Engel, Christoph; Kurschilgen, Michael

Working Paper

The jurisdiction of the man within: Introspection, identity, and cooperation in a public good experiment

Preprints of the Max Planck Institute for Research on Collective Goods, No. 2015/1

Provided in Cooperation with:
Max Planck Institute for Research on Collective Goods

Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2015/1

The Jurisdiction of the Man Within – Introspection, Identity, and Cooperation in a Public Good Experiment

Christoph Engel
Michael Kurschilgen

MAX PLANCK SOCIETY

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2015/1

The Jurisdiction of the Man Within – Introspection, Identity, and Cooperation in a Public Good Experiment

Christoph Engel / Michael Kurschilgen

December 2014

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

The Jurisdiction of the Man Within – Introspection, Identity, and Cooperation in a Public Good Experiment*

December 2014

Christoph Engel / Michael Kurschilgen

Abstract

According to Adam Adam Smith (1790), human selfishness can be restrained by introspection. We test the effect of introspection on people's willingness to cooperate in a public good game. Drawing on the concept of identity utility (George A. Akerlof and Rachel E. Kranton, 2000), we show theoretically that introspection may enhance cooperation by increasing the relative cost of deviating from one's self-image. Experimentally, we induce introspection through the elicitation of (normative) expectations. Our results show that introspection causally increases cooperation. Both home-grown idealism and the experiences with the cooperativeness of the environment predict individual cooperativeness throughout the game.

Keywords: Social Dilemma, Identity, Introspection, Expectations, Experiment

JEL-Codes: C90, D63, H41

* Helpful comments by Sebastian Goerg and Ionnna Grypari on an earlier version are gratefully acknowledged.

1. Introduction

In his Theory of Moral Sentiments, Adam Smith (1790) identified the restraint of one's 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 the collective interests of society and the selfish interests of its individual members are at odds.

Smith believed that the remedy to the curse of selfishness is to be found within the individual himself. Specifically, Smith conjectured that people can counter their selfish instincts through introspection, i.e. the conscious assessment of one's own behaviour relative to a normative standard.

This paper examines experimentally whether introspection can enhance cooperation in a social dilemma. Our study thus contributes both to the growing body of literature trying to understand the role of morality in economic interactions (Roland Bénabou and Jean Tirole, 2011, Jason Dana et al., 2007, Armin Falk and Nora Szech, 2013) 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).

We conceptualize the social dilemma as a standard repeated public good game (R. Mark Isaac et al., 1985) and induce introspection through the elicitation of (factual or normative) expectations. Drawing on the concept of identity utility (George A. Akerlof and Rachel E. Kranton, 2000), we test an outcome model and a process model. In line with the outcome model, our data show that choices come significantly closer to the normative ideal if participants privately indicate the minimum contribution to the joint project that can generally be expected. In line with the process model, our data show that inducing introspection leads to participants forming more demanding initial normative expectations. The treatment effect on contributions to the joint project is almost completely mediated by the effect of introspection on expectations. This effect comes on top of other-regarding preferences.

In the next section, we introduce the experimental design. Section three presents the game theoretic paradigm and the behavioral predictions. Section four discusses the experimental results and section five concludes the paper.

2. Experimental Design

a. Treatments

The experiment consists of a *Baseline* and four treatments. The *Baseline* is a standard linear public good game. Each member of a group of K members decides how much of her endow-

ment e she wants to contribute c_i to a public project. Each unit contributed to the project yields a benefit of μ 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 π_i for a given player is thus defined as follows:

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

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.[1] Every period of the Baseline consists of just one stage, i.e. the (incentivized) contribution decision just described.

We induce introspection 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 (see Table 1).[2]

Table 1: Experimental Treatments

Treatment	**Expectation question** (in stage 2 of every period)
Baseline	[no stage 2]
First-Order Belief (FB)	What do you believe do the other group members on average contribute to the project?
Second-Order Belief (SB)	What do you believe do the other group members think that you contribute to the project?
Normative Ideal (NI)	What do you believe should every group member contribute to the project?
Normative Minimum (NM)	What do you believe is the minimum contribution to the project that should generally be expected from every group member?

1 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.

2 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 know her answer. Subjects were informed in the instructions that answers to the second stage question are never made available to other participants.

