Aim High or Aim Low?
Self-Set Normative Goals in a Social Dilemma
Draft: May 2018

Christoph Engel# and Michael Kurschilgen§#

Can self-set normative goals restrain free-riding in a social dilemma? And is it more effective to set oneself a high – but possibly unrealistic – normative yardstick or rather a more pragmatic – but possibly more malleable – one? We test the effect of these two different types of self-set normative goals on people’s willingness to cooperate in a public good game. Our results suggest that focusing on lower, more realistic normative goals might be more effective at restraining material incentives to freeride.

Keywords: Social Dilemma, Identity, Norms, Goals, Focus, Expectations, Experiment
JEL: C90, D63, H41

German universities do not offer IRB scrutiny for behavioral experiments in the experimental economics paradigm, which is why no IRB approval has been obtained. The experiment has been funded from the regular budget of the Max Planck Institute for Research on Collective Goods, Bonn.

# Max Planck Institute for Research on Collective Goods
§ Technical University of Munich
Corresponding author: Michael Kurschilgen, Technical University of Munich, Arcisstr. 21, 80333 Munich, Germany, mkurschilgen@tum.de
1. Introduction

In his Theory of Moral Sentiments, Adam Smith (1790) identifies the restraint of one’s innate selfishness as “the perfection of human nature” (I.i.5.5). That perfection, however, seems rather difficult to attain. In fact, people’s selfishness is often the biggest obstacle to reaching better social outcomes. From environmental pollution over tax evasion, corruption and misappropriation, to doping, queuing, and bank runs, the list of situations is sheer endless in which individual incentives to free-ride lead to subpar social outcomes. This paper examines experimentally whether setting oneself normative benchmarks can restrain selfish behavior in a social dilemma. In particular, we distinguish between setting an ambitious goal of normatively ideal behavior, and a more pragmatic goal that sets a minimum standard below which one should not fall.

Our study thus contributes both to the growing body of literature trying to understand the role of social norms (Erik O Kimbrough and Alexander Vostroknutov, 2015, Erin L. Krupka and Roberto A. Weber, 2013) and morality (Roland Bénabou and Jean Tirole, 2011, Jason Dana et al., 2007, Armin Falk and Nora Szech, 2013) in economic interactions, as well as to the large literature on the determinants of human cooperation (see for instance Gabriele Camera and Marco Casari, 2009, Urs Fischbacher and Simon Gächter, 2010, Daniel Friedman and Ryan Oprea, 2012). Understanding the power of self-set normative standards is not only a matter of basic science. If it indeed reduces the gap between individual and social rationality, policy makers learn something about a tool in their toolbox. This tool might be particularly helpful where changing monetary incentives is impractical or undesirable. This holds if the cost of vigilance and enforcement is prohibitive, as is characteristic for social dilemma situations where large populations all must make a contribution. Many environmental problems are of this kind.

We conceptualize the social dilemma as a standard repeated public good game (R. Mark Isaac et al., 1985) and propose a utility function that combines an other-regarding component with a self-image component. For the latter we draw on the concept of identity utility (George A. Akerlof and Rachel E. Kranton, 2000). Experimentally, we induce participants to focus on one of two normative goals through the elicitation of normative
expectations. Our data show that selfishness is considerably restrained when subjects aim for a pragmatic normative goal, and less so when the goal is rather ambitious. By means of a path model we show that the focus on self-set normative goals mainly affects contributions to the public good on an indirect path: it affects initial normative expectations; the initial expectations affect expectations in later periods; expectations in later periods affect choices, and have an effect that is independent of experiences.

In the next section, we specify the theoretical framework, followed by the experimental design. Section four discusses the experimental results. Section five concludes with discussion.

2. Theoretical Framework
We study cooperation behavior in a standard linear public good game. Each member of a group of \( K \) members decides how much of her endowment \( e \) she wants to contribute \( c_i \) to a public project. Each unit contributed to the project yields a benefit of \( \mu \) to all group members (contributors and non-contributors alike) whereas each unit kept by a member yields a benefit of 1 to that very member only. The payoff \( \pi_i \) for a given player is thus defined as follows:

\[
\pi_i = e - c_i + \mu \sum_{k=1}^{K} c_k
\]  

