In Press, **Journal of Experimental Psychology: General**

The role of a "common is moral" heuristic in the stability and change of moral norms

Björn Lindström[1,2], Simon Jangard[1], Ida Selbing[1], Andreas Olsson[1]

1. Department of Clinical Neuroscience, Karolinska Institutet, Sweden
2. Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich, Switzerland

Corresponding author: Björn Lindström bjorn.lindstrom@ki.se / bjorn.r.lindstrom@gmail.com

Abstract: Moral norms are fundamental for virtually all social interactions, including cooperation. Moral norms develop and change, but the mechanisms underlying when, and how, such changes occur are not well-described by theories of moral psychology. We tested, and confirmed, the hypothesis that the commonness of an observed behavior consistently influences its moral status, which we refer to as the "common is moral" (CIM) heuristic. In nine experiments, we used an experimental model of dynamic social interaction that manipulated the commonness of altruistic and selfish behaviors to examine the change of peoples' moral judgments. We found that both altruistic and selfish behaviors were judged as more moral, and less deserving of punishment, when common than when rare, which could be explained by a classical formal model (Social Impact Theory) of behavioral conformity. Furthermore, judgments of common versus rare behaviors were faster, indicating that they were computationally more efficient. Finally, we used agent-based computer simulations to investigate the endogenous population dynamics predicted to emerge if individuals use the CIM heuristic, and found that the CIM heuristic is sufficient for producing two hallmarks of real moral norms; stability and sudden changes. Our results demonstrate that commonness shapes our moral psychology through mechanisms similar to behavioral conformity with wide implications for understanding the stability and change of moral norms.

Keywords: norms; morals; conformity; judgments; agent-based modeling

Moral judgments, informally codified in moral norms, are ubiquitous in all human societies (Bicchieri, 2005; Chudek & Henrich, 2011; Henrich et al., 2010). Broadly defined, moral norms[*] are shared notions about rightness and wrongness (Harms & Skyrms, 2008), and regulate social behavior in virtually all aspects of human life. Evolutionary theory provides compelling reasons why moral norms exist (DeScioli & Kurzban, 2013; Harms & Skyrms, 2008; Young, 2015), but in spite of endless treatment in philosophy, psychology and, more recently, the neurosciences, the basic proximate mechanisms behind the change of moral judgments and norms remain poorly understood (Bloom, 2010). Research in moral psychology typically construes moral convictions as stable once established (Graham et al., 2011; Kohlberg & Hersh, 1977), consistent over time (Colby, Kohlberg, & Gibbs, 1983; Graham et al., 2011), and as playing an integral role in individual self-identity (Aquino & Reed, 2002). However, norms are characterized by both long periods of stability and sudden changes (Young, 2015). Thus, the standard view on moral psychology can explain stability in moral norms by stability in moral convictions on the individual level, but lack mechanisms explaining how and why existing moral norms evolve and change (Bloom, 2010).

Here, we propose, and verify, the existence of a moral heuristic that helps to explain both the stability and change of moral norms; the tendency to infer the moral value of a social behavior from its relative frequency. We refer to this as the "common is moral" (CIM) heuristic. Heuristics are typically defined as the application of attribute substitution, where a target attribute is substituted by a more accessible attribute[†]. For example, people are prone to base judgments on the "availability heuristic", which substitutes the target attribute (for

[*] There are naturally many proposed definitions of moral norms. Some definitions focus on content (e.g., fairness; Turiel, 1983) while other's focus on expectancies, or lack thereof, about what other people do (Bicchieri, 2005). In most definitions, moral norms are closely related to, or synonymous with, injunctive norms (Cialdini, Reno, & Kallgren, 1990). Our account of the "common is moral" heuristic only assumes that people have moral values (e.g., notions about rightness/wrongness, which potentially are expressed in moral judgments) and care about what other's do (Bicchieri, 2005).

[†] Note that other definitions of heuristics do not include attribute substitution as a fundamental characteristic, but rather view the reduction of effort costs as the hallmark of a heuristic (Gigerenzer & Gaissmaier, 2011).

example city size) with a more assessable attribute (cities that come easily comes to mind) (Kahneman & Frederick, 2002). Here, we demonstrate that people substitute the target attribute “moral value” with the more assessable attribute “behavioral frequency” when making judgments about others´ social behavior, with the result that behaviors are judged as more moral, and less deserving of punishment, when common than when rare (for general discussions about moral heuristics, see Gigerenzer, 2008; Sinnott-Armstrong, Young, & Cushman, 2010; Sunstein, 2005. For complementary mechanistic theories of other aspects of moral judgments, see Crockett, 2013, 2016; Cushman, 2013). The CIM heuristic thereby provides a psychological mechanism for how descriptive norms, which describe what behaviors are common, can be transformed into moral (or injunctive) norms, which prescribe the correct way to behave in a certain context (Bicchieri, 2005; Cialdini & Goldstein, 2004). These types of norms are logically distinct, yet people often fail to differentiate between them, as famously described by David Hume as a logical fallacy (stating that it is wrong to derive “ought” from “is”) (Hume, 2003).

A moral norm central to well-functioning societies is that each individual should contribute to the public good (Ostrom, 2000). The violations of this norm is often judged as *unfair* or *wrong* in a moral sense, which motivates people to punish norm violators even when they are not directly suffering from the consequences of the violation themselves (Buckholtz & Marois, 2012; Fehr & Fischbacher, 2004b). These processes are thought to be critical for sustaining cooperation (Boyd, Gintis, Bowles, & Richerson, 2003). However, even this central moral norm can change with time: for example[‡], according to the most recent *World Value Survey*, tax evasion, a form of free riding on the public good, was in South Africa judged as 48 % more justifiable in 2014 than 1981. During the same time period, tax evasion became 35 % *less* justifiable in Mexico, while another morally relevant behavior, riding without paying on a

[‡] Explaining these particular changes is however outside the scope of the present study.

public transport, became 25 % more justifiable during the same time period. The moral statues of behaviors not directly related to public goods are also subject to change. For example, smoking has in many countries changed from being viewed as an individual preference to a violation of a moral norm (Rozin, 1999), whereas the moral value of other behaviors, such as pre-marital sex, has changed or disappeared. The influential social intuitionist account of moral psychology posits that moral intuitions, which are more or less automatic affective processes (Haidt, 2001, 2007), serve as the basis for moral judgments, and can be shaped by social factors, such as culture and persuasion (Haidt, 2001, 2007). Yet, this account does not specify the details as to when and how moral judgments are shaped by the social environment. Similarly, existing work on moral heuristics do not treat the dynamics of moral norms (Gigerenzer, 2008, 2010; Sunstein, 2005). In sum, limited theory and data exist to explain how moral judgments are influenced by the social environment, and why moral norms change (Bloom, 2010).

We hypothesized that changes of moral judgments and norms can originate from the CIM heuristic, which posits that people use frequency information as a basis for moral judgments. Furthermore, we hypothesized that the CIM heuristic rests on general social influence mechanisms (Cialdini & Goldstein, 2004; Latané, 1981). This conjecture was based on the observation made outside the field of moral psychology that people spontaneously (i.e., without any instruction to do so) extract frequency information from others' behavior in a wide variety of situations, which in turn affects their own behavior (Cialdini & Goldstein, 2004; Latané, 1981; MacCoun, 2012). Broadly, social influences arise when an individual, or a group of individuals, changes its behavior due to the presence of individuals with different behaviors (Denrell, 2008; Latané, 1981; MacCoun, 2012), and are considered fundamental for the transmission of ideas (i.e. culture) between individuals and across generations (Boyd, Richerson, & Henrich, 2011; Laland & Rendell, 2013; Richerson & Boyd, 2005). One example of social influence is conformity, where a minority adopts its behavior to that of the majority

(Asch, 1956). Importantly, social influence does not need to imply that the individual deliberatively change attitude or behavior. Instead, social influence can exert an incidental influence on ongoing behavior, as illustrated by the propensity to stop and look into the sky on a busy street if several other people already are doing so (Latané, 1981). Similarly, many species of non-human animals are prone to copy what conspecifics do (Hoppitt & Laland, 2013). Moreover, classic field studies of social influence have demonstrated that peoples' behaviors are affected by both moral norms and information about commonness (descriptive norms) in morally relevant situations involving public goods (e.g., household energy consumption) (Cialdini et al., 2006; Schultz, Nolan, Cialdini, Goldstein, & Griskevicius, 2007). These studies did however not investigate how social influence mechanisms affect moral judgments, or if people spontaneously infer moral value from what is common as posited by the CIM heuristic.

Several previous studies provide evidence consistent with this social influence account of moral judgments. Most importantly, a recent study showed that people automatically associate commonness and moral value (i.e., use descriptive norms as a basis for moral judgments) (Eriksson et al., 2014) (see also McGraw, 1985). For example, commonness and morality were mixed up in memory recall (Eriksson et al., 2014), which was interpreted as evidence for automatic associations between these concepts. In this study, the participants were provided explicit descriptions (e.g., "80% of the population does X") about common or moral behaviors in highly abstract and hypothetically-phrased vignettes (Eriksson et al., 2014). The authors suggested that the automatic association between commonness and moral values might sustain existing moral norms. In the same vein, it´s been shown that people prefer things that already are ("existence bias"), which might sustain the status quo (Eidelman, Crandall, & Pattershall, 2009). Another recent study demonstrated that a tendency to infer value from commonness is related to a heuristic reasoning process when judging vignette-based statements about how things should be (e.g., "should Americans eat pizza") (Tworek & Cimpian, 2016).