Introspection is the conscious assessment of one's own behaviour relative to a normative goal. The expectation questions of NI and NM invite subjects to think about their own conception of that goal. Smith distinguishes between two standards of comparison: "[…] 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)

The expectation question of the NI-treatment reflects Smith's first standard.[3] It asks subjects to deliberate about the normatively ideal behaviour. Smith's second standard of comparison motivates the expectation question in the NM-treatment. It is a bit more subtle than the normative ideal. Subjects are asked to think about the border between normatively tolerable and intolerable deviations from ideal behavior. This idea is a fundamental principle of legal reasoning. It is the essence of the distinction between morality and legality (Herbert Lionel Adolphus Hart, 1961).[4] In fact, Smith repeatedly alludes to the "jurisdiction of the man within" (III.2.32).

In order to connect the concept of normative expectations with the vast literature on behavioural expectations, we have two additional treatments. The expectation question of our FB-treatment is a common belief elicitation, which has been done in many other studies, though usually incentivized (Rachel T.A. Croson, 2007). The SB-treatment bridges between behavioural and normative expectations. On the one hand, the question is behavioral (i.e. empirical) in the sense that subjects are asked to guess the first-order beliefs of the other players. On the other hand, however, the question is normative as it reduces introspection to its cognitive core, i.e. viewing one's "own conduct through the eyes of other people" (Smith III.1.2). In fact, in the literature, second-order beliefs have been both been treated as behavioral (Pierpaolo Battigalli and Martin Dufwenberg, 2007) and as normative (Cristina Bicchieri, 2006).

b. Expectations and Incentives

Elicitation questions typically contain two problems: First, people might not care enough about the question and answer randomly. That would produce uninformative, noisy data. Second, 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 $\hat{a}$.

3 Smith recurrently refers to it as the "ideal man within the breast" (III.3.26).

4 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.

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, 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 $\hat{a}$ 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 $\hat{a}$. 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 have drastically changed. Even more importantly, for our research question the critical function of the expectation question is to induce introspection. Our theory expects the very process of introspection to affect individuals' choices. Reducing the question of normative appropriateness to a coordination game likely changes the participants' thought process and might well preclude introspection. For these reasons, we choose not to incentivize the expectation questions.

Our study contributes to a growing body of work in behavioural and experimental economics. Jean-Robert Tyran and Lars P. Feld (2006) had an explicit norm of full contribution, and an explicit (though imperfect) sanction. Erin L. Krupka and Roberto A. Weber (2009) had a different strategic situation, a dichotomous action space and interaction was one-shot. 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 supposed to be appropriate behavior, we ask subjects for their own standards.

c. Procedure

The experiment was run at the Bonn EconLab. 228 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. 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 $). We had 48 subjects (12 groups) in the Baseline, 44 subjects (11 groups) in the FB-treatment, 40 subjects (10 groups) in the SB-treatment, 48 subjects (12 groups) in the NI-treatment, and 48 subjects (12 groups) in the NM-treatment.[5]

3. Theoretical Framework

In the linear public good game, for the group as a whole it is best if all members 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 $\mu < 1$. A rich experimental literature (Ananish Chaudhuri, 2011, John O. Ledyard, 1995, Jennifer Zelmer, 2003) has shown that, indeed, a large fraction of participants play $c_i^* = 0$ throughout the game, and that over time, more and more subjects contribute zero. However, the 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).

5 The imbalance results from the fact that in the FB and SB-treatments too many invited participants did not show up.

a. Other-regardingness

In principle, cooperative choices of conditional cooperators could be motivated by other-regarding preferences, for instance inequity aversion, as modeled by Ernst Fehr and Klaus M. Schmidt (1999):

$$u_i = e - c_i + \mu \sum_{k=1}^{K} c_k - \frac{1}{N-1}\alpha \sum_{j=1}^{N-1} \max\{E(c_j) - c_i, 0\} - \frac{1}{N-1}\beta \sum_{j=1}^{N-1} \max\{c_i - E(c_j), 0\} \quad (2)$$

