of contributing to a public good might illicitly take the benefits in the future, should they be moralized now? There are two possible errors you could make: (1) moralize them even though they will *not* illicitly take the benefits, or (2) fail to moralize them even though they *will* illicitly take the benefits. The cost of the first error is the time and energy that moralization requires. In some cases, this cost can be quite low if moralization is spread across multiple people (Boehm, 1993), lowering the cost of moralization in the moment and lowering the likelihood that any particular moralizer will be retaliated against (one cost of moralization is the potential for its targets to retaliate; e.g., DeScioli & Kurzban, 2013; McCullough, Kurzban, & Tabak, 2013; Nikiforakis, 2008). This first error, moreover, has the potential to encourage someone to contribute when they otherwise might not, even if they would not have consumed collective benefits. (There are also cases where the costs of preventive moralization could be quite high; we return to this in the general discussion.) The cost of the second error is potentially more dire: the destruction of cooperation that occurs when free riding is allowed to proliferate, whether considered over the lifespan of any particular collective action or over evolutionary timescales. This asymmetry predicts that decision rules should be biased toward moralization in public goods, even if moralization is not "ratio-

nally" warranted given that no benefits have been illicitly consumed.

Having such a sensitive trigger would create a phenomenon of *preventive moralization*: moralization of non-contributors now because of what they might do in the future. This prevention might function in several related ways. It might cause a person who opts out to change their behavior and cooperate in the present collective action, thereby legitimizing any benefits they do consume. It might cause a person who opts out to refrain from consuming benefits of the current collective action. It might prevent them from illicitly taking benefits from future collective actions by excluding them from later cooperation. Or it might cause the moralizer or others to avoid the non-contributor in the future, indirectly preventing the non-contributor from reaping the benefits of cooperation. In all of these cases, moralization occurs without any benefits having been illicitly consumed. Although preventive moralization would have the effect of often stopping free riding before it starts, it will necessarily create another type of error: the moralization of people who are not and will not become free riders.

The Present Research

Studies 1–5: Testing Against Alternative Hypotheses

The preventive moralization hypothesis predicts that opting out of a public good engenders moralization. However, perhaps opting out of any cooperative endeavor elicits moralization, even if the benefit generated does not have the structure of a public good. Opting out may be viewed, for instance, as selfish or showing a lack of conformity, attributes that might be moralized in any group context. To show that the preventive moralization hypothesis has any unique predictive power—over and above other potential causes of moralization—we compared public goods to goods where free riding is not possible, *club goods*. Unlike public goods, by definition the benefits of club goods are available only to contributors; non-contributors cannot access them (Crosson, Orbell, & Arrow, 2004; Sandler & Tschirhart, 1997). Many early analyses of club goods were conducted on literal clubs, voluntary organizations where people can pay fees to become a member and enjoy member-only privileges. Health clubs, for example, are club goods in that they allow only paying members to access their exercise equipment. The structure of club goods (e.g., being behind locked doors), requires a person to contribute before the good can be accessed; free riding is therefore difficult or impossible.

Nonetheless, because the successful provisioning of public and club goods requires member contributions, both types of goods can be negatively impacted by, for example, selfishness or deviance. For instance, air quality in cities (a public good) is diminished if people burn proscribed materials or do not maintain their cars; a fraternal lodge (a club good) may close down if enough members fail to pay their dues. Similarly, across both goods, motivation to recruit labor is also constant: For both goods, additional contributors are likely to increase available benefits.[2] This creates an economic incentive to encourage others to contribute, and moralization may be one means to this end. But despite these parallels, free riding is only possible on public goods, so public goods should elicit preventive moralization over and above that elicited by club goods.

Thus, we conducted five studies contrasting public and club goods. To establish a wide empirical base for the preventive moralization hypothesis, we measured moralization in several ways, including implicit associations (using the Implicit Association Test; Greenwald, Nosek, & Banaji, 2003; Lane, Banaji, Nosek, & Greenwald, 2007), explicit moral judgments (using rating scales), and experimental game behavior (using costly punishment); our data collection also included a nationally representative sample of almost 1,000 adults in the United States. Measuring both implicit and explicit responses is important given the ongoing debate regarding the role of fast, intuitive mental systems versus deliberate conscious reasoning in moral judgment (Bloom, 2010; Haidt, 2001, 2007). Because preventive moralization has not been previously examined, however, there is no clear theory predicting whether implicit associations, explicit responses, and costly punishment will show intra-individual correlations. Indeed, judgments of wrongness and decisions to punish may serve somewhat different functions and thus may be dissociable (Cushman, 2008; Lieberman & Linke, 2007). So although each measure should reveal greater moralization toward non-contributors on a public good compared to a club good, we have no prior hypothesis that an individual who shows, for example, strong implicit associations would be especially likely to punish. Our goal is simply to cast a wide empirical net. Nonetheless, in an exploratory fashion, we test for correlations between measures whenever possible.

Many theoretical treatments analyze "pure" public or club goods: In a pure public good, free riders benefit just as much and just as easily as cooperators; in a pure club good, free riding is impossible. Such pure versions are unlikely to obtain in real life; most goods exist somewhere on a continuum between these two points. Although we use the terms public and club goods for convenience, we are only using them to denote relative differences and do not intend for our discussion or our experiments to generally imply pure versions.

Studies 6–9: Process and Quantitative Variation

We had several goals with Studies 6–9. First, we wanted to examine the underlying process that gives rise to preventive moralization. According to the preventive moralization hypothesis, when someone opts out of a public good moralization is elicited because there is a chance they will eventually free ride. This leads to the following model (see Figure 1): Cues to potential free riding are perceived, leading to an estimation of the likelihood of free riding by a person opting out, in turn leading to greater moralization. We sought confirmatory statistical evidence for this putative causal chain using mediation analyses.

Second, with Studies 6–9 we wanted to examine a number of different possible cues that could lead to variation in perceptions of the likelihood of free riding. Each study manipulated a different possible cue to the likelihood of free riding: non-contributors' subjective valuation of the good, non-contributors' cost if they free rode, the ability of contributors to monitor the good and prevent free riding, and the benefits non-contributors can obtain if they do free ride.

[2] Sometimes additional contributors are not beneficial (e.g., in step-level goods when a sufficient number of contributors is already reached), but this can apply equally to public and club goods.

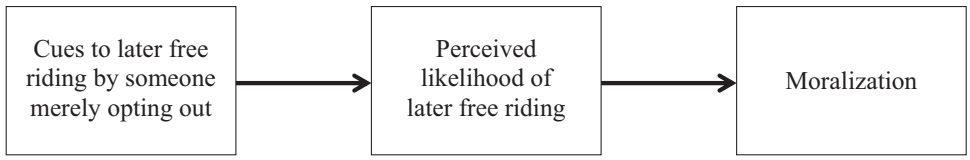

Figure 1. Theoretical model of preventive moralization.

Finally, with Studies 6–9 we wanted to examine quantitative variation in the likelihood of free riding. Studies 1–5 used stark, dichotomous contrasts: In the club goods, free riding was difficult or impossible; in the public goods, free riding was easy, free riding delivered the same benefits as contributing, and non-contributors were likely to free ride. Although this design was useful for an initial demonstration of the effect and for ruling out alternatives, it has two related drawbacks. First, the designs of Studies 1–5 treated preventive moralization as an all-or-none phenomenon. By quantitatively varying the likelihood of free riding in Studies 6–9, we investigated whether preventive moralization is a quantitative phenomenon that responds in degrees, not absolutes. Second, the designs of Studies 1–5 only allowed us to observe preventive moralization in cases where free riding was highly likely to occur. With Studies 6–9, we sought to show that preventive moralization could also happen even in situations where free riding was clearly not guaranteed.

Study 1: Moralization in Public Versus Club Goods

As an initial test of the preventive moralization hypothesis, Study 1 contrasted public and club goods and measured moralization implicitly and explicitly. Participants learned about a group of college students who had the option of participating in a carwash fundraiser for an upcoming hiking trip. In the public good condition, the carwash proceeds approximated a public good, paying group-related expenses for all hikers regardless of their carwash participation. In the club good condition, the carwash proceeds approximated a club good, being divided only among carwash participants. Importantly, no one—contributor or non-contributor—was depicted consuming benefits. Indeed, the carwash had not yet occurred; contribution was manipulated by stated intent, not a completed act.

Method

Participants and design. One hundred seven undergraduates (53 women) in introductory psychology or anthropology classes at the University of California, Santa Barbara, participated for course credit or for $10 compensation. Participants worked at private computers that randomly assigned conditions between subjects. First, participants read about a collective action—a carwash fundraiser—producing either a public or club good. Second, participants saw photos of the individuals involved in the collective action and learned who did and did not intend to contribute. Third, participants completed a memory test to ensure they accurately remembered contribution intentions. Fourth, they completed an Implicit Association Test (IAT) measuring implicit moralization. Finally, they made explicit moral wrongness ratings.

Materials and procedure.

Public good versus club good manipulation. All participants read about a group of college students who independently signed up for a hiking trip through their campus recreation center. Participants learned that the recreation center was planning a carwash fundraiser to help defray trip expenses. In the public good condition, they further read the following: "All the money generated by the carwash will then go toward paying the costs of the trip for everyone. Even if someone decided not to participate in the carwash, they would still benefit from the trip costs being paid." In the club good condition, they further read the following: "All the money generated by the carwash will be divided equally among the people who came. These people will then be able to apply their share of the profits toward paying their individual trip costs."

Learning about contributors and non-contributors. Participants next serially viewed randomly ordered photos of eight college-age white men. They learned that four intended to contribute ("this person decided to go to the carwash"), and the remainder did not ("this person decided to stay home"). Between participants, the computer randomized which photos were paired with which decisions. Because the IAT requires knowledge of contribution intentions, participants repeated a memory test (with feedback) until they correctly identified all eight men's intentions twice in a row.

Dependent measures.

Implicit Association Test (IAT). To measure moralization of non-contributors, participants completed the IAT according to standard protocol (Greenwald, Nosek, & Banaji, 2003; Lane, Banaji, Nosek, & Greenwald, 2007). The IAT measure uses reaction times in a dual categorization task to assess how strongly two dimensions are associated. In Study 1, participants simultaneously categorized faces of contributors and non-contributors as intending to contribute or not contribute (implemented with non-evaluative labels: "staying home" vs. "going to carwash") and categorized eight positively- and eight negatively-valenced morally evaluative words as good or bad. The eight positive morally evaluative words were as follows: fair, noble, upstanding, principled, honorable, trustworthy, integrity, and virtuous. The eight negative morally evaluative words were as follows: dishonest, cheater, unethical, wrongful, corrupt, sinful, shady, and crooked.

