Research Article

Psychological Science
1–23

DOI: 10.1177/09567976251398454
www.psychologicalscience.org/PS

Social Rewards Protection Theory: Why People Morally Derogate Prosocial Actors for Undisclosed Personal Benefits

Sebastian Hafenbrädl
IESE Business School

Abstract

Prosocial behavior is common and often socially rewarded (e.g., via liking, status, and trust). Yet prior research has found that if actors themselves also benefit from their prosocial behavior, then they are morally derogated: They are evaluated as worse than purely selfish actors. This *tainted-altruism effect* has been explained by the use of different counterfactuals for the evaluation of prosocial and selfish actors. Here I propose *social rewards protection theory*, which explains why evaluators use these different counterfactuals in the first place: Social rewards are treated as being reserved for costly prosocial actions. Claiming such rewards without incurring costs seems like cheating and thus deserves moral derogation. Accordingly, being transparent about the action's costs and benefits prevents such derogation. I conducted six experiments (five preregistered) with Amazon Mechanical Turk (MTurk) workers in the United States and lab participants in Spain (total $N = 4{,}732$ adults). The findings provide support for the proposed functional explanation of tainted altruism, which also sheds light on related phenomena, such as overhead aversion and hypocrisy.

Keywords

cooperation, corporate social responsibility, evolutionary psychology, hypocrisy, morality, social cognition, social evaluations, signaling, tainted altruism

Received 9/7/22; revision accepted 9/29/25

Prosocial behavior is common. People help others, volunteer, and donate money for noble causes; organizations pursue social goals and engage in corporate social responsibility. Such behavior that benefits others is often socially rewarded (Ariely et al., 2009; Bénabou & Tirole, 2006; Nowak, 2006): Prosocial actors receive praise and gratitude, are trusted and liked more, and are granted higher status (Bai et al., 2020; Flynn, 2003; Harbaugh, 1998; Hardy & Vugt, 2016; Willer, 2009). Some social rewards are tangible: Charities receive donations, and socially responsible organizations sell more products (Sen & Bhattacharya, 2001) and inspire higher motivation in their employees (Flammer & Luo, 2017). In sum, prosocial actors often receive social rewards. The expectation of these rewards inspires more prosocial behavior (Grant & Gino, 2010; McCullough et al., 2008). Humans have evolved to be prosocial, as is reflected in evolutionary game theory models of altruism (Axelrod & Hamilton, 1981; Gintis et al., 2003; Hoffman et al., 2016; Panchanathan & Boyd, 2004).

At the same time, research has shown that actors who gain personally from their prosocial acts receive fewer social rewards (Bai, Ho, & Yan, 2020; Berman & Silver, 2022; Carlson & Zaki, 2018; Cassar & Meier, 2021; Lin-Healy & Small, 2013; Makov & Newman, 2016; Raihani & Power, 2021), and they sometimes even receive punishments in the form of moral derogation (Newman & Cain, 2014): They are evaluated as worse than actors who did not engage in prosocial behavior in the first place—their altruism becomes tainted. Consider the case of Daniel Pallotta, whose fundraising company raised more than $305 million for charities. Once it became public that he personally earned a high salary, close to $400,000, he faced public outrage, his

Corresponding Author:
Sebastian Hafenbrädl, Department of Managing People in Organizations, IESE Business School, University of Navarra, Barcelona, Spain.
Email: shafenbraedl@iese.edu

company collapsed, and the donations to the charities he represented plummeted (Kristof, 2008).

Existing research proposes that this moral derogation can be explained by the accessibility of different counterfactuals: "When someone was charitable for self-interested reasons, people considered his or her behavior in the absence of self-interest, ultimately concluding that the person did not behave as altruistically *as he or she could have*. However, when someone was only selfish, people did not spontaneously consider whether the person could have been more altruistic" (Newman & Cain, 2014, p. 648). Although this account is well supported by those earlier studies (and their replication; Alcala et al., 2022), it begs for a more ultimate explanation: Why are the different counterfactuals (and, relatedly, reference points; Zlatev & Miller, 2016) more accessible in the first place?

Here I propose and test such an explanation, *social rewards protection theory*, which aims at reconciling the tainted-altruism effect with the high prevalence of prosocial behavior in society and the longstanding research tradition of evolutionary game theory. Building on two research streams, this new theory specifies the circumstances under which prosocial actors are morally derogated: when actors are seen as deserving fewer social rewards than they are seen as claiming.

The first research stream I build on is the moral-character-evaluation literature (Berman & Silver, 2022; Carlson et al., 2022; Critcher et al., 2020; Siegel et al., 2017), which explains discounted or cheapened altruism: People care more about the motivation and the moral character of the actor than about the consequences of the action. For instance, actors engaging in prosocial behavior are considered less benevolent and, in turn, are evaluated as worse when the cause personally affects them (Lin-Healy & Small, 2012). In the worst case, when prosocial behavior cannot even partially be attributed to their character, people are seen as deserving zero social rewards. What this research stream cannot explain, however, is the case of prosocial actors being morally derogated—that is, being evaluated as morally worse than actors who did not provide any prosocial benefits in the first place.

This is where the second research stream—on hypocrisy—comes into play. Hypocrisy typically stems from a divergence between words and deeds (Effron et al., 2018) and has been shown to lead to social punishment. For instance, Jordan et al. (2017) showed that moral transgressors are punished more harshly when they verbally condemn the immoral behaviors they engage in, and the authors explained these punishments with a *false-signaling* account (building on *signaling theory*; Spence, 1973). People hate hypocrites not for the inconsistency per se but for the transgressors' signal that they are more moral than they actually are. Reconceptualizing this mechanism for prosocial actors (instead of transgressors), I propose that prosocial deeds themselves serve as signals (instead of transgressors' words). The proposed explanation thereby mirrors Jordan et al.'s false-signaling account and highlights parallels in how individuals evaluate hypocrisy and prosocial behavior.

Integrating elements from these two research streams, social rewards protection theory conceptualizes the evaluation of prosocial actors as a three-step process. The first step is to determine the extent to which the actor deserves social rewards. Building on the first research stream on moral-character evaluation, this is a question of attribution. Social rewards are reserved for prosocial behavior that is attributed to the character of the actor, rather than to the situation in which the behavior occurred. If prosocial behavior occurs in a situation that is stacked against that behavior, it is costly for the actors and thus diagnostic of their prosocial motivation (Kawamura et al., 2021). They make a sacrifice to act prosocially (and people are surprisingly motivated to make such prosocial sacrifices; Kirgios et al., 2020; Olivola, 2011; Olivola & Shafir, 2013).

The second step is the evaluation of the actor's signaling: Does the actor claim social rewards for the action? By default, evaluators seem to interpret prosocial behavior as a signal to claim social rewards, unless actors actively declare that they do not claim such rewards. The third step is the comparison: Is the actor seen as claiming more social rewards than he or she is perceived to deserve?

Because social rewards are valuable, actors who claim more than they deserve are seen as deceptive and are judged as morally worse than actors who did not claim the rewards (e.g., people who did not act prosocially in the first place). Rather than an inherent feature of the actor's behavior (or motive), deceptiveness is an outcome of this evaluation process; it stems from the perceived divergence between the deserved and the claimed social rewards. Conversely, actors who make it clear that they do not deserve social rewards for their (personally beneficial) prosocial behavior are not punished with moral derogation. As the signaling step is specific to social rewards, nonsocial rewards (such as emotional or self-concept rewards) should not lead to moral derogation (but merely to "discounted altruism").

In sum, what taints prosocial actors is not the mere presence of self-interest, but the perception that actors try to reap social rewards without deserving them (i.e., without paying the price), which makes them seem deceptive. This explanation does not contradict but rather incorporates the existing counterfactuals account,

and provides a functional answer to the question of why people use these different counterfactuals in the first place: Observers evaluate whether actors who claim social rewards actually deserve them by comparing the actors with counterfactual actors who are prosocially motivated, and hence considered as deserving. On the basis of this comparison, observers either socially reward or punish the prosocial actors. If, however, actors are not seen as trying to reap these social rewards, they avoid entering this comparison and thereby prevent the potentially ensuing moral derogation.

Moving into the realm of more ultimate explanations, reserving social rewards such as praise, status, and trust only for prosocially motivated actors prevents diluting their value. This is advantageous on two levels. On the individual level, such actors are more likely to act prosocially in the future, even under potentially different circumstances, and thus are more desirable cooperation partners (Davis et al., 2023; Simpson & Willer, 2015); on the societal level, such social rewards can motivate prosocial behavior in situations in which this behavior is costly for the actors, and thus least likely to occur in the absence of these social rewards.

Research Transparency Statement

General disclosures

Conflict of interest: The author declared that there were no conflicts of interest with respect to the authorship or the publication of this article. **Funding:** This research was supported by Spanish Ministerio de Ciencia e Innovation Grant No. PID2019-103897GB-I00. **Artificial intelligence:** Except for spell-checking by Grammarly and Microsoft Word, no artificial intelligence-assisted technologies were used in this research or the creation of this article. **Ethics:** This research received approval from local ethics boards (Yale ID No. 1605017813; IESE ID No. 2023.14)

Study disclosures

Preregistration: No aspects of Study 2a were preregistered. For all other studies, the hypotheses, methods, and analysis plan were preregistered prior to data collection (Study 1: https://aspredicted.org/tgg2-bkgr.pdf; Study 2b: https://aspredicted.org/3syp-2w9v.pdf; Study 3: https://aspredicted.org/ptwd-kc66.pdf; Study 4: https://aspredicted.org/g4ck-b4dc.pdf; Study 5: https://aspredicted.org/2cjd-4rkw.pdf). Study 3 deviated from the preregistration because more participants were accidentally recruited than preregistered (see the Supplemental Material for an analysis of only the preregistered sample; all results were consistent in direction and significance). Study 5 also deviated from the preregistration, because a larger sample had to be recruited to fill all the sender positions in the sender-receiver game (while avoiding deception). In all other studies, there were no deviations from the preregistrations. **Materials:** All study materials are publicly available: https://researchbox.org/646. **Data:** All primary data are publicly available: https://researchbox.org/646. **Analysis scripts:** All analysis scripts are publicly available: https://researchbox.org/646. **Computational reproducibility:** The computational reproducibility of the results has been independently confirmed by the journal's STAR Team.

Experiment 1

Experiment 1 replicates the tainted-altruism effect, using the very scenario that initially established it (Newman & Cain, 2014), and tests the newly proposed social rewards protection theory by making the effect disappear when the actor clarifies that he does not deserve social rewards.

Experiment 1: method

Design. Participants were randomly assigned to one of four between-subjects conditions. The first two conditions were the original conditions from Newman and Cain (2014): the *homeless-shelter* condition, featuring a selfishly motivated actor engaging in prosocial behavior (volunteering in a homeless shelter), and the *coffee-shop* condition, featuring a similarly selfishly motivated actor engaging in neutral behavior (volunteering in a coffee shop). The selfish motivation is, in both conditions, to gain the affection of a woman. Aiming to replicate Newman and Cain's tainted-altruism effect, participants are expected to rate the actor who volunteers in the homeless shelter as less moral than the actor who volunteers in the coffee shop.

Note that whereas the actor in both conditions is arguably somewhat inherently deceptive, this inherent deceptiveness is held constant across these two conditions: In both conditions, the actor is driven by the same ulterior motive (to gain the woman's affection), and he engages in the same behavior (volunteering). The only difference is that the behavior yields prosocial consequences in the homeless-shelter condition, but not (or at least not as much) in the coffee-shop condition. Any difference in the actor's ascribed deceptiveness between the conditions, therefore, cannot be explained by the actor's inherent deceptiveness. Rather, to explain the differences, the new theory places the prosocial consequences at the center of its three-step social-evaluation process. In the two novel conditions

(*homeless-shelter full disclosure* and *coffee-shop full disclosure*), the actor directly discloses his selfish motivation to the woman before starting to volunteer. He thereby admits that his prosocial behavior is not driven by prosocial motivation but by self-interest, and thereby ensures that he no longer can be seen as (deceptively) pretending to deserve social rewards. These two conditions allow for testing the new theory within the original paradigm: Specifically, I hypothesize that removing the possibility that the actor can be seen as pretending to deserve social rewards will moderate the tainted-altruism effect so that there is no difference in moral evaluations between the two new conditions. This hypothesis, and these two novel conditions, parallel Jordan et al.'s (2017) insights about the psychology of hypocrisy, in particular their "honest hypocrite" hypothesis and the corresponding experimental conditions.