Yet, for inequity aversion to be the exclusive motive for conditional cooperation, one must make rather strong assumptions. Even if participants are perfectly optimistic, i.e. if they believe all other group members to contribute exactly the amount they plan to contribute themselves, this choice only gives them higher utility than defection if $\beta > (1 - \mu)$. This degree of inequity aversion is implausibly high (Mariana Blanco et al., 2011, Ernst Fehr and Klaus M. Schmidt, 1999). Moreover participants do not know with certainty how other participants are going to behave. They have to rely on their beliefs $E(c_j)$. This is particularly important since most participants will hold $\alpha > \beta$: being exploited is worse than exploiting others. Yet forming reliable beliefs about others' behavior is a difficult task in a public good game. In the initial period participants 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 necessarily indicate 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.

b. Self-regardingness

For these reasons, there is reasonable doubt that other-regarding preferences alone are able to explain the typical pattern of relatively high contributions to the public project that decay over time. 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 […] praiseworthiness" (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 re-

sponsible social being ("self-regarding"). In the context of the public good game, this additional motive may be formalized by an extension of (2):

$$u_i = e - c_i + \mu \sum_{k=1}^{K} c_k - \frac{1}{N-1}\alpha \sum_{j=1}^{N-1} \max\{E(c_j) - c_i, 0\} - \frac{1}{N-1}\beta \sum_{j=1}^{N-1} \max\{c_i - E(c_j), 0\} - \gamma \max\{\tilde{c}_i - c_i, 0\} \quad (3)$$

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. $\gamma_i \geq 0$ denotes the weight of identity utility, which Benabou & Tirole call the relative "strength of the self-esteem motive" (Roland Bénabou and Jean Tirole, 2011). In this model, people who are usually described as selfish types are characterized by having low values of γ_i. They will behave as if they were purely payoff-driven and maximize u_i by picking ${c_i}^{**} = {c_i}^{*} = 0$. But if a subject's γ_i is sufficiently large, she will maximize her overall utility by choosing ${c_i}^{**} = \tilde{c}_i$. People commonly described as altruists are characterized by having both a high $\tilde{c}_i$ and a high γ_i. While we do not want to exclude that a participant is exclusively motivated by identity, we do not deem this likely. We deem it considerably more probable that participants are simultaneously motivated by other-regarding preferences and by self-regarding preferences, i.e. identity.

The critical parameter of the model is γ_i. Yet there is still little insight with respect to the determinants of γ_i. We propose to split up the relative weight of identity utility into a personal characteristic $\omega_i \geq 0$, which can be regarded as an agent's intrinsic concern for her self-image, and a context variable $s \in [0,1]$. We further specify a simple linear relationship:

$$\gamma_i = s \cdot \omega_i \quad (4)$$

The idea is that in normatively unambiguous situations ($s = 1$), the relative weight of identity is only determined by a person's intrinsic concern for her self-image. But as the normative context becomes less clear, the definition of appropriate behavior becomes increasingly blurred and people confine their focus more and more to their selfish interests.

An important reason for this behavior seems to be self-deception. People like earning money and they like seeing themselves as *a good person*. However, often their wish to earn money entices them to deviate from what 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 $\tilde{c}_i$. 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. A recent

study by 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. The experimental evidence thus supports Adam Smith's assertion that "self-deceit, this fatal weakness of mankind, is the source of half the disorders of human life" (III.4.6).

Smith believes that in order to enhance the normative clarity of a situation, and thereby reduce the room for self-deceit, "we must become the impartial spectators of our own character and conduct" (III.2.3). This suggests that exogenously inducing participants to engage in introspection will increase s, which translates into a larger effect of identity utility γ, and increases the probability that the individual has higher utility from contributing $\tilde{c}_i$, rather than freeriding.

The model assumes the intrinsic normative standard $\tilde{c}_i$ to be unaffected by introspection. Only the normative clarity of the situation s is expected to react to the introspection manipulation. Yet arguably just asking participants to elaborate their first order beliefs about the contributions of the remaining group members (Treatment FB) is unrelated to identity and does therefore not affect the normative clarity of the situation. Asking them to elaborate their second order beliefs (SB) might already have an effect. While this question is only empirical, not normative, it might make it salient to participants that they are about to fall below their personal normative standard. Following Adam Smith, the effect should be even stronger if they are induced to directly express their personal normative standard. Yet if this normative standard is a personal ideal, participants still have an excuse. They may assuage bad conscience by the argument that circumstances do not permit living up to such a high standard. This excuse is removed if, instead, they are induced to formulate their subjective normative minimum, i.e. the level of contributions that should be upheld, however bad the environment. Consequently our theory leads to the following prediction:

Hypothesis: In a linear public good, contributions are highest if participants are induced to elaborate their subjective normative standard for the level of contributions, whatever the circumstances. Contributions are lower if they are induced to formulate their subjective normative ideal, and yet lower if they are asked to explicate second order beliefs. If participants are led to formulate first order beliefs, contributions cannot be distinguished from a baseline with no introspection manipulation.