In the linear public good game described in equation (1), the social optimum is for all members to contribute their entire endowments. But individually, each member is best off if she contributes nothing. Money-maximizing agents contribute \( c_i^* = 0 \) as long as marginal per capita rate \( \mu < 1 \). A rich literature (Ananish Chaudhuri, 2011, John O. Ledyard, 1995, Jennifer Zelmer, 2003) has shown that, indeed, a large fraction of experimental participants play \( c_i^* = 0 \) throughout the game, and that over time, more and more players contribute zero. However, this same literature also shows that there is a substantial fraction of people who contribute rather high (often close to the maximum), especially in the first rounds of the game.
One common explanation distinguishes between three types of people: selfish, altruist, and conditional cooperator. Selfish people pursue their material self-interest and altruists pursue the good of the group as a whole. Empirically, the former has been found to be a large minority of subjects, whereas altruists usually are a very small minority. The largest fraction of experimental subjects has been found to be conditional cooperators. Conditional cooperators are intermediate types. On the one hand, they have a regard for the social good, but on the other hand they strongly dislike being taken advantage of. Thus, they are willing to contribute if they believe others will also contribute, but they behave as if they were purely selfish if they doubt others’ cooperativeness (Urs Fischbacher and Simon Gächter, 2010, Urs Fischbacher et al., 2001).

a. Other-regardingness

In principle, cooperative choices of conditional cooperators could be motivated by other-regarding preferences: \( u_i = \pi_i + f(E(c_j), c_i) \). A common operationalization is inequity aversion, as modeled by Ernst Fehr and Klaus M. Schmidt (1999). In a linear public good game, an inequity averse player has the following utility:

\[
\begin{align*}
u_i &= e - c_i + \mu \sum_{k=1}^{K} c_k - \frac{1}{N-1} \alpha \sum_{j=1}^{N-1} \max\{E(c_j) - c_i, 0\} \\
&\quad - \frac{1}{N-1} \beta \sum_{j=1}^{N-1} \max\{c_i - E(c_j), 0\} \\
&= u_i \end{align*}
\]

Yet, for inequity aversion to be the exclusive motive for conditional cooperation, one must make rather strong assumptions. Even if players are perfectly optimistic, i.e. if they believe all other group members to contribute exactly the amount they plan to contribute themselves, cooperation only gives them higher utility than defection if \( \beta > (1 - \mu) \). Using the typical parameters of experimental public good games\(^1\), this degree of inequity aversion

\(^{1}\) as stated in the experimental design section.
is implausibly high (Mariana Blanco et al., 2011, Ernst Fehr and Klaus M. Schmidt, 1999). Note what $\beta > .6$ implies: A player is willing to spend 60 Cents to avoid earning 1 $ more than a member of her group. Moreover, players do not know with certainty how other players are going to behave. They have to rely on their beliefs $E(c_j)$. This is particularly important since most players will hold $\alpha > \beta$; being exploited hurts more than exploiting others. Yet forming reliable beliefs about others’ behavior is a difficult task in a public good game. In the initial period players can at best rely on some home-grown sense of a typical distribution of choices. In the subsequent periods, the best proxy they have for current choices is past choices. But high contributions in the past do not guarantee high contributions in the present. First, they might come from strategic agents who themselves hold selfish preferences but contribute a positive amount, expecting this to be a profitable investment in cooperativeness (David M. Kreps et al., 1982), which they may want to reap at any given moment. Second, high contributions might come from other conditional cooperators who might lose faith in the group before oneself.

Hence, based on other-regarding preferences alone, cooperation is the more difficult to sustain the more pronounced the uncertainty about the preferences of others. From the same model it follows that preference heterogeneity hurts. Compare the following two situations in the linear public good game described above: (A) three other players are known to all contribute 10 (of 20) tokens. An individual strongly enough motivated by other-regarding concerns will then also contribute 10 tokens, and not suffer any disutility from (advantageous or disadvantageous) inequity. Compare this with two players contributing 15, and one contributing 0. The average contribution of the three other players is again 10 tokens. Yet if now the individual contributes 10 tokens herself, she suffers some disutility from advantageous utility (she exploits the two players who contribute 15) and considerable disutility from being exploited by the one free rider.

b. Self-image

For these reasons, there is reasonable doubt that other-regarding preferences alone are able to explain high contributions in a public good game. Arguably, one needs an additional motive. The literature on identity utility assumes that people have an intrinsic sense of right
and wrong behavior, be it by means of an “instinctive feeling, [...] a conscious self-assessment” (Roland Bénabou and Jean Tirole, 2011), or via the “internalization [...] of behavioral prescriptions” (George A. Akerlof and Rachel E. Kranton, 2000), and that they derive utility from being “moral, prosocial, or cooperative” (Roland Bénabou and Jean Tirole, 2011), or, in the words of Adam Smith, “love of [...] praise-worthiness” (Adam Smith, 1790) (III.2.2). In short: individuals are not only concerned how they compare with others (“other-regarding”); they also care about their own identity as a responsible social being (“self-image”): \[ u_i = \pi_i + f(E(c_j),c_i) + g(\tilde{c}_i,c_i). \] The additional motive is captured by the final term in the utility function, which has the individual compare her actual choice \( c_i \) with her subjective norm \( \tilde{c}_i \). In the context of the public good game, this additional motive may be formalized by an extension of (2):

\[
\begin{align*}
    u_i &= e - c_i + \mu \sum_{k=1}^{K} c_k - \frac{1}{N-1} \alpha_i \sum_{j=1}^{N-1} \max\{E(c_j) - c_i, 0\} \\
    &\quad - \frac{1}{N-1} \beta_i \sum_{j=1}^{N-1} \max\{c_i - E(c_j), 0\} - \gamma_i (\tilde{c}_i - c_i) \\
\end{align*}
\]