Thus, these previous studies illuminate the cognitions underlying an association between commonness and moral values. However, hypothetical moral problems and vignettes provides only limited insight into moral heuristics in the their dynamic social context (Gigerenzer, 2008). Accordingly, these previous studies, which all exclusively used hypothetical vignettes, do not tell us if people spontaneously use the CIM heuristic in the type of dynamical situations that characterize real world social interactions (e.g., real-time observation of others' behaviors), how the CIM heuristic relates to social influences mechanisms outside of the moral domain, if the CIM heuristic affects moral behavior (in addition to moral judgments), or, crucially, how the CIM heuristic might contribute to the stability and change of moral norms.

The current study

We characterized the mechanisms underlying the CIM heuristic, and its implications for the stability and change of moral norms, using behavioral and computational methods. First (see *"The basic function and computations of the CIM heuristic"* below*),* we experimentally manipulated the commonness of observed behaviors to test if people use the CIM heuristic when making moral judgments. We conducted nine independent experiments (one in the laboratory and eight online), in which participants served as third-party observers of a sequence of one-shot Public Goods games (PGG). The participants judged the morality, and appropriate punishment, of altruistic (others investing their own resource into a public good, thereby benefitting the group) and selfish (others keeping everything for themselves) behaviors while the relative frequencies of the altruistic and the selfish behaviors were manipulated (see Figure 1). We used the PGG as a tool for eliciting moral considerations related to cooperation, because of the importance such considerations have for cooperation between non-kin (Bowles & Gintis, 2011; Boyd, Gintis, & Bowles, 2010; Haidt, 2007). Moreover, we characterized the basic computations of the CIM heuristic, and found that it uses the relative frequency of social behaviors as the basis for moral judgments (Exp. 1-7). This was possible since our experimental model, in contrast to previous studies, allowed manipulating the objective distribution of observed behaviors (e.g., instead of using vignette descriptions of what is common). We used a classic formal model of social influence (Social Impact Theory [[SIT]; Latané, 1981) to show, for the first time, that the relative frequency of social behaviors affect moral judgments through similar mechanisms as behavioral conformity. Second (see "*Judgments of common behaviors are easier than judgments of rare behaviors"* below.), we studied the psychological processes underlying the CIM heuristic using reaction time analysis, and found that judgments of common behavior were faster, and therefore likely to be more computationally efficient, than of rare behaviors. Third (see "*The CIM heuristic can explain the stability and change of group-level moral norms"* below), we investigated the potential of the CIM heuristic for explaining endogenous (i.e., without manipulating commonness) changes in moral norms, by analyzing a simple agent-based simulation model of population dynamics emerging when many individual agents use the CIM heuristic. This model, together with an empirical conformation that moral judgments predicts moral behavior (Exp. 8) and that commonness affects costly punishment (Exp. 9), showed that the CIM heuristic in principle is sufficient for explaining both the stability and change of moral norms on the population level. In summary, our contribution is both theoretical and methodological. Theoretically, we provide a novel, plausible and parsimonious answer to the important theoretical question of how morals change (Bloom, 2010) by documenting how the CIM heuristic influences moral judgments and moral behavior on the individual level, and show how this influence can help explain the dynamics of moral norms on the population level. Furthermore, we add new knowledge by demonstrating that the social influence on moral judgments rest on similar mechanisms as those underpinning behavioral conformity. Methodologically, we go beyond previous vignette-based studies of social influences on moral judgments by establishing an experimental model of social interaction, which in turn allowed us to, for the first time, use quantitative modeling to establish the link between moral judgments and behavioral conformity. Unlike previous work on social influences on moral judgments, we also establish the relationship between moral judgments and moral behavior in the experimental model (Exp. 8-9). Finally, we extend previous work by demonstrating how the empirically established CIM heuristic, a micro-level psychological mechanism, can be related to macro-level population dynamics using agent-based simulation (Smith & Conrey, 2007).

Method

Participants

The participants in Experiment 1 were recruited on the campus of Karolinska Institutet, Sweden, and were reimbursed with two movie vouchers (approximately worth 25 USD). The participants for Experiment 2-6 were recruited on Amazon´s Mechanical Turk, and were reimbursed with 1 USD. In Exp. 7, participants received an additional bonus based on their earnings in one random PGG interaction. In Exp. 8, participants received an additional 2 USD endowment which they could use for costly punishment. Participants who had taken part in any of the previous experiments, or failed control questions about the one-shot nature of the experiment were removed. All procedures were approved by the local ethics committee at Karolinska Institutet.

In total, the data set involved 473 (227 women, 4 unknown) participants. Experiment 1 included 32 participants (23 women. Mean age = 27.16, SD = 6.0); Experiment 2 included 50 participants, 24 women. Mean age = 36.6, SD=12.46); Experiment 3 included 31 participants (17 women, 1 unknown. Mean age = 36.37, SD = 12.1); Experiment 4 included 54 participants (23 women, 1 unknown. Mean age = 36.67, SD= 13.46); Experiment 5 included 75 participants (46 women, 1 unknown. Mean age = 33.7, SD = 10.16); Experiment 6 included 37 participants (24 women. Mean age = 38.6, SD = 10.85); Experiment 7 included 41 participants (16 women, 1 unknown. Mean age = 42.31, SD = 13.75); Experiment 8 included 55 participants (22 women. Mean age = 36.62, SD = 10.65); Experiment 9 included 98 participants (22 women. Mean age = 36.62, SD = 10.65). The participant numbers above refer to the final samples (see below for exclusion criteria).

In experiment 1 (lab), the experimenter verbally quizzed the participants' understanding of the task prior to the experiment. In experiment 2-4 & 6-7 (online), the participants answered two comprehension questions, which served as exclusion criteria, after the experiment ("To what extent did the participants have knowledge about each another?" and "Did the participants reappear in the experiment?", thus targeting the crucial aspect of anonymity and the one-shot nature of the Public Goods Games [PGG]). Failure to respond correctly to either of these questions led to the complete exclusion of this participant from analysis. In experiment 5 & 8-9, the participants answered control questions prior to participation, and where corrected and provided additional information if incorrect. For this reason, all participants in these experiments were included in the final samples. In experiment 9 (costly punishment), the participants in addition answered questions, prior to the experiment, to test their understanding of the punishment system: "If you decide to transfer 4 deduction points to a selected participant, how much will that reduce your own bonus?" and "If you decide to transfer 2 deduction points to a selected participant, how much will that reduce the bonus of the participant?". Based on these exclusion criteria, in Exp. 2, 10 would-be participants were excluded, in Exp. 3, 9 were excluded, in Exp. 4 16 were excluded, in Exp. 6, 40 were excluded (the unusually high number was due to attempted repeated participations), and in Exp. 7, 19 were excluded.

On Mechanical Turk, we required that would-be participants had at least a 95% approval rating, which is a standard criterion for providing good data quality (Peer, Vosgerau, & Acquisti, 2014). A number of previous studies have demonstrated that Mechanical Turk generates data of comparable quality to the lab, both in experimental psychology (Crump, McDonnell, Gureckis, Romero, & Morris, 2013) and behavioral economics (Amir, Rand, Gal, Johannesson, & List, 2012). Due to an experimenter error, Exp. 2-3 & 6 were conducted with Mechanical Turk's "international" option, while the remaining Mechanical Turk experiments only involved US citizens. Statistical control analyses showed no differences in the CIM effect between the "international" and US samples (moral judgments: $\chi2$ (1) = 2.03, p = .15. Punishment judgments: $\chi2$ (1) = 0.6, p = .44).

We used number of participants in the Exp. 1 (lab experiment) as the stopping rule for the minimum participant number (after exclusion of participants who failed the comprehension questions) in the subsequent Mechanical Turk experiments. The sample size for Exp. 1 was based on a priori power calculation using G*Power, specifying ~ 80 % power given a medium effect size in a repeated measures design. For experiments 5 and 9, we doubled this minimum sample size under the assumption that the effect sizes would be smaller than in the first set of experiments.

Experimental Design

The primary experimental design was identical across the experiments (see Figure 1). Participants were informed that these behaviors were acquired from an earlier experiment, in which the alleged previous participants could chose to keep (selfish) or invest (altruistic) a fixed sum of money in a group of four (exp.1-4) or eight (exp. 6-7) anonymous participants (i.e., a one-shot Public Goods game). The complete instructions for the basic experiment can be found in the Supplementary Information (SI): Methods. Experiment 5 had the same design, but included no instructions about the payoff consequences of the behaviors (referred to as A and B). The complete instructions for Experiment 5 can be found in the SI: Methods.

The task had a 2 (Common behavior: Altruistic/Selfish) x 2 (Rated behavior: Altruistic/Selfish) design (see Figure 1), with the order of the Common behavior factor counterbalanced across subjects. Each behavior was common for a block of 20 trial. Within this block, the common behavior had an occurrence rate of 80%, while the exact distribution varied probabilistically for each trial.

The fictional PGG players were represented by unique 2-letter "initials" on each trial, which were presented in black font on white background. On each trial, one target behavior, either the Altruistic behavior (represented by the text "Everyone") or Selfish behavior (represented by the text "Myself") was randomly selected for moral and punishment ratings, see Figure 1. These ratings were conducted using nine-point scale ranging from 1 (corresponding "morally wrong" for the moral judgments, and to "no economic reduction" for punishment ratings) to 9 (corresponding to "morally right" for moral judgments, and "total economic reduction" for punishment ratings). The location of the target stimuli varied randomly between the four corners of the monitor.

Experiment 9 (costly punishment) had the same design as exp. 1-6, but the participants only made punishment decisions (no moral judgments) about the focal PGG player. The participants could "invest" in deduction points, which were subtracted from an endowment of 2 US $. Each deduction point was multiplied with 5, and the corresponding sum was (allegedly, since the PGG players were simulated) deducted from the profit of the focal PGG player. See SI: Methods for additional information.

