I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15322

Preferences and Perceptions in Provision and Maintenance Public Goods

Simon Gächter
Felix Kölle
Simone Quercia

MAY 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15322

Preferences and Perceptions in Provision and Maintenance Public Goods

Simon Gächter
University of Nottingham, CESifo and IZA

Felix Kölle
University of Cologne

Simone Quercia
University of Verona

MAY 2022

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Preferences and Perceptions in Provision and Maintenance Public Goods

We study two generic versions of public goods problems: in Provision problems, the public good does not exist initially and needs to be provided; in Maintenance problems, the public good already exists and needs to be maintained. In five lab and online experiments (n=2,584), we document a robust asymmetry in preferences and perceptions in two incentive-equivalent versions of these public good problems. We find fewer conditional cooperators and more free riders in Maintenance than Provision, a difference that is replicable, stable, and reflected in perceptions of kindness. Incentivized control questions administered before gameplay reveal dilemma-specific misperceptions but controlling for them neither eliminates game-dependent conditional cooperation, nor differences in perceived kindness of others' cooperation. Thus, even when sharing the same game form, Maintenance and Provision are different social dilemmas that require separate behavioral analyses. Despite some inconsistencies, a theory of revealed altruism comes closest to explaining our results.

JEL Classification: C92, H41

Keywords: maintenance and provision social dilemmas, conditional cooperation, kindness, misperceptions, experiments, revealed altruism

Corresponding author:
Simon Gächter
School of Economics
University of Nottingham
Sir Clive Granger Building
University Park
Nottingham NG7 2RD
United Kingdom

E-mail: simon.gaechter@nottingham.ac.uk

1. Introduction

In this paper, we study two generic forms of voluntary cooperation: providing initially inexistent public goods and maintaining existing ones. Contributing to charities, volunteering, being a team player, or participating in collective action, are examples of voluntary cooperation that provides public goods. Shared natural resources, known as "common-pool resources" (e.g., Ostrom (1990)), but also biodiversity and a stable climate, are important public goods that nature has provided but people need to limit extraction or environmentally damaging emissions if they want to maintain them. Similarly, public goods that previous generations created, such as democracy and the rule of law, only continue existing if people limit rule-bending, rent-seeking, and corruption.

As the examples illustrate, "provision" and "maintenance" public goods differ along many dimensions. Crucially, however, for selfish players, they are all *social dilemmas*: providing or maintaining the public good is often collectively beneficial, but individual incentives are to hold back on provision and to exploit rather than to maintain the public good – the "tragedy of the commons" (Hardin (1968)). Although the comparison between these two problems has been studied both in economics and psychology (see, e.g., Sell and Son (1997) and Dufwenberg et al. (2011)), until recently, most studies only investigated cooperative *behavior*, and less the psychological *mechanisms* that produce cooperation.[1] Here, we ask whether, from the perspective of social preferences and perceptions, maintenance and provision dilemmas are psychologically different social dilemmas.

Studying *preferences* and *perceptions* as drivers of cooperation and understanding whether their impact on cooperative behavior differs across maintenance and provision dilemmas is important from both a theoretical and practical point of view. From a theoretical viewpoint, studying *preferences* and *perceptions* is interesting because it can help explain why people often cooperate even in anonymous one-shot games without communication, where mechanisms that can support cooperation, such as reputation or repeated interactions (e.g., Dal Bó and Fréchette (2018); Rand and Nowak (2013)) do not apply (see, e.g., Fischbacher and Gächter (2010); and Gächter et al. (2017) for evidence from the lab; and, e.g., Frey and Meier (2004); Alpizar et al. (2008); Rustagi et al. (2010); Fehr and Leibbrandt

[1] The behavioral evidence about cooperation in maintenance and provision dilemmas comes from largely separate literatures. For cooperation in maintenance (common-pool resource) problems, see, e.g., the surveys by Ostrom (1990) and Ostrom (2006). Evidence on cooperation in public goods provision problems is surveyed in, e.g., Ledyard (1995); Gächter and Herrmann (2009); Chaudhuri (2011) and Fehr and Schurtenberger (2018).

(2011) for evidence from the field). Do people cooperate because they misperceive their incentives in a social dilemma, or do they have a 'genuine' preference for cooperation? If people misunderstand their incentives, they may implement choices they otherwise wouldn't. Thus, observing cooperation without controlling for perception of incentives may not be a conclusive revelation of a preference for cooperation (for related arguments see, e.g., Koszegi and Rabin (2008) and Cason and Plott (2014)). Apart from potential differences in the understanding of the incentive structure, maintenance and provision might also differ in the way people perceive others' actions, e.g., in terms of kindness. Perceptions of the kindness of other players' actions are important because, in social dilemmas, they can explain why some people cooperate in the first place (e.g., Falk and Fischbacher (2006)).

The practical relevance of our research question comes from the fact that any policy intervention aimed at fostering cooperation must rest on accurate behavioral mechanisms. If these mechanisms are dilemma-specific, this would imply that maintenance and provision require different approaches to overcome the tragedy of the commons.

To answer these questions, and to provide a comprehensive understanding of the fundamental nature of cooperation in maintenance and provision dilemmas, we present the results from a series of experiments, in which we study preferences and perceptions in conjunction. We compare two versions of a linear public good game that share the same game form: A Provision game and a Maintenance game. In *Provision*, the public good initially does not exist; four players in a group are endowed with 20 tokens each and decide simultaneously how many of them to contribute to the public good. In *Maintenance*, players have no endowment, but the public good already exists because 80 tokens are invested at the outset in the public good. Players decide simultaneously how many (up to 20) tokens to withdraw from the public good. Any token contributed to the public good (in *Provision*) or not withdrawn from the public good (in *Maintenance*) is worth 1.6 money units to the group, which is then shared equally between group members; any token not contributed to the public good or withdrawn from the public good is worth 1 money unit.[2]

To ascertain the role of preferences and perceptions in influencing cooperation in these two dilemmas, we proceed in four steps that we summarize in Table 1.

[2] We focus sharply on the social dilemma dimension of provision and maintenance public goods, and abstract from technological features (e.g., resource rivalry in common-pool resources vs. non-rivalrous public goods) and institutional details (rules and regulations) that define real-world social dilemmas (e.g., Ostrom (1990); Cornes and Sandler (1996); Poppe (2005); Apesteguía and Maier-Rigaud (2006); Levin (2014)).

Table 1: Four steps to test whether Maintenance and Provision are different dilemmas

Exp.	Experiment	Number of participants	Subject pool	Purpose
Step 1: Establishing the replicability and stability of cooperative preferences (Section 3)				
0	Gächter et al. (2017)*	(*n* = 703)	Students (UoN)	These data provide a previous benchmark result, which we replicate in Experiment 1
1	Replication study	*n* = 704	US citizens (MTurk)	Assess whether Experiment 0 with students can be replicated in a non-student subject pool
2a	Temporal stability (5 months delay)	*n* = 119	Students (UoN, sampled from experiment 0)	Assess the role of stability of conditionally cooperative preferences over time
2b	Predictive power of cooperation attitudes	*n* = 116	Students (UoN, sampled from experiment 4)	Assess the predictive power of conditionally cooperative preferences plus beliefs to explain actual cooperation levels
Step 2: Measuring perceptions of kindness (Section 4)				
3	Kindness survey	*n* = 185	Students (UoN, new participants)	Measure how kind or unkind people perceive a certain cooperation level to be. Measured on a scale of -100 (=very unkind) to +100 (= very kind)
		n = 401	US citizens (MTurk, new participants)	
Step 3: Measuring game form misperceptions and controlling for them (Section 5)				
4	Dilemma-specific game-form misperceptions	*n* = 696	Students (UoN, new participants)	Measure with 8 incentivized questions about payoffs and goals how people perceive the incentives in the public good game. Control for misperceptions to test for potential differences in cooperative preferences
Step 4: Assessing theoretical explanations (Section 6)				
5	Guilt survey	*n* = 347	Students (UoN, new participants)	Measure perceptions of guilt to evaluate the explanatory power of guilt aversion to explain our results
		n = 402	US citizens (MTurk, new participants)	

* Data taken from https://doi.org/10.5061/dryad.8d9t2.

Our *first step* is to ensure that what we study is a replicable phenomenon. This is the purpose of Experiment 1. In this experiment, we replicate the one-shot results of Gächter et

al. (2017) using a diverse online subject pool (MTurk). We find that in a simultaneous one-shot game, people contribute 52% of their endowment in *Provision* compared to 39% in *Maintenance,* a difference that is highly statistically significant. Furthermore, using the Fischbacher et al. (2001) strategy-method experiment to separate beliefs from preferences, we replicate that there are systematically fewer conditional cooperators and more free riders in *Maintenance* than *Provision*. In two additional experiments (Experiments 2a and 2b in Table 1), we collect new evidence to test whether, within-participants, our measure of cooperation preferences is stable over time and, together with beliefs, predicts contribution or withdrawal decisions in one-shot games played immediately or five months after preferences were elicited.

In Steps 2 and 3 we turn to our central question of how people perceive others' behavior (Step 2) and the incentives in the dilemmas (Step 3). Our *second step* measures people's perceptions about the kindness of others' behavior. While differences in kindness perceptions could support a preference interpretation, it is possible that some people *misperceive* the game form because they do not understand the material incentives of the public good game. Such "confusion" is likely because of previous evidence (e.g., Andreoni (1995a); Houser and Kurzban (2002); Ferraro and Vossler (2010); Bayer et al. (2013)). Moreover, irrespective of the game form, cooperation preferences as measured by the strategy method might be influenced by misperceptions (Burton-Chellew et al. (2016)).

Testing for misperception of incentives is our *third step.* To this end, we designed a set of eight *incentivized* control questions that people answered after they had correctly solved ten standard understanding questions covering payoffs in the public good game. We administered the incentivized control questions *before* we measured participants' cooperation preferences using the strategy method by Fischbacher et al. (2001).[3] Controlling for people's misunderstanding will then allow us to test whether cooperation preferences continue to differ statistically significantly between *Maintenance* and *Provision*. Our *fourth step* is to discuss how various theories of social preferences, in particular revealed altruism and guilt aversion, can explain our results.

[3] To measure confusion, previous studies used various experimental designs, like changed incentive structures (Andreoni (1995a)), information conditions (Bayer et al. (2013)), or computerized players (Houser and Kurzban (2002); Ferraro and Vossler (2010); Burton-Chellew et al. (2016)). Fosgaard et al. (2017) used an incentivized post-experimental questionnaire.

Our paper offers several contributions to the literature. Our methodology of holding the nature of the social dilemma constant relates us to literatures on framing and context effects in other-regarding behavior.[4] More specifically, our focus on provision/maintenance of public goods leads us naturally to a design that is a version of what psychologists (Dawes (1980)) have called Take-some vs. Give-some dilemmas (e.g., Sell and Son (1997); van Dijk and Wilke (1997); Sonnemans et al. (1998); Messer et al. (2007); Cubitt et al. (2011a); Dufwenberg et al. (2011); Cox (2015); Cox and Stoddard (2015); Fosgaard et al. (2014); Fosgaard et al. (2017); Khadjavi and Lange (2015); Isler et al. (2021)). Our maintenance/provision design differs from designs that manipulate whether the positive externality of contributing to the public good or the negative externality of not contributing is emphasized. Papers in this line of research are Andreoni (1995b); Park (2000); and Fujimoto and Park (2010). For a comparative discussion of give/take or positive/negative externality framing effects and an overview of studies see Cartwright (2016).[5]

Different from most early literature on give-some and take-some, our analysis consists of separating preferences and perceptions as determinants of cooperation rather than focusing only on cooperation decisions. With regard to preferences, our paper joins the small literature that elicits preferences for conditional cooperation in maintenance or provision problems (e.g., Frackenpohl et al. (2016); Gächter et al. (2017); Fosgaard et al. (2014); Fosgaard et al. (2017); Isler et al. (2021)). Our experiments also measure the perception of incentives *before* we elicit preferences for cooperation. This heeds arguments by Koszegi and Rabin (2008) and Cason and Plott (2014) that measuring preferences requires controlling for the perception of incentives.[6]

Our goals relate us to Fosgaard et al. (2017). They also measure preferences for conditional cooperation and misperceptions albeit after the elicitation of preferences and with fewer questions. Theirs is a representative subject pool from Denmark (n = 2,042), whereas we present evidence from a student subject pool in the UK and online workers

[4] Our focus is on public goods games, but also relates to the importance of context effects. For instance, previous evidence from dictator games reveals that people are less willing to give if the choice set also includes the option to take away money (List (2007); Bardsley (2008); Cappelen et al. (2013); Dreber et al. (2013); Korenok et al. (2014); Bicchieri et al. (2022)). Also, the approval of egoistic behavior seems to be context-dependent too, e.g., in markets vs. non-market settings (Bartling et al. (2021)).

[5] Another dimension of framing effects is due to attaching labels to games (e.g., "Wallstreet vs Community game"; e.g., Ellingsen et al. (2012) and Dufwenberg et al. (2011)). In this paper we use neutral labels.

[6] Our focus on social preferences and (mis-)perception does not deny the possibility that cognitive ability, risk preferences and loss aversion might matter too (e.g., De Dreu and McCusker (1997); Iturbe-Ormaetxe et al. (2011)), but we leave this for future research, not least to keep this paper manageable.

(MTurk) from the general population in the US (total n = 2,854). More importantly, unlike Fosgaard et al. (2017), we report evidence on the temporal stability of preferences, perceptions of kindness and guilt, and link our results to theories of social preferences. Our four-step analysis that combines lab and online experiments, replications, between and within-subject stability tests, measurement of perceptions of kindness as well as of the incentives in the games, establishes that *Maintenance* and *Provision* are different social dilemmas even when sharing the same game form and when controlling for possible misperceptions of incentives.