4. Results

a. Treatment Effects on Contributions

Figure 1 suggests that Smithian introspection indeed increases cooperation. On average, cooperation increases from 35% in the Baseline to 45% in SB, 48% in NI, and 55% in NM. In contrast, there is virtually no change of cooperation in the FB-treatment (35%), as expected by our theory. According to a simple Mann-Whitney ranksum test over group means there is a weakly significant increase from the Baseline to the SB-treatment ($p=0.087$, $N=22$, two-sided) and a significant increase to the NM-treatment ($p=0.043$, $N=24$, two-sided). Introspection 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 the FB-treatment even in period 4. In contrast, in NI and NM cooperation only drops for good below 10 in periods 23 and 24 respectively; in the SB-treatment in period 15. 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 identity utility and material utility.

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[6] in Table 2 confirms the robust effect of NM on cooperation. In contrast, the effects of SB and NI remain descriptive. If we separately calculate marginal effects of treatment from this model for each period, we find a significant difference between the Baseline and the NM-treatment for every period between the first and the 22nd. For the remaining periods, the difference is significant at the 10% level ($p < .088$).

> *Result 1: Contributions are higher if participants are induced to formulate the normative minimum.*

6 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.

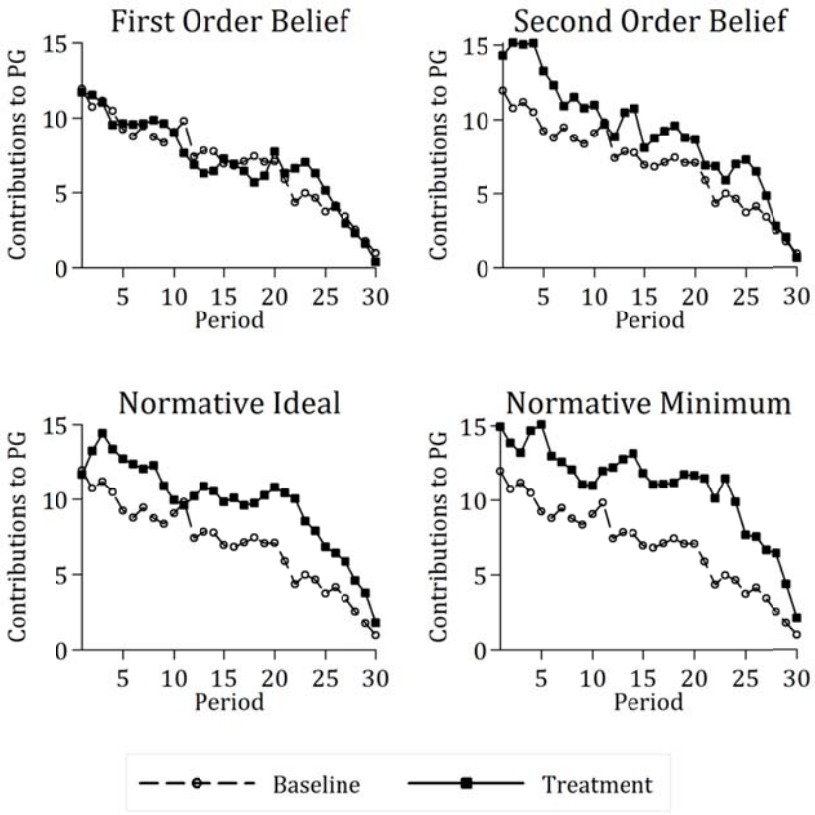

Figure 1: Contributions over time

Note: The dashed line denotes the mean contribution to the public good in the Baseline, the solid line denotes the mean contribution in the respective treatment. The Baseline has 48 subjects, FB has 44 subjects, SB 40, NI 48, and NM 48.