Statistical analysis. The statistical programming language R was used for all analyses. The lme4 package (Bates & Sarkar, 2007) was used for estimating linear and logistic mixed models (GLMM). Reported main- and interaction effects were calculated using Type II analysis of deviance based on the Wald test. The degrees of freedom (DF) underlying the Wald test are based on the difference in the number of parameters between the models being compared (e.g. whether the interaction term is included or not). P-values for individual parameter estimates were calculated based on Satterthwaite's approximation for denominator DF. All LLM´s included by-participant random intercepts and by-participant random slopes for all fixed effects.

In exp. 1, and 4-7, and the moral judgment phase of exp. 8, we defined trials with reaction times longer than the 95th percentile of the total distribution for each experiment as outliers, in an attempt to eliminate non-attentive responses (especially on Mechanical Turk). All results are qualitatively unchanged if these responses are included. Due to technical issues, reaction time data was not collected for Exp. 2-3 on Mechanical Turk.

Computational modeling. Sources were defined as the number of individuals exhibiting the behavior the participant was to judge on a given trial (i.e., the same definition as degree of consensus), and targets as N (i.e., 4 or 8) - the number of sources. The parameters were estimated by minimizing the mean square error using optimization methods. See SI: Methods for additional information.

Agent-based simulation. See the SI: ABM for model description and additional analyses.

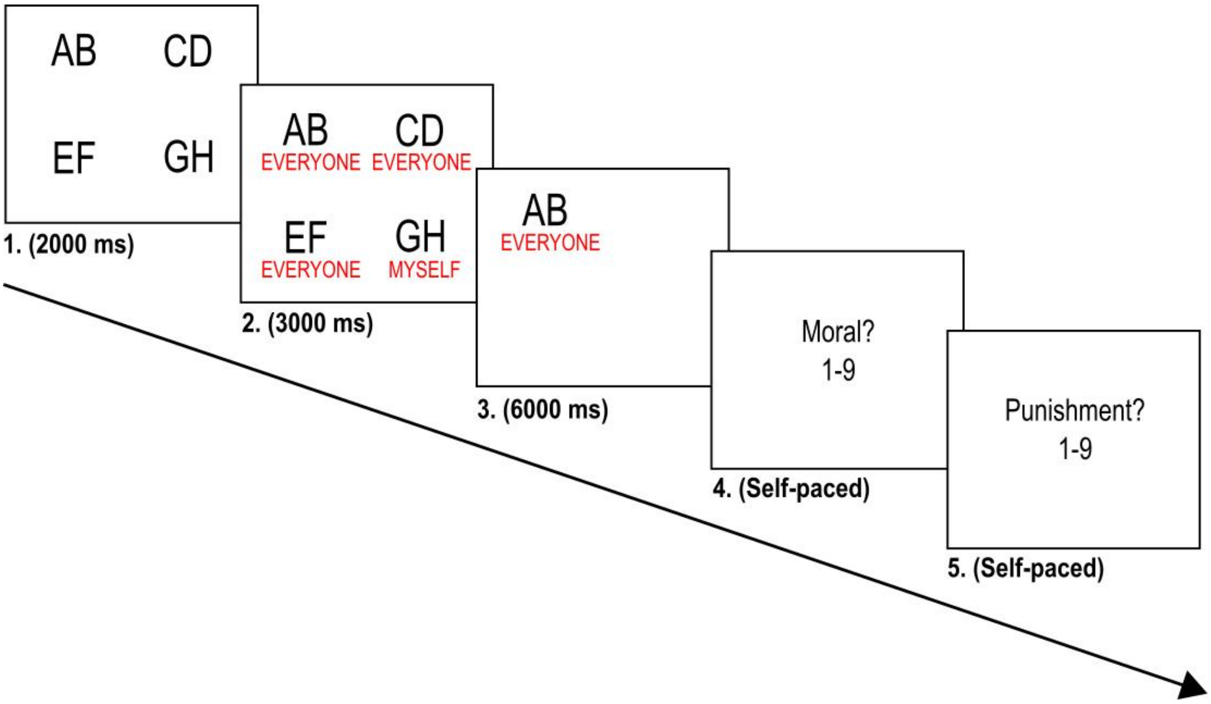

Figure 1. Overview of the primary experimental task (Exp. 1-4 & 6-7). Participants were exposed to 40 one-shot Public Goods game (PGG) interactions as third-party observers. In 20 of these, the selfish behavior (not investing in the common pool) was common (80%) and the altruistic (investing in the common pool) rare (20%). In the other 20, the altruistic behavior was common and the selfish behavior rare (order counterbalanced across participants. The order of the conditions did not affect the results. See Supplementary Information). The participants judged the moral status and appropriate punishment for the behavior of a randomly selected player from each interaction. The figure displays a trial where the altruistic behavior was common (75%). The commonness of each behavior on each trial varied probabilistically (subject to the 80/20 % constraints). See Methods for further information and verbatim phrasing of the questions. The arrow indicates the direction of time.

Results

1. The basic function and computations of the CIM heuristic

1.1 Common behaviors become more moral and less deserving of punishment

We tested our key hypothesis that commonness affects moral and punishment judgments, as described by the CIM heuristic, by analyzing the interaction between the type of behavior the participants rated (altruistic or selfish) and how common it was. In our experimental design, commonness was exogenously manipulated both on the level of blocks (where the altruistic/selfish behavior was common in one block and rare in the other, counter-balanced), and on the level of individual trials (due to probabilistic variation) (see Figure 1 and Methods). In the first section, we focus on the block-level effect of commonness, and then address trial-level variation in section 1.2.

The first experiment, conducted in the lab in Stockholm, followed by three direct replications on Mechanical Turk (Exp. 1-4, combined n = 167. The results of these experiments did not differ so we pooled the data. See the SI: Results), showed that the altruistic action was judged to be more moral on average than the selfish behavior (linear random effects regression main effect of type of behavior (χ^2 (1) = 522.21, p ≪ .001). This is important as it confirms that our participants shared the pre-existing moral norm that cooperation was considered morally good (Cubitt et al., 2011; Haidt, 2001, 2007). Confirming our hypothesis, this effect was qualified by a significant interaction with how common the behavior was (χ^2 (1) = 27.7, p ≪ .001, see Figure 2, and SI: Results and Table S1-S2 for individual experiments). Directly contrasting the moral judgments for the altruistic behavior showed that it was judged to be less moral when rare than common (contrast estimate = 0.21, SE = 0.07, z = -3.14, p = .003, corresponding to a *Cohen´s D* effect size of ~.25), while the selfish behavior was judged as more moral when common than rare (contrast estimate = 0.55, SE = 0.1, z = 4.74, p < .001, *Cohen´s D* effect size ~.37). These results show that while the pre-existing moral norm about cooperation was the strongest determinant of moral judgments, these judgments were robustly moderated by commonness.

Previous research have shown that some people are conditional cooperators, who favor cooperation over defection provided others do so as well (Fehr & Fischbacher, 2004a). Such a conditional cooperation norm could explain why the selfish behavior was judged as less immoral the more common it was, but appears insufficient for explaining why the moral status of the altruistic behavior was reduced when rare relative to common. To fully rule out distributional explanations, and to test how the CIM heuristic affects judgments of novel social behaviors devoid of moral priors, we conducted a control experiment (Exp. 5). Seventy-five new participants made judgments of two behaviors (referred to as ”A” and ”B”), with exactly the same frequencies as in Exp. 1-4, but in the absence of information about the payoff structure (but knowing that the observed behaviors carried relevance, and had monetary consequences for the observed players, see SI: Methods). If conditional cooperation or equality norms, rather than the CIM heuristic, underlie the above results, this experiment should show no effect of commonness. This was not the case; there was a strong interaction between behavior and commonness (χ^2 (1) = 15.3, p ≪ .001. See Figure S3 and SI: Results for additional analyses).

Together, these results confirmed our prediction that behavioral frequencies influence moral values, regardless of if they judged altruistic, selfish, or arbitrary behaviors, as described by the CIM heuristic. As such, they show the potency of commonness to moderate judgments of both behaviors subject to pre-existing moral norms and of novel social behaviors.

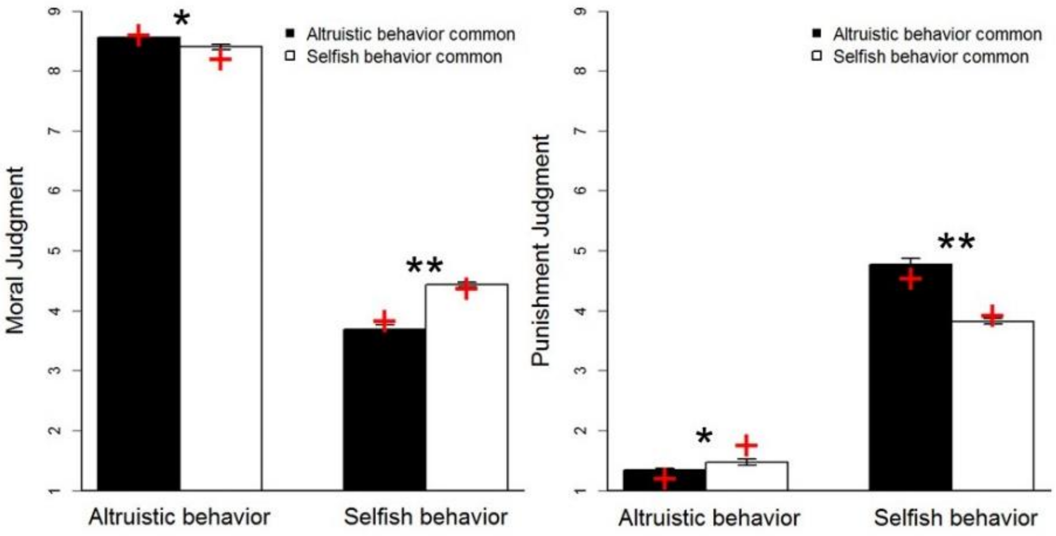

Figure 2. Moral and Punishment judgments were affected by commonness. Average moral (left) and punishment (right) judgments**.** For moral judgments, the scale was anchored with “Morally wrong” (1) and “Morally right” (9). For punishment judgments the scale was anchored with “No punishment” (1) and “Maximum punishment” (9). The influence of commonness on these judgments could be explained by a classical computational model of social influence (Social Impact theory; see *A formal model of social influence can explain the effect of commonness on moral judgments*). The red crosses denote the condition averages from the model, derived from fitting the model to each participant ratings. Error bars denote standard error of the mean. The figure includes data from experiment 1-4, n = 167. Stars indicate significant contrasts between adjacent bars. * = p < .01, ** = p < .0001