2. The basic setup and the proxy for cooperation preferences

Our setup consists of the two social dilemmas described above, *Provision* and *Maintenance*. In both conditions, participants are randomly assigned to groups of n = 4. In *Provision*, each group member i is endowed with 20 tokens, which they can either keep or (partly or fully) contribute (c_i) to a "group project". Contributions to the group project are summed up, multiplied by a factor of 1.6, and distributed equally among the four members. Equation (1) describes the material incentives of individual i:

$$\pi_i = 20 - c_i + \frac{1.6}{4}\sum_{j=1}^{4} c_j. \quad (1)$$

In *Maintenance*, 80 tokens are initially placed in a "group project". Each group member i decides about the allocation of 20 tokens, which they can either leave or (partially or fully) withdraw (w_i) from the project. Material incentives are described by equation (2):

$$\pi_i = w_i + \frac{1.6}{4}(80 - \sum_{j=1}^{4} w_j). \quad (2)$$

If people are only motivated by material incentives, (1) and (2) are incentive-equivalent social dilemmas because $c_i = 20 - w_i$. Furthermore, because the material costs of cooperation outweigh its benefits, both the *Maintenance* and *Provision* dilemma have full free-riding ($c_i = 0; w_i = 20$) as the unique Nash equilibrium in dominant strategies, no matter what other members of their group (are believed to) do.

All experiments were based on these two incentive-equivalent social dilemmas and consisted of several parts. In the first part of each experiment, participants were introduced to the basic decision situation explaining either the *Maintenance* or the *Provision* dilemma

and its incentive structure, that is, each participant only faced one of the two social dilemmas (between-subjects design). To ensure understanding, participants then had to complete a set of ten computerized control questions. Only after correctly answering all of them, participants could proceed with the experiment.

The exact design of the remaining parts differed across our experiments.[7] In most of our experiments, in the second part we implemented a strategy-method public goods game (described below) through which we measure *cooperation attitudes*, our main proxy for cooperation preferences. Some of the sessions in these experiments included a third part in which participants played a direct-response game in which they simultaneously had to state their contribution decision and belief about others' contributions. In the experiments in which we elicited game-form misperceptions, the strategy-method game in part 2 was preceded by a set of incentivized control questions. In the following, we explain how we elicited cooperation attitudes, which is our main variable of interest. All instructions and control questions are in Online Appendix A.

To elicit a proxy for cooperation preferences we used the design introduced by Fischbacher et al. (2001), which employs a variant of the strategy method (Selten (1967)). This design elicits an individual's willingness to cooperate as a function of other group members' cooperation. Participants played a one-shot version of the game and were asked to make an *unconditional* and a *conditional* contribution (or withdrawal) decision. In the unconditional decision, participants chose one contribution or withdrawal level. In the conditional decision, participants were asked to fill in a table in which they had to indicate their contribution (or withdrawal) decision for *each* possible (rounded) average contribution (or withdrawal) of the other three group members. To guarantee incentive compatibility, in each group a random mechanism selected three members for whom the unconditional decision was payoff-relevant and one member for whom the conditional decision was payoff-relevant. For this participant, the conditional decision was calculated according to the (rounded) average unconditional decision of the other three group members. The incentive-compatibly elicited attitudes are a proxy for cooperation preferences in the sense that they measure people's willingness to pay for conditional cooperation.

[7] At the beginning of the experiment, participants were told that the experiment consists of several parts, but that the details about later parts would be disclosed only after they had completed the respective parts. The different designs of the later parts could therefore not affect behavior in previous parts.

Following Fischbacher et al. (2001), we classify a participant as a (i) *conditional cooperator* if their contribution/withdrawal schedule exhibits a (weakly) monotonically increasing pattern, or if the Spearman correlation coefficient between their schedule and the others' average contribution (or withdrawal) is positive and significant at $p < 0.01$; (ii) a *free rider* if they never contribute anything or withdraw everything irrespective of how much the others contribute (or withdraw); and (iii) as *other* if none of the criteria in (i) & (ii) apply.[8]

Our data come from six experiments and three main sources: the CeDEx lab at the University of Nottingham; the online labor market platform Amazon Mechanical Turk (MTurk); and online experiments conducted with students at the University of Nottingham (see Table 1 above for an overview of our experiments). A total of 2,854 people participated in our experiments. We used z-Tree (Fischbacher (2007)) for conducting the laboratory sessions. For the online experiments on MTurk and the University of Nottingham, we used the survey software Qualtrics. For the lab and online experiments at Nottingham, we recruited student participants (average age 20.2 years; 58% female) from various disciplines at the University of Nottingham using the software ORSEE (Greiner (2015)). Students were only allowed to participate in one lab or online session. On MTurk, participants (all US residents) were 31.9 years old and 41% were female.[9] Average payments were £20.60 for lab sessions, and $2.60 for MTurk sessions (corresponding to an hourly wage of $13.00).

3. Step 1: Replicability and stability of preferences in Provision and Maintenance dilemmas

We start by summarizing the findings from our previous study (Gächter et al. (2017)). We then compare these results with an online replication study conducted on MTurk. Being able to replicate the basic phenomenon we want to study is an important first step in our analysis.[10] After that, we show that cooperation preferences are not only stable between different subject pools but are also stable over time within participants. Finally, we investigate the predictive power of the elicited cooperation preferences for simultaneous

[8] As a robustness check, we used an alternative classification method by Thöni and Volk (2018), who proposed a refinement of the criteria of Fischbacher et al. (2001). All results are qualitatively and quantitatively in line with those reported below.

[9] See Horton et al. (2011) and Arechar et al. (2018) for a detailed description of MTurk, and a comparison of MTurk versus lab experiments. Both studies as well as Snowberg and Yariv (2021) demonstrate that behavior in a variety of games is similar on MTurk and the lab.

[10] See Maniadis et al. (2014) Camerer et al. (2016) and Camerer et al. (2019) on the importance of replicability in experimental economics.

gameplay and compare it across *Maintenance* and *Provision*. All procedural details and further supporting evidence are in Online Appendix B1.

3.1. Gächter et al. (2017) and a replication on MTurk

The left panel of Figure 1 summarizes the main relevant finding for our paper from Gächter et al. (2017), which was based on a one-shot strategy method experiment as described in Section 3. Participants were significantly more likely to be conditional cooperators ($\chi^2(1) = 31.03$; $p < 0.001$) and significantly less likely to be free riders ($\chi^2(1) = 10.46$; $p = 0.001$) and others ($\chi^2(1) = 11.08$; $p = 0.001$) in *Provision* than in *Maintenance*.[11] In a one-shot direct response game played after the type elicitation, Gächter et al. (2017) further found that cooperation rates were significantly higher in *Provision* than in *Maintenance* (41% vs. 30%; two-sided t-test: $p = 0.007$).

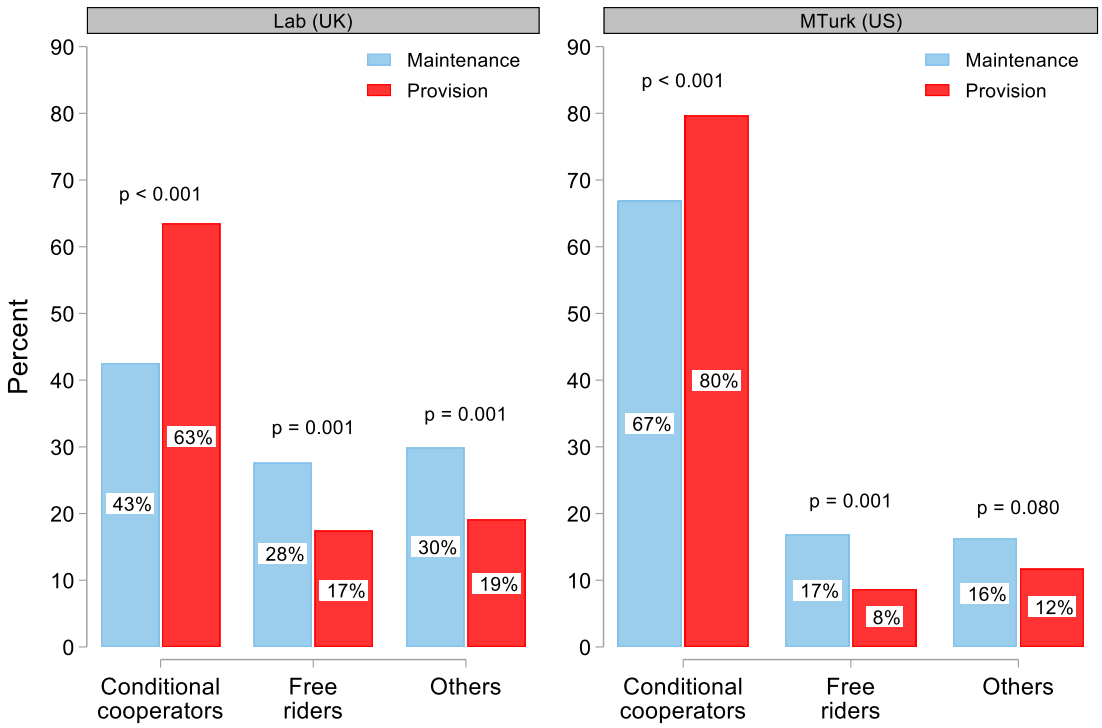


Figure 1: Distribution of cooperation types. Left panel: *n* = 704 students from the UK (source: Gächter et al. (2017)). Right panel: *n* = 703 US participants from MTurk (source: new experiments). *p*-values from χ^2-tests.

[11] The category 'others' contains "unconditional cooperators" who contribute a constant positive amount irrespective of what other group members contribute, "anti-conditional cooperators" whose cooperation depends negatively on the cooperation of other group members, "triangle cooperators" who are conditionally cooperative up to a certain level when they turn into anti-conditional and the rest. See Online Appendix B1 for further details and the relative frequencies of these subtypes.

Our replication for the purposes of this paper was conducted on MTurk with $n = 703$ US participants (instructions are in Online Appendix A).[12] The results match our previous findings. While the levels in the frequency of types are different compared to our UK student sample – we observe more conditional cooperators (73% vs. 53%, $\chi^2(1) = 62.21$; $p < 0.001$) and less free riders (13% vs. 22%, $\chi^2(1) = 23.26$; $p < 0.001$) and others (14% vs. 24%, $\chi^2(1) = 24.97$; $p < 0.001$) on MTurk - treatment differences are highly significant.[13] Specifically, as shown in the right panel of Figure 1, in line with Gächter et al. (2017), we find a significantly different distribution of types across treatments ($\chi^2(2) = 15.96$, $p < 0.001$) with a larger fraction of conditional cooperators (80% vs. 67%, $\chi^2(1) = 14.75$; $p < 0.001$), and a lower fraction of free riders (8% vs. 17%, $\chi^2(1) = 10.75$; $p = 0.001$) and others (12% vs. 16%, $\chi^2(1) = 3.07$; $p = 0.080$) in *Provision* compared to *Maintenance*.[14]

Like in our previous study, we also find that effective cooperation rates (after contributions/withdrawals), measured in a one-shot direct-response game played after the type elicitation, are significantly higher in *Provision* than in *Maintenance* (52% vs. 39%, two-sided t-test: $p < 0.001$). In both samples, we also find unconditional contributions in the strategy method to differ significantly across both dilemmas (Lab: *Provision*: 42%, *Maintenance:* 34%, two-sided t-test: $p = 0.003$; MTurk: *Provision*: 53%, *Maintenance:* 38%, two-sided t-test: $p < 0.001$).

[12] We decided to replicate the findings of Gächter et al. (2017) with a planned sample size of $n = 700$ because we were interested in the robustness of our lab results with undergraduates in a much more diverse subject pool. Based on the differences in the type distributions in the left panel of Fig. 1, a sample size of $n = 215$ would have sufficed to detect the same effect size with a power of 0.99 at $\alpha = 0.001$ (calculations based on G*Power 3.1, Faul et al. (2007)). However, given the different socio-demographic characteristics of the subject pool and the online nature of the experiment in MTurk, we decided to increase the sample to $n = 700$.

[13] The different levels of the frequency of types across our two studies is not surprising given the different cultural and sociodemographic background of the participants. We note, however, that the results from our MTurk study are very similar to Kocher et al. (2008) who elicited cooperation types among US students using a Provision public goods game: When comparing their results to ours, we find a remarkably similar distribution of types ($\chi^2(2) = 0.02$; $p = 0.992$): 81% vs. 80% conditional cooperators, 8% vs. 8% free riders, and 11% vs. 12% others. We thank M. Kocher for providing the data. Disaggregating the category 'others' in our MTurk data shows similar results than in Gächter et al. (2017). See Table B1 (Panel B) in Online Appendix B.

[14] We note that the differences across *Maintenance* and *Provision* are somewhat less pronounced in the MTurk sample compared to the student sample: the difference in the share of conditional cooperators amounts to 13 and 20 percentage points, respectively; the difference in the share of free riders is 8 and 11 percentage points, resp.; and the difference in the share of others is 4 and 11 percentage points, resp.. To test whether these differences across the two samples are significant, we run logistic regressions in which we use the different types as dependent variable, a treatment dummy, a MTurk dummy, and an interaction between the latter two as independent variables. The results, reported in Table B2 in Online Appendix B, confirm that there are significantly more conditional cooperators and significantly less free riders and others in *Provision* than *Maintenance* and on MTurk, but that there are no significant interaction effects (Provision × MTurk).

3.2 Temporal stability

To test whether the observed difference in the distribution of types is also stable *within* participants, we ran an additional experiment in which we re-invited a subset of participants from the Gächter et al. (2017) sample (left panel of Figure 1) four months after their first participation. Without knowing in advance, participants took part in sessions that were identical to the ones in which they participated before. We report results from $n = 119$ participants ($n = 65$ in *Provision* and $n = 54$ in *Maintenance*) who showed up in both waves.

At the aggregate level, cooperation preferences are remarkably stable over a period of four months; the distribution of types *within treatments* is very similar and does not significantly change between waves, neither in *Maintenance* ($\chi^2(2) = 0.51$, $p = 0.776$) nor *Provision* ($\chi^2(2) = 1.57$, $p = 0.456$). Consequently, when comparing the distribution of types *across treatments*, we find a significantly different distribution across *Maintenance* and *Provision* for both Wave 1 and Wave 2 ($\chi^2(2) = 10.32$, $p = 0.006$ and $\chi^2(2) = 11.87$, $p = 0.003$, respectively; see also Table B3 in Online Appendix B).