Table 2: Treatment Effects on Contribution Behavior

First Order Belief (**FB**)	-.538
	(4.011)
Second Order Belief (**SB**)	2.523
	(4.114)
Normative Ideal (**NI**)	5.724
	(3.924)
Normative Minimum (**NM**)	8.674*
	(3.932)
Constant	5.057^{+}
	(2.773)
# Observations	6840
thereof left-censored	2101
thereof right-censored	1490
# Individuals	228
# Groups	57

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. *** p < .001, ** 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. Most striking is the difference between NI and the three other treatments. On average, every subject adapts her statement of the normative ideal only 5.83 times over the 30 periods of interaction, by a total amount of 3.08 tokens. The median expectation is 20 in every single period. In contrast, subjects display considerably less reservations about adjusting their expectation of a normative minimum, 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, the NM-treatment has a (descriptively) stronger effect on cooperation than the NI-treatment even though the stated normative expectations in NM are both significantly lower (p=0.009, Ranksum Test, N=24, two-sided) and decay significantly faster (p=0.009, Ranksum Test, N=24, two-sided). In fact, stated expectations in NM are much more similar to FB and SB than to NI. There is neither a significant difference of level (10.31 in NM vs. 9.53 in SB and 7.46 in FB), nor of decay over time (8.34 in NM vs. 11.02 in SB and 10.25 in FB).

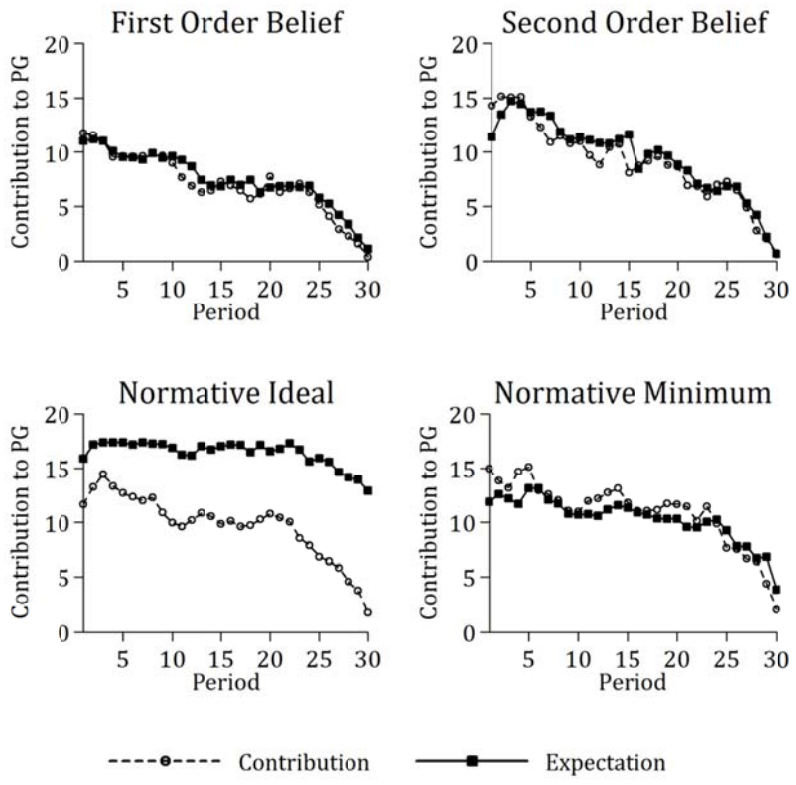

Figure 2: Expectations and Contributions over Time

Note: The dashed line denotes the mean contribution to the public good in the respective treatment, the solid line denotes the mean expectation (see Table 1) in the respective treatment. FB has 44 subjects, SB 40, NI 48, and NM 48.

To better understand the process by which introspection affects choices, we estimate a structural model (Figure 3). In line with our theory, we expect a subject's contribution c_{it} in the current period to be driven by two forces: other-regarding and self-regarding concerns. We capture other-regarding concerns by the average contribution of the remaining group members in the preceding period $\overline{c}_{j,t-1}$. This measure directly corresponds to behavior motivated by inequity aversion. We expect self-regarding concerns to result from introspection, which we manipulate with our treatments (that we represent in the path diagram by letter τ). We cannot directly observe the subjective standard $\tilde{c}_i$, the idiosyncratic weight ω_i, the perceived level of normative ambiguity s, and hence the degree of disutility γ following from violating one's subjective standard. Yet we see the reported subjective standard $\hat{c}_i$. In the initial period, this report cannot be influenced by experiences. We therefore expect it to be a function of the idiosyncratic components $\tilde{c}_i$ and ω_i and of perceived ambiguity s as influenced by treatment τ. 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, participants adjust their own *reported* standard[7] to the experiences they are making. The chain of adjustment is the perceived level of ambiguity s: the further the choices of others deviate from one's own choices, the more the situation is perceived as normatively ambiguous.