Participants completed several blocks of trials, including practice and familiarization blocks. On the critical trial blocks (used to score the IAT) participants categorized all types of stimuli, with the stimulus type (i.e., contributor, non-contributor, positive word, negative word) randomly determined each trial. Importantly, despite categorizing four types of stimuli, participants used only two response keys (the "A" and "L" keys). One block, for instance, might use a single key to respond both "staying at home" and "bad." Such trials are *consistent* because not contributing and negative moral evaluation should be associated (and similarly for contributing and positive moral evaluation). Other blocks have the opposite pairing, using a single key for "staying at home" and "good." Such trials are *inconsistent* because not contributing and positive moral evaluation should not be associated; indeed, these

responses may compete. When trials are consistent—when the stimulus types "go together"—responses should be relatively quick. When trials are inconsistent—when the stimulus types conflict—responses should be relatively slow. The IAT is based on this response time difference (along with standardization to remove within-participant response time variability; see Greenwald et al., 2003; Lane et al., 2007). Beyond being instructed that the IAT measured their ability to classify words and objects into groups, subjects were not given any further rationale for the task.

We scored the IAT according to standard protocol (Greenwald et al., 2003; Lane et al., 2007). Thus, all latencies greater than 10 s were deleted (for Studies 1 and 2, this was less than 1% of trials). In principle, participants with greater than 10% of latencies under 300 ms are removed; in this experiment, no participant met this criterion. The scoring algorithm produces an IAT D score, with greater scores indexing stronger associations between non-contributors and negative moral words and between contributors and positive moral words. We computed both overall IAT D scores, based on all four types of stimuli, and separate IAT D scores for each stimulus type. The preventive moralization hypothesis most strongly predicts differences in the non-contributor score, with greater IAT D scores in the public good condition relative to the club good condition. For completeness, we also provide the other separate IAT D scores.

Explicit rating scale. Explicit moralization of the non-contributors was assessed by the question "Some organization members decided to stay home instead of going to the carwash. How morally wrong do you think it was of them to make this choice?" with a 7-point rating scale (1 = *not at all wrong*, 7 = *very wrong*). The preventive moralization hypothesis predicts greater explicit moralization in the public good condition, relative to the club good condition.

Analysis strategy. All analyses used two-tailed *p*-values. Mean differences were tested using independent samples *t*-tests. Pearson's *r* was used to measure correlations between dependent measures. We also used Pearson's *r* as a measure of effect size for mean differences (Rosenthal, Rosnow, & Rubin, 2000). Analysis of variance (ANOVA) was used to test for interactions between independent variables. To determine whether there were any sex differences, prior to our main analyses we conducted preliminary analyses on our dependent measures using ANOVA with two factors (condition and participant sex); these analyses revealed no interactions with participant sex.

Results and Discussion

Did public goods, relative to club goods, lead to more explicit moralization? Yes: Non-contributors on a public good were rated as more morally wrong, $r = .33$, $p < .001$ (see Table 1).

Did public goods, relative to club goods, lead to more implicit moralization? Yes: As shown in Table 1, implicit associations between non-contributor–bad and between contributor–good were stronger in the public good condition. This difference in implicit associations is reflected in the overall IAT D score ($r = .29$, $p < .01$) and the non-contributor IAT D score ($r = .25$, $p = .01$).

Interestingly, associations indexed by the contributor and positive word IAT D scores showed greater positivity in the public good condition (*p*s $< .05$). There was no effect for negative moral words. We note that there was no interaction between contributor/non-contributor and good type in predicting IAT D scores ($p = .96$). Finally, all IAT D scores were significantly greater than zero in both the club and public goods conditions (all *p*s $< .05$). Although our experimental manipulation affected IAT D scores, even in the club good scenario there were significant implicit associations between non-contribution and badness and between contribution and goodness.

Were implicit and explicit moralization correlated? No, the explicit ratings did not correlate with either the overall IAT D score ($r = .03$) or the non-contributor IAT D score ($r = –.02$). We return to this in the general discussion.

Study 2: Further Constraining the Club Good

Study 1 provided preliminary evidence for the preventive moralization hypothesis: People who were potential beneficiaries of a public good but opted out of contributing to its creation were moralized. Moralization occurred more strongly in a public good, relative to a club good, despite both involving group cooperation. Thus, our effect is not attributable to causes that are present in all forms of group cooperation. For instance, group cooperation might require some level of conformity or group-directed effort. This does not differ, however, between club and public goods; creating both goods requires successfully coordinating with others and investing effort on behalf of the group. Study 2 was designed to replicate Study 1 with different stimuli and to tighten the contrast between conditions. In Study 1, the club good was ultimately privately held (money divided among contributors), whereas the

Table 1
Means (Standard Deviations) of Implicit and Explicit Moralization as a Function of Good Type in Study 1 (Carwash Fundraiser)

Measure	Public good *M* (*SD*)	Club good *M* (*SD*)	*t*	*p*	*r*
Explicit rating of wrongness	4.36 (1.19)	3.39 (1.56)	3.61	.0005	.33
IAT D score, overall	0.60 (0.29)	0.42 (0.30)	3.13	.002	.29
IAT D score, non-contributors	0.64 (0.44)	0.40 (0.48)	2.64	.01	.25
IAT D score, contributors	0.72 (0.37)	0.49 (0.44)	2.91	.004	.27
IAT D score, negative moral words	0.58 (0.48)	0.60 (0.44)	−0.27	.79	−.03
IAT D score, positive moral words	0.64 (0.41)	0.35 (0.45)	3.47	.0008	.32

Note. For all tests, $df = 105$. IAT = Implicit Association Test.

public good was communally owned (a joint fund). In Study 2, the club good was also communally owned (as is typical of club goods in institutional analysis; e.g., Ostrom, 2003; Sandler & Tschirhart, 1997). Study 2 also used a different club good exclusion mechanism. Study 1 used physical exclusion: With the money physically divided among contributors, non-contributors could not benefit without outright theft. In Study 2, exclusion from the club good relied on social pressure.

In Study 2, participants read about a village of African hunter–horticulturalists who subsist partly on fish. Villagers can participate in collective actions that create access to fish. In the public good condition, they can move trees that have fallen upriver, preventing fish from swimming past their village. With the trees removed, no one—contributor or not—can be prevented from fishing. In the club goods condition, villagers can use fallen trees to create a fishing pier near the village in full sight of the villagers, with the understanding that only contributors are to use it.

Method

Unless noted, the method is identical to Study 1.

Participants. One hundred twelve undergraduates (58 women) participated for partial course credit. Two additional participants were eliminated for having >10% IAT trials faster than 300 ms. Two further participants were eliminated for not successfully finishing the memory test.

Materials. In both conditions, participants read about the Samana, a (fictional) hunter–horticulturalist group living in Botswana, Africa. Eight computer-generated images of African men represented the Samana. Whether they were "staying home" or "going upriver" was randomized between participants.

Participants learned that the Samana subsist partly on fish from the (fictional) Pata river. The Pata runs for a great distance such that "even if someone wanted to stop another person from using the river, it would be nearly impossible. The riverbanks are far too long." Participants also learned that a recent earthquake caused several large trees to fall upstream. In the public good condition, the trees blocked fish from swimming near the village. Because of the blockage, the Samana are organizing a collective action to remove the trees. The more Samana who participate, the easier the task is for those involved. Finally, participants learned that "because the banks of the Pata are long and the fish are very easy to catch from the water's edge, all members of the Samana would benefit from the removal of the trees by the party that travels upriver, even those Samana who have decided not to travel upriver to remove the trees."

In the club good condition, participants learned that the fish tend to swim in the middle of the river, where fishing is difficult. The fallen trees will finally allow the Samana to construct a fishing pier, creating easy access to the fish. Given this opportunity, the Samana are organizing a collective action to bring the trees back to the village and build the pier. The more Samana who participate, the easier the task is for everyone involved. Finally, participants learned that "all members of the Samana understand that only villagers who go to gather the logs necessary to construct the fishing pier will be able to use it. In fact, the Samana will build the fishing pier in plain sight of the entire village: With all the other members of the village watching, no one would ever try to use the fishing pier unless they helped gather the logs to build it."

Results and Discussion

Our analysis strategy was identical to Study 1. As before, preliminary analyses to examine whether our effects were moderated by participant sex used ANOVA and revealed no interactions involving participant sex. We therefore moved on to testing our focused hypotheses.

Did public goods, relative to club goods, lead to more explicit moralization? Yes: Despite a tighter contrast between the public and club goods, not intending to contribute to a public good was rated as more wrong, $r = .33$, $p < .001$ (see Table 2).

Did public goods, relative to club goods, lead to more implicit moralization? Yes: Again despite the tighter contrast, there were greater implicit associations in the public good condition between non-contribution and badness, for both the overall IAT D score ($r = .19$, $p < .05$) and the non-contributor IAT D score ($r = .20$, $p < .05$; see Table 2). Unlike Study 1, there was no effect for the contributor IAT D score ($r = .12$). However, there was also no interaction between contributor/non-contributor and good type on IAT D scores ($p = .34$). There was a trend for both IAT D scores for words to be influenced by good type (see Table 2). As in Study 1, all IAT D scores were significantly greater than zero in both the club and public goods conditions (all ps $< .05$); even in the club good scenario there were implicit associations between non-contribution and badness and between contribution and goodness.

Were implicit and explicit moralization correlated? Unlike Study 1, here explicit ratings correlated somewhat with overall

Table 2
Means (Standard Deviations) of Implicit and Explicit Moralization as a Function of Good Type in Study 2 (Fishing)

Measure	Public good *M* (*SD*)	Club good *M* (*SD*)	*t*	*p*	*r*
Explicit rating of wrongness	4.88 (1.15)	3.93 (1.53)	3.73	.0003	.33
IAT D score, overall	0.36 (0.35)	0.21 (0.44)	2.02	.046	.19
IAT D score, non-contributors	0.41 (0.52)	0.18 (0.58)	2.16	.03	.20
IAT D score, contributors	0.42 (0.47)	0.29 (0.56)	1.32	.19	.12
IAT D score, negative moral words	0.38 (0.50)	0.19 (0.63)	1.73	.09	.16
IAT D score, positive moral words	0.37 (0.43)	0.18 (0.57)	1.94	.06	.18

Note. For all tests, $df = 110$. IAT = Implicit Association Test.

IAT D score ($r = .18$, $p = .06$) and the non-contributor IAT D score ($r = .19$, $p = .04$).

Summary. Despite a variety of changes—different photos, different scenarios, different instantiations of goods—Studies 1 and 2 both supported the preventive moralization hypothesis: Opting out of a public good, relative to a club good, elicited more moralization. But was moralization actually due to the hypothesized preventive effects or some other feature of the goods?

Study 3: Are Non-Contributors Assumed to Always Free Ride?