Regarding *individual-level stability* of cooperation preferences, we find that in *Provision* 66% of participants are classified as the same type in both waves, compared to 59% in *Maintenance*, a difference that is not statistically significant ($\chi^2(1) = 0.60$, $p = 0.438$). While these numbers indicate that the stability of types across waves is clearly not perfect, for both treatments we find it to be significantly higher than chance (which amounts to 46% and 35%, in *Provision* and *Maintenance*, resp.; t-tests, both $p < 0.001$; see Table B4 in Online Appendix B). The stability rate in *Provision* is thereby very similar to the results by Volk et al. (2012) who, using a similar setup and a gap of 2.5 months between waves, find a stability rate of 64%. No such comparison is possible for *Maintenance* because, as far as we are aware of, no previous study has investigated the stability of cooperation preferences using a maintenance game.

3.3. Predictive power of cooperation preferences

If our proxy for cooperation preferences measures something fundamental about people's attitude towards cooperation, it should be predictive of actual behavior in another comparable environment. To test this, we rely on the third part of our experiment in which a subset of participants took part in a one-shot direct-response public goods game in which they made a single contribution decision. We also elicited incentivized beliefs about the

average contribution of the other group members. Following Fischbacher et al. (2012), by combining elicited cooperation attitudes with stated beliefs we can make a point prediction about the contribution decision, $\hat{c}_i$. We then compare $\hat{c}_i$ with c_i (*i's* actual contribution in the direct-response game), delivering an individual-level measure of consistency.

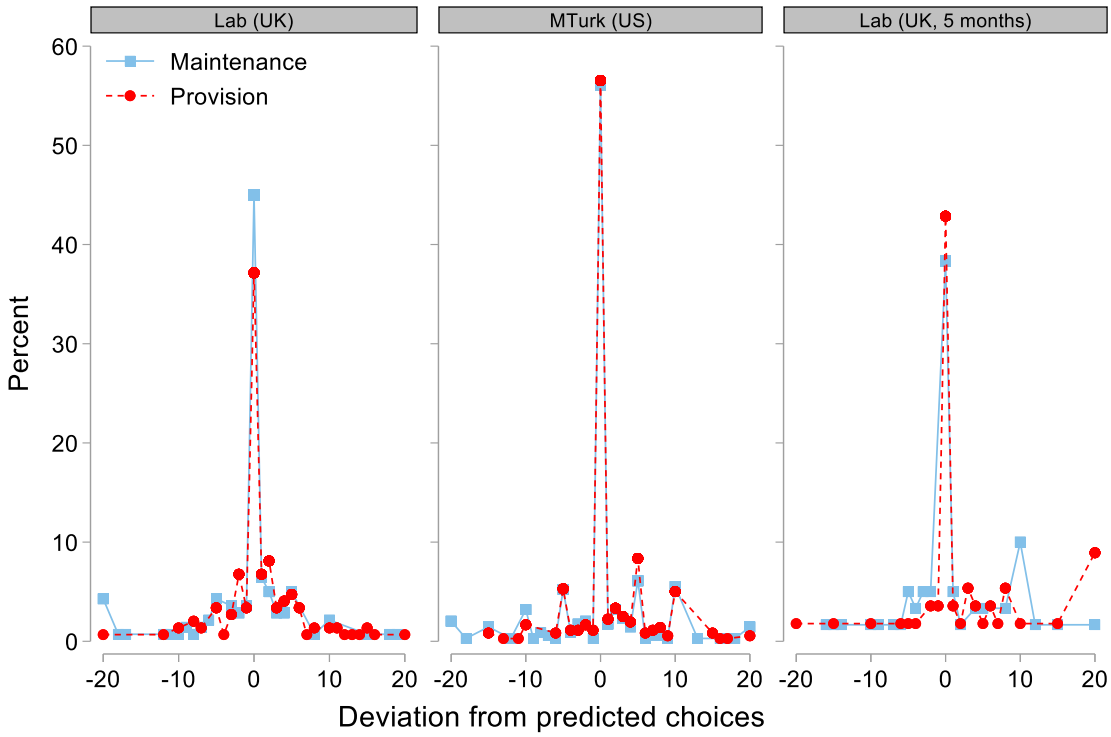


Figure 2: Deviations from predicted choices in *Maintenance* and *Provision*. Left panel: Students (n = 288). Middle panel: MTurk (n = 703). Right panel: Students who participated in the direct-response experiment five months after the preference elicitation experiment (n = 116)

In total, we have (1) n = 288 observations from our Gächter et al. (2017) sample, and (2) n = 703 observations from our MTurk experiment.[15] We further report data from (3) a set of n = 116 participants, for which the elicitation of cooperation preferences and the direct-response game took place in two separate sessions that lay five months apart. To our knowledge, this is the first paper that undertakes such a test of temporal stability. Our results are shown in Figure 2, depicting the distribution of individual deviations from predicted choices, $c_i - \hat{c}_i$, separately for *Maintenance* and *Provision* and for each of the three samples.

[15] The remaining participants from Gächter et al. (2017) played a repeated game. The results on the predicted power for the first-period contributions are similar to the one reported here (see Gächter et al. (2017)).

Figure 2 reveals that in all cases the modal and the median deviation is zero, that is, participants' contribution decision in the direct-response game is perfectly consistent with their predicted contribution from the strategy-method, even after a delay of 5 months. While not all participants are completely consistent (see Online Appendix B for further details), for none of the three samples the distribution of deviations is significantly different across *Maintenance* and *Provision* (Kolmogorov Smirnov tests; Lab: $p = 0.195$; MTurk: $p = 0.532$; Lab (5 months): $p = 0.472$). Overall, this demonstrates that the elicited attitudes are, together with elicited beliefs, an equally good predictor of actual cooperation behavior in both *Maintenance* and *Provision*.

3.4. Discussion

Consistent with the evidence from Frackenpohl et al. (2016) and Fosgaard et al. (2017), in Gächter et al. (2017) we have shown that *Maintenance* and *Provision* dilemmas elicit systematically different cooperation attitudes with significantly fewer participants behaving conditionally cooperative in the former than in the latter. In Gächter et al. (2017) we have further shown that together with differences in the beliefs about others' cooperation, this translates into different levels of cooperation in both one-shot and repeated games. We extend this prior evidence by showing that (i) differences in cooperation attitudes across *Maintenance* and *Provision* are replicable across different subject pools, (ii) elicited attitudes in both dilemmas are equally stable within participants over a period of four months, and (iii) elicited attitudes are (jointly with beliefs) an equally good predictor of actual cooperation decision in both dilemmas, even after a delay of five months. We summarize these findings in our first result:

> **Result 1**: *Maintenance and Provision evoke systematically different cooperation attitudes. Most importantly, conditional cooperation is more frequent in* Provision *than* Maintenance*. The elicited attitudes are stable within individuals and, jointly with beliefs, predictive of actual cooperation decisions.*

While replicability, stability, and predictive power are necessary conditions to interpret the effects as differences in underlying social preferences, they are not sufficient. An alternative interpretation of the observed differences is that they are due to stable and systematic misperceptions of the game form. In the next two steps (and sections), we disentangle the relative importance of social preferences and misperceptions. We start in the

next section by investigating whether the differences between *Provision* and *Maintenance* can be related to different social perceptions across the two contexts.

4. Step 2: Perceptions of kindness differ between Maintenance and Provision

A prominent psychological explanation for the existence of conditional cooperation is that individuals are reciprocal, that is, they have a desire to reward kind intentions with kindness and punish unkind intentions with unkindness (see Fehr and Schurtenberger (2018) for a review). Hence, as reciprocity is the behavioral response to perceived kindness or unkindness (Rabin (1993); Dufwenberg and Kirchsteiger (2004); Falk and Fischbacher (2006)), a crucial question is how participants evaluate actions of others in terms of (un)kindness, and whether these evaluations differ across games. If people perceive payoff-equivalent actions differently in terms of kindness across *Maintenance* and *Provision*, this could trigger game-specific reciprocal responses which, in turn, could explain the observed differences in conditional cooperation across the two setups.

To test this conjecture, we conducted two online studies in which we elicited kindness perceptions about other people's contribution behavior for both types of social dilemmas (see Falk and Fischbacher (2006) for a related exercise and Wilson (2012) for a cautionary note). In the questionnaire, we explained to participants either a *Maintenance* or a *Provision* dilemma and then asked them to evaluate the kindness of average effective contributions of three other group members on a scale from -100 to +100 (where -100 corresponds to 'very unkind' and +100 corresponds to 'very kind'). We asked participants to evaluate the kindness of a low, an intermediate, and a high effective contribution of 0, 10, and 20, respectively (see Online Appendix A3). We recruited n = 185 students from the University of Nottingham and n = 401 participants from MTurk. No participant was involved in any of our experimental sessions before.[16]

Figure 3 reports the average kindness evaluation of others' average effective contributions. The results from the two samples are remarkably similar. While low effective contributions of 0 are considered as significantly less kind in *Provision* than in *Maintenance* (two-sided t-tests, $p < 0.001$ and $p < 0.001$ for students and MTurkers, respectively), we

[16] Since we asked participants for their personal perceptions, answers were not incentivized. However, we did incentivize participation. Student participants were offered three randomly drawn prizes of £50 each. MTurkers received a flat payment of $2. According to Cubitt et al. (2011b) who studied moral judgments in social dilemmas, incentivizing participation does not affect moral judgments, making it unlikely that it affects kindness evaluations.

observe the reverse pattern for medium and high effective contributions of 10 and 20, respectively. In these cases, payoff equivalent actions are considered as unkinder in *Maintenance* compared to *Provision* (two-sided t-tests, average others' contribution = 10, $p = 0.045$ and $p = 0.001$ for students and MTurkers, respectively; average others' contribution = 20, $p = 0.007$ and $p = 0.062$ for students and MTurkers, respectively).

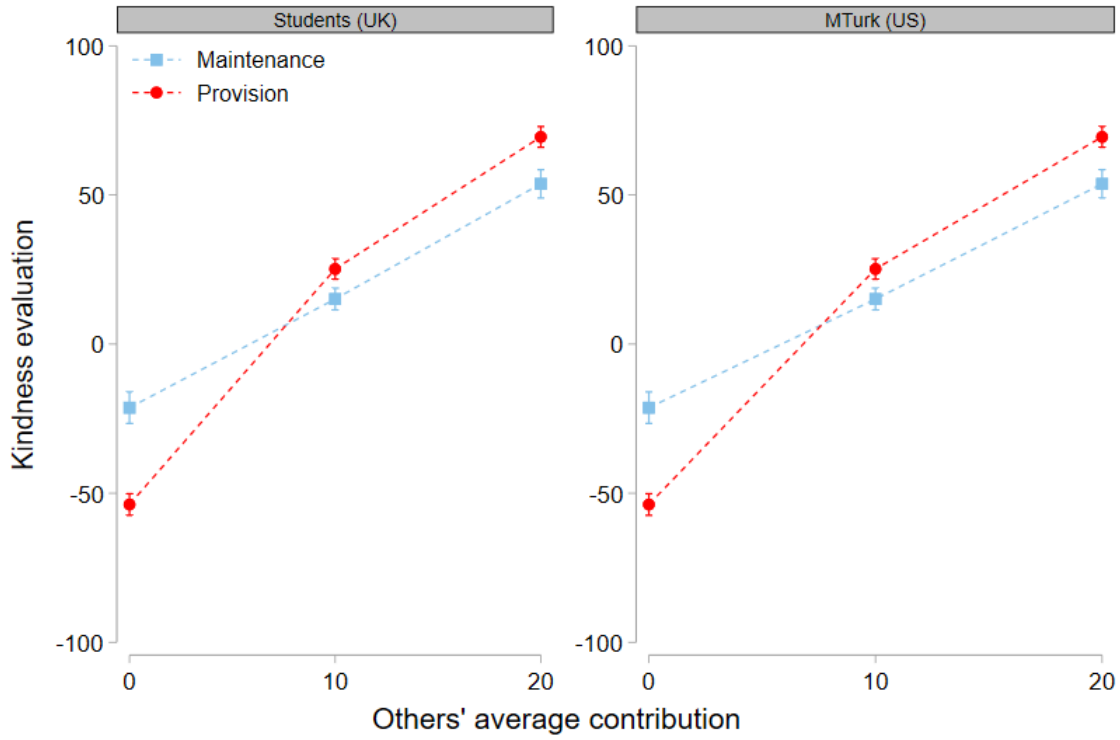


Figure 3: Kindness perceptions in *Provision* and *Maintenance* of others' effective contributions (±1 s.e.m).

Further support comes from OLS regressions in which we use kindness evaluations as the dependent variable, others' average contributions, a dummy for the framing manipulation, and an interaction term of the last two as independent variables. We run this regression separately for each subject pool. The results are in Table B5 in Online Appendix B. In line with Figure 3, we find a positive and significant coefficient for the interaction term, indicating greater responsiveness of kindness evaluations to others' contributions in *Provision* than in *Maintenance*. We summarize these findings in our second result:

> **Result 2**: *Maintenance and Provision evoke systematically different perceptions of kindness: Complete free-riding is perceived to be unkinder in Provision than in Maintenance, while positive contributions (of 10 and 20) are perceived as more kind in Provision than in Maintenance.*

Overall, Result 2 suggests that both with respect to kindness and unkindness, individuals have stronger reactions in their perception of others' contributions when contributing to, rather than withdrawing from, a public good. Result 2 is also consistent with Cubitt et al. (2011b) who found a similar pattern for moral judgments: Failing to contribute to the public good was perceived as morally worse than withdrawing everything. These stronger reactions likely trigger a stronger need to reciprocate and, hence, can explain the higher frequency of conditional cooperators in *Provision* compared to *Maintenance*. The result from our kindness survey thus favors the explanation that the differences in cooperation attitudes across treatments are rooted in differences in the underlying preferences.

5. Step 3: Measuring and controlling for game form misperceptions

One alternative explanation for our results is that participants may have systematic misperceptions of the game form and that these misperceptions may be dilemma specific. If this were the case, differences in conditional cooperation may not be due to different social preferences but due to differences in the understanding of the incentives of the game. In this section, we assess to what extent game-form misperceptions can explain our results.