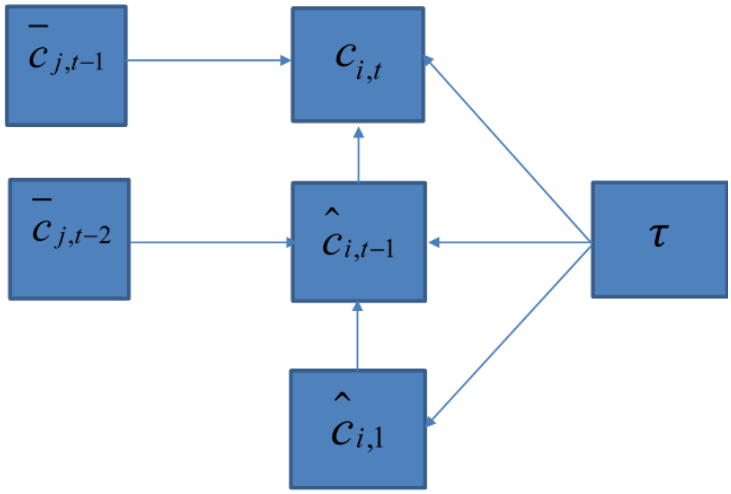

Figure 3: Structural Model

7 Not their idiosyncractic normative standard $\tilde{c}_i$.

The regression in Table 3 tests this structural model. It is best read from bottom to top. When they are asked to state first order beliefs, in the first round participants on average say they expect other group members to contribute 12 of 20 tokens (constant of the final component of the model). If participants are, instead, asked to formulate second order beliefs, statements are neither less nor more demanding (insignificant treatment effect of SB treatment). Yet the stated normative minimum is about 1 unit more demanding, and the stated normative ideal is 7.5 units more demanding.

The stated expectation in the first period, i.e. before having any information about others' behavior, predicts stated expectations in later periods. According to our model, stated normative expectations are a function of the intrinsic normative standard. Yet this effect is not very strong (controlling for experiences, the correlation between stated expectations in the first and in later periods is only around .4). By contrast, experiences (measured by the contributions of the remaining group members in the previous period) do have a strong effect. As expected, stated expectations are not only a function of idiosyncratic standards, but also of perceived normativity in the community. Conditional on both initial stated expectations and experiences, only the NI treatment stands out. The normative ideal remains high above experiences.

Finally, contributions independently respond to experiences (significant effect of the average contribution of the remaining group members in the previous period) and to expectations (significant effect of the stated expectation in the previous period). Payoff comparisons are not the only motivating force. There is a strong additional effect of identity utility. Even if we control for these two driving forces, we still find a weakly significant ($p = .060$) treatment effect of asking for the normative minimum.

With the structural model, we are also able to estimate indirect effects. If we ask participants to report their normative ideal, this influences both their initial and their current reported standards, which translates into contribution choices through the effect of the current normative ideal. If we ask them to report the normative minimum, this indirectly influences their current choices through the effect on the initial reported standard. Adding the treatment effect on choices up with these two indirect effects, we find a highly significant total effect of asking for the normative ideal, and a weakly significant ($p = .062$) total effect of asking for the normative minimum.

Table 3: Structural Model

Dependent Variables	Independent Variables	Direct effect	Once indirect effect	Twice indirect effect	Total effect
c_{it}	$\hat{c}_{i,t-1}$	.455*** (.035)			
	$\overline{c}_{j,t-1}$	1.199*** (.039)			
	SB	-.227 (1.316)	.011 (.162)	.069 (.065)	-.147 (1.324)
	NI	-1.142 (1.453)	5.203*** (.437)	1.354*** (.141)	5.414*** (1.469)
	NM	2.510^{+} (1.337)	-.200 (.156)	.204** (.066)	2.514^{+} (1.345)
	cons	-8.520*** (1.122)			
$\hat{c}_{i,t-1}$	$\hat{c}_{i,1}$	.391*** (.020)			
	$\overline{c}_{j,t-2}$	1.095*** (.022)			
	SB	.023 (.355)			
	NI	11.432*** (.380)			
	NM	-.439 (.341)			
	cons	-5.766*** (.370)			
$\hat{c}_{i,1}$	SB	.390 (.361)			
	NI	7.603*** (.361)			
	NM	1.148** (.353)			
	cons	11.996*** (.254)			
	# Observations	5400			
	thereof left-censored	2101			
	thereof right-censored	1490			
	# Individuals	228			
	# Groups	57			