Procedure. In both the original and the modified experimental conditions, participants read a short description involving either a homeless shelter or a coffee shop:

> Andy has a crush on an attractive girl named Kim.
>
> Kim spends most of her time working at the local homeless shelter/coffee shop, where she has become a manager. Only because Andy wants to impress Kim and get her to go out with him, he starts volunteering many hours per week at the same homeless shelter/coffee shop, passing out soup/coffee and making the homeless people/customers feel comfortable.

In the original homeless shelter/coffee shop conditions in Newman and Cain (2014), the description was followed by the following sentences:

> Andy does a good job and other people believe that he really enjoys volunteering there. However, Andy only does a good job in order to impress Kim. Eventually his plan is successful and he gets a date with her.

In the newly developed homeless-shelter full-disclosure/coffee-shop full-disclosure conditions, the description was followed by the following sentences instead:

> He gets invited for an interview with Kim, and when she asks him why he wants to work there, he honestly replies that he is only here because he wants to date her.
>
> Andy starts volunteering many hours per week and does a good job and other people believe that he really enjoys volunteering there. However, as Kim knows Andy only does a good job in order to impress her. Nevertheless, eventually his plan is successful and he gets a date with her.

Measures. After reading a vignette, participants evaluated Andy using Likert scales ranging from 1 to 9 (all measures were adopted from Newman & Cain, 2014).

The first measure assessed Andy's morality with three items (Cronbach's $\alpha = .90$ in this sample): How ethical was Andy's behavior? (1 = *completely unethical*, 9 = *completely ethical*); how moral was Andy's behavior (1 = *completely immoral*, 9 = *completely moral*); and to what extent do you approve or disapprove of Andy's actions? (1 = *definitely not*, 9 = *definitely so*).

The second measure assessed how deceptive Andy is perceived with four items (Cronbach's $\alpha = .85$ in this sample): How manipulative were Andy's actions? (1 = *not at all manipulative*, 9 = *very manipulative*); how honest were Andy's actions? (1 = *not at all honest*, 9 = *very honest*; reverse-coded for scale aggregation); how deceptive were Andy's actions? (1 = *not at all deceptive*, 9 = *very deceptive*); and how hypocritical were Andy's actions? (1 = *not at all hypocritical*, 0 = *very hypocritical*).

The third measure assessed the benefit of Andy's actions with two items (Cronbach's $\alpha = .95$ in this sample): How beneficial were Andy's actions? (1 = *not at all*, 9 = *very beneficial*), and to what extent Andy's actions make the world a better place? (1 = *not at all*, 9 = *very much so*).

As exploratory measures, to evaluate the generalizability of the results and to rule out alternative explanations, participants also responded to the question of how much they liked and trusted Andy and how altruistic and selfish Andy's actions were.

Attention check. After responding to the last survey question, participants responded to the following attention check: "As a final question for you, we want to know whether you are processing the information given in each question. For the question below, please click on the first choice *not at all satisfied*." In Experiment 1, the question was, "How satisfied do you think is Andy with his job?" Responses were collected on a 9-point scale ranging from 1 (*not at all satisfied*) to 9 (*very satisfied*). Consistent with our preregistration, all participants who did not respond with a 1 (*not at all satisfied*) were excluded from the sample. The rates of exclusion did not differ substantially between experimental conditions; see the Supplemental Material for details (Table S11), robustness analyses (Table S12), and figures underlining the consistency of the effects in the full data set (Figures S13-S16).

Original study and preregistered replication. For readability and space, I report below only the results of a preregistered replication study; see the Supplemental Material for a full reporting of the original study (N = 397). Results were consistent with the replication in both direction and significance, except that the interactions for deceptiveness (Table S5a), trust (Table S9a), and liking (Table S7a) did not reach statistical significance in the less well-powered original study.

Participants. I recruited participants online using MTurk. A total of 1,141 reached the attention-check question, and consistent with the preregistration, I analyzed the data of all Americans with unique IP addresses who did not fail the first attention check, resulting in a final sample of 798 participants (M_{age} = 38.5 years, 54% female). See the Supplemental Material for a detailed description of the MTurk worker requirements. This study and all subsequent studies were approved by the Yale Institutional Review Board (ID No. 1605017813).

Experiment 1: results

Figure 1a depicts the moral evaluations of the actor in the four between-subjects conditions. The first two conditions were the original conditions from Newman and Cain (2014): the homeless-shelter condition, featuring a selfishly motivated actor engaging in prosocial behavior (volunteering in a homeless shelter), and the coffee-shop condition, featuring a similarly selfishly motivated actor engaging in neutral behavior (volunteering in a coffee shop). The selfish motivation is, in both conditions, to gain the affection of a woman. Replicating Newman and Cain's tainted-altruism effect, participants rated the actor who volunteers in the homeless shelter as less moral (M = 4.63, SEM = 0.13) than the actor who volunteers in the coffee shop (M = 5.33, SEM = 0.12, $t(397) = 3.86$, $p = .0001$, $d = 0.39$). In the two novel conditions (homeless-shelter full-disclosure condition and coffee-shop full-disclosure condition), the actor directly discloses his selfish motivation to the woman before starting to volunteer (the scenario indicates that in an interview for the position, the woman asks him why he wants to work there, and "he honestly replies that he is only here because he wants to date her"). By admitting that his prosocial behavior is not driven by prosocial motivation, but by self-interest, he ensures that he no longer can be seen as pretending to deserve social rewards. These two conditions test the key prediction of the new theory within the original paradigm: Removing the possibility that the actor can be seen as pretending to deserve social rewards moderates the tainted-altruism effect: As predicted, there is no discernable difference in moral evaluations between the two new conditions—M = 5.07, SEM = 0.13; M = 5.09, SEM = 0.13; $t(397) = 0.13$, $p = .895$, $d = 0.01$; preregistered interaction with the two original conditions: $F(1, 794) = 6.78$, $p = .009$.

Ruling out alternative explanations, this interaction (as well as the effects and interactions in Experiments 2–4) is robust to including control variables, specifically judgments of the prosocial benefits the actor's behavior creates (Table S2b) and judgments of the extent to which the behavior itself is seen as altruistic or selfish (Table S10b). In other words, the interaction is not driven by participants judging that the actor's behavior leads to different levels of prosocial benefits across conditions. And it is not driven by differences in evaluations of the actor's degree of selfishness and altruism—which would be affected by the first step of the proposed three-step process, the moral-character evaluation, but which should not be affected by the second (signaling) and third (comparison) step of the process.

The pattern of moral evaluations is mirrored by judgments about the deceptiveness of the actor's behavior (see Fig. 1b) and closely tracked by more practically relevant variables, such as trusting (Figure S6) and liking (Figure S5) the actor (see the Supplemental Material for details on these and other extended analyses, as well as additional tables and figures).

Experiments 2a and 2b

Here, I test the theory's key prediction in a different context and with a different operationalization of the absence of social rewards.

Experiment 2a: method

Design. Participants were randomly assigned to one of four between-subjects conditions: *social rewards, no social rewards (no broadcasting), control, control (no broadcasting)*. In one scenario the owner of a beach resort is cleaning up the beach (a behavior with prosocial consequences), which allows the resort to make money from tourists (selfish gains). Although the actor in Experiment 1 fully discloses his selfish motivation, to prevent the impression that he is trying to reap undeserved social rewards, the actor in Experiment 2a simply does not broadcast the prosocial benefits of his behavior (i.e., he does not advertise that *he* cleaned up the beach). This manipulation is a different operationalization of how actors can prevent being seen as claiming social rewards—not by actively disclosing their selfish motivation, which could come across as inherently deceptive, but by passively avoiding the disclosure of the prosocial benefits of their behavior.

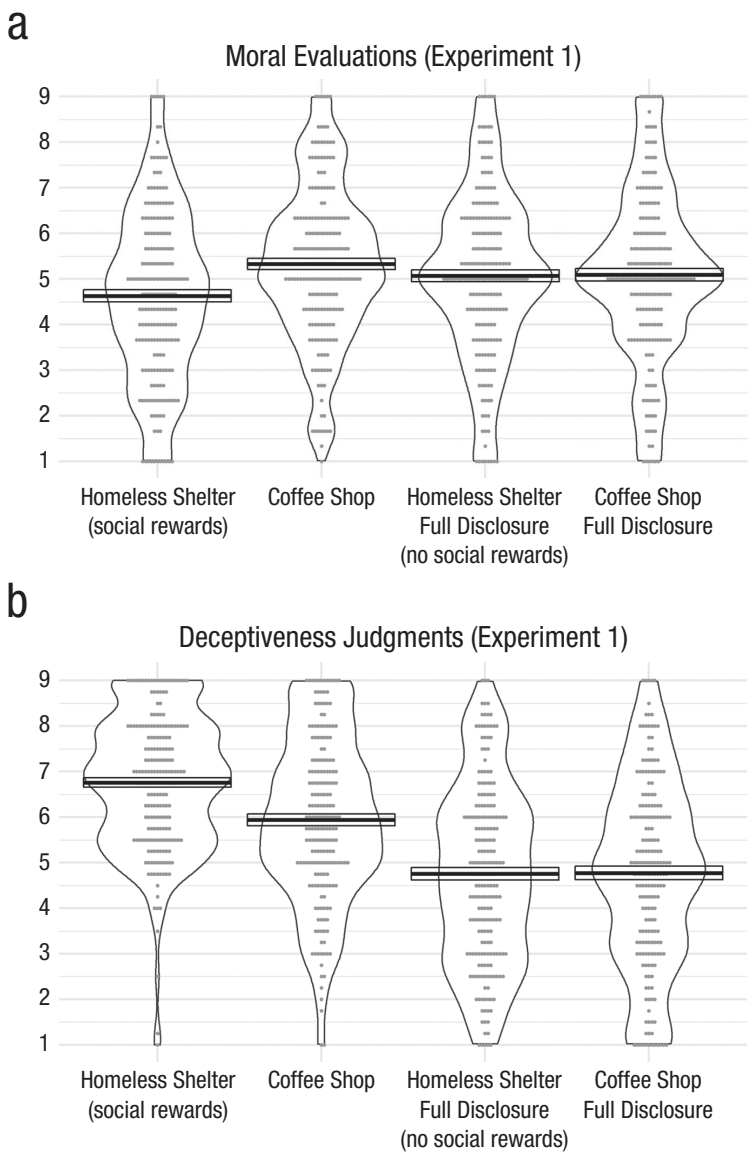

Fig. 1. Moral evaluations (a) and deceptiveness judgments (b) across the experimental conditions in Experiment 1. Crossbars display means (±*SEM*); dots represent individual responses.

Procedure. In both the social-rewards condition and the no-social-rewards condition, participants read a short description:

> Tom is the owner of a travel company that builds and operates exclusive resorts on Islands in the Maldives. He was recently building a new resort on an island that was not yet explored by tourism. While Tom and his company only care about making as much money as possible, they had to clean up several beaches on this island, as pollution and littering is widespread in the Maldives and the tourists visiting the resorts of Tom's company demand clean beaches. They spend $100,000 for the cleanup and removed 30 tons of garbage, which greatly helped wildlife in the region and benefitted the relatively poor native inhabitants of the Islands.

In the social-rewards condition, this description was followed by the following sentences:

> Tom's company documented their cleanup efforts, which they use to advertise their resorts to ecologically oriented customers. Their campaign for the new resort has been a large success, and they already make $300,000 from bookings in the first year.

In the no-social-rewards (no-broadcasting) condition, this description was followed by these sentences instead:

> Outside of a small circle of Tom's colleagues who work on building the new resort, nobody knows about the cleanup efforts. Their new resort has been a large success, and they already make $300,000 from bookings in the first year.