5.1. Conceptualization of game form misperceptions

Our conceptualization of game-form misperceptions draws on Cason and Plott (2014). They analyze the tension between standard theory, which assumes that preferences are only influenced by elements of the game form (i.e., the set of actions, the set of material consequences, and the links between actions and consequences), and non-standard theories, which postulate that preferences may depend on elements outside the game form such as how the game form is described. In their example, Cason and Plott investigate anomalous bidding behavior in the Becker et al. (1964) mechanism. While observed bids are consistent with frame-dependent preferences, Cason and Plott show that this effect is driven by a subset of participants who mistakenly perceive the situation as a first-price auction rather than a second-price auction. They conclude that in their case, the description of the decision situation affected participants' perception of the game form, which, in turn, led them to implement 'wrong' behavioral responses given their underlying preferences. Cason and Plott's general conclusion is that researchers should be careful of interpreting choices as revealed preferences, an issue that Koszegi and Rabin (2008) also point out.

The implication of Cason and Plott's argument for our context is that interpreting differences in behavioral responses across treatments as evidence for dilemma-dependent preferences might be erroneous because such differences can be due to dilemma-dependent misperceptions of the game form. If failure of correct game-form recognition is also at work in our setup, which is possible given previous evidence on confusion (see Introduction), then the observed distribution of cooperation types might not reflect participants' true cooperation preferences as some of the participants might have mistakenly implemented behavior different from their preferred one. For example, if some participants erroneously believe that to maximize their individual income, they should increase their contribution if the contributions of other group members increase, this might lead to an inflated rate of conditional cooperation. Moreover, if this type of game-form misperception is more frequent in *Provision* than in *Maintenance*, this could explain the observed treatment effect of a higher frequency of conditional cooperation and a lower frequency of free riding in the former than the latter. Some evidence for this possibility comes from Fosgaard et al. (2017) who find that many participants fail to recognize the dominant strategy of full free-riding and that this type of mistake occurs more frequently in *Provision* than in *Maintenance*. In the next subsection, we describe the details of a new experiment that was specifically designed to examine the role of misperceptions in our context.

5.2. Measurement of game form misperceptions

We measure game-form misperceptions in a new experiment with $n = 696$ Nottingham students who had not participated in any of our experiments before. The experiment followed the structure presented in Section 2, except that there was no direct-response experiment after the strategy-method experiment. Instead, after participants answered the standard set of ten control questions, we asked them two additional sets of four incentivized questions (see Table 2), paying £0.1 per correct answer.[17] After that, the experiment proceeded with the elicitation of cooperation preferences using the strategy method. We asked the incentivized questions *before* we elicited participants' preferences to ensure maximal understanding of the situation and incentives.[18]

[17] To avoid any income effects when eliciting cooperation preferences, incentives were modest, and participants were informed about the number of questions they answered correctly only at the very end of the experiment.

[18] In this methodological aspect, our approach is akin to Plott and Zeiler (2005) who used a battery of experimental tools designed to maximize understanding before eliciting WTA-WTP valuations.

The first four questions, which we label *payoff questions*, are akin to standard control questions in which participants have to calculate earnings for various contribution scenarios. We asked participants to determine their own and others' monetary earnings in case (i) they would contribute their whole endowment, but the other group members would contribute nothing, and (ii) they would contribute nothing but each of the other group members would contribute their whole endowment (20 tokens) (see Q1 – Q4 in Table 2 for the exact wording of questions). Our first measure of game-form misperception classifies a participant as misperceiving if they make at least one mistake in the payoff questions (see Bartling et al. (2015) for a similar approach in a value elicitation task).

In the other four questions (compare Q5 – Q8 in Table 2), which we label *goal questions*, we follow a similar strategy as Fosgaard et al. (2017) and ask participants what a person who wants to implement a specific goal should do. The first goal was *individual payoff maximization*; participants were asked how much a person who "wants to make as much money as possible for him/herself" should contribute given the other group members contribute either 0 or 20. The second goal was *group payoff maximization*: we asked participants how much a person who "wants that the group as a whole makes as much money as possible" should contribute given the other group members contribute either 0 or 20. We classify a participant as misperceiving if they make at least one mistake in the goal questions. This constitutes our second measure of game-form misperception. Compared to the payoff questions, which require an understanding of the incentive structure as well as sufficient calculation skills, the goal questions require the ability to put oneself into the shoes of another person that might have different objectives than oneself, a task that is arguably more difficult than just calculating payoffs.

Finally, our third measure checks whether a participant is a *mistaken conditional cooperator*, that is, whether they think that maximizing their own income requires increasing own contribution if others' contributions increase (from 0 to 20), that is if their response to Q6 is strictly higher than their response to Q5. Such a mistake could lead participants to believe that they face incentives akin to a coordination game rather than a social dilemma game. As a result, participants may then implement 'wrong' behavioral responses given their underlying preferences. That is, while the behavioral response of such a misperceiving participant in the strategy method might look like evidence of prosocial, reciprocal preferences, such behavior is also consistent with a model in which a purely selfish participant maximizes their misperceived payoffs.

Table 2: Incentivized misperception questions and percentage of correct answers across *Maintenance* (M, *n* = 320) and *Provision* (P, *n* = 376).

	% Correct answers		
Payoff Questions:	*M*	*P*	χ^2 – test
Q1: Assume that you contribute 20 tokens to the project and the other three group members contribute nothing to the project. What will your total income be?	95.3	95.5	$p = 0.917$ ($q = 0.917$)
Q2: Assume that you contribute 20 tokens to the project and the other three group members contribute nothing to the project. What will the total income of each of the other group members be?	90.9	84.3	$p = 0.009$ ($q = 0.035$)
Q3: Assume that you contribute 0 tokens to the project and each of the other three group members contributes 20 tokens to the project. What will your total income be?	94.4	92.8	$p = 0.405$ ($q = 0.649$)
Q4: Assume that you contribute 0 tokens to the project and each of the other three group members contributes 20 tokens to the project. What will the total income of each of the other group members be?	95.9	93.1	$p = 0.103$ ($q = 0.275$)
Goal questions:			
Q5: Suppose the other group members contribute on average 0 tokens to the project. How much should a person who wants to make as much money as possible for him/herself contribute to the project?	92.8	95.0	$p = 0.239$ ($q = 0.478$)
Q6: Suppose the other group members contribute on average 20 tokens to the project. How much should a person who wants to make as much money as possible for him/herself contribute to the project?	93.8	85.6	$p = 0.001$ ($q = 0.004$)
Q7: Suppose the other group members contribute on average 0 tokens to the project. How much should a person who wants that the group as a whole makes as much money as possible contribute to the project?	79.1	77.1	$p = 0.539$ ($q = 0.664$)
Q8: Suppose the other group members contribute on average 20 tokens to the project. How much should a person who wants that the group as a whole makes as much money as possible contribute to the project?	93.1	92.0	$p = 0.581$ ($q = 0.664$)
Total*	91.9	89.4	$p = 0.030$

Notes: Shown are the questions in *Provision*. The questions for *Maintenance* were formulated equivalently. *q*-values correspond to *p*-values corrected for multiple comparisons using the Benjamini and Hochberg (1995) false discovery rate procedure. * For testing the total effect (last row) we use logistic regressions with standard errors clustered at the individual level.

5.3. Misperceptions are dilemma specific

Table 2 summarizes the percentages of correct answers separately for each question and for *Maintenance* and *Provision*. It reveals that, at the aggregate level, in both treatments there is an overall very low level of misperception. With a few exceptions, the percentage of correct answers is above 90% for every single question and treatment. On average, participants answer 91% of the questions correctly, 92% in *Maintenance* (7.35 out of 8) and 89% in *Provision* (7.15 out of 8). Despite the overall treatment differences being small, they

are statistically significant at the 5% level (p = 0.030). When comparing the fraction of correct answers between *Maintenance* and *Provision* for each question separately, we find significant differences for two out of the eight questions, Q2 (p = 0.009) and Q6 (p = 0.001).

Next, we turn to an individual-level analysis of mistakes by applying our three measures of game-form misperceptions as described above. As shown in Table 3, for all measures, we find that misperceptions are significantly more frequent in *Provision* than in *Maintenance*; the number of people misperceiving is between 8 and 9 percentage points higher in *Provision* than in *Maintenance* (χ^2 – tests, all $p < 0.023$).

Table 3: Percent of participants classified as misperceiving in *Maintenance* and *Provision*

	Maintenance [$n = 320$]	*Provision* [$n = 376$]	χ^2 *- test*
Measure 1 – At least one mistake in the payoff questions Q1-Q4	13% [$n = 40$]	22% [$n = 84$]	$p = 0.001$
Measure 2 – At least one mistake in the goal questions Q5-Q8	29% [$n = 93$]	37% [$n = 140$]	$p = 0.023$
Measure 3 – Mistaken conditional cooperation	5% [$n = 17$]	13% [$n = 47$]	$p = 0.001$

We summarize these findings in our third result:

Result 3: *Provision dilemmas cause significantly higher levels of misperceptions of the game form than Maintenance dilemmas.*

5.4. Misperceptions and cooperation attitudes

In the following, we analyze the connection between misperceptions and the elicited cooperation attitudes. If there was none, i.e., if mistakes were randomly distributed across types, then the different degrees in the level of misperception across *Maintenance* and *Provision* should not affect the distribution of types. If instead, the likelihood of game-form misperception is correlated with displaying a certain cooperation type, this could explain the differences in conditional cooperation we observed across our two treatments.

Table 4 reports the fraction of misperceiving participants conditional on type classification. We report these numbers separately for *Maintenance* and *Provision* and our

three measures of misperception. Table 4 reveals that the null hypothesis of no relationship between types and misperceptions can be rejected for both *Maintenance* and *Provision* according to Measure 2 (goal questions, χ^2 – tests, both $p < 0.007$) and Measure 3 (mistaken conditional cooperation, χ^2 – tests, both $p < 0.003$), but not for Measure 1 (payoff questions, χ^2 – tests, both $p > 0.065$). On top of that, our results reveal that the way misperceptions interact with the elicited attitudes is treatment-specific. In *Maintenance* we observe mainly the participants classified as others who display some form of misperceptions; compared to free riders their odds of displaying misperceptions are increased by a factor of 2.3 (Measure 1) up to 5.9 (Measure 3) In *Provision*, in contrast, we find that mainly the group of conditional cooperators exhibiting misperceptions; compared to free riders their odds of displaying misperceptions is increased by a factor of 1.6 (Measure 1) up to 8.3 (Measure 3).

Table 4: Fraction of misperceiving participants in Maintenance and Provision by type

	Maintenance			*Provision*		
	Measure 1	Measure 2	Measure 3	Measure 1	Measure 2	Measure 3
Conditional Cooperators	0.103	0.252	0.028	0.247	0.447	0.200
Free Riders	0.095	0.230	0.024	0.165	0.262	0.029
Others	0.195	0.425	0.126	0.241	0.337	0.072
χ^2 *- tests*	*p = 0.066*	*p = 0.005*	*p = 0.002*	*p = 0.247*	*p = 0.006*	*p < 0.001*

These results show that the two types of dilemmas not only affect the overall level of misperception but also how perceptions interfere with preferences. This provides a strong case for the need of controlling for misperceptions before interpreting behavioral differences as dilemma-dependent preferences. In the following, we therefore test whether accounting for the different types of misperceptions can explain the observed treatment differences in the distribution of cooperation preferences. We report this analysis in Table 5.

Panel A of Table 5 shows the distribution of types in the full sample (without controlling for misperceptions). In line with our results from Section 3, we find again a highly significant difference in the distribution of preferences across treatments ($\chi^2(2) = 21.23, p < 0.001$), with

significantly fewer conditional cooperators ($\chi^2(1) = 20.65$, $p < 0.001$) and significantly more free riders ($\chi^2(1) = 11.24$, $p = 0.001$) in *Maintenance* than in *Provision*.[19]

Table 5: Distribution of cooperation preferences in *Maintenance* and *Provision* after controlling for different types of misperceptions

	Panel A: Full sample			**Panel B: No mistake in payoff questions**		
Type	Maintenance (n = 320)	Provision (n = 376)	χ^2-test	Maintenance (n = 280)	Provision (n = 292)	χ^2-test
Conditional Cooperators	34%	51%	$p < 0.001$	34%	49%	$p < 0.001$
Free Riders	39%	27%	$p = 0.001$	41%	29%	$p = 0.005$
Others	27%	22%	$p = 0.118$	25%	22%	$p = 0.332$
χ^2-test	$p < 0.001$			$p = 0.001$		
	Panel C: No mistake in goal questions			**Panel D: No mistaken conditional cooperation**		
Type	Maintenance (n = 227)	Provision (n = 236)	χ^2-test	Maintenance (n = 303)	Provision (n = 329)	χ^2-test
Conditional Cooperators	35%	45%	$p = 0.042$	34%	46%	$p = 0.002$
Free Riders	43%	32%	$p = 0.019$	41%	31%	$p = 0.007$
Others	22%	23%	$p = 0.743$	25%	23%	$p = 0.623$
χ^2-test	$p = 0.050$			$p = 0.006$		

[19] The relative frequency of types is somewhat different compared to the one found in our initial student sample as reported in Section 3.1. Specifically, we find a significant change in the distribution of types in both *Provision* ($\chi^2(2) = 14.26$, $p = 0.001$) and *Maintenance* ($\chi^2(2) = 11.01$, $p = 0.004$), with more free riders (*Provision*: 27% vs. 17%, $\chi^2(1) = 10.43$, $p = 0.001$, *Maintenance* 39% vs. 28%, $\chi^2(1) = 10.44$, $p = 0.001$) and fewer conditional cooperators (*Provision*: 63% vs. 51%, $\chi^2(1) = 12.50$, $p < 0.001$, *Maintenance* 33% vs. 43%, $\chi^2(1) = 5.84$, $p = 0.016$) in the new experiment. We believe that the reason for this result is that we administered the incentivized control questions before the elicitation of cooperation attitudes. This might have made the incentive structure of the social dilemma situation even clearer, which, in turn, might have corrected some 'mistaken' conditional cooperation. Alternatively, it could be that the incentivized questions increased the salience of material incentives, thereby priming participants to be more self-interested. While we cannot rule out neither of these channels, we can ascertain whether the shifts in the distribution of types had any effect on the differences across treatments. To this end, similar to our analysis in which we compared our student with the MTurk sample, we run logistic regressions in which we use the different types as dependent variable, a treatment dummy, a sample dummy, and an interaction between the latter two as independent variables. The results, reported in Table B6 in Online Appendix B, show that there are no significant interaction effects between the treatment and the sample, indicating that adding the additional questions at the beginning of the experiment had no systematic effect on our treatment comparison.