We also asked the participants to judge how much they would like to hypothetically reduce the monetary profit (i.e., punish) of the focal PGG player. These punishment judgments consistently matched the moral judgments (data pooled for Exp.1-4. See Figure S2 in the SI for the individual experiments): the selfish behavior was on average punished more strongly than the altruistic behavior (main effect: χ^2 (1) = 154.83, p ≪ .001), and mirroring the moral judgments, this difference was modulated by how common the behaviors were (χ^2 (1) = 37.37, p ≪ .001, see Figure 2). The selfish behavior was judged to deserve less punishment when common than when rare (contrast estimate = -0.76, SE = 0.13, z = - 5.77. p ≪ .001, *Cohen´s D* effect size ~.45), and the altruistic to deserve more punishment when rare than common (contrast estimate = 0.18, SE = 0.07, z = 2.5, p = 0.01, *Cohen´s D* effect size ~.19). The same effect of commonness on punishment judgments was also observed for arbitrary behaviors in Exp. 5 (χ^2(1) = 7.7, p = .005, Figure S3). These findings relate moral judgments to the phenomenon of third-party punishment, by demonstrating that moral judgments and perceived justness of punishment as seen by a third party goes hand in hand (Fehr & Fischbacher, 2004b), and most importantly, that both types of judgments are shaped by the CIM heuristic. In section 3.2, we confirm that commonness not only affects punishment judgments, but also costly punishment.

In the SI: Results, we in addition present control analyses which show that the CIM effect on both moral and punishment judgment is evident even if only the five first trials are included in the analysis (thus turning the analysis into a between subjects comparison). The purpose of these analyses was to verify that the observed CIM effect was not caused by repeated judgments.

1.2 Commonness affects moral judgments through the relative frequency of the behavior

Our first set of experiments showed that people behave as predicted by the CIM heuristic but left unresolved whether the *absolute* or *relative* (to the group size) number of people exhibiting a behavior is the basis for the social influence on moral judgments. The answer to this question is important for predicting when peoples' moral judgments will be shaped by others' behavior. If the absolute number is the driving factor for the CIM heuristic (e.g., counting the number of people exhibiting a particular behavior), people will be more influenced in large than small group settings. Alternatively, if the relative frequency, or degree of consensus, is most important (e.g., comparing the number of people exhibiting a particular behavior to the number of people displaying an alternative behavior), people will be equally affected in small group interactions. Thus, identifying the basic input to the CIM heuristic is necessary to characterize its algorithmic nature.

To address this question, we ran two additional experiments (Exp. 6-7, combined n = 76) where we doubled (from 4 to 8) the number of players in the PGG the participants observed, while keeping degree of consensus identical to the first set of experiments. These experiments (the data were pooled, see Figure S1-S2) again replicated the effect of commonness on judgments of moral (χ^2 (1) = 15.49, p < .001) and punishment (χ^2 (1) = 20.41, p < .001) ratings. However, the absolute number of PGG players affected neither moral nor punishment judgments (tested by adding absolute number, 4 vs 8, as a between-participants predictor interacting with the experimental factors to data pooled across experiments 1-4 and 6-7: p = .73 and p = .48. See Figure S1-S2.).

Next, we analyzed how the relative frequency (which we refer to as degree of consensus) affected judgments by calculating the proportion of PGG players on each trial exhibiting the behavior the participant was to judge (the trial-by-trial proportions of each behavior were probabilistic in all experiments, see Methods and Figure 1). The degree of consensus was a highly significant predictor of moral (β = 0.71, SE = 0.14, t (240.48) = 5.1, p ≪ .001) and punishment judgments (β = -0.94, SE = 0.14, t (246.47) = -6.8, p ≪ .001). These effects did not differ for groups composed of four or eight players (see SI: Results). In summary, our analyses demonstrated that the degree of consensus, or relative frequency, drives the CIM heuristic.

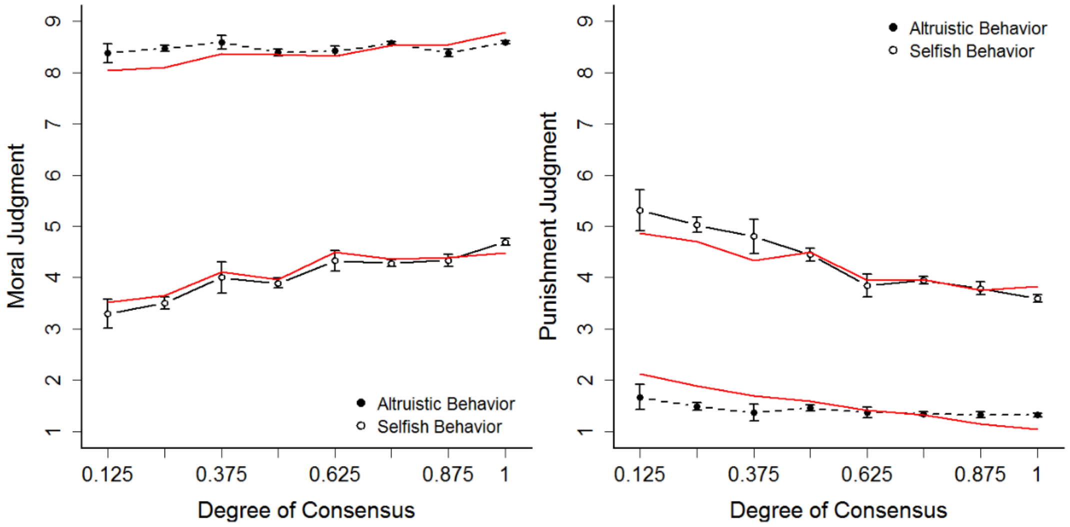

Figure 3. The degree of consensus shaped moral and punishment judgments. Moral (left) and punishment (right) judgments were shaped by the trial-by-trial degree of consensus. For moral judgments, the scale was

anchored with “Morally wrong” (1) and “Morally right” (9). For punishment judgments the scale was anchored with “No punishment” (1) and “Maximum punishment” (9). The influence of degree of consensus could be explained by a classical formal model of social influence (Social Impact theory; see *A formal model of social influence can explain the effect of commonness on moral judgments*). The red lines denote the condition averages from the model, derived from fitting the model to each participant ratings. Error bars denote standard error of the mean. The figure includes data from experiment 1-4 and 6-7, n = 245.

1.3 A formal model of social influence can explain the effect of commonness on moral judgments

The robust effect of consensus on moral and punishment judgments indicates that the computational mechanisms underlying the CIM heuristic, as predicted, are similar to social influences on behavior outside of the domain of moral judgments, such as behavioral conformity (Latané, 1981; MacCoun, 2012). Social influence research has identified a number of strong regularities in the influence of a majority (termed *sources*) on the behavior of a minority (termed *targets*), leading target individuals to change attitude or behavior. A classical formal model of social influence is *Social Impact Theory* (SIT) (Latané, 1981), which accounts for social influences across an extraordinary wide array of phenomena, for example, social loafing, helping behaviors, and diffusion of responsibility (Latané, 1981). The SIT can be described as a psychosocial law (in references to psychophysical laws), which specifies the impact of the sources on the targets as “social force field”, whose strength depends on the quantity of the sources. In practice, SIT is expressed as a power law that gives a precise quantitative account of how these factors together affect behavior.

We fitted a version of the SIT model to each participant’s trial-by-trial moral judgments (see SI: Method) to directly test the hypothesis that the CIM heuristic is based on similar computational mechanisms as social influences outside the moral domain (Latané, 1981; MacCoun, 2012). To our knowledge, this is the first application of SIT to moral judgments. The model had the form:

$$Moral\,Judgement(t) = k + mA + \ s * \ (Sources(t)/\,(Sources(t) + Targets(t))^{x} \quad [1]$$

where *k* is a constant (i.e., intercept), *mA* the baseline difference in the moral judgment of the altruistic behavior (i.e., simple effect. Set to 0 if a selfish behavior was judged), *s* determines the strength of the social influence ($0 \leq s \leq 10$), and *x* ($0 \leq x \leq 1$) determines the shape of the influence function, and *t* is the trial index (see SI: Methods for additional details).

The SIT model provided a good account of the moral judgment data, with a mean R^2 of .98. In fact, 236 of 245 participants had an individual R^2 of at least .9 (see Figures 2-3). Removing the two social influence parameters (*s* and *x*) led to a drop in mean R^2 to .72. Although this reduction in explained variance seems relatively modest, removal of these two parameters led the model to completely miss the empirical interaction (see Figure 2) between behavior and commonness (mixed effects regression with the model predicted moral judgments as dependent variable, and the experimental factors as fixed effects: χ^2 (1) = 0.032, p = .858).