In some club goods, free riding is possible but difficult to accomplish. For instance, in Study 2, it was possible for non-contributors to use the pier generated by collective action, but it would be socially costly for them to do so. Although this is the realistic case, it creates a potential alternative hypothesis. Perhaps participants' minds treated both public and club goods as if all non-contributors had consumed the collective benefits. If so, this would not be consistent with an assumption of the preventive moralization hypothesis. The preventive moralization hypothesis assumes the mind can encode differences between public and club goods in the likelihood of free riding; this difference, in turn, should cause differences in moralization.

If all non-contributors were treated as free riders, moreover, we might still find the results predicted by the preventive moralization hypothesis without the hypothesis being correct: Perhaps there is greater moralization of free riding on public goods relative to club goods. This could occur, for instance, because public goods are social dilemmas. In social dilemmas, deciding whether to contribute involves making a tradeoff between self-interest and group interest. Perhaps more moralization is required in public goods to motivate others to make this tradeoff in favor of the group.

We test against this alternative by (1) testing whether participants perceive free riding as more likely on public goods (as assumed by the preventive moralization hypothesis) and (2) testing whether there are differences between public and club goods in the moralization of free riding (as predicted by the alternative hypothesis).

Method and Results

One hundred and three participants read either about a public or a club good (materials from Study 2). Using 5-point scales, they then separately rated the likelihood of contributors and non-contributors consuming the collective benefit (access to fish). The scales indicated frequency of fishing from the side of the river (public good) or from the pier (club good) (anchors were 1 = *never*, 2 = *rarely*, 3 = *sometimes*, 4 = *often*, and 5 = *very often*). Whereas there was no difference between public and club goods in the rated likelihood of contributors consuming benefits (both *M*s = 4.3, $SD_{Public} = 0.8$, $SD_{Club} = 1.0$), non-contributors on a public good, relative to a club good, were rated as more likely to consume benefits ($M_{Public} = 3.3$ vs. $M_{Club} = 2.6$, $SD_{Public} = 1.1$, $SD_{Club} = 0.9$; independent samples $t(101) = 3.7$, $p < .001$). The greater likelihood of consumption by non-contributors in the public, relative to the club, good is consistent with the assumption of the preventive moralization hypothesis.

Participants also rated separately on 5-point scales the wrongness of contributors and non-contributors consuming collective benefits—with participants explicitly asked to assume that benefits were actually consumed by everyone. (The anchors were 1 = *not at all wrong*, 2 = *a little wrong*, 3 = *moderately wrong*, 4 = *very wrong*, and 5 = *extremely wrong*.) There were no moralization differences between public and club goods for contributors ($M_{Public} = 1.3$ vs. $M_{Club} = 1.4$, $SD_{Public} = 0.8$, $SD_{Club} = 0.7$; $t(101) = 0.5$, $p = .64$) or for non-contributors ($M_{Public} = 3.3$ vs. $M_{Club} = 3.0$, $SD_{Public} = 0.9$, $SD_{Club} = 1.2$; $t(101) = 1.0$, $p = .32$). Regardless of good type, participants similarly moralized free riding.

Study 3 confirms the preventive moralization assumption: Non-contributors were seen as more likely to consume public good benefits. Study 3 also casts doubt on the alternative hypothesis that the effects of Studies 1 and 2 are due to differences in moralization of free riding in public and club goods (e.g., because public goods are social dilemmas that require a tradeoff of personal for group welfare): When non-contributors were explicitly described as having taken collective benefits, they were equally moralized regardless of the good.

Study 4: Experimental Public and Club Goods Games

Study 4 sought to conceptually replicate the preventive moralization effect with a substantially different method. Instead of having participants take a third-person perspective and complete consequence-free dependent measures, Study 4 involved participants in real groups, from a first-person perspective, with consequential behavior. To do so, we used experimental economic games with the possibility of consequential moralization: costly punishment of non-contributors.

Although experimental games have the drawback of being relatively abstract and artificial compared to real-world situations, they offer important benefits (for some early relevant examples, see, e.g., Dawes, 1980; Dawes, Mctavish, & Shaklee, 1977; Yamagishi, 1986). First, they use behavioral decisions with monetary consequences. Second, because they use money, decisions can be precisely quantified. Third, game structure can be precisely controlled, guaranteeing that situations are public or club goods games. Indeed, in Study 4 they are "pure" public and club goods: It is cost-free to free ride in the public good game and impossible to free ride in the club good game.

All participants played a single public good game and a single club good game, each in a different, anonymous, four-person group. Each game had only a single contribution round and a single punishment round. Given that the games were one-shot and anonymous, economically rational income maximization predicts no punishment in any game because punishment is costly and it cannot affect future payoffs. Thus, if there are differences across goods in punishment, these cannot be due to a rational calculation that differs by good; in other words, if we observe a difference by condition, it cannot be ascribed to this process of domain-general rationality. The games were played for tokens. At the experiment's conclusion a random person from each game could redeem their tokens for cash. Unlike standard games, participants were allowed to redeem as few or as many tokens as they desired. Thus, non-contributors in the public goods game had not (yet) free rode: They did not have to accept cash in exchange for the tokens created by others' contributions.

Method

Participants. Sixty-five undergraduates (49 women) participated for partial course credit. Some also received money (approximately $5–$10) based on their group members' decisions.

Procedure. Each experimental session had 7–10 participants. Each participant was a member of two anonymous, four-person groups, one for the public good game and one for the club good game. For a given participant, their two groups each contained a different random subset of the participants at their session. When the number of participants was not divisible by four, decisions from other participants in the session were used to complete the final group; thus, a participant's decisions might be used in two groups. Participants were not informed whether they or other members of their group had decisions used in multiple groups.

Half the sessions completed the public good game first, the other half the club good game first; the order was randomly determined. Participants learned the rules and made their decisions for one game before moving on to the second. Each game comprised a single contribution round followed by a single punishment round. Only after the entire session did participants learn what other members of their groups chose in either game.

Participants received scripted oral instructions and followed along with their own copy. Participants learned that they would earn tokens based on their own decisions and the decisions of other group members (specifics below). The experimenter quizzed the participants as a group to ensure complete understanding. Participants could not communicate with each other, only with the experimenter. The experimenter would only answer questions about the rules (e.g., questions about the number of tokens earned given a particular set of decisions), not about strategies (e.g., questions about how to maximize earnings). At the end of the session the experimenter randomly selected one person from the club good game and one person from the public good game; these participants could redeem their tokens at a rate of $0.25 per token. Participants were explicitly informed at the beginning of the session that these people could redeem as many or as few tokens as they desired.

Public good game contribution round. The public good game had a single contribution round (followed by a punishment round). Participants received 10 tokens from the experimenter and chose one of two options: non-contribution (keeping all 10 tokens for themselves) or contribution (transferring all 10 tokens into a group account). If they chose to contribute to the group account, these 10 tokens were multiplied into 28 tokens and divided equally among the group members (a share of 7 for each person). Thus, if only one participant contributed, that participant ended the round with only 7 tokens; the remaining three participants ended the round with 17 tokens (their initial 10 and the 7 from their share of the group account.) If two participants contributed, they each ended the round with 14 tokens ($= 7 \times 2$); the remaining participants each ended the round with 24 tokens (their initial 10 plus 14 created by the contributors). If three participants contributed, they each ended the round with 21 tokens ($= 7 \times 3$); the remaining participant ended the round with 31 tokens (the initial $10 + 21$). Finally, if everyone contributed, everyone ended the round with 28 tokens.

Therefore, the game presented a social dilemma: When more people contributed everyone earned more, but it was always payoff maximizing for any individual to not contribute regardless of others' choices. Note that so long as two people contributed, contributors earned more than their initial 10 tokens; of course, non-contributors earned even more.

Club good game contribution round. The club good game also had a single contribution round (followed by a punishment round). As with the public good game contribution round, participants received 10 tokens and could choose to keep them or to contribute them all to the group account. Unlike the public good game, however, only participants who contributed to the group account received any tokens from the group account. Specifically, every contributor received $7 \times n$ tokens, where n is the number of people who chose to contribute their tokens. Thus, if only one participant contributed, that participant ended the contribution round with 7 tokens ($= 7 \times 1$), and other members of the group ended with their initial 10 tokens. If two participants contributed, each ended the contribution round with 14 tokens ($= 7 \times 2$); the other two members again ended the round with just their initial 10 tokens. Finally, if everyone contributed, everyone ended the contribution round with 28 tokens ($= 7 \times 4$). Thus, the game is a club good: Only contributors received benefits from the group account. As with the public good game, note that so long as two people contributed, contributors earned more than their initial 10 tokens; in the club good game, however, non-contributors always ended the contribution round with just 10 tokens.

In our club good game, contributors always earned seven tokens per contributor regardless of the number of contributors. Thus, a contribution of 10 tokens did not generate a fixed amount of tokens to be shared; the amount generated depended on the number of others contributing. An alternative game structure could have had each contribution of 10 tokens become a fixed, larger number of tokens (e.g., 28); these tokens would then be divided among whoever contributed. In this latter structure, the fewer people who contribute, the more benefits there are for everyone per contribution. Either structure would be consistent with a club good. The club good structure we used in Study 4 has the benefit of keeping the game identical in almost all respects to the public good game. This structure, moreover, follows the logic of a number of club goods important in human sociality, such as coalition prestige, where one member's consumption does not substantially affect another's (Delton & Cimino, 2010).

Punishment and explicit wrongness rating. Both games had a single punishment round. We used a standard technique from experimental economics, the strategy method, for the punishment round. In the standard game method, participants would learn the actual choices their group members made before deciding whether to punish. In the strategy method, however, participants do not learn what their group members did; instead, they commit to a punishment decision for each possible choice their group members could have made. There is some debate about whether the strategy method substantively affects behavior compared to standard game methods (Brandts & Charness, 2000; Brosig, Weimann, & Yang, 2003). The benefit of the strategy method, however, is that it allows much denser data collection. This makes the strategy method useful when researchers are studying responses to low frequency behaviors. For instance, given the structure of the club good game most people are likely to contribute; thus, there would be few opportunities to observe punishment of non-contributors. By using the strategy method, we can learn for every participant

whether they would punish non-contributors in both public and club goods.

The punishment round occurred after the contribution round and proceeded as follows: Participants were randomly assigned another member of their group (hereafter the "target"). They did not learn the target's identity and these assignments were not necessarily reciprocal (i.e., if Person A is assigned to Person B, Person B might be assigned to Person C). Participants were then asked to assume that their target had not contributed. Participants were given 5 additional tokens, and, on the assumption that their target had not contributed, they decided how many tokens to spend on punishment. They could spend any number in whole token increments, including 0. Every 1 token spent on punishment subtracted 2 tokens from their target. If the target had, in fact, not contributed, then the participants' choices were automatically realized: Both players lost money according to the participants' choices. If the target had actually contributed, they were not punished and the participants kept their 5 tokens. Given this design, participants could punish and be punished by at most one other person per game and no one could earn less than zero tokens.