In panels B, C, and D, we compare the distribution of types across *Maintenance* and *Provision* after dropping participants who are classified as misperceiving according to Measures 1, 2, and 3, respectively. The results reveal that our main result of different distributions of cooperation preferences across *Maintenance* and *Provision* is robust to the exclusion of participants who do not fully understand the game form. That is, the distribution of types is significantly different across *Maintenance* and *Provision* across all subsamples of non-confused participants (χ^2 – tests, all $p < 0.05$).

We observe significantly more conditional cooperators (χ^2 – tests, all $p < 0.05$) and significantly fewer free riders (χ^2 – tests, all $p < 0.02$) in *Provision* than in *Maintenance* (the difference in others is never significant, χ^2 – tests, all $p > 0.117$; see also Table B7 in Online Appendix B). Notably, however, we find that once we control for misperceptions, the differences in the distribution of types become smaller compared to the full sample. The percentage difference in conditional cooperators decreases from 17 percentage points in the full sample to 10 to 15 percentage points depending on the misperception measure. The difference in the fraction of free riders, in contrast, remains stable, varying between 10 and 12 percentage points. This demonstrates that while misperceptions can account for some of the observed differences across treatments, even after controlling for misperceptions, we observe substantial and significantly different degrees of conditional cooperation in *Maintenance* and *Provision.* We summarize these findings in our fourth result:

> **Result 4**: *Even after accounting for the different degrees of misperceptions across Maintenance and Provision, we find significantly more conditional cooperators and fewer free riders in Provision than in Maintenance.*

5.5. Misperceptions do not affect perceptions of kindness

The results above already strongly suggest that the observed differences in cooperation types across *Maintenance* and *Provision* are rooted in differences in the underlying social preferences. As we have argued above, these differences can be explained by differences in the perceived kindness of others' actions across the two treatments. As a final test of this argument, we provide evidence that the different perceptions of kindness that we presented in Section 4 are not a consequence of game-form misperceptions. This is important because if differences in kindness perceptions are indeed the main trigger of differences in conditional cooperation across the two social dilemmas, we should observe that perceptions of kindness still differ when controlling for misperceptions. If instead, the different kindness

perceptions across *Maintenance* and *Provision* disappear when controlling for misperceptions, then the elicited kindness perceptions may not be considered a relevant explanation for the differences in conditional cooperation across the two dilemmas.

To test this, in some sessions of the misperception experiment reported in the previous two subsections, we included the kindness questionnaire at the end of the experiment before participants received feedback about the outcome of the game. Hence, while the kindness results reported in Section 4 (see Figure 3) were elicited using non-involved participants, we can now test whether the results hold when participants have experienced the decision situation. Furthermore, we can test whether the differences in kindness perceptions across *Maintenance* and *Provision* are robust to the exclusions of participants who exhibit some misperception. Our sample comprises $n = 200$ participants, $n = 80$ in *Maintenance* and $n = 120$ in *Provision*. The results are shown in Figure 4.

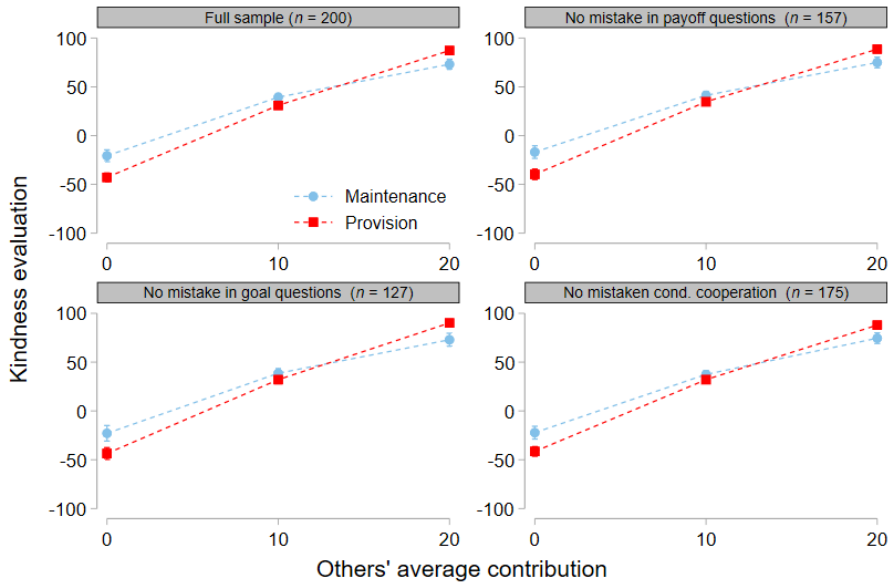


Figure 4: Kindness perceptions in *Provision* and *Maintenance* of participants who experienced the decision situation (±1 s.e.m).

The first (upper left) panel of Figure 4 depicts the comparison between the kindness schedules between *Maintenance* and *Provision* for the full sample. The second, third, and fourth panel show the same data for the subset of participants without misperceptions according to our three measures. Strikingly, the figure shows that the differences in kindness schedules across *Maintenance* and *Provision* are not only similar across the four panels but also similar to the ones reported in Figure 3. This indicates that the differences in kindness perceptions across *Maintenance* and *Provision* are robust to having experienced the decision

situation beforehand and, more importantly, to the exclusion of participants who exhibit some form of game-form misperception.[20]

We summarize these findings in our fifth result:

Result 5: *Even after accounting for the different degrees of misperceptions across Maintenance and Provision, we find that the two dilemmas evoke systematically different perceptions of kindness that can explain the differences in cooperation attitudes with more conditional cooperators and fewer free riders in Provision than in Maintenance.*

6. Step 4: Which theory of social preferences can explain our results?

In our *fourth step,* we investigate which of the existing models of social preferences can reconcile the observation of a higher share of conditional cooperators and fewer free riders in *Provision* than in *Maintenance* as found in our experiments as well as in some previous research (Fosgaard et al. (2014); Frackenpohl et al. (2016); Gächter et al. (2017); Isler et al. (2021)). We emphasize that our aim here is not to conduct a horse race between different models of social preferences (see, e.g., Miettinen et al. (2020), for such an analysis in a related context), but to provide a discussion about why and under which assumptions existing theories of social preferences can explain *dilemma-dependent* conditional cooperation (note that all theories we discuss here can explain conditional cooperation in a given social dilemma). In the following, we only describe the main arguments, more details and formal analyses can be found in Online Appendix C. At the end of the section, we discuss the models also in relation to our results on kindness perceptions (Results 2 and 5).

We start our discussion with theories of *distributional preferences* (Fehr and Schmidt (1999); Bolton and Ockenfels (2000); Charness and Rabin (2002)). While all these theories can explain conditional cooperation, e.g., if inequity aversion is strong enough, they do not predict any dilemma-specific conditional cooperation. The reason is that these theories are only based on payoff consequences, which are identical across the two incentive-equivalent

[20] Parametric estimates further corroborate these results. Using regression analyses in which we regress kindness evaluations on others' average contributions, a *Provision* dummy, and an interaction term between the last two, we find that the *Provision* dummy is significantly negative and the interaction term between the *Provision* dummy and others' average contributions is positive and significant. The size of these effects, which we report in Table B8 in Online Appendix B, is similar to the ones reported in Section 4 where we analyze the kindness evaluations of uninvolved participants (compare also Table B5).

social dilemmas of *Maintenance* and *Provision.*[21] Note that this prediction hinges on the original assumption of these theories that individual preference parameters are game-independent. If one would be willing to relax this assumption by allowing preference parameters to differ across contexts, then these models could potentially rationalize our findings. For example, if one assumes, using Fehr and Schmidt (1999) preferences, that agents are more advantageous inequity-averse when facing a *Provision* rather than a *Maintenance* dilemma, then this could explain why we observe more conditional cooperators and fewer free riders in the former than in the latter. We note, however, that while such shifts in advantageous inequity aversion might be empirically relevant, there is no psychological mechanism in the formal assumptions of Fehr and Schmidt (1999) that postulate such shifts.

Next, we consider theories of *reciprocity* (Rabin (1993); Dufwenberg and Kirchsteiger (2004); Falk and Fischbacher (2006)), in which agents' motivations derive from their material payoff as well as a psychological payoff that depends on their first- or second-order beliefs about others' actions. These theories can also explain conditional cooperation (Dufwenberg et al. (2011)) because agents want to reward kind actions with kindness and punish hostile actions with unkindness. Kindness is thereby assumed to be evaluated relative to a reference point that is the midpoint between the maximum and minimum possible payoff (see Online Appendix C and Dufwenberg and Kirchsteiger (2019) for a recent discussion). Because the payoff sets are the same across *Maintenance* and *Provision* the reference point must be the same. Therefore, beliefs are the only channel through which simultaneous gameplay may differ across *Maintenance* and *Provision*. Since in our strategy-method experiment (i) first-order beliefs are fixed because participants condition their contributions on all possible average contributions of others, and (ii) as argued by Dufwenberg et al. (2011), in a linear public goods game the evaluation of others' kindness only depends on first-order but not on second-order beliefs, these models do not predict game-dependent conditional cooperation. This prediction hinges on the assumption that the concern for reciprocity as measured by a reciprocity parameter does not vary across dilemmas. If one would be willing to relax this assumption and instead assume that the reciprocity parameter is stronger under *Provision* than *Maintenance*, then reciprocity theory could also rationalize our finding that conditional cooperation is more frequent in *Provision* than *Maintenance*.

[21] A similar argument holds for models of altruism and fairness such as those by Levine (1998) and Cox et al. (2007).

Note, however, that also in this theory there is no psychological mechanism in the formal assumptions that postulates such a shift.

Next, we consider theories of *guilt aversion*. Specifically, we rely on a model of simple guilt by Battigalli and Dufwenberg (2007) which assumes that an agent's utility depends on her material payoff as well as her second-order beliefs, that is, what she believes the other players believe she will do. Applied to the context of a public goods game, guilt aversion predicts that player *i* will suffer guilt if *i* contributes less than what *i* thinks the other three group members expect *i* to contribute (on average). If the disutility from guilt becomes large enough, player *i* has an incentive to contribute whatever she thinks others expect her to contribute. Since our strategy-method experiment only fixes first-order beliefs, differences in second-order beliefs could reconcile differences in cooperation preferences: If more participants in *Provision* than in *Maintenance* have second-order beliefs that others expect them to reciprocate their contributions, the perceived guilt from not matching others' contributions is stronger, which, in turn, can lead to a higher fraction of conditional cooperators in *Provision* than in *Maintenance.* Although guilt aversion theory does not explicitly model a mechanism for why there should be any difference in second-order beliefs between *Maintenance* and *Provision,* in contrast to the theories above, it does not require the preference parameter (sensitivity to guilt) to be context-dependent to predict a difference across the two dilemmas.

Since in our experiments we did not elicit second-order beliefs (because this is too complex in our current strategy method design), we cannot provide a direct test of the theory. As a proxy alternative, however, and following previous literature (see, e.g., Chang et al. (2011); Bellemare et al. (2019)) we conducted an online survey in which we elicited *ex-post* feelings of guilt. In the survey, we asked participants on a scale from 0 to 100 (where 0 corresponds to "not guilty at all" and 100 corresponds to "very guilty") to assess how guilty they would feel if as a response to others' contributions (withdrawals) of 0, 10, or 20, they would contribute 0 (withdraw 20) tokens.

Our results based on responses from n = 347 students from the University of Nottingham and n = 402 participants on MTurk reveal that, if anything, free riding on others' contributions generates stronger feelings of guilt in *Maintenance* than in *Provision*. Based on this, if feelings of guilt would be the only driver behind the differences in cooperation preferences across the two dilemmas, we should observe *more* conditional cooperation in *Maintenance* than in *Provision*, which is the opposite of what we find (see Online Appendix

D for a full description of our survey design and results). As a caveat, we acknowledge that our results are only suggestive and future research should elicit second-order beliefs in a suitably simplified design to provide a more stringent test of guilt aversion.[22]

Finally, we review the theory of *revealed altruism* by Cox et al. (2008). This is an axiomatic theory of social preferences that allows the status quo to influence the strength of reciprocity and hence, in our games, predicts dilemma-specific conditional cooperation. Applying Cox et al.'s reciprocity axiom, Axiom R, to our games, the model predicts that participants will perceive higher contributions by the other group members as more generous towards them, and, therefore, they will be more altruistic towards the others. In our strategy-method experiment, this will be manifested in a positive slope of the contribution schedule (see Cox et al. (2013) for a related analysis).

In a second axiom, Axiom S, Cox et al. (2008) assume that based on the psychological asymmetry behind omission and commission (see, e.g., Spranca et al. (1991)), generous actions that change the status quo trigger stronger reciprocity than generous actions that just uphold the status quo. That is, Axiom S strengthens or weakens the effect of Axiom R depending on the status quo of where resources are allocated initially. Applied to our strategy-method experiment, Axiom S predicts that second movers will be less altruistic towards first movers in *Maintenance* than in *Provision.* The reason is that in *Provision* any positive contributions by the other three group members increase the payoff opportunities of the second mover compared to the status quo where nothing is contributed to the public good. In *Maintenance,* in contrast, where all resources are initially allocated to the public good, any withdrawal by the other three group members reduces the payoff opportunities for the second mover. This asymmetry triggers a stronger preference for reciprocity in *Provision* compared to *Maintenance,* which can explain our finding of more conditional cooperators and fewer free riders in *Provision* than in *Maintenance* (see Online Appendix C for further details).

Next, given the robust and replicable evidence on different perceptions of kindness between *Maintenance* and *Provision* (Results 2 and 5), we discuss which theories could reconcile this evidence. First, notice that models of distributional preferences are only based on payoff consequences and therefore do not incorporate any evaluation of others' actions.

[22] These results are potentially also relevant for inequity aversion because there is evidence by Beranek et al. (2015) of a strong correlation between guilt proneness and advantageous inequity aversion. Since we find stronger guilt in *Maintenance* than *Provision*, this would map into higher advantageous inequity aversion in *Maintenance* than *Provision* which should lead to more conditional cooperation in *Maintenance* than *Provision*, but we find the opposite.