In both the control and the control (no-broadcasting) conditions, participants read this short description:

> Tom is the owner of a travel company that builds and operates exclusive resorts on mountains in Alaska, deep in the mountains and far off from civilization. He was recently building a new resort in a remote valley that is not yet explored by tourism. While Tom and his company only care about making as much money as possible, they had to build a new road to the resort, as the tourists visiting the resorts of Tom's company demand high accessibility.
>
> They spend $100,000 on the road. Because the road only led to the resort, it did not provide any benefits (but also no problems) for others than the resort's personnel and visitors.

In the control condition, this description was followed by the following sentences:

> Tom's company documented their road-building efforts, which they use to advertise their resorts to customers who care about good accessibility. Their campaign for the new resort has been a large success, and they already make $300,000 from bookings in the first year.

In the control (no-broadcasting) condition, this description was followed by these sentences instead:

> Outside of a small circle of Tom's colleagues who work on building the new resort, nobody knows about the road building efforts. Their new resort has been a large success, and they already make $300,000 from bookings in the first year.

Measures. After reading a scenario, participants evaluated Tom on Likert scales ranging from 1 to 9 (all measures are adopted from Newman & Cain, 2014). Specifically, I used the same moral evaluations (Cronbach's α = .87 in this sample), deceptiveness judgments (Cronbach's α = .88 in this sample), and benefit judgments (Cronbach's α = .77 in this sample) as in Experiment 1. Participants also responded again to the two exploratory questions about how altruistic and how selfish they perceived Tom's behavior to be.

Attention check. After responding to the last survey question, participants responded to the following attention check: "As a final question for you, we want to know whether you are processing the information given in each question. For the question below, please click on the second choice from your left. How satisfied to you think Tom is with his business?" Responses were collected on a 9-point scale ranging from 1 (*not at all satisfied*) to 9 (*very satisfied*). All participants who did not respond with a 2 were excluded from the analysis. The rates of exclusion did not differ substantially between experimental conditions; see the Supplemental Material for details (Table S18), robustness analyses (Table S19), and figures underlining the consistency of the effects in the full data set (Figures S20-S21).

Participants. I recruited participants online using MTurk. A total of 355 people reached the attention-check question. I excluded participants who had no unique IP address and who did not respond correctly to the attention check; this produced a final sample of 215 participants (no demographic variables were collected).

Experiment 2a: results

Results support the new theory that the protection of social rewards drives the tainted-altruism effect, rather than the mere presence of selfish gains. As depicted in Figure 2a, participants rated the travel-company owner as less moral when the company disclosed its cleanup effort to customers, $M = 6.68$, $SEM = 0.24$, than when the company did not disclose its cleanup efforts, $M = 7.56$, $SEM = 0.19$, $t(109) = -2.78$, $p = .0065$, $d = 0.53$ (keeping the selfish gains, the amount of money the company makes, constant). In the two control conditions, broadcasting that the company engaged in a selfish action (building a road to a resort), $M = 6.09$, $SEM = 0.27$, or not, $M = 5.86$, $SEM = 0.24$, $t(102) = 0.62$, $p = .54$, $d = 0.12$, interaction: $F(1, 211) = 5.31$, $p = .022$, did not affect moral evaluations, ruling out the possibility that the results are explained by actors simply receiving a moral-reputation boost when they do not disclose their efforts. As in Experiment 1, the pattern of moral evaluations was mirrored by deceptiveness judgments (depicted in Fig. 2b): Actors who are seen as pretending to deserve social rewards are seen as more deceptive, $M = 4.82$, $SEM = 0.27$, than actors who keep information about the prosocial consequences of

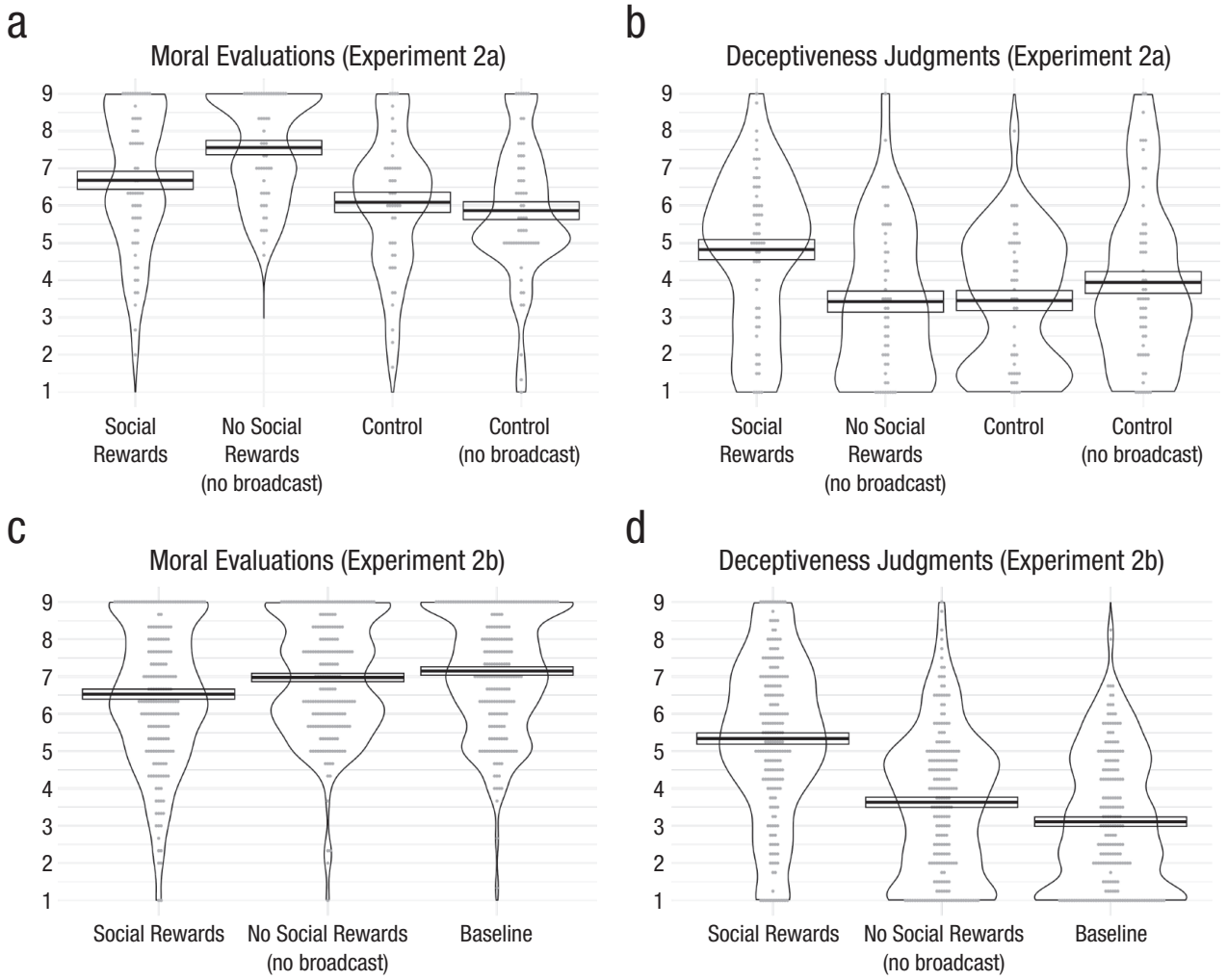

Fig. 2. Moral evaluations and deceptiveness judgments across experimental conditions in Experiments 2a and 2b. Experiment 2a's evaluations and judgments are illustrated in (a) and (b); Experiment 2b's evaluations and judgments are illustrated in (c) and (d). Crossbars display means (±*SEM*); dots represent individual responses.

their actions to themselves, $M = 3.43$, $SEM = 0.28$, $t(109) = 3.56$, $p = .0006$, $d = 0.68$. In the control conditions, when there are no prosocial consequences, disclosure ($M = 3.46$, $SEM = 0.27$) or nondisclosure ($M = 3.94$, $SEM = 0.29$) had no significant effect on deceptiveness, $t(102) = -1.19$, $p = .24$, $d = 0.24$, interaction: $F(1, 211) = 2.59$, $p = .11$.

Tom's selfishly motivated behavior in both control conditions (when he builds a road to the resort) is evaluated as less moral than in the conditions in which he creates prosocial benefits (when he cleans the beach). Of course, whether or not the classic tainted-altruism effect occurs, in which the tainted altruistic behavior is seen as worse than a neutral behavior, depends on the specific choice of the comparison point of the neutral-behavior control condition (and the benefits this behavior creates). Indeed, statistically controlling for the benefits Tom's behavior creates (see Table S14, Model 2, in the Supplemental Material), not only reveals that the overall interaction reported above is robust, $F(1, 210) = 4.81$, $p = .03$, but also that there is a classic tainted-altruism effect—the coefficients for either of the two control conditions (control: $b = 0.99$, $SE = 0.33$, $p = .003$; control (no broadcast): $b = 0.84$, $SE = 0.32$, $p = .009$) are larger than the coefficient for the social-rewards condition. In other words, when statistically removing differences in perceived benefits between the experimental conditions (i.e., controlling for the benefits judgments), then Tom in the prosocial conditions is seen as less moral than Tom in the control conditions.

In Experiment 2b, I made several changes to the scenario, aiming to use the same manipulation (i.e., broadcasting vs. not broadcasting the prosocial

behavior) and to replicate the original tainted-altruism effect without statistically controlling for the benefits. To be specific, I modified the scenarios to make the control condition with the neutral behavior as similar as possible to the social-rewards condition and the no-social-rewards condition, to keep benefits more balanced.

Experiment 2b: method

Design. Participants were randomly assigned to one of three between-subjects conditions: social rewards, no social rewards (no broadcasting), and baseline. They read a scenario about the owner of a beach resort who, in order to make money from tourists (selfish gains), is either cleaning up the beach (a behavior with prosocial consequences) or renovating the resort's kitchen (a behavior without prosocial consequences). Whereas the actor in Experiment 1 fully discloses his selfish motivation to prevent the impression that he is trying to reap undeserved social rewards, the actor in Experiment 2b simply does not broadcast the prosocial benefits of his behavior (i.e., not advertising that he cleaned up the beach).

Procedure. In both the social-rewards condition and the no-social-rewards condition, participants read a short initial description:

> Tom is the owner of a travel company that builds and operates exclusive resorts on islands in Southern Europe. He recently bought and renovated a resort on the Mediterranean coastline. While Tom and his company only care about making as much money as possible, they had to clean up a nearby beach, as pollution and littering is widespread in the area and the tourists visiting the resorts of Tom's company demand clean beaches.

In the baseline condition, participants read this short initial description:

> Tom is the owner of a travel company that builds and operates exclusive resorts on islands in Southern Europe. He recently bought and renovated a resort on the Mediterranean coastline. While Tom and his company only care about making as much money as possible, they had to thoroughly refurbish and modernize the entire kitchen and pool area to meet hygiene and operational standards, as the tourists visiting the resorts of Tom's company demand clean, functional, and up-to-date facilities.

These short initial descriptions stayed on the screen while participants had to correctly respond to the following two comprehension-check questions:

> *Please indicate which of the following two statements is correct or incorrect.*
>
> *Tom's company builds exclusive resorts in Asia*
>
> *Tom and his company only care about making as much money as possible*

In both the social-rewards condition and the no-social-rewards condition, the following sentence was added on the next screen:

> They spent $20,000 on the cleanup and removed 1.5 tons of garbage, which helped wildlife in the region and benefited the locals in the surrounding villages.

In the social-rewards condition, this sentence was followed by the following text:

> Tom's company documented their cleanup efforts, which they use to advertise their resorts to ecologically oriented customers. Their campaign for the new resort has been a large success, and they have already made $300,000 from bookings in the first year.

In the no-social-rewards (no-broadcasting) condition, the following text was added instead:

> Outside of a small circle of Tom's colleagues who work on building the new resort, nobody knows about the cleanup efforts. Their new resort has been a large success, and they have already made $300,000 from bookings in the first year.