We fit the same SIT model to the trial-by-trial punishment judgments, where the mean R^2 was .97 (Figure 2-3). Out of 245 participants 229 had a $R^2 > .9$. Removing the two social influence parameters (*s* and *x*) led to a drop in mean R^2 to .85, which again led to a failure to predict the empirical interaction (Figure 2) (mixed effects regression with the model predicted punishment judgments as dependent variable, and the experimental factors as fixed effects: χ^2 (1) = 0.12, p = .726).

We validated the robustness of the SIT model by using the average estimated parameters to fit the entire data set of 9472 judgments (separately for moral and punishment judgments), thereby reducing 245 (participants)*4 free parameters to just 4. For moral judgments, R^2 was then .92, and for punishment judgments, R^2 was .91. The same results are obtained if the 4 parameters are directly fitted to the entire data set (i.e., not distinguishing between different subjects).

In summary, the good fit of the SIT model strongly suggests that moral and punishment judgments share common underlying computational mechanisms with other social influence situations (Claidière & Whiten, 2012; Latané, 1981; MacCoun, 2012). In section 3.3, we use the SIT model to provide a mechanistic implementation of the CIM heuristic in our agent-based simulation model of the stability and change of population-level moral norms.

2. Judgments of common behaviors are easier than judgments of rare behaviors.

What psychological process mechanisms underpin the CIM heuristic? A standard test of the efficiency of a computational process is its speed (Gershman, Horvitz, & Tenenbaum, 2015). If people directly use commonness as a proxy for moral value, as suggested by CIM, judgments of common behaviors should be faster than of rare behaviors, and the higher the degree of consensus, the faster (because it is easier) should the judgment be. Put differently, the more evidence the social environment provides for substituting the target attribute moral value with behavioral frequency (i.e., applying the CIM heuristic), the faster should the judgment be.

Both moral (Common * Behavior: $\chi2$ (1) = 8.3, p = .004) and punishment (Common * Behavior: $\chi2$ (1) = 50.86, p < .001) judgments were faster when the behavior was common than rare (see Figure 4 and SI: Results). In testing this, we controlled statistically for the participants' actual moral preferences by adding the (grand mean-centered) judgments as covariates to the regression (mean-centering for individual participants give very similar results). Controlling for preferences is crucial for drawing conclusions about computational efficiency based on reaction times, because reaction times otherwise mainly are driven by the difference between decision options (Krajbich, Bartling, Hare, & Fehr, 2015). Thus, these results show that judgments of common behaviors were on average easier and more efficient than judgments of rare behaviors. This suggests that people do directly use commonness as a proxy for moral value (Sinnott-Armstrong et al., 2010). The influence of commonness on reaction times (RTs) appear analog to how differences in value or difficulty affect reaction times according to sequential sampling (e.g., drift diffusion) models of choice and RTs (Krajbich, Hare, Bartling, Morishima, & Fehr, 2015).

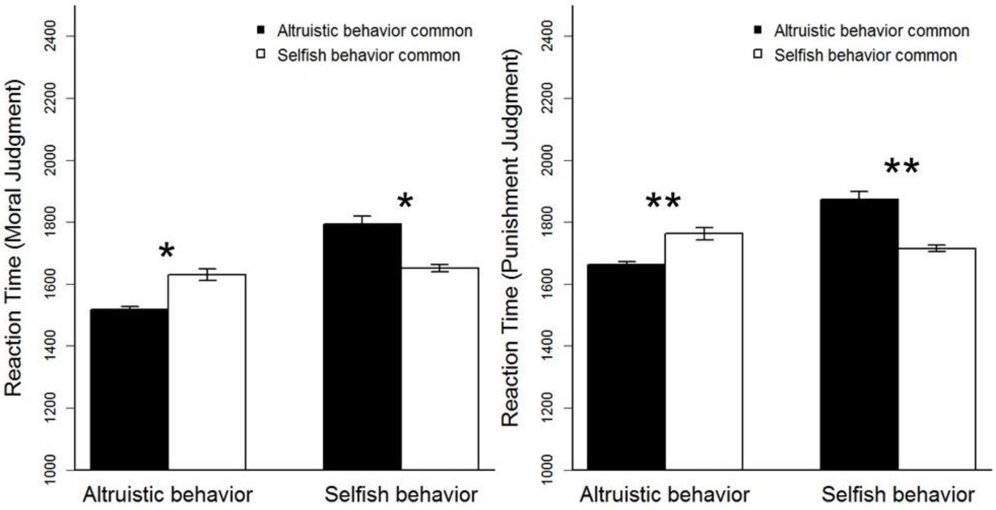

Figure 4. Judgments of common behaviors are faster than of rare behaviors. The means are adjusted for the judgment covariate (see section 2). Error bars denote standard error of the mean. The figure includes data from Exp. 1, 4 & 6-7 (n = 164). Stars indicate significant contrasts between adjacent bars. * = p < .05, ** = p < .01

3. The CIM heuristic can explain the stability and change of group-level moral norms.

Is the use of the CIM heuristic sufficient for changing moral norms on the group-level? To provide a proof of principle answer to this question, we constructed a simple agent-based model, which allowed us to explore the theoretical population level effects of CIM (section 3.2, see also SI: ABM) (Smith & Conrey, 2007). To provide additional realism to this model, we first investigated the relationship between CIM and two behavioral mechanisms that typically are seen as fundamental for moral norms: internalization (section 3.1), and moralistic punishment (section 3.2). We then incorporated these mechanisms into the simulation model of moral norms (section 3.3).

3.1. Moral judgments predict behavior

The human tendency to internalize moral norms, by acting in accordance with the norm in the absence of external influences, is an important mechanism for large-scale human cooperation (Bowles & Gintis, 2011). Nonetheless, the link between moral judgments and moral behavior is contested. Some laboratory experiments report a high degree of moral hypocrisy (Batson & Thompson, 2001), while others show a clear relationship between judgments and behaviors (Reynolds & Ceranic, 2007). We investigated the relationship between moral values and behavior in a context where reputational concerns and punishment were ruled out, to verify that people acted in concordance with internalized moral norms in our experimental model. This is important for verifying that the judgments in the first set of experiments carried real moral relevance.

As in the previous experiments, participants (n = 55) first made moral and punishment judgments about selfish and altruistic PGG behaviors (rating phase), although in Exp.8, these behaviors were equally frequent. In the subsequent game phase, the same participants took part in ten incentivized one-shot PGG games (against computer controlled "players", who played each action with p = .5). Thus, this experiment tested if internalized moral values (as expressed in judgments) predicted subsequent moral behavior, which is an assumption underlying our agent-based simulation (below, section 3.3). The average moral and punishment judgments of the selfish action in the rating phase were strong predictors of the subsequent probability to choose the selfish action in the game phase (random effects logistic regression of moral judgment: β = 0.86, SE = 0.38, z = 2.28, p = .02. Punishment: β = -0.83, SE= 0.31, z = -2.66, p = .007, Figure 5 A-B). These results confirm that moral and punishment judgments when serving as a third-party and subsequent behavior in the PGG are closely related, and indicates that the experimental model used in Exp. 1-7 is relevant for real moral behaviors.

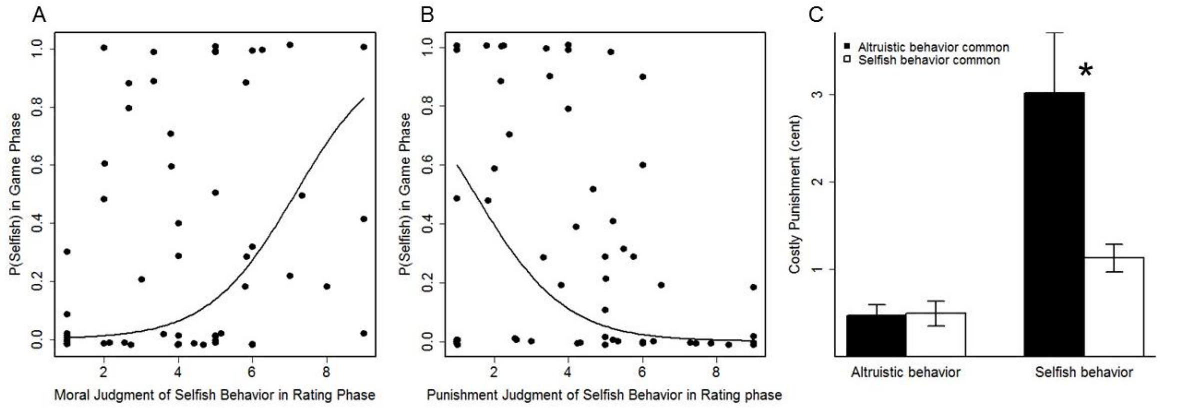

Figure 5. **Judgments predict behavior** (A-B) Average moral and punishment judgments of the selfish behavior in the rating phase predicted the % selfish behaviors during the game phase in Exp. 8. Lines show the fixed effects estimate from random effects logistic regressions. Data points are slightly jittered for visibility. **CIM affects costly moralistic punishment** (C) Costly punishment (mean number of cents invested in punishment per experimental trial) was affected by commonness in Exp. 9. Error bars denote standard error of the mean. Stars indicate significant contrasts between adjacent bars. * = p < .05.

3.2. Commonness affects costly moralistic punishment

Given the importance of costly moralistic punishment for cooperation(Boyd & Gintis, 2003; Fehr & Gächter, 2002; Henrich et al., 2006; Jordan, Hoffman, Bloom, & Rand, 2016), we sought to corroborate the influence of CIM on third-party punishment by running a new experiment (Exp. 9, n = 98). Instead of making hypothetical punishment judgments as in Exp. 1-7, the participants could invest in “deduction points” (Fehr & Gächter, 2002) for real money (see Methods).The prediction from standard economic theory is simple: punishment should be non-existent. A large literature has however shown that people are willing to pay a cost to punish (Henrich et al., 2006), also as unaffected third parties (Fehr & Fischbacher, 2004b). In our experiment, 48 % participants paid to punish at least once, and these participants on average spent 41 % of the total endowment of 2 $ on punishment. There was an interaction between the behavior to be judged and commonness in the use of costly punishment ($\chi2$ (1) = 4.93, p = .026, see Figure 5 C). The effect of commonness was particularly pronounced for the selfish behavior. However, given the reliable effect of commonness on punishment *judgments* for both behavior (see Figure 2), we interpret this as floor effect for the altruistic behavior rather than an effect specific for the selfish behavior. Experiment 9, together with Experiment 8, confirms that moral judgments in our experiments are closely tied to moral behavior.