Guilt aversion is also mute with respect to others' kindness as players evaluate their *own* action with respect to the distance from their second-order belief but make no evaluation about *others*' actions. Reciprocity models such as those by Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006) are natural candidates to explain our kindness results as these theories incorporate the perception of other's (un)kindness as a central element into their models. According to these models, however, kindness is evaluated with respect to a reference point that is only based on material payoffs. Given that material payoffs are identical across *Maintenance* and *Provision*, for a given effective contribution perceived kindness is predicted to be the same across *Maintenance* and *Provision*.

Finally, we consider the revealed altruism model by Cox et al. (2008) that also explicitly incorporates kindness. However, like the reciprocity models, also this model does not predict any differences in kindness perceptions across the two dilemmas. Instead, it postulates that the different status-quo allocation in *Maintenance* and *Provision* directly affect the reciprocity parameter (see Axiom S above). If one would be willing to adjust the model by allowing Axiom S to also affect perceptions of kindness, then every effective contribution in *Provision* should be perceived as more generous than the corresponding effective contribution in *Maintenance*.[23] The results from our kindness survey reveal, however, that while this is indeed true for average effective contributions of 10 and 20, for effective contributions of 0 we find the opposite as contributing 0 is perceived as unkinder than withdrawing everything.

In sum, while none of the theories discussed above is perfectly consistent with our findings., it seems that the theory of revealed altruism by Cox et al. (2008) comes closest to explaining our results as it is the only theory that postulates an explicit mechanism for why one should expect stronger conditional cooperation in *Provision* than in *Maintenance* – which our Results 1 and 4 confirm. However, as we have seen in our results on kindness perceptions (Results 2 and 5), some discrepancies of the theory with our data remain here, too. Setting up an experiment that is explicitly designed to test and differentiate between these different theories in a context like ours is left for future research.

[23] Formally, Axiom S would need to be modified to allow the status quo allocation to affect not only the *MAT* (more altruistic than) partial ordering but also the *MGT* (more generous than) partial ordering (see Online Appendix C for further details).

7. Discussion and Conclusion

In this paper, we provided a comprehensive behavioral analysis of two generic and incentive-equivalent social dilemmas of voluntary cooperation: providing and maintaining public goods. We first established a lower fraction of conditional cooperators (and higher fraction of free riders) in *Maintenance* than *Provision* (Result 1). We then focused on two fundamental dimensions: social preferences and (mis)perceptions of others' intentions and the game form. We reported two important asymmetries. First, regarding *perceptions*, we found that perceptions of the kindness of others' actions differ between *Maintenance* and *Provision* because withdrawing everything from the public good is seen as less unkind than failing to contribute to the public good; and contributing everything is considered kinder in *Provision* than in *Maintenance* (Result 2). Regarding perceptions of the game form, we found that misperceptions are game-specific: misunderstandings are more likely in *Provision* than *Maintenance* (Result 3). Second, even after controlling for misperceptions, we observe substantial and significantly different degrees of conditional cooperation in *Maintenance* and *Provision* (Result 4); perceptions of kindness remain unaffected by misperceptions (Result 5). Hence, conditional cooperation is not just mistaken cooperation (as argued, e.g., by Burton-Chellew et al. (2016)), but a true preference that is less frequently found in *Maintenance* than in *Provision* dilemmas.

Our Result 4 somewhat differs from the results by Fosgaard et al. (2017). Like them, we find that the differences in conditional cooperation across dilemmas become smaller once accounting for misperceptions. Unlike us, in their case the differences become statistically insignificant while in ours they remain economically and statistically significant. One possible explanation for the different findings is that Fosgaard et al. (2017) asked their misperception questions, which are similar to our Measure 2, only at the end of the experiment rather than *before* the elicitation of cooperation preferences as we do. Another reason could be the different subject pools used across the two studies – our results are based on a UK student sample while theirs is based on a representative sample of the Danish population - which could also explain why the overall level of misperceptions is much higher in Fosgaard et al.: In their sample, 41% and 51% of participants exhibit some form of misperception in *Maintenance* and *Provision,* respectively, compared to 29% and 37% in our case. Our finding that kindness perceptions remain different across dilemmas also after removing misperceiving subjects reinforces further a preference explanation for the difference between *Maintenance* and *Provision*.

Our Results 1 – 5 also suggest an important general lesson: the revealed preference approach, that is, using choices to infer social preferences and/or dilemma-specific effects, requires controlling for perceptions.[24] This includes potential misperceptions of the game form to ensure measurement of preferences over clearly understood alternatives. Administering simple understanding questions at the beginning of experiments is nowadays quite common in experimental economics. However, it might not be enough. Our evidence on the existence of misperceived conditional cooperation is a point in case.

Our results also have implications for future literature. Hitherto, behavioral investigations of public goods provision and common pool resource problems have largely been conducted in independent literatures, in particular with regards to conditional cooperation, which was mostly studied in the context of linear public goods provision games (see, e.g., Chaudhuri (2011); Fehr and Schurtenberger (2018); and Thöni and Volk (2018)). Our comparative analysis of preferences and perceptions in maintenance and provision problems with identical social dilemma incentives is only a first step in bringing these two literatures closer together.

Finally, our results are not only of theoretical significance but have some potential policy implications. If many people are conditional cooperators, any factor that shifts beliefs about others' cooperativeness will shift cooperation – a fact that can be used for policy interventions (e.g., Gächter (2007)). The observation that conditional cooperation, even after being corrected for misperceptions, is weaker in *Maintenance* than *Provision* suggests that policy proposals that reckon with conditional cooperation (e.g., MacKay et al. (2015)) need to take into account that the extent of it is dilemma-specific. Some of the most pressing challenges for mankind such as stopping global warming and sustaining natural resources and biodiversity concern mainly *Maintenance* dilemmas (e.g., Fehr-Duda and Fehr (2016)). Our results suggest that the power of conditional cooperation may be limited in maintenance problems, at least in comparison with provision dilemmas. Other solutions such as punishment (Gächter et al. (2017); Ramalingam et al. (2019)) or incentives may instead be needed.

Acknowledgments

This work was supported by the European Research Council [grant numbers ERC-AdG 295707 COOPERATION and ERC-AdG 101020453 PRINCIPLES] and the Economic and Social Research

[24] Alternatively, when it is not possible to measure misperceptions directly (e.g., in representative surveys), econometric techniques such as those proposed by Goldin and Reck (2020) can be applied to correct for potential measurement error *ex post*.

Council [grant number ES/K002201/1]. The research reported in this paper was approved by the Research Ethics Committee of the Nottingham School of Economics. The authors declare they have no relevant or material financial interests that relate to the research described in this paper. We thank Ben Beranek for excellent research support and Abigail Barr, Tim Cason, Gary Charness, Jim Cox, Robin Cubitt, Martin Dufwenberg, Urs Fischbacher, Maria Garcia-Vega, Werner Güth, Friederike Mengel, Charles Noussair, Elena Manzoni, Vjollca Sadiraj, Maroš Servátka, Chris Starmer, Robert Sugden, and referees and participants from various seminars and conferences for helpful comments.

Data availability

The data and analysis code of this paper are available at https://osf.io/3jpgh/

References

Alpizar, F., Carlsson, F., Johansson-Stenman, O., 2008. Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica. Journal of Public Economics 92, 1047-1060.

Andreoni, J., 1995a. Cooperation in public-goods experiments - kindness or confusion? American Economic Review 85, 891-904.

Andreoni, J., 1995b. Warm glow versus cold prickle - the effects of positive and negative framing on cooperation in experiments. Quarterly Journal of Economics 110, 1-21.

Apesteguía, J., Maier-Rigaud, F. P., 2006. The role of rivalry. Public goods versus common-pool resources. Journal of Conflict Resolution 50, 646-663.

Arechar, A. A., Gächter, S., Molleman, L., 2018. Conducting interactive experiments online. Experimental Economics 21, 99-131.

Bardsley, N., 2008. Altruism or Artefact? A Note on Dictator Game Giving. Experimental Economics 11, 122–133.

Bartling, B., Engl, F., Weber, R. A., 2015. Game form misconceptions are not necessary for a willingness-to-pay vs. willingness-to-accept gap. Journal of the Economic Science Association 1, 72-85.

Bartling, B., Fehr, E., Özdemir, Y., 2021. Does Market Interaction Erode Moral Values? The Review of Economics and Statistics 1-32.

Battigalli, P., Dufwenberg, M., 2007. Guilt in games. The American Economic Review 97, 170-176.

Bayer, R.-C., Renner, E., Sausgruber, R., 2013. Confusion and learning in the voluntary contributions game. Experimental Economics 16, 478-496.

Becker, G. M., DeGroot, M. H., Marschak, J., 1964. Measuring utility by a single-response sequential method. Behavioral Science 9, 226-232.

Bellemare, C., Sebald, A., Suetens, S., 2019. Guilt aversion in economics and psychology. Journal of Economic Psychology 73, 52-59.

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) 57, 289-300.

Beranek, B., Cubitt, R., Gächter, S., 2015. Stated and revealed inequality aversion in three subject pools. Journal of the Economic Science Association 1, 43-58.

Bicchieri, C., Dimant, E., Gächter, S., Nosenzo, D., 2022. Social proximity and the erosion of norm compliance. Games and Economic Behavior 132, 59-72.

Bolton, G. E., Ockenfels, A., 2000. ERC: A theory of equity, reciprocity, and competition. American Economic Review. 90, 166-193.

Burton-Chellew, M. N., El Mouden, C., West, S. A., 2016. Conditional cooperation and confusion in public-goods experiments. Proceedings of the National Academy of Sciences 113, 1291-1296.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., Wu, H., 2016. Evaluating replicability of laboratory experiments in economics. Science 351, 1433-1436.

Camerer, C. F., Dreber, A., Johannesson, M., 2019. Replication and other practices for improving scientific quality in experimental economics. In Schram, A., Ule, A., (Eds.), Handbook of Research Methods and Applications in Experimental Economics. Edward Elgar Publishing, Cheltenham, pp. 83-102.

Cappelen, A. W., Nielsen, U. H., Sørensen, E. Ø., Tungodden, B., Tyran, J.-R., 2013. Give and take in dictator games. Economics Letters 118, 280-283.

Cartwright, E., 2016. A comment on framing effects in linear public good games. Journal of the Economic Science Association 2, 73-84.

Cason, T. N., Plott, C. R., 2014. Misconceptions and game form recognition: Challenges to theories of revealed preference and framing. Journal of Political Economy 122, 1235-1270.

Chang, Luke J., Smith, A., Dufwenberg, M., Sanfey, Alan G., 2011. Triangulating the Neural, Psychological, and Economic Bases of Guilt Aversion. Neuron 70, 560-572.

Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Quarterly Journal of Economics 117, 817-869.
Chaudhuri, A., 2011. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. Experimental Economics 14, 47-83.
Cornes, R., Sandler, T., 1996. The theory of externalities, public goods and club goods. Cambridge University Press, Cambridge.
Cox, C. A., 2015. Decomposing the effects of negative framing in linear public goods games. Economics Letters 126, 63-65.
Cox, C. A., Stoddard, B., 2015. Framing and Feedback in Social Dilemmas with Partners and Strangers. Games 6, 394-412.
Cox, J. C., Friedman, D., Gjerstad, S., 2007. A tractable model of reciprocity and fairness. Games and Economic Behavior 59, 17-45.
Cox, J. C., Friedman, D., Sadiraj, V., 2008. Revealed altruism. Econometrica 76, 31-69.
Cox, J. C., Ostrom, E., Sadiraj, V., Walker, J. M., 2013. Provision versus Appropriation in Symmetric and Asymmetric Social Dilemmas. Southern Economic Journal 79, 496-512.
Cubitt, R., Drouvelis, M., Gächter, S., 2011a. Framing and free riding: emotional responses and punishment in social dilemma games. Experimental Economics 14, 254–272.
Cubitt, R., Drouvelis, M., Gächter, S., Kabalin, R., 2011b. Moral Judgments in Social Dilemmas: How Bad is Free Riding? Journal of Public Economics 95, 253-264.
Dal Bó, P., Fréchette, G. R., 2018. On the Determinants of Cooperation in Infinitely Repeated Games: A Survey. Journal of Economic Literature 56, 60-114.
Dawes, R. M., 1980. Social Dilemmas. Annual Review of Psychology 31, 169-193.
De Dreu, C. K. W., McCusker, C., 1997. Gain–loss frames and cooperation in two-person social dilemmas: A transformational analysis. Journal of Personality and Social Psychology 72, 1093-1106.
Dreber, A., Ellingsen, T., Johannesson, M., Rand, D., 2013. Do people care about social context? Framing effects in dictator games. Experimental Economics 16, 349-371.
Dufwenberg, M., Gächter, S., Hennig-Schmidt, H., 2011. The framing of games and the psychology of play. Games and Economic Behavior 73, 459-478.
Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games and Economic Behavior 47, 268-298.
Dufwenberg, M., Kirchsteiger, G., 2019. Modelling kindness. Journal of Economic Behavior & Organization 167, 228-234.
Ellingsen, T., Johannesson, M., Mollerstrom, J., Munkhammar, S., 2012. Social framing effects: Preferences or beliefs? Games and Economic Behavior 76, 117-130.
Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior 54, 293-315.
Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods 1-17.
Fehr, E., Leibbrandt, A., 2011. A field study on cooperativeness and impatience in the Tragedy of the Commons. Journal of Public Economics 95, 1144-1155.
Fehr, E., Schmidt, K. M., 1999. A theory of fairness, competition, and cooperation. Quarterly Journal of Economics 114, 817-868.
Fehr, E., Schurtenberger, I., 2018. Normative foundations of human cooperation. Nature Human Behaviour 2, 458-468.
Fehr-Duda, H., Fehr, E., 2016. Game human nature: finding ways to adapt natural tendencies and nudge collective action is central to the well-being of future generations. Nature 530, 413-416.
Ferraro, P. J., Vossler, C. A., 2010. The source and significance of confusion in public goods experiments. The BE Journal of Economic Analysis & Policy 10, 1-42.
Fischbacher, U., 2007. z-Tree: Zurich toolbox for readymade economic experiments. Experimental Economics 10, 171-178.
Fischbacher, U., Gächter, S., 2010. Social preferences, beliefs, and the dynamics of free riding in public good experiments. American Economic Review. 100, 541–556.