The explicit moral wrongness item was as in previous studies, with appropriate rewording.

Results and Discussion

In the public good game, 65% of the sample contributed; in the club good game, all but two participants contributed (of these two, one contributed in the public good game, and the other did not). Because moralization is typically undertaken by contributors, contributors and non-contributors to the public good were analyzed separately. We included in our analyses the two participants who did not contribute to the club good; the results are unchanged if they are excluded. Our analysis strategy for Study 4 was similar to Studies 1 and 2, except we used paired-samples *t*-tests given the within-subjects design. Given the small number of men, tests involving participant sex would be unreliable. Recall, however, that Studies 1 and 2 showed no interactions involving sex.

Did public goods, relative to club goods, lead to more explicit moralization? Yes: Contributors to the public good explicitly judged non-contribution to be more morally wrong in the public good game than in the club good game, $r = .34$, $p < .05$ (see Table 3). Interestingly, non-contributors to the public good judged non-contribution to be less morally wrong in the public good game than in the club good game ($r = .47$, $p < .05$), perhaps attempting to justify their behavior or out of belief that social dilemmas do not require contribution.

Did public goods, relative to club goods, lead to more costly punishment? Yes: Contributors in the public good game engaged in more costly punishment in the public good game than in the club good game, $r = .41$, $p < .01$ (see Table 3). This finding conceptually replicates the basic results of Studies 1 and 2 but with consequential behavior. These data support the hypothesis that preventive moralization is due to evolved decision rules and, given the one-shot and anonymous nature of the games, are inconsistent with the alternative hypothesis that preventive moralization is due to rational calculation. Note that costly punishment by non-contributors to the public good was not affected by the type of good ($r = .04$).

Were moral wrongness ratings correlated with costly punishment? No, none of the four correlations were significant when disaggregating by public/club good and whether the participant did/did not contribute to the public good, ps $> .05$, although all were positive (rs ranged from .08 to .37).

Study 5: A Nationally Representative Sample

With Study 5, we sought to conceptually replicate the preventive moralization hypothesis with a broader sample. Whereas the previous studies used convenience samples of college students, Study 5 surveyed a nationally representative sample of almost 1,000 U.S. adults.

We also sought to extend the previous results in several ways. First, instead of using experimental manipulation, Study 5 measured perceptions that a good was a public good. Measuring perceptions is important because ultimately it is the interpretation of games, not their objective features, that determine behavior (e.g., Halevy, Chou, & Murnighan, 2012; Kelley et al., 2003; Kelley & Thibaut, 1978; Kiyonari et al., 2000). Second, we sought to test against a potential alternative. Perhaps the mind heuristically associates public goods with greater personal benefits. After all, some of the largest benefits of modern life—clean water, public health, transportation infrastructure—are often organized as public goods. If this association does exist, then perceptions of greater personal benefits—not perceptions of a good being public—may have driven the previous results. This could occur if an expectation of greater personal benefits increases the incentive to moralize and thereby encourage contribution from others. To test against this alternative hypothesis, we examined whether percep-

Table 3
Means (Standard Deviations) of Costly Punishment and Explicit Ratings of Moral Wrongness as a Function of Good Type in Study 4 (Experimental Game)

Measure	When playing public good *M* (*SD*)	When playing club good *M* (*SD*)	*t*	*p*	*r*
Explicit ratings of wrongness of non-contribution made by:					
Contributors to public good	3.79 (1.96)	3.26 (1.95)	2.36	.02	.34
Non-contributors to public good	1.54 (0.98)	2.42 (1.56)	−2.56	.02	.47
Tokens spent on costly punishment of a non-contributor by:					
Contributors to public good	2.42 (2.21)	1.70 (2.02)	2.89	.01	.41
Non-contributors to public good	1.58 (1.95)	1.50 (1.62)	0.19	.86	.04

Note. Contributors' $df = 42$, non-contributors' $df = 23$.

tions of a good being public affect moralization when controlling for perceptions of personal benefit. Third, we tested for specificity in the use of preventive moralization. Perceptions of a good being public should drive moralization of non-contributors involved with that particular good; they should not drive moralization of non-contributors more generally. In other words, viewing one collective action to be a public good should not increase moralization of non-contributors involved in other collective actions. To test for this specificity, we contrasted preventive moralization of ingroup non-contributors with moralization of outgroup non-contributors (see also Shinada, Yamagishi, & Ohmura, 2004).

Study 5 asked participants about a hypothetical, upcoming presidential election. We measured their perceptions of how much better off (a) they personally and (b) the nation as a whole would be if their preferred candidate won; the latter measures the degree to which the election is perceived as generating a public good. To assess moralization, we measured their anger at (a) an ingroup non-contributor and at (b) an outgroup non-contributor. Non-contribution was operationalized as neither voting nor spending any time or energy to get the non-contributor's preferred candidate elected. The ingroup member's preferred candidate was the same as the participant's; the outgroup member's was an opponent of the participant's. As before, the collective action had not been completed: No election had yet taken place, and thus benefit consumption was not possible. The preventive moralization hypothesis predicts that, when controlling for perceptions of personal benefit, perceptions of national benefit should independently predict anger at an ingroup non-contributor; moreover, perceptions of national benefits for the participant's candidate winning should not predict anger at an outgroup non-contributor.

Method

Participants and data collection. Nine hundred twenty-three people responded over the Internet. The survey was administered by Time-Sharing Experiments in the Social Sciences (www.tessexperiments.org). Technical aspects and participant recruiting were handled by Knowledge Networks (www.knowledgenetworks.com/ganp). The present data are part of a larger project; for a fuller project description and details of Knowledge Networks methods, see www.tessexperiments.org/data/delton363.html. Data from 32 additional participants were excluded because they did not answer >50% of the 13 questions composing our larger survey.

Fifty percent of the sample was women. Based on the categories provided by Knowledge Networks, the ethnic/racial breakdown was as follows: 77% white, 9% black, 9% Hispanic, 2% two or more races, 3% other. The modal education was high school graduate or equivalent (29%), followed closely by some college or bachelor's degree (23% and 22%, respectively). Average age was 47 years ($SD = 16$). Average income was approximately $38,000/year (ranging from approximately $5,000 to $175,000). Seventy percent belonged to a political party; the rest were independents.

Materials and measures. Participants were asked to imagine an upcoming, hypothetical presidential election. The current president did not share the participants' views. Because of these diverging views, participants and people who held similar opinions as the participants' were "worried and angry" and wanted "to win back control of the White House." All questions used scales ranging from 0 to 100, with anchors at 0, 50, and 100, and responses allowed in multiples of 5. Personal benefit: "How much better off would *you personally* expect to be if your candidate is elected and his opponent is defeated?" National benefit: "How much better off would you expect *the nation as a whole* to be if your candidate is elected and his opponent is defeated?" (Anchors for both were as follows: *not at all better off*, *moderately better off*, *very much better off*.) Anger at ingroup and outgroup non-contributors: This voter "did not spend any time, effort, or money" trying to get their (the voter's) preferred candidate elected and "did not even bother to go out and vote . . . how *angry* would you feel towards [him/her]?" (Anchors were as follows: *not at all angry*, *moderately angry*, *very angry*.) The elided parts described how the participants and the voters were members of the same or different political parties (if the participant belonged to a political party) or were of the same or different political views (if the participant was an independent). Ingroup non-contributors supported the same candidate as the participant; outgroup non-contributors an opposing candidate. The non-contributors' sex was randomized.[3]

Analysis strategy. Approximately 3% of responses were missing across the two items measuring anger and the two items measuring perceptions of benefits. To analyze the data despite these missing values, we used Amos statistical software (Arbuckle, 2010). Amos is a structural equation modeling program that uses maximum likelihood methods. These methods allow the estimation of parameters even with incomplete data. All analyses use the demographic variables described above as covariates. We allowed covariates and predictors to freely covary with each other; when a model had multiple dependent measures, we allowed their residual variances to freely covary. To test whether regression weights differed from each other, we used χ^2 tests on nested models (Kline, 2010): In the initial model, each of the two focal weights was allowed to have a unique estimate; in a second model, the focal weights were constrained to have identical estimates. A significant χ^2 test would show that the second model fits the data more poorly and that separate parameters are required; in other words, a significant test shows that the focal weights are not identical.

Results and Discussion

Did perceptions of a good being public predict preventive moralization? Yes. Based on a maximum likelihood analysis, the more that participants viewed the election of their candidate as a national benefit, the more angry they were at an ingroup non-contributor (standardized regression weight = .26, $p < .001$). Studies 1–4 showed that experimental manipulations of a good being public (vs. club) led to moralization; the present result shows that participants' quantitative perceptions of a good being public, not just experimental manipulation, also predicted greater moralization.

Did perceptions of a good being public continue to predict moralization when expected personal benefit was controlled? Yes, when personal benefit was also included in the analysis, personal and national benefit both independently predicted anger at an ingroup non-contributor (standardized regression weights =

[3] Although beyond the scope of this article, the survey manipulated the likelihood that the participants' preferred candidate would win and whether the hypothetical voters expected a large or small benefit if their preferred candidate won. These variables were also used as covariates.

.23 and .11, respectively; $ps < .05$). Note that the correlation between personal and national benefit is .70. Thus, although the standardized regression weights are not large, we think it is telling that both variables had independent effects on anger despite their high correlation with each other. Perceptions of a good being public, over and above perceptions of personal benefit, predicted greater moralization. This result suggests that, even if there is a heuristic association between public goods and greater personal benefits, this potential association cannot explain the totality of our results.

Was preventive moralization absent for outgroup non-contributors? Yes. To test for the absence of an effect for outgroup non-contributors, we created a maximum likelihood model that contained (a) anger at an ingroup member and anger at an outgroup member as dependent measures and (b) personal and national benefits as predictors. The preventive moralization hypothesis predicts a larger effect of national benefits on anger toward an ingroup non-contributor relative to anger at an outgroup non-contributor; this was true, $\chi^2(1) = 8.03$, $p < .01$. Indeed, national benefit did not predict anger toward outgroup non-contributors whatsoever, even after controlling for personal benefits ($\beta = -.02$, $p = .70$). The lack of an effect is especially striking given that members of the outgroup are likely to perceive their candidate's election to be a public good. Importantly, however, it was a separate collective action from the one generating participants' perceptions. Perceptions of whether a good is public seem to drive moralization of non-contributors involved with that good, but not moralization of non-contributors generally.