The last term captures identity utility, closely following George A. Akerlof and Rachel E. Kranton (2005), where \( \tilde{c}_i \) is individual i’s behavioral prescription, i.e. a specific normative goal she would like to adhere to. If the contribution falls below the subjective standard \( (c_i < \tilde{c}_i) \), the subject suffers disutility. This disutility increases in the distance between \( \tilde{c}_i \) and \( c_i \). In contrast, exceeding the normative standard \( (c_i > \tilde{c}_i) \) yields additional utility. \( \gamma_i \geq 0 \) denotes the weight of identity utility, which Benabou & Tirole call the relative “strength of the self-esteem motive” (Roland Bénabou and Jean Tirole, 2011).

Extant literature has shown \( \gamma_i \) to be rather malleable. An important reason for this seems to be self-deception. People like both earning money and seeing themselves as a good person. However, often their wish to earn money entices them to deviate from what (their own conception of) a good person would do. In those situations, a lack of normative clarity
comes in handy as a legitimate excuse (to oneself) and reduces the bad conscience from not complying with \( \hat{c}_i \). Jason Dana, Roberto A. Weber and Xi Kuang (2007) show that many “people exploit such ‘moral wiggle room’ in order to behave self-interestedly”. In their experiment the share of selfish choices rises substantially when deciders in a dictator game are given a tool to blur the responsibility for a selfish decision. Armin Falk and Nora Szech (2013) suggests that markets have a similar effect. In their experiment, participants choose between the life of a mouse and a monetary amount. In a market setting, a substantially higher share of people is willing to sacrifice the mouse than in a non-market setting. The authors conclude that people “seem to ignore their moral standards when acting as market participants” and ascribe it to the fact that markets reduce a person’s perceived responsibility for a morally undesired outcome and thus her feelings of guilt.

However, there is still little understanding as to how people set their normative goals \( \hat{c}_i \), and whether they feel more intensely bound to certain goals than to others, i.e. how \( \hat{c}_i \) and \( \gamma \) interact. In this paper, we distinguish two types of \( \hat{c}_i \). The first, \( \hat{c}_i^{\text{high}} \), refers to the socially ideal type of behavior, i.e. the type of behavior that if followed by everybody would generate the greatest benefit to all. The second standard, \( \hat{c}_i^{\text{low}} \), is a bit more subtle. It refers to the border between normatively tolerable and intolerable deviations from ideal behavior. In fact, we can view the two standards as an upper bound and a lower bound of the normatively tolerable spectrum. They relate to Adam Smith’s distinction between “absolute perfection” (Smith I.i.5.9) and “ordinary perfection” (Smith IV.iii.27)\(^2\), as well as to a fundamental principle of legal reasoning, the distinction between morality and legality (Herbert Lionel Adolphus Hart, 1961).\(^3\) The following experiment tests how focusing on these different

\(^2\) “[…] when we are determining the degree of blame or applause which seems due to any action, we very frequently make use of two different standards. The first is the idea of complete propriety and perfection […]. The second is the idea of that degree of proximity or distance from this complete perfection, which the actions of the greater part of men commonly arrive at.” (I.i.5.9)

\(^3\) Examples include minimum safety requirements, minimum environmental standards as well as the difference between (the legally sanctioned concept of) fraud and (the morally deplorable but legally unsanctioned) lie.
normative goals affects people's ability to restrain individual incentives to free-ride in a public good game.

3. Experimental Design

a. Treatments

In line with the large majority of the experimental public good literature, we set \( e = 20, K = 4, \mu = .4 \) (see the metastudy by Jennifer Zelmer, 2003). We repeat the game over 30 periods with fixed group composition. From the second period on, participants may check out on their computer screens a graph informing them about contributions by the other group members during all preceding periods.\(^4\) Every period of the Baseline consists of just one stage, i.e. the (incentivized) contribution decision just described.

We induce participants to focus on the two normative goals via the (non-incentivized) elicitation of expectations. For that purpose, every period the treatments have an additional second stage. After participants have decided how much to contribute to the public project, but before receiving feedback about the current period’s contributions of the remaining group members, they are privately asked on their computer screens to state their expectations. The expectation questions of High and Low invite subjects to think about their own conception of the respective normative goal (see Table 1).\(^5\)

\(^4\) This last design feature contrasts with most repeated public good games in the literature. Usually, studies have two layers of anonymity: The first consists of not revealing subjects’ true identity (i.e. one’s real name) but instead replacing it with a playing identity (i.e. a number). The second consists of randomizing subjects’ playing identity from period to period so that group members cannot track one another over time. We stick to the first layer but remove the second because information about variance, and about the development of individual contributions over time may be critical for the formation of normative expectations. This design feature is constant across all treatments.