Fischbacher, U., Gächter, S., Fehr, E., 2001. Are people conditionally cooperative? Evidence from a public goods experiment. Economics Letters 71, 397-404.
Fischbacher, U., Gächter, S., Quercia, S., 2012. The behavioral validity of the strategy method in public good experiments. Journal of Economic Psychology 33, 897-913.
Fosgaard, T. R., Hansen, L. G., Wengström, E., 2014. Understanding the nature of cooperation variability. Journal of Public Economics 120, 134-143.
Fosgaard, T. R., Hansen, L. G., Wengström, E., 2017. Framing and Misperception in Public Good Experiments. The Scandinavian Journal of Economics 119, 435-456.
Frackenpohl, G., Hillenbrand, A., Kube, S., 2016. Leadership effectiveness and institutional frames. Experimental Economics 19, 842-863.
Frey, B. S., Meier, S., 2004. Social comparisons and pro-social behavior. Testing 'conditional cooperation' in a field experiment. American Economic Review. 94, 1717-1722.
Fujimoto, H., Park, E.-S., 2010. Framing effects and gender differences in voluntary public goods provision experiments. Journal of Socio-Economics 39, 455-457.
Gächter, S., 2007. Conditional cooperation: Behavioral regularities from the lab and the field and their policy implications. In Frey, B. S., Stutzer, A., (Eds.), Psychology and Economics: A Promising New Cross-Disciplinary Field (CESifo Seminar Series). The MIT Press, Cambridge, pp. 19-50.
Gächter, S., Herrmann, B., 2009. Reciprocity, culture, and human cooperation: previous insights and a new cross-cultural experiment. Philosophical Transactions of the Royal Society B – Biological Sciences 364, 791-806.
Gächter, S., Kölle, F., Quercia, S., 2017. Reciprocity and the tragedies of maintaining and providing the commons. Nature Human Behaviour 1, 650-656.
Goldin, J., Reck, D., 2020. Revealed-Preference Analysis with Framing Effects. Journal of Political Economy 128, 2759-2795.
Greiner, B., 2015. Subject pool recruitment procedures: organizing experiments with ORSEE. Journal of the Economic Science Association 1, 114-125.
Hardin, G., 1968. The tragedy of the commons. Science 162, 1243-1148.
Horton, J. J., Rand, D. G., Zeckhauser, R. J., 2011. The online laboratory: conducting experiments in a real labor market. Experimental Economics 14, 399-425.
Houser, D., Kurzban, R., 2002. Revisiting Kindness and Confusion in Public Goods Experiments. American Economic Review 92, 1062-1069.
Isler, O., Gächter, S., Maule, A. J., Starmer, C., 2021. Contextualised strong reciprocity explains selfless cooperation despite selfish intuitions and weak social heuristics. Scientific Reports 11, 13868.
Iturbe-Ormaetxe, I., Ponti, G., Tomás, J., Ubeda, L., 2011. Framing effects in public goods: Prospect Theory and experimental evidence. Games and Economic Behavior 72, 439-447.
Khadjavi, M., Lange, A., 2015. Doing good or doing harm: experimental evidence on giving and taking in public good games. Experimental Economics 18, 432-441.
Kocher, M. G., Cherry, T., Kroll, S., Netzer, R. J., Sutter, M., 2008. Conditional cooperation on three continents. Economics Letters 101, 175-178.
Korenok, O., Millner, E. L., Razzolini, L., 2014. Taking, giving, and impure altruism in dictator games. Experimental Economics 17, 488-500.
Koszegi, B., Rabin, M., 2008. Choices, situations, and happiness. Journal of Public Economics 92, 1821-1832.
Ledyard, J. O., 1995. Public goods: a survey of experimental research. In Roth, A. E., Kagel, J. H., (Eds.), The Handbook of Experimental Economics. Princeton University Press, Princeton, pp. 111-181.
Levin, S. A., 2014. Public goods in relation to competition, cooperation, and spite. Proceedings of the National Academy of Sciences 111, 10838-10845.
Levine, D. K., 1998. Modeling Altruism and Spitefulness in Experiments. Review of Economic Dynamics 1, 593-622.
List, J. A., 2007. On the interpretation of giving in dictator games. Journal of Political Economy 115, 482-493.

MacKay, D. J. C., Cramton, P., Ockenfels, A., Stoft, S., 2015. Price carbon — I will if you will. Nature 526, 315-316.
Maniadis, Z., Tufano, F., List, J. A., 2014. One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects. American Economic Review 104, 277-290.
Messer, K. D., Zarghamee, H., Kaiser, H. M., Schulze, W. D., 2007. New hope for the voluntary contributions mechanism: The effects of context. Journal of Public Economics 91, 1783-1799.
Miettinen, T., Kosfeld, M., Fehr, E., Weibull, J., 2020. Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions. Journal of Economic Behavior & Organization 173, 1-25.
Ostrom, E., 1990. Governing the commons. The evolution of institutions for collective action. Cambridge University Press, Cambridge.
Ostrom, E., 2006. The value-added of laboratory experiments for the study of institutions and common-pool resources. Journal of Economic Behavior & Organization 61, 149-163.
Park, E.-S., 2000. Warm-glow versus cold-prickle: a further experimental study of framing effects on free-riding. Journal of Economic Behavior and Organization 43, 405-421.
Plott, C. R., Zeiler, K., 2005. The willingness to pay-willingness to accept gap, the "endowment effect," subject misconceptions, and experimental procedures for eliciting valuations. American Economic Review 95, 530-545.
Poppe, M., 2005. The specificity of social dilemma situations. Journal of Economic Psychology 26, 431-441.
Rabin, M., 1993. Incorporating fairness into game-theory and economics. American Economic Review 83, 1281-1302.
Ramalingam, A., Morales, A. J., Walker, J. M., 2019. Peer punishment of acts of omission versus acts of commission in give and take social dilemmas. Journal of Economic Behavior & Organization 164, 133-147.
Rand, D. G., Nowak, M. A., 2013. Human cooperation. Trends in Cognitive Sciences 17, 413-425.
Rustagi, D., Engel, S., Kosfeld, M., 2010. Conditional Cooperation and Costly Monitoring Explain Success in Forest Commons Management. Science 330, 961-965.
Sell, J., Son, Y., 1997. Comparing public goods and common pool resources: three experiments. Social Psychology Quarterly 60, 118-137.
Selten, R., 1967. Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes. In Sauermann, H., (Ed.), Beiträge zur experimentellen Wirtschaftsforschung. J.C.B. Mohr (Paul Siebeck), Tübingen, pp. 136-168.
Snowberg, E., Yariv, L., 2021. Testing the Waters: Behavior across Participant Pools. American Economic Review 111, 687-719.
Sonnemans, J., Schram, A., Offerman, T., 1998. Public good provision and public bad prevention: The effect of framing. Journal of Economic Behavior & Organization 34, 143-161.
Spranca, M., Minsk, E., Baron, J., 1991. Omission and commission in judgment and choice. Journal of Experimental Social Psychology 27, 76-105.
Thöni, C., Volk, S., 2018. Conditional Cooperation: Review and Refinement. Economics Letters 171, 37-40.
van Dijk, E., Wilke, H., 1997. Is It Mine or Is It Ours? Framing Property Rights and Decision Making in Social Dilemmas. Organizational Behavior and Human Decision Processes 71, 195-209.
Volk, S., Thöni, C., Ruigrok, W., 2012. Temporal stability and psychological foundations of cooperation preferences. Journal of Economic Behavior & Organization 81, 664-676.
Wilson, B. J., 2012. Contra Private Fairness. American Journal of Economics and Sociology 71, 407-435.

ONLINE APPENDIX to

Preferences and Perceptions in Provision and Maintenance Public Goods

Simon Gächter[a,b,c], Felix Kölle[d], and Simone Quercia[e]

April 27, 2022

[a] School of Economics, University of Nottingham, University Park, Nottingham NG7 2RD, UK, e-mail: simon.gaechter@nottingham.ac.uk; [b] CESifo Munich; [c] IZA Bonn. [d] University of Cologne, Albertus Magnus Platz, 50923 Cologne, e-mail: felix.koelle@uni-koeln.de. [e] Economics Department, University of Verona, Via Cantarane 24, 37129 Verona, e-mail: simone.quercia@univr.it.

ONLINE APPENDIX A – Experimental Instructions

A.1 – Laboratory experiments

In the following we present the original instructions subjects received in our PROVISION treatment. Differences in the MAINTENANCE instructions are reported in square brackets and highlighted in italics. We also report instructions for the one-shot game conducted in part 3 and used to evaluate predictive power in Section 3.3.

Part 1

Instructions

You are participating in a study in which you will earn some money. The amount will depend on the outcome of a game you will play. The amount of money which you earned with your decisions will be paid to you in cash at the end of the experiment. We will not speak of Pounds during the experiment, but rather of points. At the end, the total number of points you have earned will be converted to Pounds at the following rate:

1 point = £0.2

These instructions are solely for your private information. **You are not allowed to communicate during the experiment.** If you have any questions, please raise your hand. A member of the experimental team will come to you and answer them in private.

All participants will be divided into groups of four members. **Only the experimenters will know who is in which group.**

The decision situation

We first introduce you to the basic decision situation. Then, you will complete a pre-study questionnaire on the screen in front of you, which is intended to help you understand the decision situation.

In each group, every member has to decide the allocation of 20 tokens. You can put these 20 tokens into your **private account** or you can put some or all of them into a **project**. [*In each group, there are 80 tokens in a project. You can withdraw up to 20 tokens from the **project** and put them into your **private account** or you can leave them fully or partially in the **project**.*] The other three members of your group have to make the same decision.

Your income from the private account

You will earn 1 point for each token you put into your private account. For example, if you put all 20 tokens into your private account, your income from your private account would be 20 points. If you put 6 tokens into your private account, your income from this account would be 6 points. **No one except you earns anything from tokens you put in your private account.**

Your income from the project

Each group member will profit equally from the amount you or any other group member put into [*leave in*] the project. The income for each group member from the **project** will be determined as follows:

Income from the project = *0.4 × (sum of contributions)* [*0.4 × (80 - sum of all tokens withdrawn from the project)*]

If, for example, the sum of all contributions to the project [*tokens withdrawn from the project*] by you and your other group members is 60 [*20*] tokens, then you and each other member of your group would earn 60 [*80-20*] × 0.4 = 24 points out of the project. If the four members of the group contribute [*withdraw*] a total of 10 [*70*] tokens to [*from*] the project, you and the other members of your group would each earn 10 [*80-70*] × 0.4 = 4 points.

Total income

Your total income is the sum of your income from your private account and from the project:

Your Total Income = *Income from your private account + Income from the project*

=20 – your contribution to the project + 0.4 × sum of all contributions to the project

[= Tokens withdrawn from the project by you + 0.4 × (80-sum of all tokens withdrawn from the project)

Please answer all the following questions, to help you understand the determination of your income.

1. Each group member has 20 tokens. Assume that none of the four group members (including you) contributes anything to the project. [*There are 80 tokens in the project. Assume that everyone in your group withdraws 20 tokens from the project.*]

What will your total income (in points) be?

What will the total income (in points) of each of the other group members be?

2. Each group member has 20 tokens. You contribute 20 tokens in the project. Each of the other three members of the group also contributes 20 tokens to the project. [*There are 80 tokens in the project. You withdraw 0 tokens from the project. Each of the other three members of the group also withdraws 0 tokens from the project.*]

What will your total income (in points) be?

What will the total income (in points) of each of the other group members be?

3. Each group member has 20 tokens. The other three members contribute a total of 30 tokens to the project. [*There are 80 tokens in the project. The other three members withdraw 30 tokens from the project.*]

a) What will your total income (in points) be, if - in addition to the 30 tokens contributed by others - you contribute 0 tokens to the project? [*What will your total income (in points) be, if - in addition to the 30 tokens withdrawn by others - you withdraw 20 tokens from the project?]*

b) What will your total income (in points) be, if - in addition to the 30 tokens contributed by others - you contribute 8 tokens to the project? [*What will your total income (in points) be, if - in addition to the 30 tokens withdrawn by others - you withdraw 12 tokens from the project?*]

c) What will your total income (in points) be, if - in addition to the 30 tokens contributed by others - you contribute 15 tokens to the project? [*What will your total income (in points) be, if - in addition to the 30 tokens withdrawn by others - you withdraw 5 tokens from the project?]*

4. Each group member has 20 tokens. Assume you invest 8 tokens to the project. [*There are 80 tokens in the project. Assume you withdraw 12 tokens from the project.*]

a) What will your total income (in points) be, if the other group members - in addition to your 8 tokens - contribute another 7 tokens to the project? [*What will your total income (in points)? be, if the other group members - in addition to your 12 tokens - withdraw another 53 tokens from the project.*]

b) What will your total income (in points) be, if the other group members - in addition to your 8 tokens - contribute another 12 tokens to the project? [*What will your total income (in points) be, if the other group members - in addition to your 12 tokens - withdraw another 48 tokens from the project?*]

c) What will your total income (in points) be, if the other group members - in addition to your 8 tokens - contribute another 22 tokens to the project? [*What will your total income (in points) be, if the other group members - in addition to your 12 tokens - withdraw another 38 tokens from the project?*]

Part 2

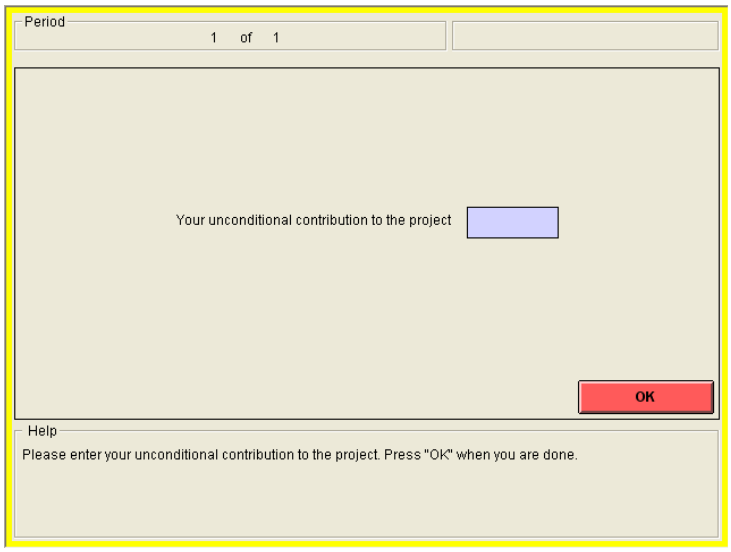
The Experiment

The experiment is based on the decision situation just described to you, conducted **once.** You will enter your decisions in the screen in front of you.