In the baseline condition, the following sentence was added on the screen instead, after the comprehension-check questions:

> They spent $20,000 on the modernization, which greatly helped the resort's visual appeal and overall functionality while preserving its authenticity and character, benefiting both the resort's signature ambiance and its smooth operations.
>
> Their new resort has been a large success, and they have already made $300,000 from bookings in the first year.

Measures. After reading a scenario, participants evaluated Tom using Likert scales ranging from 1 to 9 (all measures are adopted from Newman & Cain, 2014). Specifically, I used the same moral evaluations (Cronbach's $\alpha = .88$ in this sample), deceptiveness judgments (Cronbach's $\alpha = .90$ in this sample), and benefit judgments (Cronbach's $\alpha = .63$ in this sample) as in Experiment 1 and 2a. Participants also responded again to the two exploratory questions about the extent to which they saw Tom's behavior as altruistic and selfish.

Attention check. Before being assigned to experimental conditions, participants had to correctly respond to two attention-check questions (randomly selected from a larger pool) to be admitted to the study.

Participants. I recruited participants online using MTurk using the CloudResearch panel. A total of 615 people completed the survey. Excluding participants without a unique IP address resulted in a final sample of 601 participants (M_{age} = 46.4 years, 49.25% female, 49.42% male, 0.67% nonbinary, 0.67% other/do not wish to disclose).

Experiment 2b: results

As in Experiment 2a, the results support the new theory: rather than the mere presence of selfish gains, it is the protection of social rewards that drives the tainted-altruism effect. Consistent with the preregistration (and as depicted in Fig. 2c), participants rated the travel-company owner as less moral when the company disclosed its cleanup effort to customers, $M = 6.53$, $SEM = 0.13$, than when the company did not disclose its cleanup efforts, $M = 6.98$, $SEM = 0.11$, $t(399) = 2.55$, $p = .011$, $d = 0.25$, and as less moral than in the baseline condition, in which he did not engage in any prosocial behavior in the first place, $M = 7.15$, $SEM = 0.11$, $t(397) = 3.53$, $p < .001$, $d = 0.35$. Regression analysis reveals that this pattern of results is robust toward controlling for benefit judgments (see Table S20 in the Supplemental Material). The selfish gains (the amount of money Tom's company made) were held constant across all three conditions. There was no significant difference between the no-social-rewards condition and the baseline condition, $t(400) = -1.09$, $p = .28$, $d = -0.11$.

As in Experiments 1 and 2a, deceptiveness judgments (depicted in Fig. 2d) inversely tracked the pattern of moral evaluations: actors who disclose the prosocial consequences of their actions, and who therefore are seen as pretending to deserve social rewards, are rated as more deceptive, $M = 5.34$, $SEM = 0.15$, than actors who keep information about the prosocial consequences of their actions to themselves, $M = 3.63$, $SEM = 0.14$, $t(399) = 8.41$, $p < .001$, $d = 0.84$, and as actors in the baseline condition, $M = 3.11$, $SEM = 0.13$, $t(397) = 11.34$, $p < .001$, $d = 1.14$. There was also a significant difference between the no-social-rewards condition and the baseline condition, $t(400) = 2.81$, $p = .005$, $d = 0.28$, indicating that participants perceive actors who keep the prosocial action secret as somewhat deceptive, but much less so than actors who claim more social rewards than they are seen as deserving.

Experiment 3

Experiment 3 compares eight different reward types to test the prediction that claiming undeserved social rewards (praise, liking, trust, or status), but not other nonsocial rewards (emotional or self-concept rewards), leads to moral derogation.

Experiment 3: method

Design. To explicitly compare evaluations of actors engaging in prosocial behaviors because they were motivated by different social and nonsocial rewards, I had each participant respond to eight different scenarios of an actor engaging in prosocial behavior, with eight different rewards, randomly paired with the scenarios. The order of the scenarios was randomized. Specifically, there were eight within-subject conditions. The first four conditions represented four types of social rewards: praise, liking, trust, and status. The next two conditions represent two types of nonsocial rewards: emotional rewards and self-concept rewards (O'Connor et al., 2020). To put these rewards into perspective, an *altruistic-rewards* condition and a *no-rewards* control condition were added.

To examine the link between the rewards or punishments for prosocial behavior and the expectation of future prosocial behavior of actors, this experiment included an additional dependent variable: participants' predictions of the actor's future prosocial behavior.

Procedure. The experiment was embedded in a lab session that also included several other tasks (in randomized order). When starting the experiment, participants were first shown an introduction screen that read, "On the following screens, you will read ten different scenarios, and answer a few questions about each of them." Participants then were presented with 10 short vignettes describing an actor and were asked to evaluate the actor on several dimensions.

The first and the last vignette were neutral behaviors, to familiarize participants with the task and the measures, and to reduce and potentially quantify order effects. The neutral behavior vignettes were "Rory purchased two classic novels in a bookstore" and "Aiden went on a hike in nature" (order counterbalanced). The

eight vignettes describing prosocial behavior (several of them inspired by vignettes used by Carlson & Zaki, 2018) were:

1. "Alex signed up to volunteer for the Red Cross once a week."
2. "Robin helped organize a fundraiser for the American Cancer Society."
3. "Carey helped the owner of a small flower shop collect his flowers from the sidewalk after a gust of wind had knocked over several buckets."
4. "Andy helped his neighbor unload a heavy set of power tools from a moving truck."
5. "Chris made a significant donation to a charity that provides insecticidal nets to fight against malaria."
6. "Charlie jumped on the train tracks to grab a briefcase that an older person just dropped, and returns it to them."
7. "Wyatt donated blood at a local clinic."
8. "Noah found a wallet on the sidewalk, looked inside to find the owner's name and address, and returns it to them."

The eight different benefits were as follows:
Social rewards:

Praise: "because he wants others to praise him."

Liking: "because he wants others to like him more."

Trust: "because he wants to appear more trustworthy to his friends and acquaintances."

Status: "because he wants to impress a group of colleagues at work, and to thereby increase his status in the workplace."

Nonsocial rewards:

Emotional rewards: "because he thinks this would make him feel very good."

Self-concept rewards: "because he wants to think of himself as a highly moral person, almost a role model."

Altruistic rewards: "because he wants to help people in need."

Control: "." (In the control condition, the sentence ended after the vignette that describes the behavior, without mentioning any reward.)

The vignettes were randomly paired with the rewards, resulting in 8 vignettes × 8 rewards = 64 combinations. For example, vignette 1 paired with the social rewards of praise would read "Alex signed up to volunteer for the Red Cross once a week, because he wants others to praise him." Every participant saw each vignette, and each reward, only once.

Measures. After reading each vignette, participants evaluated the behavior using Likert scales ranging from 1 to 9 (all measures were adopted from Newman & Cain, 2014). I used the same morality (Cronbach's $\alpha = .89$ in this sample), deceptiveness (Cronbach's $\alpha = .95$ in this sample), and benefit measures (Cronbach's $\alpha = .76$ in this sample) as in Experiments 1, 2a, and 2b.

In addition, three items were added to measure participants' predictions about the actor's future prosocial behavior: (a) How likely is Alex [the actor's name] to donate blood in the future? (b) How likely is Alex to give to a homeless person in the future? (c) How likely is Alex to donate used clothes in the future?

Participants. Participants were recruited in the behavioral lab of a large European business school as part of a longer session including several tasks (M_{age} = 22.7 years, 59% female). For readability and space, I report below the results from the full sample (401 participants) of the preregistered lab study; this includes 19 participants who erroneously participated before the submission of the preregistration and 187 participants who participated above and beyond the preregistered sample size of 200. (See the Supplemental Material Table S23 for a separate reporting of the preregistered subsample; results were consistent in both direction and significance with the full sample reported below.) This study received additional approval from the IESE Institutional Review Board for Research in Social Sciences and Humanities (ID No. IESE.2023.14).

Experiment 3: results

The results from Experiment 3, depicted in Figure 3, support the preregistered prediction that actors engaging in prosocial behavior to reap social rewards (such as praise, liking, trust, or status) are evaluated as being worse than neutral actors who did not engage in any prosocial behavior in the first place ($b = 1.44$, $p < .0001$) and as being worse than actors who were motivated by nonsocial selfish rewards ($b = 1.7$, $p < .0001$), such as emotional rewards and self-concept rewards. Even actors who were motivated by trust, the social reward that led to the highest moral evaluations, were seen as significantly less moral than actors motivated by their self-concept, the nonsocial reward that led to the lowest moral evaluations, $F(1, 382) = 32.38$, $p < .0001$.

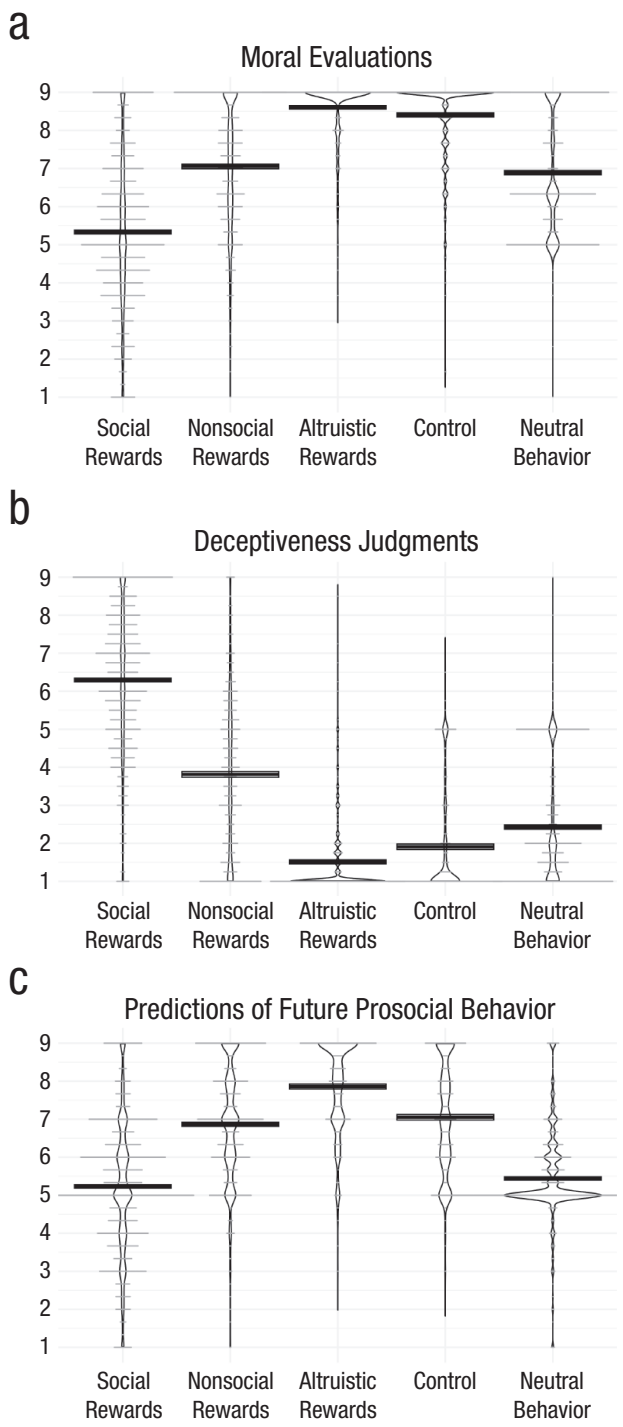

Fig. 3. Moral evaluations (a), deceptiveness judgments (b), and predictions of future prosocial behavior (c) across aggregated experimental conditions in Experiment 3. Crossbars display means (±*SEM*); dots represent individual responses.