\(^5\) Every subject is completely free in choosing her answer to the expectation question. Neither will the answer affect her payoffs, nor will another participant ever learn her answer. Subjects are informed in the instructions that answers to the second stage question are never made available to other participants. The instructions
Table 1: Experimental Treatments

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Expectation question (in stage 2 of every period)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>[no stage 2]</td>
</tr>
<tr>
<td>High</td>
<td>What do you believe should every group member contribute to the project?</td>
</tr>
<tr>
<td>Low</td>
<td>What do you believe is the minimum contribution to the project that should generally be expected from every group member?</td>
</tr>
</tbody>
</table>

b. Related Experimental Literature

Our study contributes to a growing body of work in behavioural and experimental economics. Jean-Robert Tyran and Lars P. Feld (2006) have an explicit norm of full contribution, and an explicit (though imperfect) sanction. Erin L. Krupka and Roberto A. Weber (2009) have “focusing” treatments where they ask participants to either focus on what they think others have done in a previous experiment, or what they think others have said in that experiment that should be done. Yet in their experiment, participants do not make a strategic choice; they are dictators. The experiment tests whether thinking about others induces prosocial choices, while we investigate whether this manipulation suffices to overcome a social dilemma. Simon Gächter and Elke Renner (2010) find that merely asking participants for their first order beliefs does not increase contributions to the public good. Tibor Neugebauer et al. (2009) play a repeated linear public good and elicit beliefs, which are incentivized in each period. Beliefs are significantly correlated with contributions in that same period. Contributions are lower if feedback on payoffs and beliefs is given. Ernesto Dal Bó and Pedro Dal Bó (2009) provide participants in a public good game with messages that define moral behaviour and find that, if subjects are told that full contribution is moral, their cooperation rates increase but still fall over time. Whereas Ernesto Dal Bó and Pedro Dal Bó (2009) tell subjects what is inform about the complete design. Hence when making contribution choices in the first period, participants know what they will be asked in the second stage of this period.

---

6 We discuss the related theoretical literature in section 4.
supposed to be appropriate behavior, we ask subjects for their own personal standards. Erik O Kimbrough and Alexander Vostroknutov (2015) introduce a “rule following task” to sort participants in several experimental games. They find that rule followers display substantially higher levels of cooperation in a public good game when matched with other rule followers but not when matched with rule breakers.

c. Design Choices

Two of our design choices warrant explanation. First, we have not incentivized the elicitation questions. Elicitation typically faces two challenges: On the one hand, one wants to avoid that people do not care enough about the question and answer randomly. That would produce uninformative, noisy data. On the other hand, people might care for the wrong reasons, and deliberately choose a specific, untruthful answer $a'$ instead of the truthful answer $\hat{a}$. Both problems arise whenever respondents derive too little intrinsic utility from answering truthfully. Appropriate incentivization provides subjects with a material reason to answer truthfully. The higher the incentive, the more costly it is for the respondent to not answer $\hat{a}$.

Incentivizing expectations, however, reduces the question to a mere gamble in which there are “profitable” and “unprofitable” answers. For behavioural expectations this is defendable since they are empirical by nature. One’s first order belief is correct if it matches the average person’s actual behavior. One’s second order belief is correct if it matches the average person’s first order belief. In contrast, normative expectations are not purely empirical but involve a judgment. Erin L. Krupka and Roberto A. Weber (2013) propose an incentive mechanism where the stated normative expectation yields a material payoff if it matches the modal response given by all other participants. That incentivization has quite some appeal when one looks for a “shared understanding” of normative appropriateness in a one-shot setting, as Erin L. Krupka and Roberto A. Weber (2013) do.

In contrast, when looking for subjects’ individual normative judgments as this paper does, an incentivization in the spirit of Erin L. Krupka and Roberto A. Weber (2013) would raise concerns. First, it would induce coordination on certain focal answers that do not necessarily coincide with subjects’ truthful individual judgments. A respondent might truthfully believe $\hat{a}$ to be the normatively appropriate choice, but choose rather to answer $a'$
because it is more focal. Even if everybody believed ́a to be the normatively appropriate choice, they might still all choose ́a’, only because they believe it more probable to coordinate on ́a’ than on ́a. Second, in a repeated game, incentivization artificially introduces inertia. Once people have coordinated on a certain answer in the early periods of a game, the incentive meant to elicit the true norm will prevent them from departing from the reached equilibrium in later periods, even if their actual judgments had drastically changed. Even more importantly, for our research question the critical function of the expectation question is to induce participants to focus on their own individual conception of a normative standard. Tying the question of normative appropriateness to a monetary coordination game would likely change the participants’ thought process.