Non-linear (Tobit) structural model, with lower censoring at 0 and upper censoring at 20. Random effects for individuals nested in groups. We exclude the Baseline, because in the Baseline we do not elicit expectations. SB, NI and NM are treatment dummies. First indirect effect (column 4): effect of stated expectation in previous period on contribution * effect of treatment on stated expectation in previous period. Twice indirect effect (column 5): effect of stated expectation in previous period on contribution * effect of stated expectation in first period on stated expectation in previous period * effect of treatment on stated expectation in first period. Total effect (column 6) = first indirect effect + twice indirect effect + effect of treatment on contributions. All significance tests of indirect and total effects are calculated using the delta method (Stata command nlcom). *** $p < .001$, ** $p < .01$, * $p < .05$, $^{+}$ $p < .1$.

5. Conclusion

Inspired by Adam Smith's idea that human selfishness can be restrained by becoming "the impartial spectators of our own character and conduct" (III.2.3), this paper has tested the extent to which introspection can reduce selfish behavior and thus increase cooperativeness in a public good game. Theoretically, we have shown that other-regardingness alone is not able to explain typical cooperation patterns. The concept of identity utility provides a framework to rationalize why people may behave socially out of a self-regarding motivation. In this framework, introspection enhances cooperation by increasing subjects' consciousness of their own normative goals and thus the relative weight of identity utility.

To test the Smithian conjecture we have proposed an experimental design that exogenously varies normative consciousness without imposing any specific normative goals ourselves. Our experimental results yield strong support both to the theoretical concept of identity utility and to the behavioral effect of introspection. Between treatments, we show that being nudged to actively deliberate about one's own normative goals causally increases cooperation.

Using a structural model, we show how experiences and expectations independently explain choices. This speaks against the idea that conditional cooperators exclusively care about the behavior of other group members. The (other-regarding) desire not to outperform others is not strong enough to explain cooperation, particularly when paired with the desire not to be outperformed by others. We show that the necessary additional impulse for cooperative behavior may come from the (self-regarding) desire to live up to one's own normative expectations.

This sense of normativity can be enhanced by inducing participants to use introspection. The effect of the introspection manipulation on stated expectations is strongest if participants are asked to elaborate their normative ideal. Yet the effect of the introspection manipulation on contributions is strongest if participants are, instead, asked to explicate the normative minimum. This result supports the idea that the self-regarding motive is not just there. It is sensitive to the normative ambiguity of the situation. If one considers the normative ideal, one is forced to elaborate the conflict between one's ideal world and the less ideal reality, as experienced in the behavior of others with whom one shares the context. By contrast if one considers the normative minimum, one has already translated the intrinsic ideal into a potentially more modest, realistic goal. Whether or not other community members live up to this standard is less relevant for one's behavior. The "jurisdiction of the man within" is more likely to take precedence.

References

Akerlof, George A. and Rachel E. Kranton. 2000. "Economics and Identity." *The Quarterly Journal of Economics*, 115(3), 715-53.

____. 2005. "Identity and the Economics of Organizations." *Journal of Economic Perspectives*, 19, 9-32.

Battigalli, Pierpaolo and Martin Dufwenberg. 2007. "Guilt in Games." *American Economic Review*, 97(2), 170-76.

Bénabou, Roland and Jean Tirole. 2011. "Identity, Morals, and Taboos: Beliefs as Assets." *The Quarterly Journal of Economics*, 126(2), 805-55.

Bicchieri, Cristina. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge UK: Cambridge University Press.

Blanco, Mariana; Dirk Engelmann and Hans-Theo Normann. 2011. "A within-Subject Analysis of Other-Regarding Preferences." *Games and Economic Behavior*, 72, 321-38.