Engaging in prosocial behavior motivated by nonsocial rewards, such as emotional rewards and self-concept rewards, was evaluated as more moral than or comparable to neutral behaviors—emotional rewards: $F(1, 387) = 18.28$, $p < .0001$; self-concept rewards: $F(1, 387) = 0.04$, $p = .8487$—and as worse than engaging in prosocial behavior motivated by altruistic rewards—emotional rewards: $F(1, 387) = 202.46$, $p < .0001$; self-concept rewards: $F(1, 387) = 403.02$, $p < .0001$—and as worse than actors in a control condition in which no information about the underlying motivation for the prosocial behavior was provided—emotional rewards: $F(1, 387) = 134.66$, $p < .0001$; self-concept rewards: $F(1, 387) = 333.21$, $p < .0001$.

The pattern of moral evaluations is mirrored by judgments about the deceptiveness of the actor's behavior (Table S26), robust to including judgments about the social benefits the actor's behavior creates as control variables (Table S25), and closely tracked by more practically relevant variables such as predictions of the actor's future prosocial behavior (see the Supplemental Material for details on these and other extended analyses, as well as additional tables and figures).

The within-subject design also allows for classifying participants on the basis of their answer patterns. These classifications require the assumption that only the benefits to the actors, but not the different scenarios (which were randomly assigned to rewards and randomized in their order), affected the moral evaluations.

To classify participants as answering as predicted by the new theory, I used two criteria: The first classification uses the strict criterion that each moral evaluation of actors motivated by social rewards would have to be lower than each of the moral evaluations of actors motivated by nonsocial rewards. The second classification uses the less strict criterion that the average moral evaluation of actors motivated by social rewards would have to be lower than the average moral evaluation of actors motivated by nonsocial rewards. About one third (30.7%) of participants satisfied the strict criterion that all their moral evaluations of actors who engaged in prosocial behavior motivated by social rewards were lower than all their moral evaluations of actors motivated by nonsocial rewards, and 89% of participants satisfied the less strict criterion that, on average, their moral evaluations of actors who engaged in prosocial behavior motivated by social rewards were lower than their moral evaluations of actors motivated by nonsocial rewards.

To classify participants as answering in line with the tainted-altruism effect, I also used the strict and the less strict criteria: 35.9% of participants satisfied the strict criterion that all their moral evaluations of actors who engaged in prosocial behavior motivated by social rewards were lower than all their moral evaluations of neutral actors, and 76.3% of participants satisfied the less strict criterion that, on average, their moral evaluations of actors who engaged in prosocial behavior

motivated by social rewards were lower than their moral evaluations of neutral actors.

Experiment 4

Experiment 4 directly manipulated counterfactuals to examine whether the new theory can explain why people rely on different counterfactuals for prosocial and selfish actors in the first place.

Experiment 4: method

Design. To explicitly examine which counterfactual people are thinking about, I experimentally manipulated whether or not people were reminded of a selfish counterfactual (following Newman & Cain, 2014, Experiment 3). Specifically, participants were randomly assigned to one of six between-subjects conditions. The first two conditions allowed for testing the key proposition again in a new context—an investment scenario—thereby replicating the pattern from Experiments 1, 2a, 2b, and 3. In the *social-rewards* condition, an actor engages in prosocial behavior for selfish motives without disclosing her selfish motives, whereas in the *no-social-rewards* condition the actor engages in the same action for the same motives but avoids being seen as pretending to deserve social rewards by fully disclosing her selfish motives.

In the next two conditions, the counterfactual that the actor could have pursued her selfish motives without engaging in any prosocial behavior is added at the end of the scenario, resulting in the *social-rewards counterfactual* condition and the *no-social-rewards counterfactual* condition.

If participants form their moral evaluations by intuitively comparing the actor with a prosocially motivated actor (i.e., an actor who qualifies for the social rewards), then reminding them that the actor could have been purely selfish should make their moral evaluations more favorable. If, however, participants already use a purely selfish actor as a counterfactual (because the actor already clarified that he or she does not claim any social rewards), then adding this purely selfish counterfactual should not affect the moral evaluations. In other words, if reminding participants of this counterfactual increases their morality ratings in the social-rewards counterfactual condition, but not in the no-social-rewards counterfactual condition, this would support the proposed social rewards protection theory. It would imply that in the no-social-rewards condition, people already rely on this selfish counterfactual, whereas in the social-rewards condition, people use a different counterfactual: They compare the focal actor with prosocially motivated others who qualify for the social rewards.

The remaining two conditions, the *purely selfish* condition and the *purely altruistic* condition, allow for understanding the evaluations in context. Specifically, the comparison with the purely selfish condition allows for replicating the classic tainted-altruism effect, and the purely altruistic condition allows for alleviating concerns about a potential ceiling effect.

Procedure. Participants were presented with vignettes describing an actor and were then asked to evaluate the actor on several dimensions. All vignettes included several sentences about the economic risk of the actor's behavior, to preemptively rule out any potential alternative explanation that would involve differences in the perceived riskiness of the actor's behavior across the experimental conditions. Moreover, by clarifying that the economic performance of the actor's company would be irrelevant for investors who care only about the economic risk of their investment, the scenario reduces the inherent deceptiveness of withholding the information about the predicted profits. Rather, when the social-rewards condition increases the perceived deceptiveness in the absence of a selfish counterfactual but does not increase the deceptiveness in the presence of a selfish counterfactual, then this deceptiveness judgment is the predicted outcome of the three-step process at the heart of the new theory. What causes the deceptiveness is not so much the omission of one piece of information but the signal that the actor deserves social rewards for the actions (i.e., by omitting that he or she will personally benefit). This signaling leads observers to spontaneously consider a prosocial counterfactual (and not a selfish counterfactual); this is why mentioning the selfish counterfactual reduces the moral derogation. In the absence of this signaling (in the no-social-rewards conditions), mentioning the very same counterfactual does not increase moral evaluations, consistent with the idea that participants already consider this counterfactual spontaneously.

For consistency between conditions, the vignettes were constructed from text blocks. The first four conditions (social rewards, no social rewards, social-rewards counterfactual, no-social-rewards counterfactual) all start with the same general-description text block. This text block is followed by either the social-rewards description text block or the no-social-rewards description text block, depending on the condition. In the two counterfactual conditions (social-rewards counterfactual, no-social-rewards counterfactual), the counterfactual description text block was added at the end. The remaining two conditions (purely selfish, purely altruistic) consist of a modified version of the general-description text block followed by a modified version of the social-rewards-description/no-social-rewards-description text block.

Specifically, the general-description text block read as follows:

> Alex is the owner of a small pharma company.
>
> She developed a few patents for a new drug, and wrote up a business plan. She predicts that this drug will generate a handsome profit of $7 million in the next 3 years. At the same time, she developed a plan to distribute this drug to poor people in developing countries for free. It will save several hundred thousand people from severe medical symptoms, and save at least a few thousand lives per year.
>
> To fund the necessary steps in the development of the drug, Alex needs to raise money. To do so, she invites a small circle of seven rich investors to a conference. While some of these investors are mainly interested in the financial consequences of their investments, others might also care about the social consequences their investments bring about.
>
> The investment is risk-free, as Alex inherited several real estate objects that she provides as a guarantee. Investors don't get a share of the company's profits. Rather, they lend Alex the money for a fixed interest rate. In other words, investors don't risk their money, and their return on investment is not affected by whether or not Alex's company becomes successful and makes profits.

The social-rewards-description text block reads as follows:

> When Alex pitches the investment opportunity, she stresses to all potential investors the plan to distribute the drug for free to poor people in developing countries (and how the investments would help achieving that). At the same time, she does not mention the profits she predicts the drug will generate for her company. In the end, Alex receives the full funding she needs, as three of the potential investors decide to lend her money.

The no-social-rewards-description text block reads as follows:

> When Alex pitches the investment opportunity, she gives all potential investors a detailed overview of the profits the drug will generate for her company. At the same time, she stresses the plan to distribute the drug for free to poor people in developing countries (and how the investments would help [in] achieving that). In the end, Alex receives the full funding she needs, as three of the potential investors decide to lend her money.

The counterfactual-description text block reads as follows:

> Keep in mind that if Alex wanted to, she could have planned the future of her pharma company without thinking about and developing the plan to distribute her drug to poor people in developing countries for free.

In the purely selfish condition, the general-description and no-social-rewards-description text blocks were modified as follows:

> Alex is the owner of a small pharma company.
>
> She developed a few patents for a new drug, and wrote up a business plan. She predicts that this drug will generate a handsome profit of $7 million in the next 3 years.
>
> To fund the necessary steps in the development of the drug, Alex needs to raise money. To do so, she invites a small circle of seven rich investors to a conference. While some of these investors are mainly interested in the financial consequences of their investments, others might also care about the social consequences their investments bring about.
>
> The investment is risk-free, as Alex inherited several real estate objects that she provides as a guarantee. Investors don't get a share of the company's profits. Rather, they lend Alex the money for a fixed interest rate. In other words, investors don't risk their money, and their return on investment is not affected by whether or not Alex's company becomes successful and makes profits.
>
> When Alex pitches the investment opportunity, she gives all potential investors a detailed overview of the profits the drug will generate for her company. In the end, Alex receives the full funding she needs, as three of the potential investors decide to lend her money.

In the purely altruistic condition, the general-description and social-rewards-description text blocks were modified as follows:

> Alex is the owner of a small pharma company.
>
> She developed a few patents for a new drug, and a plan to distribute this drug to poor people in developing countries for free. It will save several hundred thousand people from severe medical symptoms, and save at least a few thousand lives per year.
>
> To fund the necessary steps in the development of the drug, Alex needs to raise money. To do so, she invites a small circle of seven rich investors to a conference. While some of these investors are mainly interested in the financial consequences of their investments, others might also care about the social consequences their investments bring about.
>
> The investment is risk-free, as Alex inherited several real estate objects that she provides as a guarantee. Investors don't get a share of the company. Rather, they lend Alex the money for a fixed interest rate. In other words, investors don't risk their money, and their return on investment is not affected by whether or not Alex's company becomes successful. When Alex pitches the investment opportunity, she gives all potential investors a detailed overview of the plan to distribute the drug for free to poor people in developing countries. In the end, Alex receives the full funding she needs, as three of the potential investors decide to lend her money.

Measures. After reading a vignette, participants evaluated Alex using Likert scales ranging from 1 to 9 (all measures were adopted from Newman & Cain, 2014). Specifically, I used the same morality (Cronbach's $\alpha = .95$ in this sample), deceptiveness (Cronbach's $\alpha = .92$ in this sample), and benefit measures (Cronbach's $\alpha = .80$ in this sample) as in Experiments 1, 2a, 2b, and 3.

The same exploratory measures were also included (liking, trust, selfishness, altruism). In addition, three items were added to measure participants' beliefs about (a) whether Alex's success in acquiring the funding was influenced by her prosocial plan to distribute the drug for free to people who would need it, but likely could not afford it, as well as (b) the extent to which Alex's investors cared about social consequences and (c) the extent to which Alex's investors cared about the financial consequences of their investments. See the Supplemental Material for the exact question wording and detailed analyses of these measures.

Attention check. After responding to the last survey question, participants responded to the following attention check: "As a final question for you, we want to know whether you are processing the information given in each question. For the question below, please click on the first choice *not at all satisfied*. How satisfied do you think is Alex with her business?" Responses were collected using a 9-point scale ranging from 1 (*not at all satisfied*) to 9 (*very satisfied*). As preregistered, all participants who did not respond with a 1 (*not at all satisfied*) were excluded from the sample. The rates of exclusion did not differ substantially between experimental conditions. See the Supplemental Material for details (Table S39), robustness analyses (Table S40), and figures underlining the consistency of the effects in the full data set (Figures S46-S49).

Participants. For readability and space, I report below only the results of a preregistered replication study; see the Supplemental Material for a full reporting of the original study ($N = 466$; results are consistent in direction and significance with the replication reported below except that some of the differences between the no-social-rewards condition and the purely selfish and the purely altruistic control conditions did not reach statistical significance in the less well-powered original study).

I recruited participants online using MTurk. A total of 1,813 people reached the attention-check question, and consistent with the preregistration I analyzed the data of all participants with unique IP addresses who did not fail the first attention check. This resulted in a final sample of 1,195 participants (M_{age} = 38.9 years, 59% female).