3.3. CIM can underlie the stability and change of simulated moral norms

Norms are characterized on the group-level by both long periods of stability, and sudden changes, or tipping points, where one norm is replaced by another (Young, 2015). Together, these characteristics result in a *punctuated equilibrium* pattern. We hypothesized that CIM could help explaining both the stability and rapid change of moral norms. Stability of moral norms could be generated by the positive social feedback created by CIM, as CIM makes already popular behaviors moral, and thus more popular (as individuals tend to behave according to their moral values). However, if there is variation between and within people in the influence of moral values on behavior, the CIM feedback mechanism might result in sudden norm changes. If a sufficient number of individuals by chance exhibit the norm violating behavior at a given time-point, CIM might push the norm over a tipping point; CIM will cause observers to assign the higher moral value to the norm violating behavior, which then subsequently will become more popular as people prefer acting according to their moral values, and hence come to replace the old norm. We evaluated the role of CIM for the stability and endogenous change of moral norms using a simple agent-based simulation model (Bonabeau, 2002; Smith & Conrey, 2007). The purpose of agent-based simulation is typically to understand how the interaction of many agents, each behaving according to simple rules (e.g., the CIM heuristic), can result in complex and emergent phenomena in larger groups and longer time scales than typically available in the lab (Bonabeau, 2002; Smith & Conrey, 2007). Such models thereby provide a way to connect the micro-level of psychological mechanisms to the macro-level of the group or population, and are widely used both to provide "proofs of principle" and for analysis and prediction throughout biology and parts of the social sciences (Smith & Conrey, 2007), but has of yet been rarely used in psychology (Gray et al., 2014; Lindström & Olsson, 2015; Lindström, Selbing, & Olsson, 2016; Luhmann & Rajaram, 2015). The structure of the model closely resembled the set of experiments we present above, while abstracting away from any particular strategic structure, costs of punishment, sub-groups, and other important features of real world moral norms (in the SI: Agent-based simulation we describe how the model can be modified to represent a specific strategic interaction). Thus, the model was not intended to be a complete account of moral norms, but rather to understand the explanatory value of CIM in isolation.

In the agent-based model, a random subset of n virtual agents, who was selected from the larger group of N (= 100 in all simulations) agents, interacted by choosing one of two behaviors $i \in \{A, B\}$ based on their moral values $m \in \{M_A, M_B\}$ on each time-step. We assumed that the utility of behavior i at time t was equivalent to its moral value (i.e., $U_i = M_i$), so that the only systematic influence on behavior was the agent's internalized moral values. Although an extreme simplification, this assumption reflects how internalized moral values, which changed through CIM (Exp. 1-9), predicted behavior in Exp. 8. The agents selected the response corresponding to the behavior with the highest subjective moral value (i.e., utility) with high probability, and the other response with low probability. The probability of each behavior was based on a standard uniform error model (Young, 1993) (also known as a ε-greedy decision function (Sutton & Barto, 1998)), where the behavior corresponding to the highest moral value was selected with probability 1–ε, and the other behavior with probability ε. We refer to ε as the decision noise parameter. The parameter ε captures idiosyncratic variation in the influence of moral values on decision making, such as variations in payoff sensitivity, interpretation of the situation, exploration tendency or actual mistakes. Thus, decision noise represents behavioral motivations beyond moral values that are not explicitly modelled (Bowles, 2004). Importantly, we observed such decision noise in Exp.8, where judgments and behavior were strongly, but imperfectly, related. This imperfect relationship between moral judgments and decision making suggests that decision making was influenced by idiosyncratic influences beyond the moral values. The remaining agents (N-n), who were not actively making decisions, observed the interacting agents as third parties, and, in analogy to our experiments, updated the moral values $\{M_A, M_B\}$ of the observed behaviors based on the CIM heuristic. The CIM heuristic was in the agent-based model implemented with the SIT model, which we found could explain the social influence on moral and punishment ratings in our experimental data (Eq. 1, section 1.3). In other words, the implementation of the CIM heuristic in the agent-based model closely matched its empirical characteristics. In simulations with punishment, the observing agents could punish the interacting agents, which decreased by half the utility of the punished behavior. It is important to note that in contrast to the experiments, the agent-based model involved no exogenous manipulation of commonness, as the purpose was to study how the CIM heuristic endogenously contributes to the stability and change of moral norms.

We evaluated the explanatory value of the CIM heuristic for the stability and change of moral norms by simulating the population dynamics where moral norm A initially was completely established ($M_A = 1$, $M_B = 0$) and then calculated how often the alternative norm (B) “invaded” the population (defined as % $B > 80$ %) as a function of the degree of decision noise and the existence of punishment. In accordance with our hypothesis, CIM could generate both stability and sudden norm changes, given some decision noise (see figure 6. see also SI: ABM for additional analyses of how the observability of the behaviors affects norm changes). We also found that punishment amplified the feedback-effect of CIM, and thus resulted in a higher likelihood of sudden norm changes. In summary, the agent-based simulation model demonstrated that the CIM heuristic can function as a mechanism underlying both the stability and change of moral norms on the group level.

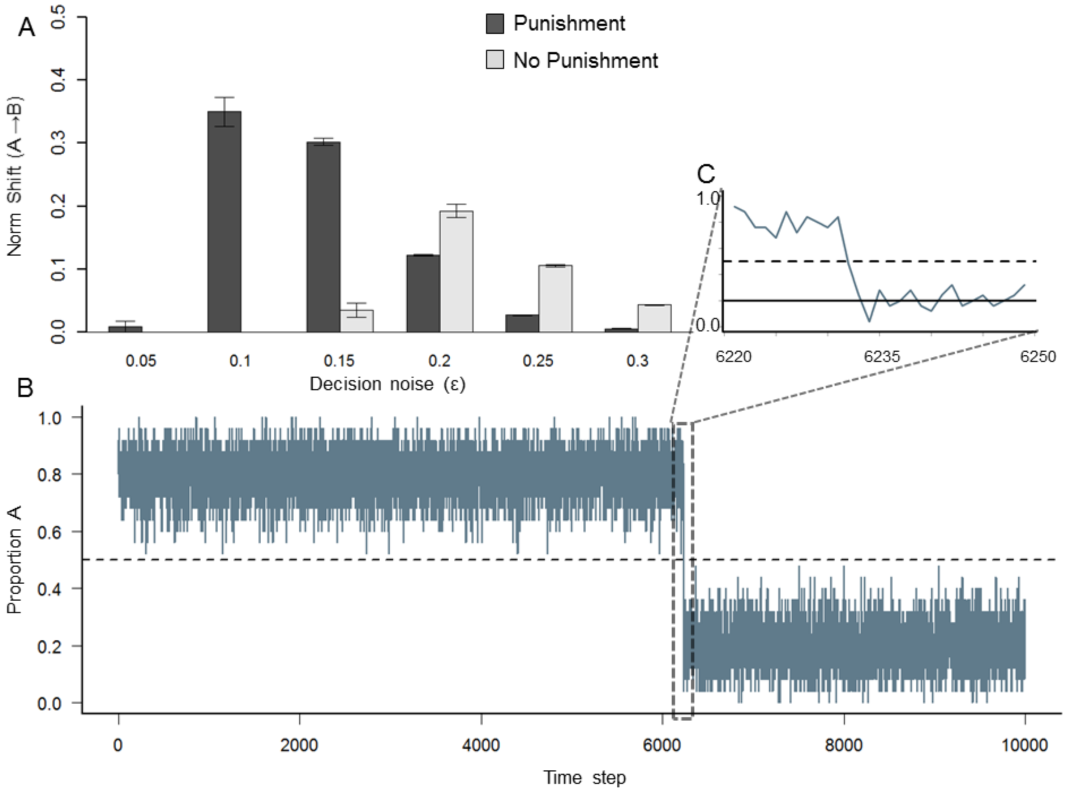

Figure 6. Simulated population dynamics resulting from the CIM heuristic. The simulation involved a group of N = 100 virtual agents, where n = 25 randomly selected agents interacted on each time step, while the remaining agents observed the interaction. The moral values $\{M_A, M_B\}$ of the observed behaviors $\{A, B\}$ were updated according to the SIT model. At the outset of the simulation, there was a norm for *A* ($M_A = 1$, $M_B = 0$). (A) The influence of decision noise on the probability of a norm shift, averaged over 100 runs of 10000 time steps of the model. SIT parameters were set to $s = 1$, $x = 1$. The dotted line represents indicates the hypothetical norm free distribution (.5) (B) Representative run, with $\varepsilon = .2$, $s = 1$, $x = 1$, and no punishment. (C) Zoom in on the sudden

norm shift in B. The lower black line represents the invasion criterion (% B > .8) See SI: ABM for additional information and analyses.

Discussion

We used an experimental model of dynamic social interactions to test if peoples' moral and punishment judgments are shaped by what is common. Across nine independent experiments, we found that this was indeed the case. A selfish behavior that was common was judged as more moral than when rare, and an altruistic behavior that was rare was judged as less moral than when common. In the same way, the behaviors were judged to be deserving of more severe punishment when rare than when common. Thus, commonness moderated pre-existing moral norms. The same pattern also held for arbitrary behaviors without pre-existing moral value (Exp.5). We refer to this pattern as the CIM (common is moral) heuristic.