Second, we do not ask for factual or normative assessment before participants choose how much to contribute, but after they have made a choice. We do so because we believe this is the harder test. Had we first asked for the (normative) expectation this might have set a strong anchor before the subsequent contribution decision. In contrast, in our design subjects first set a (binding and incentivized) contribution decision level and subsequently answer an expectation question. Since we do not incentivize the expectation question, it is actually costless for subjects to answer the expectation question in a self-legitimizing way in the light of the preceding contribution choice, unless subjects perceive some intrinsic disutility from answering the expectation question dishonestly. Of course, from the second period on, this difference becomes subtle; participants are likely to take into consideration what they have stated after last period’s choice when deciding what to choose in the current period. This is why, in our data analysis, we will pay special attention to the effect of statements in the first period.\footnote{See the path model below.}

d. Procedure

The experiment was run at the Bonn EconLab. Student participants were randomly selected from a pool of approximately 5000 subjects, using the software ORSEE (Ben Greiner, 2004).
About half of the participants were female. We had 48 subjects (12 groups) in the Baseline, 48 subjects (12 groups) in the High-treatment, and 48 subjects (12 groups) in the Low-treatment. Mean age was 23 years. Students held various majors. The experiment was computerized using the software zTree (Urs Fischbacher, 2007). Before playing the game, participants read experimental instructions and answered a set of control questions (see Appendix). The latter were identical for all treatments. Participants spent about 60 minutes in the lab and earned on average 12.40 € (approximately 16.70 $).

4. Results

a. Treatment Effects on Cooperation

Figure 1 suggests that focusing on self-set normative goals indeed increases cooperation. On average, cooperation increases from 35% in the Baseline to 48% in High, and 55% in Low. According to a simple Mann-Whitney ranksum test over group means, there is a significant increase to the Low-treatment (p=0.043, N=24, two-sided). In contrast, the increase of cooperation in High is not significantly different from the Baseline.

The elicitation of normative standards also appears to substantially delay the erosion of cooperation over time. In the Baseline, mean contributions drop irrevocably below 10 tokens already in period 5. In contrast, in High and Low cooperation only drops for good below 10 in periods 23 and 24, respectively. The substantial endgame effects, even in the treatments with high cooperation rates, refute the idea of subjects naively following a self-set normative goal and supports the idea of a tradeoff between self-image concerns and selfish motives.

In accordance with results from previous public good studies (see for instance Fehr and Gächter 2000), our data show that, even though the action space comprises 21 possible levels of cooperation, participants predominantly choose between the two extremes. The
parametric estimation\textsuperscript{8} in Table 2 confirms the robust effect of \textit{Low} on cooperation. In contrast, the effect of \textit{High} remains insignificant. If we separately calculate marginal effects of treatment from model 2 for each period, we find a significant difference between the \textit{Baseline} and the \textit{Low}-treatment for every period between the first and the 22\textsuperscript{nd}. For the remaining periods, the difference is significant at the 10\% level (p < .088): the more the endgame effect plays itself out, the less pronounced the difference between the two treatments.

![Figure 1: Contributions](image)

\textbf{Figure 1: Contributions}

*Note:* The solid line denotes the mean contribution in the respective \textbf{Treatment} (left: \textit{High}; right: \textit{Low}), the dashed line denotes the mean contribution to the public good in the \textbf{Baseline}. The Baseline has 48 subjects, \textit{High} 48, and \textit{Low} 48.\textsuperscript{8}

\textsuperscript{8} Over all five experimental conditions 53\% of individual choices were either 0 or 20. This is reflected by a Tobit model. We have data from choices nested in individuals nested in groups, which we account for by a mixed effects model.

If we explain a dummy that is 1 if the participant does not contribute anything with treatment (in a mixed effects logit model with standard errors for choices nested in individuals nested in groups), we do not find any significant effects. We do, however, find a significant effect of both the \textit{High} and the \textit{Low} treatment if we explain a dummy that is 1 if the participant contributes her full endowment of 20 with treatment. These additional regressions are available upon request.
### Table 2: Treatment Effects on Contribution Behavior

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>5.638</td>
<td>5.469</td>
</tr>
<tr>
<td></td>
<td>(4.241)</td>
<td>(4.240)</td>
</tr>
<tr>
<td>Low</td>
<td>8.607**</td>
<td>10.538**</td>
</tr>
<tr>
<td></td>
<td>(4.249)</td>
<td>(4.259)</td>
</tr>
<tr>
<td>Period</td>
<td></td>
<td>-.546***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.032)</td>
</tr>
<tr>
<td>High*period</td>
<td></td>
<td>-.013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.046)</td>
</tr>
<tr>
<td>Low*period</td>
<td></td>
<td>-.127**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.047)</td>
</tr>
<tr>
<td>Constant</td>
<td>5.116*</td>
<td>13.607***</td>
</tr>
<tr>
<td></td>
<td>(2.773)</td>
<td>(2.773)</td>
</tr>
</tbody>
</table>

Mixed effects Tobit model, allowing for lower censoring at 0 and upper censoring at 20. Standard errors for choices nested in individuals nested in groups in parenthesis. *** p < .01, ** p < .05, * p < .1.