Studies 6–9: Quantitatively Manipulating Likelihood of Future Free Riding

Studies 1–5 were designed to provide initial confirmatory support for the preventive moralization hypothesis and to rule out a series of related alternative hypotheses, including alternatives based on labor recruitment, group conformity and related constructs, and rational income maximization. We designed Studies 6–9 to address two additional issues. These studies encompassed a diverse sample of almost 700 U.S. adults.

Our first goal with Studies 6–9 was to provide evidence for the process predicted by the preventive moralization hypothesis. The preventive moralization hypothesis proposes that one source of moralization is the possibility that a person who merely opts out may become a free rider in the future. Thus, preventive moralization should be mediated by perceptions of the likelihood of free riding (see Figure 1). Although we cannot test this mediation hypothesis directly (because we cannot directly manipulate the internal perception of likelihood of free riding), in Studies 6–9 we can test for statistical mediation that is at least consistent with this causal model.

A second goal was to empirically examine a wider range of likelihoods to free ride. Studies 1–4 used stark contrasts: Free riding was trivially easy and highly likely in the public goods and virtually impossible in the club goods. In Studies 6–9, we quantitatively manipulated the likelihood of free riding through five levels. This allowed us to test whether preventive moralization occurs outside of cases where free riding is all but guaranteed. It also allowed us to test whether preventive moralization is a simple all-or-none phenomenon or whether it quantitatively tracks the likelihood of free riding.

Our third goal was to explore a variety of potential cues that could serve as input to perceptions of the likelihood of free riding (i.e., possible inputs in the leftmost box of Figure 1). Each study manipulated a different possible cue to the likelihood of free riding: non-contributors' subjective valuation of the good, non-contributors' cost if they free rode, the ability of contributors to monitor the good and prevent free riding, and the benefits non-contributors can obtain if they do free ride.

Method

Participants. Six hundred sixty-eight people participated (280 were female). Average age was 30 years ($SD = 10$). Average income was approximately $33,000/year (ranging from approximately $15,000 to $150,000). Participants were randomly assigned to experiments and to conditions within experiments.

To recruit participants, we used the Mechanical Turk platform, a subsidiary of Amazon.com. Participants were paid between $0.10 and $0.20, consistent with standard rates on Mechanical Turk for a survey this length. Mechanical Turk has recently emerged as a promising tool for the behavioral sciences, allowing data collection in surveys or simple economic games from a diverse sample of participants (Buhrmester, Kwang, & Gosling, 2011). Although Mechanical Turk does not guarantee a nationally representative sample (as in our Study 5), it does allow us to test predictions using a broader pool than psychology undergraduates.

Materials. In all four studies, participants were asked to imagine that they were members of a gardening co-op. They read that: "Every year, the owners of the co-op organize a 'bag drive.' Co-op members who are willing each contribute $300 to fund the purchase of high quality fertilizer that can be used throughout the year. The bags of fertilizer will be delivered approximately two weeks from today." Thus, free riding was not yet possible because no benefits were available to take.

Cooperation was synergistic: "The greater the number of members who contribute, the cheaper it is to buy the fertilizer in bulk. In other words, the more people who contribute, the more fertilizer there is for *each* member of the co-op."

Next, they read that the bags would be stored in a common area that everyone had access to. Thus, even non-contributors could free ride by taking bags.

Co-op members were described as having received a form where they could mark whether they opted in to the bag drive. On this form, members also rated their perceptions of the usefulness of the fertilizer from 0 (*not useful at all*) to 10 (*extremely*). Co-op members who opted in were described as choosing 10 on average; participants were asked to assume they opted in and rated the usefulness as 10. Ratings by the target who opted out served as the manipulation in Study 6; in Studies 7–9, the target who opted out always also chose 10.

Participants in all studies read about a member, Alan, who opted out of the bag drive. Further details about Alan or the actions available to him differed by study.

Study 6: Subjective valuation. Participants in this study learned how Alan rated the usefulness of the bags from 0 (*not at all useful*) to 10 (*extremely useful*). His ratings differed by condi-

tion: 0, 2.5, 5, 7.5, or 10. Because participants were asked to assume they rated the bags at 10, Alan therefore rated them as 0%, 25%, 50%, 75%, or 100% as useful as participants did.

Study 7: Cost. Participants in this study learned that the co-op had several tracts of land. Alan gardened at and lived very near one of their locations, but the bags would be stored at a different location. The drive to get to the location with the bags varied by condition: 5 min, 30 min, 1 hr, 2 hr, or 4 hr.

Study 8: Monitoring. Participants in this study learned that volunteers staffed the common area to monitor whether people trying to take bags had actually contributed. The percentage of time there were volunteers available varied by condition: 100%, 75%, 50%, 25%, or 0% of the time.

Study 9: Benefits. Participants in this study learned that: "At the insistence of the co-op owners, and despite objections from some of the members who contributed, *everyone* who wants fertilizer can come and get bags of fertilizer if they choose." Those who contributed would get approximately 100 bags (the specific number depending on how many people ultimately contributed). The number of bags available to non-contributors varied by condition: about 100 bags, 75 bags, 50 bags, 25 bags, or 1 bag.

Measures. Participants answered three dependent measures, each on 7-point scales. To measure perceptions of the likelihood of free riding, participants were asked the following: "How likely is Alan to take some fertilizer despite not contributing to the fertilizer drive?" (1 = *not at all likely*, 7 = *very likely*). To measure moralization, they were asked two questions: "How morally wrong is it that Alan did not contribute to the fertilizer drive?" (1 = *not at all wrong*, 7 = *very wrong*), and "How angry are you at Alan?" (1 = *not at all angry*, 7 = *very angry*). Finally, participants answered demographic questions about age, income, and sex.

Analysis strategy. Our primary question was whether within each study there was an indirect statistical effect of our manipulation (valuation, cost, monitoring, or benefits) on preventive moralization (moral wrongness and anger) through perceived likelihood of free riding. To test for this effect, we used a bootstrapping approach to statistical mediation and indirect effects (Preacher & Hayes, 2008). The experimental manipulations were coded 1 through 5; they were always ordered so that 1 was the condition predicted to be least likely to induce free riding, and 5 was the condition predicted to be most likely to induce free riding. We report estimates of the bootstrapped indirect effects and their bias-corrected and accelerated 95% confidence intervals. If the estimate is positive and the confidence intervals do not include zero, then the data indicate significant support for an indirect effect. All bootstrapping analyses used 20,000 bootstrap samples and age, income, and sex as covariates. Estimates are reported in their raw metric. For instance, an indirect effect of .5 means that a one unit increase in the independent variable (e.g., subjective valuation) will cause a half-unit increase in ratings of moralization (e.g., anger) as mediated by perceived likelihood of free riding. We emphasize that inferences about causal mediation are necessarily indirect. Although we manipulated cues to future free riding, we were only able to measure moralization and perceptions of future free riding. Thus, our results can only speak directly to statistical mediation.

Results

Study 6: Valuation—Does perceived likelihood of free riding statistically mediate the relationship between a non-contributor's valuation of the good and moralization of that non-contributor? Yes, there were significant indirect effects of the manipulation of fertilizer valuation on both moral wrongness and anger, and these indirect effects were statistically mediated by perceptions of the likelihood of free riding. For moral wrongness, the bootstrapped indirect effect estimate was .39 with a 95% confidence interval from .24 to .57. For anger, the bootstrapped indirect effect estimate was .41 with a 95% confidence interval from .26 to .57. Neither confidence interval includes zero, which supports the hypotheses that the manipulation of valuation would lead to preventive moralization and that such effects would be statistically mediated by the perceived likelihood of free riding.

We also conducted additional regression analyses to probe the components of this indirect effect; these analyses used the same covariates as the bootstrapping analyses. First, these analyses showed that manipulation of valuation had a significant impact on the mediator, the perceived likelihood of free riding (estimate = .58, $p < .001$). The mediator of perceived likelihood, moreover, significantly predicted both moral wrongness and anger (estimates = .68 and .70, ps $< .001$). With the mediator not in the model, manipulation of fertilizer valuation had a significant total effect on both moral wrongness and anger (estimates = .48 and .44, respectively; ps $< .001$). With the mediator included, the effects of valuation on both wrongness and anger were no longer significant (estimates = .08 and .03, ps $> .4$). These analyses reinforce the bootstrapping results and additionally show that the mediator statistically accounts almost entirely for the correlations between valuation and preventive moralization.

Study 7: Costs—Does the perceived likelihood of free riding statistically mediate the relationship between costs of future free riding and moralization of a potential free rider? Yes, there were significant indirect effects of the manipulation of costs on both moral wrongness and anger and these indirect effects were statistically mediated by perceptions of the likelihood of free riding. For moral wrongness, the bootstrapped indirect effect estimate was .19 with a 95% confidence interval from .09 to .33. For anger, the bootstrapped indirect effect estimate was .19 with a 95% confidence interval from .09 to .32. Neither confidence interval includes zero, supporting the hypotheses that manipulations of cost would lead to preventive moralization and that such effects would be statistically mediated by the perceived likelihood of free riding.

Additional regression analyses showed that the manipulation of costs had a significant impact on the mediator, the perceived likelihood of free riding (estimate = .33, $p < .001$). The mediator of perceived likelihood, moreover, significantly predicted both moral wrongness and anger (estimates = .58 and .56, ps $< .001$). With the mediator not in the model, manipulations of cost had a significant total effect on both moral wrongness and anger (estimates = .30 and .32, respectively; ps $< .01$). With the mediator included, the effects of cost on both wrongness and anger were no longer significant (estimates = .11 and .13, ps $> .15$). These analyses reinforce the bootstrapping results and additionally show

that the mediator statistically accounts almost entirely for the correlations between valuation and the measures of moralization.

Study 8: Monitoring—Does perceived likelihood of free riding statistically mediate the relationship between the availability of monitoring and moralization of a potential free rider? Yes, there were significant indirect effects of the manipulation of monitoring on both moral wrongness and anger and these indirect effects were statistically mediated by the perceived likelihood of free riding. For moral wrongness, the bootstrapped indirect effect estimate was .14 with a 95% confidence interval from .03 to .27. For anger, the bootstrapped indirect effect estimate was .14 with a 95% confidence interval from .03 to .26. Neither confidence interval includes zero, supporting the hypotheses that the manipulation of monitoring would lead to preventive moralization and that such effects would be statistically mediated by the perceived likelihood of free riding.