We characterized the mechanisms underlying the CIM heuristic, and its implications for the stability and change of moral norms, using behavioral and computational methods. Our findings can be summarized in three points. First, the effect of commonness on moral and punishment judgments was driven by the trial-by-trial degree of consensus in the PGG´s the participants observed. This effect could be explained by a classical formal model originally developed to explain social influence situations outside the domain of moral judgments (Social Impact Theory (Latané, 1981)), such as conformity in behavior. These findings represent, to our knowledge, the first clear demonstration of how the objective commonness of social behaviors directly affects moral judgments and decisions to punish. Previous research have shown that people sometimes associate commonness with moral value, but these studies have exclusively used vignettes and hypothetical judgments (Eriksson et al., 2014; Tworek & Cimpian, 2016), which circumscribe their implications. In contrast, we used an experimental model of social interaction involving the observation of dynamic social behaviors (Exp. 1-9), and showed that moral judgments in this model are related to incentivized moral behaviors (Exp. 8-9). Our results demonstrate that people are not only conformist on a behavioral level, but also on the level of moral values. Second, judgments of common behaviors were faster than of rare behaviors. This might reflect an intuitive basis for the CIM heuristic, which would link it to other social heuristics, such as intuitive altruistic decision making (Eriksson et al., 2014; Rand et al., 2014; Rand, Greene, & Nowak, 2012; Sinnott-Armstrong et al., 2010). We note however that we did not directly investigate if the CIM heuristic operates automatically or unconsciously, which represents an important question for future research. Third, and finally, we modeled the population dynamics implicated by use of the CIM heuristic, and found that it could give rise to the punctuated equilibrium effect, which is typical of social norms; long periods of stability, followed by sudden norm changes (Young, 2015). These analyses show that the CIM heuristic can parsimoniously help explain both how moral values and norms are sustained and change. The mechanisms underlying the change of moral values and norms have hitherto been obscure in theories of moral psychology (Bloom, 2010).

The consistent social influence on moral judgments is in support of the social intuitionist perspective on moral psychology, which holds that moral judgments are based in affective, intuitive processes that can be influenced by social factors, such as persuasion (Graham et al., 2011; Haidt, 2001, 2007). Our findings add qualitative and quantitative detail to both the social intuitionist account of how moral intuitions are shaped by social factors (Haidt & Bjorklund, 2008; Haidt, 2001) by demonstrating the direct relationship between moral and behavioral conformity, and the general proposal that fast and frugal heuristics are important in the moral domain (Gigerenzer, 2008, 2010). For example, Gigerenzer (2010) noted that a "copy the majority" heuristic might contribute to moral behavior, but did not provide any evidence in support of this conjecture. It is also important to note that we found that social influences affected the moral value (as revealed in judgments) of observed behaviors directly, while, in contrast, behavioral conformity might be observed even in the absence of individual conviction (Cialdini & Goldstein, 2004). In general, theories based on moral deliberation would be hard pressed to explain why the frequency of a behavior affected moral judgments in our studies and why judgments of common behaviors were faster than of rare behaviors.

On the group-level, the agent-based simulations showed (see Figure 6) that the CIM heuristic is sufficient for recreating key characteristics of moral norms; stability and rapid changes (punctuated equilibrium) (Young, 2015). The underlying mechanism is the positive feedback created by CIM, which stabilizes moral norms, but that together with decision noise (e.g., heterogeneous utility functions) can result in rapid norm changes if a sufficient number of individuals at a given time behave in the non-normative manner (Figure 6). However, when also other systematic influences on behavior are involved, the process might be less dramatic than predicted by the agent-based simulation. For example, moral judgment based in perceived fairness are thought to at least partially be of evolutionary origin (Haidt, 2001, 2007; Zaki & Mitchell, 2013), which might prevent a complete reversal of moral status of the altruistic and selfish behaviors (see SI: ABM for an example of how the model can be extended to include intrinsic valuations). While this makes our experimental demonstration that social influences affect even judgments of these behaviors within the short time span of an experiment even more surprising, it should be noted that the degree of change was small relative to the pre-experimental moral value of the altruistic and selfish actions (see Figure 2). However, relatively small psychological effects might non the less have important practical and population-level consequences (Mesoudi, 2009), especially given the stakes involved in norm-governed behaviors, such as cooperation.

Real world examples of shifts in moral judgments reflecting the same mechanism as in our experiments and simulations might include variation in tax evasion: when tax evasion momentarily becomes more common, for whatever reasons, it also likely to be seen as less morally abhorrent, which in turn is bound to make it more common, and so on (Eriksson et al., 2014). Naturally, we make no claims that the CIM heuristic is the only possible mechanism to explain the change of moral norms. For example, harm, moral reasoning, model learning, and active social persuasion are complementary mechanisms (Bandura & McClelland, 1977; Bloom, 2010; Haidt & Bjorklund, 2008; Schein & Gray, 2016). It will be of importance to study how these different mechanisms interact to shape moral values.

Because we exclusively focused on moral judgments related to cooperation in a social dilemma (and of unknown social behaviors, Exp. 5), it remains unknown how the CIM heuristic generalizes to other moral domains. For example, the social intuitionist framework has identified a number of moral foundations other than fairness in cooperation (Graham et al., 2011), such as sanctity (related to disgust). It is possible that CIM plays role in the large cross-cultural variability also in such moral judgments. For example, the practice of eating dog meat is rare, and subject to disgust and condemnation in the West, but not in some areas of East Asia (Haidt, Koller, & Dias, 1993). Such correlations between commonness and moral judgments are in line with predictions from the CIM heuristic, which highlights the importance of empirically testing how our theoretical framework extends to moral norms beyond social dilemmas.

Although not a direct focus of the present study, it is also plausible that the CIM heuristic can contribute to the emergence of new moral norms (“moralization”; Rozin, 1999). Additional simulations of the agent-based model show that if two behaviors have initially neutral moral values, one of these behaviors will quickly come to dominate the population (due to the positive feedback mechanism of CIM), and thus establish a novel moral norm (simulation results available from first author). The negative moralization of left-handedness (Mandal & Dutta, 2001) might represent a norm that has emerged directly from relative frequency in this way. However, it is clear that not all common behaviors are seen as moral, or all rare behaviors as immoral (Eriksson et al., 2014), indicating that additional mechanisms typically contribute. Disgust (Rozin, 1999) and harm (Schein & Gray, 2016) considerations are thought to often contribute to the moralization of a behavior (e.g., smoking). It is plausible that the CIM heuristic can form a feedback loop with disgust and harm (e.g., if a behavior initially is seen as mildly disgusting or harmful, the CIM heuristic might make it rarer, and consequentially more negatively moralized, and so on). Directly investigating the role of the CIM heuristic in the emergence of new moral norms represents an important route for further research.

The proposal that the evolutionary function of moral judgments is to coordinate side-taking in conflicts, where being on the wrong side might be associated with severe costs (DeScioli & Kurzban, 2013), might provide a possible ultimate explanation for the CIM heuristic. The CIM heuristic could contribute to side-taking, by endowing the more common behavior with moral value, which could efficiently allow observing agents to coordinate on one action. Our finding that moral judgments of more common behaviors were faster than of rare behaviors goes hand in hand with this suggestion, as it implies that side taking would be faster the larger the majority (and thereby the likely cost of being on the wrong side of the conflict).

To conclude, using a multi-method approach, we demonstrated that people behave as described by the CIM heuristic: they judge others´ behaviors as more moral when they are common than when they are rare in social environment. This process was mediated by computational mechanisms similar to other social influence phenomena, such as conformity, which modulated moral intuitions. Simulation modeling showed that CIM can help to parsimoniously explain both the stability and change of moral norms. Our findings have implications for our understanding of how social dynamics shape moral norms and norm governed behaviors, with bearing for a wide variety of social situations, including attitudes to public goods and law abidance.

ACKNOWLEDGMENTS. We thank Philippe Tobler, Philip Pärnamets, and Lars Hall for helpful comments on an earlier version of the manuscript. We thank Pontus Strimling, and Kimmo Eriksson for valuable contributions to study design and interpretation. This research was supported by a grant from the Riksbankens Jubileumsfond (Grant P11-1017), an Independent Starting Grant (284366; *Emotional Learning in Social Interaction* project) from the European Research Council, and a grant from the Knut and Alice Wallenberg Foundation (KAW 2014.0237) to Andreas Olsson. Björn Lindström was supported by Forte (COFAS2: 2014-2785 FOIP).

References

Amir, O., Rand, D. G., Gal, Y. K., Johannesson, M., & List, J. (2012). Economic Games on the Internet: The Effect of $1 Stakes. *PLoS ONE*, *7*(2), e31461. doi:10.1371/journal.pone.0031461

Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423–40.

Asch, S. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*.

Bandura, A., & McClelland, D. (1977). Social learning theory.

Bates, D., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes.

Batson, C. D., & Thompson, E. R. (2001). Why Don't Moral People Act Morally? Motivational Considerations. *Current Directions in Psychological Science*, *10*(2), 54–57. doi:10.1111/1467-8721.00114

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge, UK: Cambridge University Press.

Bloom, P. (2010). How do morals change? *Nature*, *464*(7288), 490. doi:10.1038/464490a

Bonabeau, E. (2002). Agent-based modeling: methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America, 99 Suppl 3,* 7280–7. doi:10.1073/pnas.082080899

Bowles, S. (2004). *Microeconomics : behavior, institutions, and evolution*. Russell Sage Foundation.

Bowles, S., & Gintis, H. (2011). *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton University Press.