#### b. Driving Forces

Our experimental design operationalizes introspection through a private expectation question. However, as Figure 2 shows, the relationship between a person’s answer in the expectation question and her contribution behavior is not uniform but critically depends on the nature of the expectation question. The difference between High and Low is striking. On average, every subject adapts her statement of the High standard only 5.83 times over the 30 periods of interaction, by a total amount of 3.08 tokens. The median is 20 in every single period. In contrast, subjects display considerably less reservations about adjusting their statement of the Low standard, which is adapted 11.54 times and drops on average by 8.34 tokens over the course of the game, from a median of 10 to a median of 0. Remarkably, however, despite the stated normative expectations in Low being both significantly lower than in High (p=0.009, Ranksum Test, N=24, two-sided) and eroding significantly faster (p=0.009, Ranksum Test, N=24, two-sided), contributions in Low are not lower than in High but, if anything, higher.
Figure 2: Contributions vs. Expectations
Note: The solid line denotes the mean contribution to the public good in the respective Treatment (left: High; right: Low), the dashed line denotes the mean expectation (see Table 1) in the respective Treatment (left: High; right: Low). High has 48 subjects, and Low 48.

To shed some light on this surprising finding, we estimate a path model (Figure 3). We allow a subject’s contribution $c_{it}$ in the current period to be driven by two forces: other-regarding and self-image concerns. We capture other-regarding concerns by a subject’s reaction to the average contribution of the other group members in the preceding period $\bar{c}_{j,t-1}$. We expect differences in self-image concerns to result from the particular type of normative goal ($\tau$). We expect the initial self-reported expectation $\hat{c}_{i,1}$ to be a function of $\tau$ and an idiosyncratic component. In later periods, these three components should in principle still be decision-relevant. We therefore expect reported subjective standards in later periods $\hat{c}_{i,t>1}$ to be a function of the reported standard in the first period $\hat{c}_{i,1}$. Yet arguably, subjects adjust their own reported standard to the experiences they are making $\bar{c}_{j,t-2}$. 


Model 1 of Table 3 is a direct test of the path model, with *Low* as the reference treatment. Panel A shows that when asked to state their initial normative expectations $\hat{c}_{i,1}$, participants in the *Low* treatment say that group members should be expected to contribute *at least* in the order of 13 tokens. If participants are, instead, asked to formulate how much one should *ideally* contribute, the stated expectation is almost 7 tokens more demanding, i.e. virtually at the upper limit of the action space.

The stated expectation in the first period $\hat{c}_{i,1}$, i.e. before having any information about others’ behavior, predicts stated expectations in later periods (Panel B). According to our path model, later expectations are a function of initial expectations and of subsequent experiences. The correlation between stated expectations in the first and in later periods is around .66. Recent experiences $\bar{c}_{j,t-2}$ translate slightly stronger than 1 to 1 into updated expectations. Conditional on both initial expectations and experiences, the stated normative expectations in the *High* treatment remain about 13 tokens above the statements in the *Low* treatment.

---

9 Technically, coefficients are of course for the latent, uncensored variable. But in a Tobit model, the relationship between the latent variable and model prediction is fairly direct. Only the intercept and the slope are adjusted to censoring. For the ease of presentation, we work with this shortcut.
Finally, contributions independently respond to experiences and to expectations (Panel C). Payoff comparisons (other-regarding concerns) are not the only motivating force. There is a strong additional effect of stated expectations (self-image concerns). Note however that participants match their stated expectations only imperfectly (the coefficient of $\hat{c}_{l,t-1}$ is far away from 1). Strikingly, even if we control for these two driving forces, we still find a (weakly) significant difference between Low and High.