Experiment 4: results

The results from Experiment 4, depicted in Figure 4, again replicate the main findings from Experiments 1, 2a, 2b, and 3, and clarify how the prior explanation for the tainted-altruism effect (different counterfactuals) relates to social rewards protection theory. As predicted and preregistered, Alex was rated as less moral when she kept her selfish benefits to herself and thus could be seen as pretending to deserve social rewards for her behavior (in the social-rewards condition, $M = 6.32$, $SEM = 0.15$) than when she acts in a purely selfish way (purely selfish condition, $M = 7.45$, $SEM = 0.11$), $t(398) = 6.14$, $p < .001$, $d = 0.61$, again replicating the classic tainted-altruism effect.

Also as predicted and preregistered, when Alex discloses her selfish benefits (in the no-social-rewards

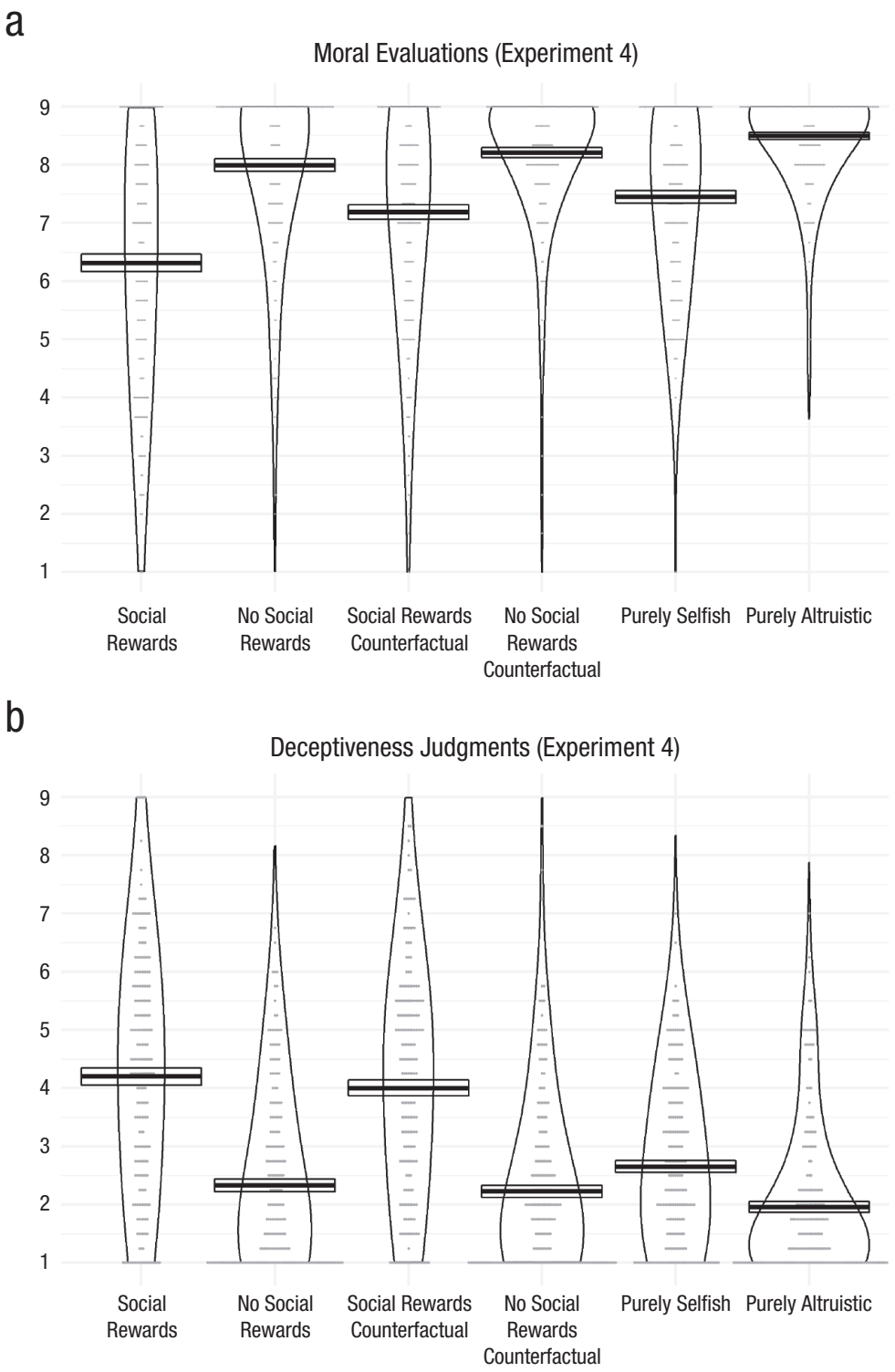

Fig. 4. Moral evaluations (a) and deceptiveness judgments (b) across experimental conditions in Experiment 4. Crossbars display means (±*SEM*); dots represent individual responses.

condition, M = 7.99, SEM = 0.11), then the tainted-altruism effect goes away. Here it even reverses, and Alex is rated as more moral than in the purely selfish condition, $t(394) = 3.64$, $p = .0003$, $d = 0.37$. Furthermore, again as predicted and preregistered, adding the counterfactual reduced the tainted-altruism effect: If

Alex is seen as pretending to deserve social rewards for her prosocial behavior, by stressing the prosocial benefits to investors with prosocial preferences (social-rewards condition: $M = 6.32$, $SEM = 0.15$), then people will compare this actor to a counterfactual actor with a prosocial motivation. In this situation, reminding people that the actor could have acted in a purely selfish way (social-rewards counterfactual condition: $M = 7.19$, $SEM = 0.13$) increases moral evaluations, $t(398) = 4.40$, $p < .001$, $d = 0.44$, conceptually replicating the (proximal) counterfactuals mechanisms proposed by Newman and Cain (2014). In contrast, when Alex is not seen as pretending to deserve social rewards (because she does not stress the prosocial consequences of the company's new drug; no-social-rewards condition, $M = 7.99$, $SEM = 0.11$), then people have no reason to compare her with a prosocially motivated counterfactual actor. Consequently, reminding people of the purely selfish counterfactual (no-social-rewards counterfactual condition: $M = 8.21$, $SEM = 0.09$) has no significant effect, $t(394) = 1.59$, $p = .11$, $d = 0.16$, because people likely already have such a counterfactual in mind. Experiment 4 strongly supports the preregistered interaction hypothesis that the effect of the counterfactual is moderated by the actor being seen as pretending to deserve social rewards—interaction: $F(1, 1189) = 8.79$, $p = .0031$.

As in Experiment 1, participants also responded to several exploratory measures (for detailed analyses, figures, and regression tables, see the Supplemental Material). The pattern of moral evaluations again extends to the measures of liking (Table S34b) and trusting (Table S36b), again underlining the generalizability of the results to these more practically relevant dimensions. The only difference emerged for the interaction between the presence of social rewards and of the selfish counterfactual. To be specific, the presence of the counterfactual seems to increase the like and trust ratings not just in the social-rewards conditions but also in the no-social-rewards conditions. This pattern of a significant interaction on moral evaluations but nonsignificant interactions for like and trust ratings was consistent across the original study and the preregistered replication. Future research could explore this divergence. One might speculate that like and trust ratings could be more subjective than moral evaluations, leading participants to feel a stronger demand effect from the counterfactual manipulation, which then could have operated also in the no-social-rewards condition.

Moreover, as in Experiments 1, 2a, 2b, and 3, the results from the pattern of moral evaluations between the experimental conditions are robust toward including the exploratory measures of the extent to which her actions were rated as selfish and altruistic as control variables. This supports the hypothesis that what drives the moral evaluations is whether Alex is seen as pretending to deserve social rewards for her actions, rather than the selfishness or altruism of her actions alone.

When reading the scenarios of Experiment 4, participants likely formed beliefs about whether Alex's success in acquiring the funding was influenced by her prosocial plan to distribute the drug for free to people who would need it but likely could not afford it. Similarly, they likely formed beliefs about the extent to which Alex's investors cared about the social and financial consequences of their investments. Participants were asked explicitly about these three beliefs (see the Supplemental Material for the exact question wording, detailed analyses, figures, and regression tables), with the goal of ruling out the possibility that participants in the different experimental conditions understood the scenario in different ways and thus formed different beliefs, and that these different beliefs (rather than the new social rewards protection theory) could explain the differences in moral evaluations between experimental conditions. All results are robust toward the inclusion of these variables as control variables (Table S38), successfully ruling out this potential alternative explanation.

Experiment 5

Extending this investigation, Experiment 5 moves from moral evaluations to downstream consequences such as trust, and from hypothetical scenarios to monetary stakes in an incentivized trust game.

Experiment 5: method

Design. Instead of measuring and comparing the moral evaluation of prosocial actors who received or did not receive social rewards and who fully disclosed or did not disclose these rewards (as in Experiments 1–4), Experiment 5 focuses on tangible downstream consequences. Therefore, the main dependent measure in Experiment 5 is how much senders trust receivers, using a trust game (also known as a sender-receiver game). Specifically, senders received a budget of \$0.30 and made a decision about how much, if any, of that money to send to receivers. The amount of money the senders sent was the main dependent variable, as it operationalizes how much the senders trust the receivers. This amount of money was tripled, and the receivers decided how much of it to return. By putting financial stakes behind the trust decision, Experiment 5 tested the predictions of the new theory on consequential, fully incentivized decisions. Specifically, before senders decided how much they trusted receivers, they learned about the earlier behavior of the receivers (i.e., which type the receiver they were paired with resembles).

To create different types of receivers that would represent different experimental conditions, receivers made a decision about a donation in a first step and then made a decision about a message to send to the senders in a second step. In the first step, receivers are given a bonus payment and offered the opportunity to donate this bonus payment to a charity. If they decline (and thus decide to keep the money), some receivers are randomly assigned to be offered another (larger) bonus payment if they revise their initial decision and make the donation. This leads to three behavioral types of receivers: *donation decliners* (those keep the initial bonus), *donators* (who donate the initial bonus right away), and *selfish donators* (who initially decline to donate, but donate after being offered the bonus).

In the second step, receivers were asked to select a message they wanted to send to the sender with whom they were paired. Donation decliners could choose between a simple message wishing the sender a nice day and a full-disclosure message mentioning that they declined to make the donation. Donators could choose between the same simple message and a message announcing that they had made a donation. Selfish donators could choose between the same donation-announcement message and a full-disclosure message mentioning that they had made the donation to receive the bonus.

On the basis of their donation and message decisions, receivers were assigned to be presented to senders in one of the six following ways, which correspond to six experimental conditions for the senders:

1. Social rewards: Receivers are selfish donators who sent the donation announcement message.
2. No social rewards (full disclosure): Receivers are selfish donators who sent the full disclosure message.
3. Purely selfish: Receivers are donation decliners and sent the simple message wishing the sender a nice day.
4. Selfish (full disclosure): Receivers are donation decliners and sent the full disclosure message.
5. Purely altruistic: Receivers are donators who sent the donation-announcement message.
6. Baseline: No donation information about the receiver is given, and receivers send the simple message wishing the sender a nice day.

Procedure. After reading the instructions to the sender-receiver game, correctly responding to comprehension checks, and responding to a question measuring initial trust (how much of their \$0.30 bonus they were willing to send to the receivers in the absence of any information about them), senders were randomly assigned to one of these six experimental conditions. Then they were informed about the donation decision and the message from the receiver, dependent on their assigned experimental condition. They were then asked another set of comprehension checks about this information (see the Supplemental Material for the full text of the instructions and comprehension checks).

Measures. The main dependent variable is how much the senders' trust in the receivers changes on the basis of the donation information and the message. This measure is calculated by subtracting the baseline trust (i.e., the number of cents sent to the receivers in the absence of any information about them) from the informed trust (i.e., the number of cents sent to the receivers after being informed about the receivers' donation decision and message).