Additional regression analyses showed that manipulations of monitoring availability had a significant impact on the mediator, the perceived likelihood of free riding (estimate = .26, $p < .01$). The mediator of perceived likelihood, moreover, significantly predicted both moral wrongness and anger (estimates = .55 and .53, respectively; $ps < .001$). With the mediator not in the model, the manipulation of monitoring did not have a significant total effect on either moral wrongness or anger (estimates = .05 and .17, respectively; $ps > .14$). With the mediator included, the effects of monitoring on both wrongness and anger were even smaller or descriptively in the wrong direction (estimates = –.09 and .03, respectively; $ps > .4$). Some approaches to statistical mediation require that there be a significant total effect (i.e., when the mediator is removed from the model) as a preliminary condition for testing mediation, other approaches view a significant indirect effect as most probative (Preacher & Hayes, 2008; Shrout & Bolger, 2002). Although our results do not meet the former, more stringent version of mediation, we believe the results of Study 8 are mostly supportive of the hypothesis: There is a significant causal effect of manipulations in the independent variable on the mediator, there are significant correlations of the mediator with the dependent variables, and there are significant statistical indirect effects of the independent variable on the dependent variables through the mediator.

Again we note that inferences about causal mediation in Studies 6–8 are indirect. Both the mediator and the dependent variable were merely measured, not manipulated. Thus, we can only provide direct evidence for statistical mediation that is consistent with the causal model in Figure 1.

Study 9: Benefits—Does perceived likelihood of free riding statistically mediate the relationship between the size of the benefits available to free riders and moralization of a potential free rider? No, the indirect effects were small for both moral wrongness and anger, and both confidence intervals included zero (estimates of both = −.02; 95% CIs [−.08, .02] and [−.1, .03], respectively). Regression analyses revealed no effects involving the manipulation of benefits (all $ps > .2$).

One reason for these null results may be ceiling effects. As shown in Tables 4, 5, and 6, the means of perceived likelihood of free riding, moral wrongness, and anger were are all higher in Study 9 compared to the other three studies. The measure of likelihood was especially high, with means approaching the scale maximum of 7. Collapsing across condition and based on a series of independent *t*-tests, all three measures in Study 9 received higher ratings than any other study (for all nine comparisons, $ps < .001$). Comparing Studies 6–8 against each other revealed essentially no differences (eight of nine $ps > .15$; one comparison had $p = .011$). Although Study 9's data do not support the hypothesis of statistical mediation, we think Study 9 does illustrate just how strongly preventive moralization can operate. Consider that a free rider who is caught is likely to face sanctioning and reputational damage. In one condition of Study 9, free riders could only gain a single bag of fertilizer—little gain for potentially large negative repercussions. Yet, participants believed a non-contributor in this situation was nonetheless very likely to free ride, rating the likelihood an average of 6.3 out of 7; preventive moralization was acting quite strongly even in a case where free riding might be objectively unlikely.

Finally, note that in Study 9 as in the Studies 6–8, the perceived likelihood of free riding predicted both moral wrongness and anger (estimates = .38 and .49, respectively; $ps < .01$). Thus, despite the experimental manipulation failing in this study, the correlational data still show the pattern predicted by the preventive moralization hypothesis: greater perceived likelihood of free riding is associated with greater moralization.

Does preventive moralization occur only when free riding is virtually guaranteed? One could argue that although preventive moralization is a real phenomenon, it is not robust. Perhaps preventive moralization only occurs when free riding is virtually guaranteed. For instance, in Studies 1–4, one might expect that

Table 4
Means (Standard Deviations) of Ratings of Perceived Likelihood That a Person Who Opts Out Will Later Consume Collective Benefits (Studies 6–9)

	Ratings of the perceived likelihood of later benefit consumption				
	Ordinal ranking of conditions				
	←Free riding least likely			Free riding most likely→	
	1	2	3	4	5
Manipulation	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Valuation (Study 6)	3.61 (1.96)	3.17 (1.46)	4.30 (1.92)	5.41 (1.19)	5.55 (1.16)
Costs (Study 7)	3.85 (1.94)	3.81 (1.64)	4.32 (1.42)	4.42 (1.43)	5.15 (1.63)
Monitoring (Study 8)	3.87 (1.76)	5.20 (0.91)	5.26 (1.48)	5.43 (1.73)	4.80 (1.76)
Benefits (Study 9)	6.34 (0.94)	6.41 (1.07)	6.49 (0.80)	6.33 (1.02)	6.24 (1.06)

Table 5
Means (Standard Deviations) of Ratings of Moral Wrongness of Opting Out of a Public Good (Studies 6–9)

	Ratings of moral wrongness of opting out				
	Ordinal ranking of conditions				
	←Free riding least likely			Free riding most likely→	
	1	2	3	4	5
Manipulation	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Valuation (Study 6)	2.83 (2.10)	2.67 (1.86)	3.05 (2.39)	4.16 (1.80)	4.42 (1.87)
Costs (Study 7)	3.18 (2.16)	3.53 (1.70)	3.75 (2.03)	4.07 (2.02)	4.38 (2.13)
Monitoring (Study 8)	3.26 (1.90)	4.16 (1.84)	4.00 (2.00)	3.21 (2.04)	3.84 (2.39)
Benefits (Study 9)	4.62 (1.92)	4.92 (1.52)	5.14 (1.72)	4.93 (1.86)	5.14 (1.46)

people opting out would almost certainly consume the benefits. If preventive moralization only occurs in these limiting cases, then it would have little relevance in explaining real world behavior. Studies 6–9 were designed to test against this alternative by manipulating the likelihood of free riding quantitatively, rather than using the dichotomous contrasts of Studies 1–4.

One way to examine this is to use the unintended difference between Study 9 and Studies 6–8. Although Study 9 did not show any effects of our experimental manipulation, compared to Studies 6–8, Study 9's ratings of the likelihood of free riding and ratings of anger and wrongness were uniformly high (see results above). The ratings for likelihood of free riding were especially high compared to the other studies, suggesting that, at least for Studies 6–8, participants did not view the potential free rider as guaranteed to free ride. Nonetheless, in Studies 6–8 participants did engage in preventive moralization.

Because Study 9 did not show any effects of the manipulation, we further focus only on Studies 6–8. First, note that descriptively most measures of likelihood, moral wrongness, and anger tended to increase as the level of the cue increased (however, this increase was not perfectly monotonic; see Tables 4–6). Second, to formally test that preventive moralization occurs outside the boundary case of guaranteed free riding, we tested whether there were linear effects of our manipulations on measures of likelihood, wrongness, and anger even when excluding the condition within each study most likely to induce free riding. Of these nine tests, seven were significant or marginally so (for those seven, $ps \leq .063$). The only two that did not show an effect were anger and wrongness in Study 8 ($ps > .12$); recall that this study showed no effects on these variables even if all levels of the independent variable were included. Thus, whenever our manipulation was successful, it was successful even outside the narrow case of a condition most likely to induce free riding.

General Discussion

The social mind is not unitary but encompasses multiple domains, each with their own proprietary logic and decision rules (Bugental, 2000; Fiske, 1992; Kenrick et al., 2003, 2010). One important domain of human sociality is collective action: multiple people working together to produce a shared benefit. One important set of decision rules for collective action are those that regulate moralization. Because moralization decisions are inherently uncertain, we applied the logic of error management (Haselton et al., 2009; Haselton & Nettle, 2006) to develop the preventive moralization hypothesis: merely opting out of helping to create a public good can make that person a target of moralization because that person might (but also might not) free ride in the future.

Across nine studies, using diverse methods and measures and including a nationally representative sample of almost 1,000 U.S. adults, we showed that non-contributors to public goods were preventively moralized. Despite not actively taking any benefits, their behavior was viewed as morally wrong (Studies 1–4, 6–9), they were targets of anger (Studies 5–9), they were implicitly associated with negative moral traits (Studies 1–2), and they were punished (Study 4). In other words, they were treated as free riders despite not free riding—they were preventively moralized. Moreover, preventive moralization correlated with perceptions that a

Table 6
Means (Standard Deviations) of Ratings of Anger at a Person Who Opts Out of a Public Good (Studies 6–9)

	Ratings of anger at a person who opts out				
	Ordinal ranking of conditions				
	←Free riding least likely			Free riding most likely→	
	1	2	3	4	5
Manipulation	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Valuation (Study 6)	2.75 (1.89)	2.63 (1.88)	3.05 (2.31)	3.84 (1.75)	4.26 (1.62)
Costs (Study 7)	2.94 (1.98)	3.00 (1.77)	3.64 (2.02)	3.84 (1.90)	4.06 (2.14)
Monitoring (Study 8)	3.16 (2.09)	3.56 (1.66)	4.00 (1.79)	3.61 (2.08)	3.92 (2.20)
Benefits (Study 9)	4.28 (1.85)	4.62 (1.80)	4.54 (1.83)	4.50 (1.83)	4.83 (1.87)

non-contributor was likely to free ride (Studies 6–9), perceptions of the likelihood of free riding were quantitatively modulated by a variety of cues (Studies 6–8), and the effect of these cues on moralization was statistically mediated by perceptions of the likelihood of free riding (Studies 6–8). Preventive moralization was also a robust phenomenon, appearing even in cases where free riding was not certain (Studies 6–9).

Preventive moralization did not occur because moralization creates more labor, thereby producing more benefits—the potential to create labor was held constant across public and club goods (Studies 1–3 and especially Study 4). It was not because the social dilemma of public goods elicits more moralization or because non-contributors are confused with people who have illicitly taken benefits (Study 3). It was not because moralization can be rationally calculated to be more useful in public goods (Study 4). It was not because the mind heuristically associates public goods with greater personal benefits (Study 5). And it was not because non-contributors could objectively be viewed as selfish or deviant (Studies 1–4). Instead, preventive moralization appears tailored to solve the problem of whether to moralize non-contributors now on the chance they become free riders in the future.

Future Directions

It remains for future research to examine how moralization of people who simply opt out compares to moralization of people who have actually taken collective benefits. The preventive moralization hypothesis does not make any strong predictions here. Free riders could receive equal moralization, more moralization, or even less moralization than those who simply opt out. Although we think the latter is unlikely, there may be situations where moralizing already-completed free riding has little utility. Instead, moralizing people who might still contribute could have larger associated gains.

Future research could also further delve into the hypothesis that preventive moralization is a product of design for error management. Generally, an error management framework predicts that decision making should sometimes be "biased" away from objectively correct responding and thereby respond more adaptively (e.g., even if 99% of the time foliage does not hide a predator, the mind might respond as if it usually does). Our studies examined one aspect of "bias" in moralization. Most people would report that others should not be sanctioned for a wrong they have not committed. Yet, contra this normative stance, our studies showed that people explicitly and implicitly sanctioned others simply for opting out, despite those others not illicitly taking benefits. A logically separable error management effect could proceed from the fact that only some people who opt out will eventually take benefits illicitly. Are people who merely opt out moralized out of proportion to the number of people who would eventually take benefits illicitly? This could be tested by an experiment comparing punishment of free riders—a situation where punishers know who has actually taken benefits—to punishment of people who merely opt out—a situation with uncertainty regarding who will take benefits. If there is more punishment in the latter condition, this would suggest the operation of an error management strategy, a strategy that serves to catch all people who will eventually free ride even if not all actually do free ride.