Table 3: Path Model

<table>
<thead>
<tr>
<th>Dependent Variables</th>
<th>Independent Variables</th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A: Initial Expectations</td>
<td>$\hat{c}_{l1}$</td>
<td>constant</td>
<td>13.36***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.266)</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>6.815***</td>
<td>6.815***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.395)</td>
<td>(0.395)</td>
</tr>
</tbody>
</table>

| Panel B: Expectations over time | $\hat{c}_{l,t-1}$ | constant | -10.01*** | -11.41*** |
|                                 |                   |           | (0.741)   | (0.851)   |
|                                 | $\hat{c}_{l1}$   | 0.655***  | 0.304***  |
|                                 |                   | (0.040)   | (0.048)   |
|                                 | $\bar{c}_{f,t-2}$| 1.172***  | 1.660***  |
|                                 |                   | (0.043)   | (0.055)   |
|                                 | High               | 12.98***  | 14.15***  |
|                                 |                   | (0.558)   | (1.238)   |
|                                 | $\hat{c}_{l1}$ * High | 0.697*** | 0.697***  |
|                                 |                   | (0.074)   | (0.074)   |
|                                 | $\bar{c}_{f,t-2}$ * High | -1.141*** | -1.141*** |
|                                 |                   | (0.073)   | (0.073)   |

| Panel C: Contributions over time | $c_{lt}$ | constant | -6.577*** | -5.244*** |
|                                 |         |           | (1.359)   | (1.455)   |
|                                 | $\hat{c}_{l,t-1}$ | 0.385*** | 0.427***  |
|                                 |           | (0.055)   | (0.076)   |

10 We do not work with stated expectations in the current period since, in every period, we asked participants for their expectations after they had decided how much to contribute to the joint project.
Model 2 of Table 3 allows for additional interaction effects between τ and the other explanatory variables and thus for more fine-grained comparisons between Low and High. Panel B illustrates a key difference between the two normative standards. Statements about the Low standard (the reference treatment) are fairly malleable whereas statements about the High standard are rather robust. The correlation between initial and later statements in Low is only .3, significantly smaller than 1 (p<0.001). By contrast, the statement in High is fairly stable over time (the aggregate effect of $\hat{c}_{i,t-1}$ and $\hat{c}_{i,t-1} \cdot High$ is virtually equal to 1). The reverse holds for the effect of experiences. The main effect of $\overline{c}_{j,t-2}$ on $\hat{c}_{i,t-1}$ is significantly larger than 1, i.e. statements of the Low standard are highly sensitive to recent experiences. By contrast the interaction between $\overline{c}_{j,t-2}$ and High is strongly negative. There is still an effect of experiences (the net effect is not 0) but it is much smaller.

If the High normative standard is more robust to bad experiences, how come it does not perform better at preventing contribution to decline than the more malleable Low standard? The reason seems to be that the High standard is perceived as less binding than the Low standard. Holding normative statements constant, contributions in High are significantly more sensitive to (bad) experiences than in Low.

5. Conclusion
This paper has tested the extent to self-setting normative goals can reduce selfish behavior and thus increase cooperativeness in a public good game. In particular, we have tested two
distinct normative yardsticks: a High standard of socially ideal behavior, and a Low standard of minimally acceptable behavior.

Theoretically, we explain why the (other-regarding) desire not to outperform others is unlikely to be strong enough to explain cooperation, particularly when paired with the desire not to be outperformed by others. Uncertainty and heterogeneity even increase the problem. We show that the necessary additional impulse for cooperative behavior may come from the (self-image) desire to live up to one’s own normative expectations. The concept of identity utility provides a framework to rationalize why people may behave socially out of a self-image motivation. In this framework, a self-set standard can enhance cooperation by increasing a person’s consciousness of their own normative goals and thus the relative weight of identity utility.

Our experimental results show that being nudged to actively state one's minimally acceptable Low standard causally increases cooperation whereas the more ambitious High goal of behaving according to the normative ideal fails to improve cooperation significantly. This is particularly striking given that the former not only sets a substantially lower moral (and cognitive) anchor but also explicitly invites to being adapted in the light of the experiences. Our data show (a) while adaptation to experiences takes place, it is not excessive, and (b) that violating the more modest, more realistic goal apparently is way more painful than violating an anyhow unattainable normative ideal. This finding resonates well with fundamental principles of legal reasoning, which critically distinguishes between (desirable) morality and (binding) legality. Moreover, it casts new light on our understanding of a social norm as drawing a line between normatively tolerable and intolerable actions, reflecting both moral norms and common practice.

From a policy perspective, the motivating effect of self-image concerns is most relevant if (a) changing monetary incentives is technically impossible, prohibitively costly or normatively unwanted and (b) perceived uncertainty about the behavior of others is pronounced, or the heterogeneity of others’ choices is strong. Condition (a) implies that policy makers cannot afford to ignore the likely character of addressees’ utility functions. Condition (b) implies that other-regarding concerns alone are also not likely to lead to the socially desired outcome.
There are many policy problems that (i) constitute a social dilemma and (ii) are affected by preference heterogeneity and/or preference uncertainty. An obvious illustration are large-scale environmental problems, like waste management. Many countries desire waste separation in households. This prevents toxic fractions from polluting ground water, and saves natural resources. Enforcing waste separation at the household level by command and control regulation is impractical. Social pressure is also at best imperfect. As a practical matter, policy makers must rely on the environmental consciousness of individual households. Waste management policies become more resilient if grounded on self-image concerns, as individual behavior becomes more robust to uncertainty about what neighbors are likely to do.