Boyd, R., & Gintis, H. (2003). The evolution of altruistic punishment. *Proceedings of the ….*

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science (New York, N.Y.)*, *328*(5978), 617–20. doi:10.1126/science.1183665

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(6), 3531–5. doi:10.1073/pnas.0630443100

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, *108 Suppl* , 10918–25. doi:10.1073/pnas.1100290108

Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, *15*(5), 655–61. doi:10.1038/nn.3087

Chudek, M., & Henrich, J. (2011). Culture-gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences*, *15*(5), 218–26. doi:10.1016/j.tics.2011.03.003

Cialdini, R. B., Demaine, L. J., Sagarin, B. J., Barrett, D. W., Rhoads, K., & Winter, P. L. (2006). Managing social norms for persuasive impact. *Social Influence*, *1*(1), 3–15. doi:10.1080/15534510500181459

Cialdini, R. B., & Goldstein, N. J. (2004). Social Influence: Compliance and Conformity. *Annual Review of Psychology*, *55*(1), 591–621. doi:10.1146/annurev.psych.55.090902.142015

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015–1026. doi:10.1037/0022-3514.58.6.1015

Claidière, N., & Whiten, A. (2012). Integrating the study of conformity and culture in humans and nonhuman animals. *Psychological Bulletin*.

Colby, A., Kohlberg, L., & Gibbs, J. (1983). A longitudinal study of moral judgment. *Monographs of the ….*

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–6. doi:10.1016/j.tics.2013.06.005

Crockett, M. J. (2016). How Formal Models Can Illuminate Mechanisms of Moral Judgment and Decision Making. *Current Directions in Psychological Science*, *25*(2), 85–90. doi:10.1177/0963721415624012

Crump, M. J. C., McDonnell, J. V., Gureckis, T. M., Romero, J., & Morris, S. (2013). Evaluating Amazon’s Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, *8*(3), e57410. doi:10.1371/journal.pone.0057410

Cubitt, R., Drouvelis, M., Gächter, S., & Kabalin, R. (2011). Moral judgments in social dilemmas: How bad is free riding? *Journal of Public Economics*.

Cushman, F. (2013). Action, outcome, and value: a dual-system framework for morality. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc*, *17*(3), 273–92. doi:10.1177/1088868313495594

Denrell, J. (2008). Sociology. Indirect social influence. *Science (New York, N.Y.)*, *321*(5885), 47–8. doi:10.1126/science.1157667

DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin*, *139*(2), 477–496. doi:10.1037/a0029065

Eidelman, S., Crandall, C. S., & Pattershall, J. (2009). The existence bias. *Journal of Personality and Social Psychology*, *97*(5), 765–775. doi:10.1037/a0017058

Eriksson, K., Strimling, P., & Coultas, J. C. (2014). Bidirectional associations between descriptive and injunctive norms. *Organizational Behavior and Human Decision Processes*, *129*, 59–69. doi:10.1016/j.obhdp.2014.09.011

Fehr, E., & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190. doi:10.1016/j.tics.2004.02.007

Fehr, E., & Fischbacher, U. (2004b). Third-party punishment and social norms. *Evolution and Human Behavior*.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140. doi:10.1038/415137a

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science (New York, N.Y.)*, *349*(6245), 273–8. doi:10.1126/science.aac6076

Gigerenzer, G. (2008). Moral intuition= fast and frugal heuristics? *The Cognitive Science of Morality: Intuition and …*.

Gigerenzer, G. (2010). Moral satisficing: rethinking moral behavior as bounded rationality. *Topics in Cognitive Science*, *2*(3), 528–54. doi:10.1111/j.1756-8765.2010.01094.x

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*(1), 451–482. doi:10.1146/annurev-psych-120709-145346

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–85. doi:10.1037/a0021847

Gray, K., Rand, D. G., Ert, E., Lewis, K., Hershman, S., & Norton, M. I. (2014). The emergence of "us and them" in 80 lines of code: modeling group genesis in homogeneous populations. *Psychological Science*, *25*(4), 982–90. doi:10.1177/0956797614521816

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–34.

Haidt, J. (2007). The new synthesis in moral psychology. *Science (New York, N.Y.)*, *316*(5827), 998–1002. doi:10.1126/science.1137651

Haidt, J., & Bjorklund, F. (2008). Social intuitionists answer six questions about morality.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, *65*(4), 613–628. doi:10.1037/0022-3514.65.4.613

Harms, W., & Skyrms, B. (2008). Evolution of Moral Norms. In M. Ruse (Ed.), *Oxford Handbook on the Philosophy of Biology*. Oxford University Press.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., … Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science (New York, N.Y.)*, *327*(5972), 1480–4. doi:10.1126/science.1182238

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., … Ziker, J. (2006). Costly punishment across human societies. *Science (New York, N.Y.)*, *312*(5781), 1767–70. doi:10.1126/science.1127333

Hoppitt, W., & Laland, K. (2013). Social Learning: An Introduction to Mechanisms, Methods, and Models.

Hume, D. (2003). *A Treatise of Human Nature*. Courier Corporation.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473–476. doi:10.1038/nature16981

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *... and Biases: The Psychology of Intuitive …*.

Kohlberg, L., & Hersh, R. H. (1977). Moral development: A review of the theory. *Theory Into Practice*, *16*(2), 53–59. doi:10.1080/00405847709542675

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications*, *6*, 7455. doi:10.1038/ncomms8455

Krajbich, I., Hare, T., Bartling, B., Morishima, Y., & Fehr, E. (2015). A Common Mechanism Underlying Food Choice and Social Decisions. *PLoS Computational Biology*, *11*(10), e1004371. doi:10.1371/journal.pcbi.1004371

Laland, K. N., & Rendell, L. (2013). Cultural memory. *Current Biology : CB*, *23*(17), R736–40. doi:10.1016/j.cub.2013.07.071

Latané, B. (1981). The psychology of social impact. *American Psychologist*.

Lindström, B., & Olsson, A. (2015). Mechanisms of social avoidance learning can explain the emergence of adaptive and arbitrary behavioral traditions in humans. *Journal of Experimental Psychology. General*, *144*(3), 688–703. doi:10.1037/xge0000071

Lindström, B., Selbing, I., & Olsson, A. (2016). Co-Evolution of Social Learning and Evolutionary Preparedness in Dangerous Environments. *PloS One*, *11*(8), e0160245. doi:10.1371/journal.pone.0160245

Luhmann, C. C., & Rajaram, S. (2015). Memory Transmission in Small Groups and Large Networks: An Agent-Based Model. *Psychological Science*, *26*(12), 1909–1917. doi:10.1177/0956797615605798

MacCoun, R. J. (2012). The burden of social proof: shared thresholds and social influence. *Psychological Review*, *119*(2), 345–72. doi:10.1037/a0027121

Mandal, M., & Dutta, T. (2001). Left handedness: Facts and figures across cultures. *Psychology & Developing Societies*.

McGraw, K. M. (1985). Subjective probabilities and moral judgments. *Journal of Experimental Social Psychology*, *21*(6), 501–518. doi:10.1016/0022-1031(85)90022-8

Mesoudi, A. (2009). How cultural evolutionary theory can inform social psychology and vice versa. *Psychological Review*, *116*(4), 929–52. doi:10.1037/a0017062

Ostrom, E. (2000). Collective Action and the Evolution of Social Norms. *Journal of Economic Perspectives*, *14*(3), 137–158. doi:10.1257/jep.14.3.137

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*(4), 1023–1031. doi:10.3758/s13428-013-0434-y

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427–30. doi:10.1038/nature11467

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*, 3677. doi:10.1038/ncomms4677

Reynolds, S. J., & Ceranic, T. L. (2007). The effects of moral judgment and moral identity on moral behavior: an empirical examination of the moral individual. *The Journal of Applied Psychology*, *92*(6), 1610–24. doi:10.1037/0021-9010.92.6.1610

Richerson, P., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*. Chicago: University of Chicago Press.

Rozin, P. (1999). The Process of Moralization. *Psychological Science*, *10*(3), 218–221. doi:10.1111/1467-9280.00139

Schein, C., & Gray, K. (2016). Moralization and Harmification: The Dyadic Loop Explains How the Innocuous Becomes Harmful and Wrong. *Psychological Inquiry*, *27*(1), 62–65. doi:10.1080/1047840X.2016.1111121

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, *18*(5), 429–34. doi:10.1111/j.1467-9280.2007.01917.x

Sinnott-Armstrong, W., Young, L., & Cushman, F. (2010). Moral Intuitions as Heuristics. In J. Doris, S. Harman, J. Nichols, W. Prinz, Sinnott-Armstrong, & S. Stich (Eds.), *The Oxford Handbook of Moral Psychology*. Oxford University Press.

Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: a new approach for theory building in social psychology. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc*, *11*(1), 87–104. doi:10.1177/1088868306294789

Sunstein, C. R. (2005). Moral heuristics. *The Behavioral and Brain Sciences*, *28*(4), 531–42; discussion 542–73. doi:10.1017/S0140525X05000099

Sutton, R., & Barto, A. (1998). Introduction to reinforcement learning.

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

Tworek, C. M., & Cimpian, A. (2016). Why Do People Tend to Infer &quot;Ought&quot; From &quot;Is&quot;? The Role of Biases in Explanation. *Psychological Science*, *27*(8), 1109–22. doi:10.1177/0956797616650875

Young, H. P. (1993). The Evolution of Conventions. *Econometrica*, *61*(1), 57. doi:10.2307/2951778

Young, H. P. (2015). The Evolution of Social Norms. *Annual Review of Economics*, *7*(1), 359–387. doi:10.1146/annurev-economics-080614-115322

Zaki, J., & Mitchell, J. (2013). Intuitive prosociality. *Current Directions in Psychological Science*.