Although Studies 6–9 were designed to examine potential mediators of preventive moralization, another topic for future research is to examine potential moderators. One potential moderator is the number of moralizers available. Evolutionary modeling work shows that moralization is easier to sustain when multiple parties coordinate to do it (Boyd, Gintis, & Bowles, 2010), and ethnographic research shows that moralization is more likely with coordination (Boehm, 1993). A second potential moderator is social or physical power. People with greater power are more likely to be able to successfully implement moralization and so might be more likely to engage in it (Maner & Mead, 2010; Sell, Tooby, & Cosmides, 2009). A third potential moderator is the nature of the relationship between the potential moralizer and the person being moralized. When a relationship is likely to continue into the future, it may be more beneficial to change the behavior of a person who opts out and encourage them to cooperate in the future (Krasnow, Cosmides, Pedersen, & Tooby, 2012).

Implicit and Explicit Responses to Non-Contributors

Throughout this article, we have relied on a multi-method approach using implicit, explicit, and behavioral measures. Generally, these measures failed to significantly correlate with each other, although the direction of correlation was usually positive. Implicit and explicit items often have moderate correlations (see meta-analysis by Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005). What might explain these effects in our studies? One possibility is that different measures tap mechanisms with somewhat different functions, leading to weak associations (Cushman, 2008; Lieberman & Linke, 2007). A related possibility is that moral judgments are produced by multiple intuitive systems. This possibility is consistent with theoretical proposals suggesting that social and moral decision making is underwritten by a number of subsystems, many designed to deliver inferences in different domains (Haidt, 2007; Lieberman, Tooby, & Cosmides, 2003; Tooby & Cosmides, 2010; Tybur et al., 2009), some of which may be unrelated to moral judgment per se (DeScioli & Kurzban, 2009, 2013). The relative weighting of different inference engines, moreover, may vary across time, space, and context (Cohen & Rozin, 2001; Graham, Haidt, & Nosek, 2009; Yamagishi, Hashimoto, & Schug, 2008). In some cases, one's moral intuitions may have little or no causal relationship with consciously reasoned moral judgments (Haidt, 2001). In recent work, for example, men's upper body strength (one cause of interpersonal competitive success during human evolution; Sell, Hone, & Pound, 2012) predicted attitudes about the acceptability of force in international conflict and about public policy regarding redistribution—despite the impossibility of personal strength having any meaningful impact on these issues (Petersen, Sznycer, Sell, Tooby, & Cosmides, in press; Sell et al., 2009).

Another possibility is that explicit declarations of wrongness may function to recruit others to moralize non-contributors and implicit judgments may function as personal assessments to avoid a non-contributor in the future. The extent to which these correlate may vary by a person's social power (see above) or general agreeableness. Thus, our explicit measure might be tapping mechanisms involved with changing others' behavior and our implicit measure might be tapping mechanisms involved with changing one's own behavior.

The possible disconnect between explicit, implicit, and behavioral measures of costly punishment also relates to an issue raised in the introduction. In some cases, there can be substantial costs to preventive moralization. Moralizing a long-term associate for a single instance of opting out might destroy an otherwise profitable relationship. Indeed, it might even invite costly retaliation (Nikiforakis, 2008). On the other hand, moralizing within low stakes relationships (such as with recently met strangers) has fewer costs; even if the relationship ends, little is lost (Sznycer et al., 2012). In cases involving close others and little opting out, the primary moralization response might be internal, simply registering that steps may need to be taken in the future should the behavior continue. Alternatively, moralization might consist of avoidance or subtle exclusion instead of outright punishment. In this way the moralizer avoids being free ridden on without inviting retaliation. Overt, costly punishment might only be used when the stakes are low (such as our Study 4), when there is a consistent pattern of opting out, or when many people can be coordinated to punish and lessen the chance of costly retaliation.

The Social Cognition of Cooperation and Coalitions

This research adds to a growing literature on the evolutionary social cognition of cooperation and coalitions (Kurzban & Neuberg, 2005). Past work has, for example, investigated dedicated machinery for detecting alliances (Cosmides, Tooby, & Kurzban, 2003; Kurzban, Tooby, & Cosmides, 2001), for engaging in leader–follower relationships (Van Vugt, Hogan, & Kaiser, 2008), for integrating new members into enduring coalitions (Cimino & Delton, 2010; Delton & Cimino, 2010), and for managing intergroup relationships (Ackerman et al., 2006; McDonald et al., 2012; Miller, Maner, & Becker, 2010).

Intense focus has been placed on characterizing mechanisms for detecting free riders and for responding to them with punishment or other sanctions (e.g., Delton et al., 2012; Kiyonari & Barclay, 2008; Lieberman & Linke, 2007; Price, 2005; Price et al., 2002; Shinada & Yamagishi, 2007; Shinada et al., 2004; Tooby et al., 2006). One open question is whether there are distinct functions served by second-party sanctioning (sanctioning by a directly affected party) and third-party sanctioning (sanctioning by an apparently disinterested observer; see DeScioli & Kurzban, 2009, 2013). Whereas second-party sanctioning is widely thought to be targeted at changing an offender's behavior (McCullough et al., 2013), third-party sanctioning might be designed, in part, as a signal to others (Kurzban, DeScioli, & O'Brien, 2007). Our results have the potential to blur this distinction: In one sense, people who preventively moralize have not had their labor exploited; thus, they are engaging in third-party sanctioning because they have not been directly affected. In another sense, on the preventive moralization hypothesis, their minds act as if or implicitly assume that someone who opts out has exploited collective benefits, activating anti-free rider responses; thus, the design of the mechanism is a design for second-party sanctioning.

Finally, we note that differing intuitions about preventive moralization have the potential to create political and moral disagreement (cf. Haidt & Graham, 2007). For example, some may view a given collective action as producing a public good—a benefit to everyone in the community—while others may view it as a net drain—a waste of everyone's time and energy. Those who believe a public good is being produced may become angry and sanction the uninterested—quite rightfully from their perspective. Yet, the uninterested may view the other side as illicitly forcing their labor or resources—again, rightfully from their perspective. Thus, a reasonable difference of opinion about the nature of a good can become a moral or political battleground (Tooby & Cosmides, 2010). Moreover, by framing goods as public goods or club goods, people might strategically attempt to bring support to favored projects and prevent the completion of disfavored projects. Such disagreements and strategic framing might play a role in political polarization.

Conclusion

Humans cooperate, are altruistic, and give generously to others. Such achievements are often contrasted with less vaunted aspects of human nature, such as homicide and warfare. But cooperation requires vigilance and, sometimes, anger and punishment. However, the mind does not only moralize people who commit a clear wrong—sometimes the mind moralizes non-contributors who have taken no benefits at all.

References

Ackerman, J. M., Shapiro, J. R., Neuberg, S. L., Kenrick, D. T., Becker, D. V., Griskevicius, V., . . . Schaller, M. (2006). They all look the same to me (unless they're angry): From out-group homogeneity to out-group heterogeneity. *Psychological Science, 17,* 836–840. doi:10.1111/j.1467-9280.2006.01790.x

Arbuckle, J. (2010). *Amos 17.0 user's guide*. Crawfordville, FL: Amos Development Corporation.

Bloom, P. (2010, March 25). How do morals change? *Nature, 464,* 490. doi:10.1038/464490a

Boehm, C. (1993). Egalitarian behavior and reverse dominance hierarchy. *Current Anthropology, 34,* 227–254. doi:10.1086/204166

Boyd, R., Gintis, H., & Bowles, S. (2010, April 30). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science, 328,* 617–620. doi:10.1126/science.1183665

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology, 13,* 171–195. doi:10.1016/0162-3095(92)90032-Y

Bradfield, M. (1973). *A natural history of associations: A study in the meaning of community*. London, England: Duckworth.

Brandts, J., & Charness, G. (2000). Hot vs. cold: Sequential responses and preference stability in experimental games. *Experimental Economics, 2,* 227–238. doi:10.1007/BF01669197

Brosig, J., Weimann, J., & Yang, C.-L. (2003). The hot versus cold effect in a simple bargaining experiment. *Experimental Economics, 6,* 75–90. doi:10.1023/A:1024204826499

Bugental, D. B. (2000). Acquisition of the algorithms of social life: A domain-based approach. *Psychological Bulletin, 126,* 187–219. doi:10.1037/0033-2909.126.2.187

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5. doi:10.1177/1745691610393980

Cimino, A., & Delton, A. W. (2010). On the perception of newcomers: Toward an evolved psychology of intergenerational coalitions. *Human Nature, 21,* 186–202. doi:10.1007/s12110-010-9088-y

Cohen, A. B., & Rozin, P. (2001). Religion and the morality of mentality. *Journal of Personality and Social Psychology, 81,* 697–710. doi:10.1037/0022-3514.81.4.697

Cosmides, L., Tooby, J., & Kurzban, R. (2003). Perceptions of race. *Trends in Cognitive Sciences, 7,* 173–179. doi:10.1016/S1364-6613(03)00057-3

Crosson, S., Orbell, J., & Arrow, H. (2004). "Social poker": A laboratory test of predictions from club theory. *Rationality and Society, 16,* 225–248.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108,* 353–380. doi:10.1016/j.cognition.2008.03.006

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology, 31,* 169–193. doi:10.1146/annurev.ps.31.020180.001125

Dawes, R. M., Mctavish, J., & Shaklee, H. (1977). Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation. *Journal of Personality and Social Psychology, 35,* 1–11. doi:10.1037/0022-3514.35.1.1

Delton, A. W., & Cimino, A. (2010). Exploring the evolved concept of newcomer: Experimental tests of a cognitive model. *Evolutionary Psychology, 8,* 317–335.