Appendix: Experimental Instructions and Control Questionnaire

[The shaded areas only appear in the corresponding treatments]

General instructions for the participants

Welcome to our experiment!

If you read the following explanations carefully, you will be able to earn a substantial sum of money, depending on the decisions you make. It is therefore crucial that you read these explanations carefully.

During the experiment there shall be absolutely no communication between participants. Any violation of this rule means you will be excluded from the experiment and from any payments. If you have any questions, please raise your hand. We will then come over to you.

During the experiment we will not calculate in euro, but instead in taler. Your total income is therefore initially calculated in taler. The total number of taler you accumulate in the course of the experiment will be transferred into euro at the end, at a rate of

1 Euro = 60 Taler

At the end you will receive from us the cash sum, in euro, based on the number of taler you have earned.

The experiment consists of 30 periods, and each period consists of 3 stages. Participants are randomly divided into groups of four. Apart from yourself, your group therefore has 3 further members. During these 30 periods, the constellation of your group of four remains unchanged. Hence, you are with the same people in the same group for 30 periods. At the beginning, each group member is allocated a random number between 1 and 4. This number remains unchanged for the entire 30 periods.

Stage 1:

At the beginning of each period, each participant is given 20 taler to work with, referred to henceforth as endowment. Your task is to decide upon how to use your endowment. You must decide how many of the 20 taler you wish to pay into a common project, and how many you wish to keep for yourself. The consequences of this decision are explained in more detail below.

Your endowment hence consists of 20 taler in each period. You make a decision on your payments by typing whole numbers between 0 and 20 in the input field on your screen. Once you have keyed in your amount, press Continue. As soon as you have done this, you may no longer reverse your decision for this period.
Once all group members have made their decisions, you are told how much each individual group member has contributed to the project.

**Your total income** (in taler) therefore consists of two parts: (1) the taler income from the common project and (2) the taler you have retained.

\[
\text{Total income (in taler)} = \text{Income from the common project} + \text{Taler retained}
\]

The **income from the common project** is calculated as the total sum of all contributions to the project (within your group of four) times 0.4.

\[
\text{Income from the common project} = \text{total sum of all contributions to the project (within your group of four) \times 0.4}
\]

**Example:**

If the sum of contributions from all group members to the common project is 60 taler, you and each other group member receive an income from the project of 0.4\times60 = 24 taler. If the group members have contributed a total of 9 taler to the project, you and each other group member receive a taler income from the project of 0.4\times9 = 3.6.

If you contribute one taler from your endowment to the group project, the sum of contributions to the common project increases by 1 taler, and your income from the project increases by 0.4\times1 = 0.4 taler. However, this also means that each individual other group member’s income increases by 0.4 taler, so that the total income of the group increases by 0.4\times4 = 1.6 taler. The other group members therefore also earn something from your contribution to the project. On the other hand, you profit from the contributions made by the other group members. For each taler contributed to the project by another group member, you earn 0.4\times1 = 0.4 taler. Hence, if each member of your group of four contributes 1 taler to the project, each of you receives 0.4\times1\times4=1.6 taler as income from the project.

**Stage 2**

In **Stage 2**, you will see a screen requesting you to answer the following question:
**[High]** What do you believe should every group member contribute to the project?

**[Low]** What do you believe is the minimum contribution to the project that should generally be expected from every group member?

From the second period onwards, you will receive information on the behavior of individual group members in past periods. In order to receive this, you will have to click on an appropriate **button** on your screen. This can be done as often as you like.

- Button "**contributions**": how much have the individual group members contributed to the common project?

---

**Control Questionnaire**

1. Each group member has an endowment of 20 taler. Nobody (including you) contributes any taler to the project. What is:
   a. Your income from the common project? ........
   b. Your total income?........

2. Each group member has an endowment of 20 taler. You contribute 20 taler to the project. All other group members contribute 20 taler each to the project. What is:
   a. Your income from the common project? ........
   b. Your total income?........

3. Each group member has an endowment of 20 taler. You contribute 0 taler to the project. The other three group members contribute together a total of 30 taler to the project. What is:
   a. Your income from the common project? ........
   b. Your total income?........

4. Each group member has an endowment of 20 taler. You contribute 15 taler to the project. The other three group members contribute together a total of 5 taler to the project. What is:
a. Your income from the common project? ........

b. Your total income?........

5. After Stage 1 you have a total income of 30. Then you distribute 2 points to group member 1 and 3 points to group member 2. You also receive from the members of your group a total of 4 points. What is your total income now?