Camera, Gabriele and Marco Casari. 2009. "Cooperation among Strangers under the Shadow of the Future." *American Economic Review*, 99(3), 979-1005.

Chaudhuri, Ananish. 2011. "Sustaining Cooperation in Laboratory Public Goods Experiments. A Selective Survey of the Literature." *Experimental Economics*, 14, 47-83.

Croson, Rachel T.A. 2007. "Theories of Commitment, Altruism and Reciprocity. Evidence from Linear Public Goods Games." *Economic Inquiry*, 45, 199-216.

Dal Bó, Ernesto and Pedro Dal Bó. 2009. "" Do the Right Thing." The Effects of Moral Suasion on Cooperation," National Bureau of Economic Research,

Dana, Jason; Roberto A. Weber and Xi Kuang. 2007. "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory*, 33(1), 67-80.

Falk, Armin and Nora Szech. 2013. "Morals and Markets." *Science*, 340, 707-11.

Fehr, Ernst and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, 114, 817-68.

Fischbacher, Urs. 2007. "Z-Tree. Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics*, 10, 171-78.

Fischbacher, Urs and Simon Gächter. 2010. "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Good Experiments." *American Economic Review*, 100, 541-56.

Fischbacher, Urs; Simon Gächter and Ernst Fehr. 2001. "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment." *Economics Letters*, 71, 397-404.

Friedman, Daniel and Ryan Oprea. 2012. "A Continuous Dilemma." *American Economic Review*, 102(1), 337-63.

Gächter, Simon and Elke Renner. 2010. "The Effects of (Incentivized) Belief Elicitation in Public Goods Experiments." *Experimental Economics*, 13, 364-77.

Greiner, Ben. 2004. "An Online Recruiting System for Economic Experiments," K. Kremer and V. Macho, *Forschung Und Wissenschaftliches Rechnen 2003*. Göttingen: 79-93.

Hart, Herbert Lionel Adolphus. 1961. *The Concept of Law*. Oxford: Clarendon Press.

Isaac, R. Mark; Kenneth F. McCue and Charles R. Plott. 1985. "Public Goods Provision in an Experimental Environment." *Journal of Public Economics*, 26(1), 51–74.

Kreps, David M.; Paul R. Milgrom; John Roberts and Robert B. Wilson. 1982. "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma." *Journal of Economic Theory*, 27, 245-52.

Krupka, Erin L. and Roberto A. Weber. 2009. "The Focusing and Informational Effects of Norms on Pro-Social Behavior." *Journal of Economic Psychology*, 30, 307-20.

____. 2013. "Identifying Social Norms Using Cordination Games. Why Does Dictator Game Sharing Vary?" *Journal of the European Economic Association*, 11(3), 495-524.

Ledyard, John O. 1995. "Public Goods. A Survey of Experimental Research," J. H. Kagel and A. E. Roth, *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press, 111-94.

Neugebauer, Tibor; Javier Perote; Ulrich Schmidt and Malte Loos. 2009. "Selfish-Biased Conditional Cooperation. On the Decline of Contributions in Repeated Public Goods Experiments." *Journal of Economic Psychology*, 30(1), 52-60.

Smith, Adam. 1790. *The Theory of Moral Sentiments, or, an Essay Towards an Analysis of the Principles by Which Men Naturally Judge Concerning the Conduct and Character, First of Their Neighbours, and Afterwards of Themselves. To Which Is Added, a Dissertation on the Origin of Languages*. London: Strahan.

Tyran, Jean-Robert and Lars P. Feld. 2006. "Achieving Compliance When Legal Sanctions Are Non-Deterrent." *Scandinavian Journal of Economics*, 108, 135-56.

Zelmer, Jennifer. 2003. "Linear Public Goods. A Meta-Analysis." *Experimental Economics*, 6, 299-310.

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.

Total income (in taler)	=	Income from the common project	+	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.

Income from the common project	=	total sum of all contributions to the project (within your group of four)	×	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 0.4×60 = 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 0.4×9 = 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 0.4×1 = 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 0.4×4 = 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 0.4×1 = 0.4 taler. Hence, if each member of your group of four contributes 1 taler to the project, each of you receives 0.4×1×4=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:

[FB] What do you believe do the other group members on average contribute to the project?
[SB] What do you believe do the other group members think that you contribute to the project?
[NI] What do you believe should every group member contribute to the project?
[NM] 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?