Delton, A. W., Cosmides, L., Guemo, M., Robertson, T. E., & Tooby, J. (2012). The psychosemantics of free riding: Dissecting the architecture of a moral concept. *Journal of Personality and Social Psychology, 102,* 1252–1270. doi:10.1037/a0027026

Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). The evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences, USA, 108,* 13335–13340. doi:10.1073/pnas.1102131108

Delton, A. W., & Robertson, T. E. (2012). The social cognition of social foraging. *Evolution and Human Behavior, 33,* 715–725. doi:10.1016/j.evolhumbehav.2012.05.007

DeScioli, P., & Kurzban, R. (2009). Mysteries of morality. *Cognition, 112,* 281–299. doi:10.1016/j.cognition.2009.05.008

DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin, 139,* 477–496. doi:10.1037/a0029065

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review, 90,* 980–994. doi:10.1257/aer.90.4.980

Fiske, A. P. (1992). The four elementary forms of sociality: A framework for a unified theory of social relations. *Psychological Review, 99,* 689–723. doi:10.1037/0033-295X.99.4.689

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96,* 1029–1046. doi:10.1037/a0015141

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85,* 197–216. doi:10.1037/0022-3514.85.2.197

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108,* 814–834. doi:10.1037/0033-295X.108.4.814

Haidt, J. (2007, May 18). The new synthesis in moral psychology. *Science, 316,* 998–1002. doi:10.1126/science.1137651

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20,* 98–116. doi:10.1007/s11211-007-0034-z

Halevy, N., Chou, E. Y., & Murnighan, J. K. (2012). Mind games: The mental representation of conflict. *Journal of Personality and Social Psychology, 102,* 132–148. doi:10.1037/a0025389

Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., & Moore, T. (2009). Adaptive rationality: An evolutionary perspective on cognitive bias. *Social Cognition, 27,* 733–763. doi:10.1521/soco.2009.27.5.733

Haselton, M. G., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review, 10,* 47–66. doi:10.1207/s15327957pspr1001_3

Hauert, C., De Monte, S., Hofbauer, J., & Sigmund, K. (2002, May 10). Volunteering as Red Queen mechanism for cooperation in public goods games. *Science, 296,* 1129–1132. doi:10.1126/science.1070582

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31,* 1369–1385. doi:10.1177/0146167205275613

Kelley, H. H., Holmes, J. G., Kerr, N. L., Reis, H. T., Rusbult, C. E., & Van Lange, P. A. M. (2003). *An atlas of interpersonal situations*. New York, NY: Cambridge University Press.

Kelley, H. H., & Thibaut, J. W. (1978). *Interpersonal relations: A theory of interdependence*. New York, NY: Wiley.

Kenrick, D. T., Li, N. P., & Butner, J. (2003). Dynamical evolutionary psychology: Individual decision rules and emergent social norms. *Psychological Review, 110,* 3–28. doi:10.1037/0033-295X.110.1.3

Kenrick, D. T., Neuberg, S. L., Griskevicius, V., Becker, D. V., & Schaller, M. (2010). Goal-driven cognition and functional behavior: The fundamental motives framework. *Current Directions in Psychological Science, 19,* 63–67. doi:10.1177/0963721409359281

Kiyonari, T., & Barclay, P. (2008). Cooperation in social dilemmas: Free riding may be thwarted by second-order reward rather than by punishment. *Journal of Personality and Social Psychology, 95,* 826–842. doi:10.1037/a0011381

Kiyonari, T., Tanida, S., & Yamagishi, T. (2000). Social exchange and reciprocity: Confusion or a heuristic? *Evolution and Human Behavior, 21,* 411–427. doi:10.1016/S1090-5138(00)00055-6

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Komorita, S. S., & Parks, C. D. (1995). Mixed-motive interaction. *Annual Review of Psychology, 46,* 183–207. doi:10.1146/annurev.ps.46.020195.001151

Krasnow, M. M., Cosmides, L., Pedersen, E. J., & Tooby, J. (2012). What are punishment and reputation for? *PLoS ONE, 7*(9), e45662. doi:10.1371/journal.pone.0045662

Krasnow, M. M., Delton, A. W., Tooby, J., & Cosmides, L. (2013). Meeting now suggests we will meet again: Implications for debates on the evolution of cooperation. *Scientific Reports, 3,* 1747. doi:10.1038/srep01747

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior, 28,* 75–84. doi:10.1016/j.evolhumbehav.2006.06.001

Kurzban, R., & Neuberg, S. L. (2005). Social exclusion, stigmatization, and discrimination: Managing group relationships. In D. M. Buss (Ed.), *Handbook of evolutionary psychology* (pp. 653–675). Hoboken, NJ: Wiley.

Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences, USA, 98,* 15387–15392. doi:10.1073/pnas.251541498

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the implicit association test: IV: What we know (so far) about the method. In B. Witttenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 59–102). New York, NY: Guilford Press.

Lieberman, D., & Linke, L. (2007). The effect of social category on third party punishment. *Evolutionary Psychology, 5,* 289–305.

Lieberman, D., Tooby, J., & Cosmides, L. (2003). Does morality have a biological basis? An empirical test of the factors governing moral sentiments relating to incest. *Proceedings of the Royal Society B: Biological Sciences, 270,* 819–826. doi:10.1098/rspb.2002.2290

Maner, J. K., Gailliot, M. T., & Miller, S. L. (2009). The implicit cognition of relationship maintenance: Inattention to attractive alternatives. *Journal of Experimental Social Psychology, 45,* 174–179. doi:10.1016/j.jesp.2008.08.002

Maner, J. K., Kenrick, D. T., Becker, D. V., Robertson, T. E., Hofer, B., Neuberg, S. L., . . . Schaller, M. (2005). Functional projection: How fundamental social motives can bias interpersonal perception. *Journal of Personality and Social Psychology, 88,* 63–78. doi:10.1037/0022-3514.88.1.63

Maner, J. K., & Mead, N. L. (2010). The essential tension between leadership and power: When leaders sacrifice group goals for the sake of self-interest. *Journal of Personality and Social Psychology, 99,* 482–497. doi:10.1037/a0018559

Maner, J. K., Miller, S. L., Rouby, D. A., & Gailliot, M. T. (2009). Intrasexual vigilance: The implicit cognition of romantic rivalry. *Journal of Personality and Social Psychology, 97,* 74–87. doi:10.1037/a0014055

Masclet, D., Noussair, C., Tucker, S., & Villeval, M. C. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review, 93,* 366–380. doi:10.1257/000282803321455359

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences, 36,* 1–15. doi:10.1017/S0140525X11002160

McDonald, M. M., Navarrete, C. D., & Van Vugt, M. (2012). Evolution and the psychology of intergroup conflict: The male warrior hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367,* 670–679. doi:10.1098/rstb.2011.0301

Miller, S. L., Maner, J. K., & Becker, D. V. (2010). Self-protective biases in group categorization: Threat cues shape the psychological boundary between "us" and "them". *Journal of Personality and Social Psychology, 99,* 62–77. doi:10.1037/a0018086

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics, 92,* 91–112. doi:10.1016/j.jpubeco.2007.04.008

Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups*. Cambridge, MA: Harvard University Press.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. doi:10.1017/CBO9780511807763

Ostrom, E. (2003). How types of goods and property rights jointly affect collective action. *Journal of Theoretical Politics, 15,* 239–270. doi:10.1177/0951692803015003002

Panchanathan, K., & Boyd, R. (2004, November 25). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature, 432,* 499–502. doi:10.1038/nature02978

Petersen, M. B., Sznycer, D., Sell, A., Tooby, J., & Cosmides, L. (in press). The ancestral logic of politics: Upper body strength regulates men's assertion of self-interest over income redistribution. *Psychological Science*.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40,* 879–891. doi:10.3758/BRM.40.3.879

Price, M. E. (2005). Punitive sentiment among the Shuar and in industrialized societies: Cross-cultural similarities. *Evolution and Human Behavior, 26,* 279–287. doi:10.1016/j.evolhumbehav.2004.08.009

Price, M. E., Cosmides, L., & Tooby, J. (2002). Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior, 23,* 203–231. doi:10.1016/S1090-5138(01)00093-9

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge, England: Cambridge University Press.

Rozin, P. (1999). The process of moralization. *Psychological Science, 10,* 218–221. doi:10.1111/1467-9280.00139

Sandler, T., & Tschirhart, J. (1997). Club theory: Thirty years later. *Public Choice, 93,* 335–355. doi:10.1023/A:1017952723093

Sasaki, T., & Uchida, S. (2013). The evolution of cooperation by social exclusion. *Proceedings of the Royal Society B: Biological Sciences, 280,* 20122498. doi:10.1098/rspb.2012-2498

Sell, A., Hone, L. S., & Pound, N. (2012). The importance of physical strength to human males. *Human Nature, 23,* 30–44. doi:10.1007/s12110-012-9131-2

Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences, USA, 106,* 15073–15078. doi:10.1073/pnas.0904312106

Shinada, M., & Yamagishi, T. (2007). Punishing free riders: Direct and indirect promotion of cooperation. *Evolution and Human Behavior, 28,* 330–339. doi:10.1016/j.evolhumbehav.2007.04.001

Shinada, M., Yamagishi, T., & Ohmura, Y. (2004). False friends are worse than bitter enemies: "Altruistic" punishment of in-group members. *Evolution and Human Behavior, 25,* 379–393. doi:10.1016/j.evolhumbehav.2004.08.001

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7,* 422–445. doi:10.1037/1082-989X.7.4.422

Sznycer, D., Takemura, K., Delton, A. W., Sato, K., Robertson, T. E., Cosmides, L., & Tooby, J. (2012). Cross-cultural differences and similarities in proneness to shame: An adaptationist and ecological approach. *Evolutionary Psychology, 10,* 352–370.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). New York, NY: Oxford University Press.

Tooby, J., & Cosmides, L. (2010). Groups in mind: The coalitional roots of war and morality. In H. Høgh-Olesen (Ed.), *Human morality and sociality: Evolutionary and comparative perspectives* (pp. 191–234). New York, NY: Palgrave Macmillan.

Tooby, J., Cosmides, L., & Price, M. E. (2006). Cognitive adaptations for n-person exchange: The evolutionary roots of organizational behavior. *Managerial and Decision Economics, 27,* 103–129. doi:10.1002/mde.1287

Tybur, J. M., Lieberman, D., & Griskevicius, V. (2009). Microbes, mating, and morality: Individual differences in three functional domains of disgust. *Journal of Personality and Social Psychology, 97,* 103–122. doi:10.1037/a0015474

Van Lange, P. A. M., Liebrand, W. B. G., Messick, D. M., & Wilke, H. A. M. (1992). Introduction and literature review. In W. B. G. Liebrand, D. M. Messick, & H. A. M. Wilke (Eds.), *Social dilemmas* (pp. 3–28). Oxford, England: Pergamon Press.

Van Vugt, M., Hogan, R., & Kaiser, R. B. (2008). Leadership, followership, and evolution: Some lessons from the past. *American Psychologist, 63,* 182–196. doi:10.1037/0003-066X.63.3.182

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51,* 110–116. doi:10.1037/0022-3514.51.1.110

Yamagishi, T., Hashimoto, H., & Schug, J. (2008). Preferences versus strategies as explanations for culture-specific behavior. *Psychological Science, 19,* 579–584. doi:10.1111/j.1467-9280.2008.02126.x

Yamagishi, T., Terai, S., Kiyonari, T., Mifune, N., & Kanazawa, S. (2007). The social exchange heuristic: Managing errors in social exchange. *Rationality and Society, 19,* 259–291. doi:10.1177/1043463107080449

Received April 13, 2012
Revision received January 29, 2013
Accepted May 14, 2013 ■