Attention checks. Two attention checks were randomly selected from two separate pools of attention checks. The first attention check was distributed across two pages. On the first page, participants were asked to answer the question on the following page by responding with one of six different possible responses (either given verbatim or indirectly; e.g., "the number of days in April"). The question on the second page asked for their favorite book, but referenced back to the first page. The second attention check asked participants to identify elements in a group of words that did not refer to animals in one case, or did not name American states in the other case (with false responses constructed out of elements of existing state names). Participants had to pass both attention checks to enter the survey. Participants who failed at least one attention check were asked to return the Human Intelligence Task on MTurk ($N = 250$).

Participants (senders). A total of 1,536 people completed the survey in the sender role. Consistent with the preregistration, all data from participants with unique IP addresses (who did not fail one or both of the two attention checks; see above) was analyzed, resulting in a final sample of 1,522 participants (M_{age} = 40.6 years, 57% female). Senders' sample size was preregistered at 200 per condition, but because of operational challenges with recruiting the corresponding receivers who could be assigned to some sender conditions (on the basis of their decisions), I had to recruit more senders (see the Supplemental Material for details).

Experiment 5: results

Figure 5 shows how much more or less money (out of a budget of \$0.30) participants (senders) sent to others (receivers) about whom they obtained some

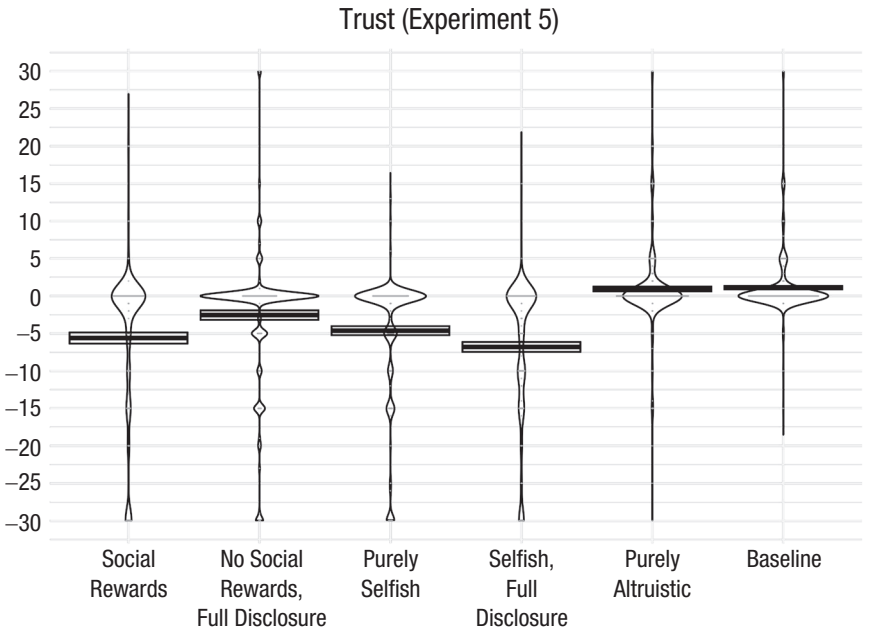

Fig. 5. Trust across experimental conditions in Experiment 5. Crossbars display means (±*SEM*); dots represent individual responses.

information (relative to how much money they sent to a receiver before they obtained this information). As predicted, and consistent with the preregistration, senders trusted receivers who donated only after being offered bonus money for doing so and who mentioned their donation in the message (in the social-rewards condition: $M = -5.61$, $SEM = 0.74$) less than receivers about whom they had no donation information (in the baseline condition: $M = 1.11$, $SEM = 0.23$), $t(527) = -10.36$, $p < .0001$, $d = 0.93$, and less than receivers who declined to make the donation (in the purely selfish condition: $M = -4.63$, $SEM = 0.6$), $t(415) = -1.04$, $p = .3$, $d = 0.1$. Although the latter difference (with the purely selfish condition) was not statistically significant, equivalence testing showed that receivers in the social-rewards condition were trusted equally or less than receivers in the purely selfish condition, by ruling out that they were trusted more: A difference of one cent or more could be rejected ($t = 2.093$, $p = .0185$, using the two one-sided tests procedure). Taken together, these results replicate the classic tainted-altruism effect: Receivers who made a donation but received selfish gains for doing so were trusted less than receivers about whom senders had no information (concerning their donation behavior), and as much as or less than receivers for whom the senders knew that they decided against making the donation. Outside of the lab, people rarely have information that allows them to place others on the spectrum ranging from actively deciding against an action with prosocial consequences to passively not thinking of taking such an action in the first place. The two comparisons (i.e., comparing the social-rewards condition with the purely selfish and baseline conditions) seem to capture a large part of this spectrum. The size of the tainted-altruism effect might therefore depend on situation-specific beliefs about the likelihood that others' failure to engage in behavior causing prosocial consequences was driven by an active decision (purely selfish condition) or not (baseline condition).

The key prediction of social rewards protection theory—that fully disclosing the selfish gains removes the possibility that actors are seen as pretending to deserve social rewards for which they do not qualify—receives strong support: Consistent with the preregistration, receivers who make a donation only after being offered bonus money for doing so are trusted more if they fully disclose their bonus (no-social-rewards full-disclosure condition: $M = -2.55$, $SEM = 0.63$) than if they do not fully disclose it (social-rewards condition: $M = -5.61$, $SEM = 0.74$), $t(398) = -3.13$, $p = .002$, $d = 0.31$.

To test the possibility that full disclosure in itself causes the increase in trust, and that therefore this increase in trust does not stem from removing the possibility of being seen as pretending to deserve social rewards for which one does not qualify, a comparison between the trust in receivers who declined to donate (purely selfish condition: $M = -4.63$, $SEM = 0.6$) and receivers who declined to donate and fully disclosed their decision (selfish full-disclosure condition: $M = -6.8$,

SEM = 0.66) was preregistered. As predicted, in this case, full disclosure did not increase, but rather decreased, trust, $t(431) = 2.44$, $p = .02$, $d = 0.23$, ruling out the possibility that people simply receive a reputational boost from full disclosure itself and that such a boost drives the tainted-altruism effect. This preregistered interaction (selfish vs. social rewards × full disclosure), $F(1, 1516) = 23.02$, $p < .0001$, strongly supports the key hypothesis that preventing people from being seen as deserving social rewards (through full disclosure) eliminates the tainted-altruism effect.

The purely altruistic condition was added to put the results in context. Moreover, in order to run the trust game without deception, there would of course be receivers who decide to donate, and, therefore, the corresponding senders would have to be recruited in any case. Given the strategic nature of the trust game about which both senders and receivers were informed first, it is perhaps unsurprising that compared to the baseline condition (in which senders were not informed about the donation opportunity in the first place: $M = 1.11$, $SEM = 0.23$), senders did not trust receivers more when they decided to donate (in the purely altruistic condition: $M = 0.93$, $SEM = 0.32$), $t(687) = 0.45$, $p = .65$, $d = 0.03$.

Discussion

Results from six experiments using personal and organizational contexts, as well as third-party evaluations and dyadic trust measured through incentivized games, support the proposed social rewards protection theory. Because people reserve social rewards for costly prosocial behavior, they see actors who claim such rewards without incurring costs as deceptive, and they morally derogate them. In contrast, actors who receive other nonsocial rewards, such as emotional or self-concept rewards, are not punished. A priori, prosocial actors (including selfishly motivated ones) are seen as claiming social rewards, yet when they clarify that they do not claim social rewards—for instance, by creating transparency about their behavior's benefits and thus their likely motivation—they are no longer punished (but they are also not rewarded as much as actors who engage in costly prosocial behavior).

The new theory provides a functional, more ultimate explanation for the tainted-altruism effect; establishes under which circumstances it occurs; and clarifies that attributions of self-interest, and the potentially ensuing moral derogation, are not insurmountable obstacles to prosocial behavior. Absent this explanation, it may seem puzzling that people engage in prosocial behavior, because prior research has shown that people tend to attribute self-interested motivations to almost any behavior (Miller, 1999)—including prosocial acts (Critcher & Dunning, 2011; Heyman et al., 2014)—and longstanding scholarly and philosophical traditions question whether acts of pure altruism even exist (Bentham, 1789; Hobbes, 1651; Kant, 1785; Nietzsche, 1878; but see Batson, 2011).

The practical implications are straightforward: Prosocial behavior does not bear any risk of derogation, as long as its motivation is transparent. There is also no quick fix for the likes of Dan Pallotta, and no easy way out of the nonprofit starvation cycle (Gregory & Howard, 2009), as long as the same expenses—seen as necessary overhead by charities themselves—are perceived as selfish gains (to actors working at the charities) by potential donors (Gneezy et al., 2014). This explains the great lengths to which leaders go to convince themselves and others that their (and their organizations') prosocial behavior is authentic and not driven by ulterior motives (Gershon et al., 2020; Savary et al., 2020; Silver et al., 2021; Wagner et al., 2009; Yoon et al., 2006).

Rather than being a psychological bias of donors, the tendency of seeing social rewards as reserved for actors engaging in costly prosocial behavior seems adaptive for promoting future prosocial behavior; it also seems aligned with evolutionary game theory models of altruism (see also Burum et al., 2020). In addition to reconciling the tainted-altruism effect with such models (and with the high prevalence of altruism in society), the proposed theory also builds on and bears potential for integrating related phenomena. For instance, cooperators who are spontaneous (Jordan et al., 2016), emotion-driven (Levine et al., 2018), and ignore information about costs and benefits (Hoffman et al., 2015) might be socially rewarded for signaling that they engage in prosocial behavior regardless of its costs, as they forgo calculated cost-benefit evaluations.

The reliance on U.S. MTurk workers and European lab participants might limit the generalizability of the present experiments. Consequently, an important future research direction is to investigate to what extent the presumption that prosocial actors are fishing for social rewards varies across cultures. In other cultures, prosocial actors might not be seen as claiming social rewards unless they actively ask for them.

Conclusion

Questions about the nature of altruism are timeless, and often traverse the boundaries between academic disciplines. Not just scientists and philosophers, but also the public at large thinks about what motivates prosocial actors, and, in turn, makes decisions on how to treat such actors. How people make such decisions has long captured the attention of psychologists. They were quick to point out some counterintuitive peculiarities that make the existence of prosocial behavior almost

seem puzzling, such as the tainted altruism effect and overhead aversion. This paper develops an integrative theory that accounts for these counterintuitive peculiarities, while reconciling them with longstanding research traditions—such as evolutionary game theory—as well as with the high prevalence of altruistic behavior in society. At its core, this integrative theory holds that selfish altruism becomes tainted because selfish altruists are seen as claiming social rewards that are reserved for pure altruists. Six experiments supported this theory, showing that selfish altruism is no longer tainted when selfish actors clarify that they are not claiming social rewards.

Transparency

Action Editor: Mark Brandt
Editor: Patricia J. Bauer
Author Contributions

Sebastian Hafenbrädl: Conceptualization; Data curation; Funding acquisition; Investigation; Methodology; Project administration; Visualization; Writing – original draft; Writing – review & editing.

Declaration of Conflicting Interests

The author declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was supported by Spanish Ministerio de Ciencia e Innovation Grant No. PID2019-103897GB-I00.

Artificial Intelligence

Except for spell-checking by Grammarly and Microsoft Word, no artificial intelligence-assisted technologies were used in this research or the creation of this article.

Ethics

This research received approval from a local ethics board (Yale ID No. 1605017813; IESE ID No. 2023.14).

Open Practices

Open practices for this article are described in the Research Transparency Statement section, which appears at the end of the Introduction section in the main text.

ORCID iD


Sebastian Hafenbrädl https://orcid.org/0000-0002-5148-766X

Acknowledgments

I am deeply grateful to Associate Editor Mark Brandt for his exceptional guidance and to the three anonymous reviewers for their helpful comments throughout the review process. I am indebted to Daylian M. Cain for enlightening discussions that provided the initial spark for this project, and to Thomas Fischer, Ulrich Hoffrage, Daniel Waeger, and Jan K. Woike for insightful feedback on earlier versions. I also benefited from comments from seminar participants at the University of Plymouth, the Universitat de les Illes Balears, IESE, and the ABC Workshop in Rome. Finally, I thank Eugenia Bajet, Mateja Drev, Carla Morales, and Nuria Saez Valle for excellent research assistance.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976251398454

References

Alcala, V., Johnson, K., Steele, C., Wu, J., Zhang, D., & Pashler, H. (2022). The tainted altruism effect: A successful pre-registered replication. *Royal Society Open Science*, *9*(1), Article 211152. https://doi.org/10.1098/rsos.211152

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*, *99*(1), 544–555. https://doi.org/10.1257/aer.99.1.544

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396. https://doi.org/10.1126/science.7466396

Bai, F., Ho, G. C. C., & Yan, J. (2020). Does virtue lead to status? Testing the moral virtue theory of status attainment. *Journal of Personality and Social Psychology*, *118*(3), 501–531. https://doi.org/10.1037/pspi0000192

Batson, C. D. (2011). *Altruism in humans*. Oxford University Press.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, *96*(5), 1652–1678. https://doi.org/10.1257/aer.96.5.1652

Bentham, J. (1948). *An introduction to the principles of morals and legislation*. Hafner. (Original work published 1789).

Berman, J. Z., & Silver, I. (2022). Prosocial behavior and reputation: When does doing good lead to looking good? *Current Opinion in Psychology*, *43*, 102–107. https://doi.org/10.1016/j.copsyc.2021.06.021

Burum, B., Nowak, M. A., & Hoffman, M. (2020). An evolutionary explanation for ineffective altruism. *Nature Human Behaviour*, *4*(12), 1245–1257. https://doi.org/10.1038/s41562-020-00950-4

Carlson, R. W., Bigman, Y. E., Gray, K., Ferguson, M. J., & Crockett, M. J. (2022). How inferred motives shape moral judgements. *Nature Reviews Psychology*, *1*(8), 468–478. https://doi.org/10.1038/s44159-022-00071-x

Carlson, R. W., & Zaki, J. (2018). Good deeds gone bad: Lay theories of altruism and selfishness. *Journal of Experimental Social Psychology*, *75*(Suppl. C), 36–40. https://doi.org/10.1016/j.jesp.2017.11.005

Cassar, L., & Meier, S. (2021). Intentions for doing good matter for doing well: The negative effects of prosocial incentives. *The Economic Journal*, *131*(637), 1988–2017. https://doi.org/10.1093/ej/ueaa136

Critcher, C. R., & Dunning, D. (2011). No good deed goes unquestioned: Cynical reconstruals maintain belief in the power of self-interest. *Journal of Experimental Social Psychology*, *47*(6), 1207–1213. https://doi.org/10.1016/j.jesp.2011.05.001

Critcher, C. R., Helzer, E. G., & Tannenbaum, D. (2020). Moral character evaluation: Testing another's moral-cognitive machinery. *Journal of Experimental Social Psychology*, *87*, Article 103906. https://doi.org/10.1016/j.jesp.2019.103906

Davis, I., Carlson, R., Dunham, Y., & Jara-Ettinger, J. (2023). Identifying social partners through indirect prosociality: A computational account. *Cognition*, *240*, Article 105580. https://doi.org/10.1016/j.cognition.2023.105580

Effron, D. A., O'Connor, K., Leroy, H., & Lucas, B. J. (2018). From inconsistency to hypocrisy: When does "saying one thing but doing another" invite condemnation? *Research in Organizational Behavior*, *38*, 61–75. https://doi.org/10.1016/j.riob.2018.10.003

Flammer, C., & Luo, J. (2017). Corporate social responsibility as an employee governance tool: Evidence from a quasi-experiment. *Strategic Management Journal*, *38*(2), 163–183. https://doi.org/10.1002/smj.2492

Flynn, F. J. (2003). How much should I give and how often? The effects of generosity and frequency of favor exchange on social status and productivity. *Academy of Management Journal*, *46*(5), 539–553. https://doi.org/10.5465/30040648

Gershon, R., Cryder, C., & John, L. K. (2020). Why prosocial referral incentives work: The interplay of reputational benefits and action costs. *Journal of Marketing Research*, *57*(1), 156–172. https://doi.org/10.1177/0022243719888440

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior*, *24*(3), 153–172. https://doi.org/10.1016/S1090-5138(02)00157-5

Gneezy, U., Keenan, E. A., & Gneezy, A. (2014). Avoiding overhead aversion in charity. *Science*, *346*(6209), 632–635. https://doi.org/10.1126/science.1253932

Grant, A. M., & Gino, F. (2010). A little thanks goes a long way: Explaining why gratitude expressions motivate prosocial behavior. *Journal of Personality and Social Psychology*, *98*(6), 946–955. https://doi.org/10.1037/a0017935

Gregory, A. G., & Howard, D. (2009). The nonprofit starvation cycle. *Stanford Social Innovation Review*, *7*(4), 49–53. https://doi.org/10.48558/6K3V-0Q70

Harbaugh, W. T. (1998). The prestige motive for making charitable transfers. *American Economic Review*, *88*(2), 277–282.

Hardy, C. L., & Vugt, M. V. (2016). Nice guys finish first: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin*, *32*(10), 1402–1413. https://doi.org/10.1177/0146167206291006

Heyman, G., Barner, D., Heumann, J., & Schenck, L. (2014). Children's sensitivity to ulterior motives when evaluating prosocial behavior. *Cognitive Science*, *38*(4), 683–700. https://doi.org/10.1111/cogs.12089

Hobbes, T. (1651). *Leviathan; or, the matter, forme, and power of a common-wealth, ecclesiasticall and civill.* Andrew Crooke.

Hoffman, M., Yoeli, E., & Navarrete, C. D. (2016). Game theory and morality. In T. K. Shackelford & R. D. Hansen (Eds.), *The evolution of morality* (pp. 289–316). Springer International Publishing. https://doi.org/10.1007/978-3-319-19671-8_14

Hoffman, M., Yoeli, E., & Nowak, M. A. (2015). Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences*, *112*(6), 1727–1732. https://doi.org/10.1073/pnas.1417904112

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, *113*(31), 8658–8663. https://doi.org/10.1073/pnas.1601280113

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Science*, *28*(3), 356–368. https://doi.org/10.1177/0956797616685771

Kant, I. (1785). *Groundwork of the metaphysic of morals*. [Grundlegung zur Metaphysik der Sitten]. Johann Friedrich Hartknoch.

Kawamura, Y., Ohtsubo, Y., & Kusumi, T. (2021). Effects of cost and benefit of prosocial behavior on reputation. *Social Psychological and Personality Science*, *12*(4), 452–460. https://doi.org/10.1177/1948550620929163

Kirgios, E. L., Chang, E. H., Levine, E. E., Milkman, K. L., & Kessler, J. B. (2020). Forgoing earned incentives to signal pure motives. *Proceedings of the National Academy of Sciences*, *117*(29), 16891–16897. https://doi.org/10.1073/pnas.2000065117

Kristof, N. (2008, December 24). The sin in doing good deeds. *The New York Times*. https://www.nytimes.com/2008/12/25/opinion/25kristof.html

Levine, E. E., Barasch, A., Rand, D., Berman, J. Z., & Small, D. A. (2018). Signaling emotion and reason in cooperation. *Journal of Experimental Psychology: General*, *147*(5), 702–719. https://doi.org/10.1037/xge0000399

Lin-Healy, F., & Small, D. A. (2012). Cheapened altruism: Discounting personally affected prosocial actors. *Organizational Behavior and Human Decision Processes*, *117*(2), 269–274. https://doi.org/10.1016/j.obhdp.2011.11.006

Lin-Healy, F., & Small, D. A. (2013). Nice guys finish last and guys in last are nice: The clash between doing well and doing good. *Social Psychological and Personality Science*, *4*(6), 692–698. https://doi.org/10.1177/1948550613476308

Makov, T., & Newman, G. E. (2016). Economic gains stimulate negative evaluations of corporate sustainability initiatives. *Nature Climate Change*, *6*, 844–846. https://doi.org/10.1038/nclimate3033

McCullough, M. E., Kimeldorf, M. B., & Cohen, A. D. (2008). An adaptation for altruism: The social causes, social effects, and social evolution of gratitude. *Current Directions in*

Psychological Science, *17*(4), 281–285. https://doi.org/10.1111/j.1467-8721.2008.00590.x

Miller, D. T. (1999). The norm of self-interest. *American Psychologist*, *54*(12), 1053–1060. https://doi.org/10.1037/0003-066X.54.12.1053

Newman, G. E., & Cain, D. M. (2014). Tainted altruism: When doing some good is evaluated as worse than doing no good at all. *Psychological Science*, *25*(3), 648–655. https://doi.org/10.1177%2F0956797613504785

Nietzsche, F. (1878). *Human, all too human: A book for free spirits* (Orig. publ. as Menschliches, Allzumenschliches: Ein Buch für freie Geister). Ernst Schmeitzner.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560–1563. https://doi.org/10.1126/science.1133755

O'Connor, K., Effron, D. A., & Lucas, B. J. (2020). Moral cleansing as hypocrisy: When private acts of charity make you feel better than you deserve. *Journal of Personality and Social Psychology*, *119*(3), 540–559. https://doi.org/10.1037/pspa0000195

Olivola, C. Y. (2011). When noble means hinder noble ends: The benefits and costs of a preference for martyrdom in altruism. In D. M. Oppenheimer & C. Y. Olivola (Eds.), *The science of giving: Experimental approaches to the study of charity* (pp. 49–62). Psychology Press.

Olivola, C. Y., & Shafir, E. (2013). The martyrdom effect: When pain and effort increase prosocial contributions. *Journal of Behavioral Decision Making*, *26*(1), 91–105. https://doi.org/10.1002/bdm.767

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

Raihani, N. J., & Power, E. A. (2021). No good deed goes unpunished: The social costs of prosocial behaviour. *Evolutionary Human Sciences*, *3*, Article e40. https://doi.org/10.1017/ehs.2021.35

Savary, J., Li, C. X., & Newman, G. E. (2020). Exalted purchases or tainted donations? Self-signaling and the evaluation of charitable incentives. *Journal of Consumer Psychology*, *30*(4), 671–679. https://doi.org/10.1002/jcpy.1157

Sen, S., & Bhattacharya, C. B. (2001). Does doing good always lead to doing better? Consumer reactions to corporate social responsibility. *Journal of Marketing Research*, *38*(2), 225–243. https://doi.org/10.1509/jmkr.38.2.225.18838

Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition*, *167*, 201–211. https://doi.org/10.1016/j.cognition.2017.05.004

Silver, I., Newman, G., & Small, D. A. (2021). Inauthenticity aversion: Moral reactance toward tainted actors, actions, and objects. *Consumer Psychology Review*, *4*(1), 70–82. https://doi.org/10.1002/arcp.1064

Simpson, B., & Willer, R. (2015). Beyond altruism: Sociological foundations of cooperation and prosocial behavior. *Annual Review of Sociology*, *41*(1), 43–63. https://doi.org/10.1146/annurev-soc-073014-112242

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, *87*(3), 355–374. https://doi.org/10.2307/1882010

Wagner, T., Lutz, R. J., & Weitz, B. A. (2009). Corporate hypocrisy: Overcoming the threat of inconsistent corporate social responsibility perceptions. *Journal of Marketing*, *73*(6), 77–91. https://doi.org/10.1509/jmkg.73.6.77

Willer, R. (2009). Groups reward individual sacrifice: The status solution to the collective action problem. *American Sociological Review*, *74*(1), 23–43. https://doi.org/10.1177/000312240907400102

Yoon, Y., Gürhan-Canli, Z., & Schwarz, N. (2006). The effect of corporate social responsibility (CSR) activities on companies with bad reputations. *Journal of Consumer Psychology*, *16*(4), 377–390. https://doi.org/10.1207/s15327663jcp1604_9

Zlatev, J. J., & Miller, D. T. (2016). Selfishly benevolent or benevolently selfish: When self-interest undermines versus promotes prosocial behavior. *Organizational Behavior and Human Decision Processes*, *137*, 112–122. https://doi.org/10.1016/j.obhdp.2016.08.004