As you know, you will have 20 tokens at your disposal. You can put them into a private account or into a project. [*As you know, there are 80 tokens in a project. You can withdraw tokens from the project which will be automatically placed into your private account or you can leave them in the project.*] Each subject has to make **two types** of decisions in this

experiment, which we will refer to below as the **"unconditional contribution [*withdrawal*]"** and the **"contribution [*withdrawal*] table"**.

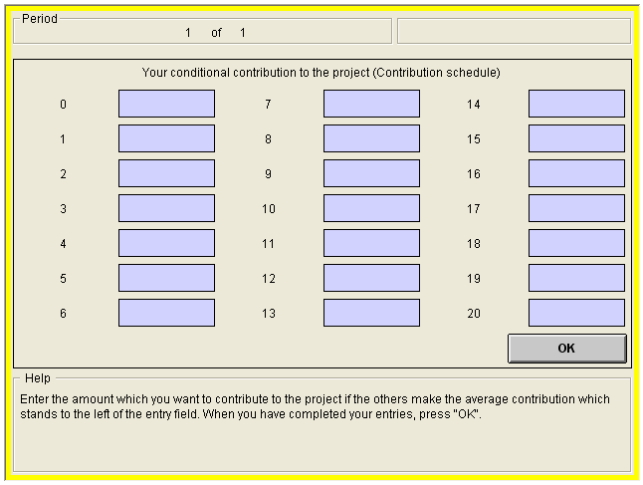
- In the **unconditional** contribution [*withdrawal*] you simply decide how many of the 20 [*80*] tokens you want to put in [*withdraw from*] the **project**. Please indicate your contribution [*withdrawal*] in the following screen (*screenshot taken from the PROVISION treatment*):



After you have determined your unconditional contribution [*withdrawal*], please click "OK".

- Your second task is to fill in a "contribution [*withdrawal*] table" where you indicate how many tokens **you want to contribute [*withdraw*]** to [*from*] the project **for each possible average contribution [*withdrawal*] of the other group members** (rounded to the next integer). Here, you can condition your contribution [*withdrawal*] on that of the other group members. This will be immediately clear to you if you *Maintenance* a look at the following table.

  This table will be presented to you in the experiment (*screenshot taken from the Provision treatment*):



The numbers to the left of the blue cells are the possible (rounded) average contributions [*withdrawals*] of the **other** group members to the project. You have to insert how many tokens you want to contribute to [*withdraw from*] the project into each input box – conditional on the indicated average contribution [*withdrawal*] by the other members of your group. **You must enter a number between 0 and 20 inclusive in each input box**. For example, you have to indicate how much you contribute to [*withdraw from*] the project if the others contribute [*withdraw*] 0 tokens on average to [*from*] the project; how much you contribute [*withdraw*] if the others contribute [*withdraw*] 1, 2, or 3 tokens on average; etc. Once you have made an entry in each input box, click "OK".

After all participants of the experiment have made an unconditional contribution [*withdrawal*] and have filled in their contribution [*withdrawal*] table, a random mechanism will select one member from every group. For **this** group member, it is his **contribution** [***withdrawal***] **table** that will determine his actual contribution [*withdrawal*]; whereas, for the **other three** group members, it is their **unconditional contributions** [***withdrawals***] that will determine their actual contributions [*withdrawals*]. You will not know whom the random mechanism will select when you make your unconditional contribution [*withdrawal*] and fill in your contribution [*withdrawal*] table. You must therefore think carefully about both decisions because either could determine your actual contribution [*withdrawal*]. Two examples should make this clear.

EXAMPLE 1: Suppose that **the random mechanism selects you;** and that the other three group members made unconditional contributions [*withdrawals*] of 0, 2, and 4 [*20, 18, and 16*] tokens, respectively. The average contribution [*withdrawal*] of these three group members is, therefore, 2 [*18*] tokens. If you indicated in your contribution [*withdrawal*] table that you will contribute [*withdraw*] 1 [*19*] token[*s*] if the others contribute [*withdraw*] 2 [*18*] tokens on average, then the total contribution to the project is given by 0+2+4+1=7 [*the total number of tokens left in the project is given by 80-(20+18+16+19)=7*] tokens. Each group member would, therefore, earn 0.4×7=2.8 points from the project plus their respective income from their own private account. If, instead, you indicated in your contribution [*withdrawal*] table that you would contribute [*withdraw*] 19 tokens [*1 token*] if the others contribute [*withdraw*] 2 [*18*] tokens on average, then the total contribution of the group to the project would be given by 0+2+4+19=25 [*the total number of tokens left in the project would be given by 80-(20+18+16+1)=25*] tokens. Each group member would earn 0.4×25=10 points from the project plus their respective income from their own private account.

EXAMPLE 2: Suppose **that the random mechanism does not select you;** and that your unconditional [*withdrawal*] contribution is 16 [*4*] tokens, while those of the other two group members not selected by the random mechanism are 18 [*2*] and 20 [*0*] tokens, respectively. Your average unconditional contribution [*withdrawal*] and that of these two other group members is, therefore, 18 [*2*] tokens. If the group member whom the random mechanism did select indicates in her contribution [*withdrawal*] table that she will contribute [*withdraw*] 1 [19] token[s] if the other three group members contribute [*withdraw*] on average 18 [*2*] tokens, then the total contribution of the group to the project is given by 16+18+20+1=55 [*the total number of tokens left in the project is given by 80-(4+2+0+19)=55*] tokens. Each group member will therefore earn 0.4×55=22 points from the project plus their respective income from their own private account. If, instead, the randomly selected group member indicates in her contribution [*withdrawal*] table that she contributes [*withdraws*] 19 [*1*] if the others contribute [*withdraw*] on average 18 [*2*] tokens, then the total contribution of the

group to the project is 16+18+20+19=73 [*the total number of tokens left in the project is 80-(4+2+0+1)=73*] tokens. Each group member would therefore earn 0.4×73=29.2 points from the project plus their respective income from their own private account.

The random selection of the group member whose contribution [***withdrawal***] **table will determine his actual contribution** [***withdrawal***] will be made as follows. Each group member is assigned a **Group Member ID** between 1 and 4, which denote his/her number inside his group. Moreover, participant number 2 was randomly selected at the very beginning of the experiment. This participant will draw a ball from an urn **after** all participants have made their unconditional contribution [*withdrawal*] and have filled out their contribution [*withdrawal*] table. Each ball in the urn has a different colour and each colour corresponds to a **Group Member ID**: orange=1, blue=2, yellow=3, green=4. The resulting number will be entered into the computer. If participant 2 draws the Group Member ID that was assigned to you, then your contribution [*withdrawal*] table will determine your contribution [*withdrawal*] and their unconditional contributions [*withdrawals*] will determine the contribution [*withdrawals*] of the other group members. Otherwise, your unconditional [*withdrawal*] contribution determines your contribution [*withdrawal*].

Part 3

Instructions

You are now taking part in a second experiment. The money you earn in this experiment will be added to what you earned in the first one. As before, we will not speak of Pounds during the experiment, but rather of points. At the end, the number of points you have earned will be converted to Pounds at the following rate:

1 point=£0.2

As in the previous experiment you are in a group composed by 4 people. However, the composition of the group is entirely new. None of the participants who were in your group in the second experiment will be in your group in this experiment.

The decision situation is the same as the one described on the first instruction sheet of the previous experiment. Each member of the group has to decide about the usage of the 20 tokens. [*In each group there are 80 tokens in a project.]* You can put these 20 tokens into your private account or you can put them fully or partially into a project. [*You can withdraw up to 20 tokens from the project or you can leave them fully or partially in the project.*] Each token you do not put into the project [*withdraw from the project*] is automatically placed into your private account. Your income will be determined in the same way as before.

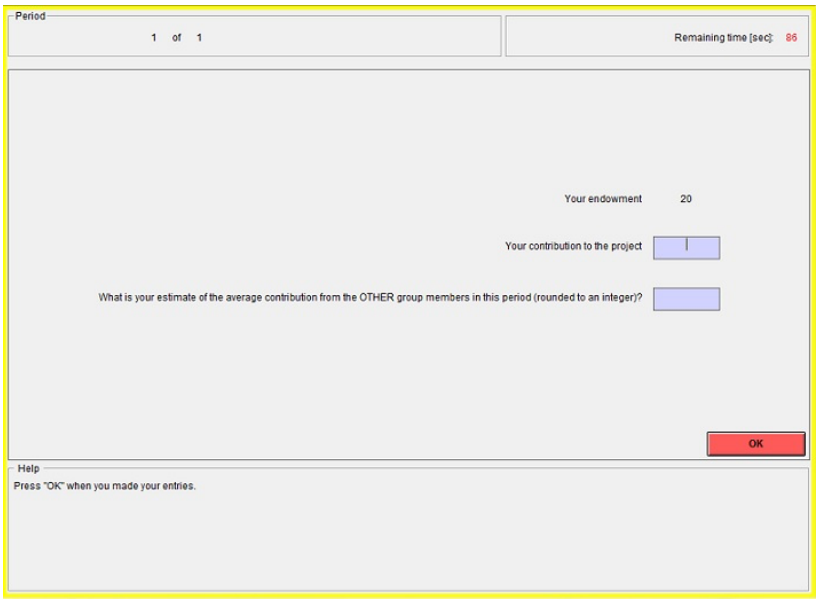
Reminder:

Your Total Income *= Income from your private account + Income from the project*

=20 – your contribution to the project + 0.4 × sum of all contributions to the project

[=Tokens withdrawn from the project by you +0.4 × (80-sum of all tokens withdrawn from the project)]

The decision screen looks like this (*screenshot taken from the PROVISION treatment*):



1. First you have to **decide on your contribution to [*withdrawal from*] the project**, that is, you have to decide how many of the 20 tokens you want to contribute to the project, and how many tokens you want to put into your private account. [*you have to decide how many of the 80 tokens you want to withdraw from the project and put into your private account*.] Each other member of your group has to make the corresponding decision. This is the only contribution [*withdrawal*] decision that you or they make in this experiment. There is **no contribution** [***withdrawal***] **table**.

2. Afterwards you have to estimate the average contribution to [*withdrawal from*] the project (rounded to an integer) of the other three group members. You will be paid for the accuracy of your estimate:
   - If your estimate is exactly right (that is, if your estimate is **exactly** the same as the actual average contribution [*withdrawal*] of the other group members), you will get **3 points** in addition to your other income from the experiment.
   - If your estimate deviates by one point from the correct result, you will get 2 additional points.
   - A deviation by 2 points still earns you 1 additional point.
   - If your estimate deviates by 3 or more points from the correct result, you will not get any additional points.

A.2 – Online experiments (MTurk)

In this decision problem you will form a group with three other people from MTurk. To determine your bonus payment, we will first record your earnings in points and then exchange the sum of points you earned into a dollar amount for your bonus payment.

Your **bonus** in Dollars will be determined as follows: **Earnings in Dollars = Earnings in Points / 20.**

In each group, every group member has an endowment of 20 tokens. You can put these 20 tokens into your **private account** or you can contribute them fully or partially to a **project**. [*In each group, there are 80 tokens in a project. You can withdraw up to 20 tokens from the **project** and put them into your **private account** or you can leave them fully or partially in the **project**.]* The other three members of your group have to make the same decision.

You will earn an income from your private account and from the project.

Your income from the private account

You will earn 1 point for each token you put into your private account. For example, if you put 20 tokens into your private account, your income from your private account is 20 points. If you put 6 tokens into your private account, your income from this account is 6 points. **No one except you earns anything from tokens you put into your private account.**

Your income from the project

Each group member will profit equally from the amount you or any other group member contributes to [*leaves into*] the project. All tokens contributed to [*left in*] the project will be **increased by 60 percent (a factor of 1.6) and split equally** among the four group members. That is, for every token contributed [*left*] by any group member, you and all three other group members will receive: 1 x 1.6 / 4 = 0.4 points each.

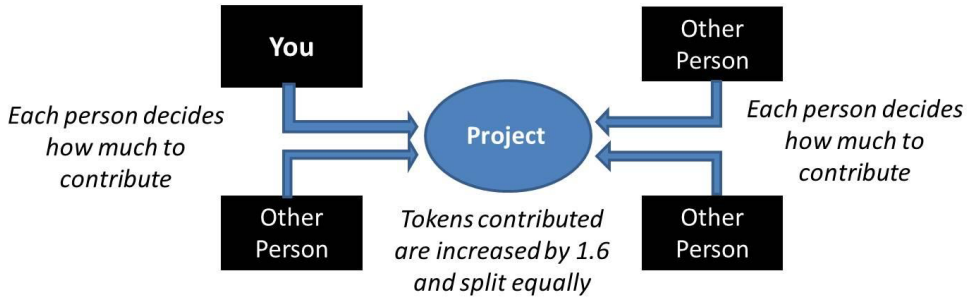
If, for example, the sum of all tokens contributed to [*left into*] the project by you and your other three group members is 60 tokens, then you and each other member of your group would earn 60 × 1.6 / 4 = 60 × 0.4 = 24 points.

If the four members of the group contribute [*leave*] a total of 10 tokens in the project, you and the other three members of your group would each earn 10 × 1.6 / 4 = 10 × 0.4 = 4 points.

Total income

Your total income is the sum of your income from your private account and from the project.

The graphic below shows a summary of the interaction (*figure from the Provision treatment only*):



Income: point earnings from private account + point earnings from project

Please answer the following questions to check your understanding of the situation.

Assume that all four group members (including you) contribute 0 [*withdraw 20*] tokens each to [*from*] the project. What will your total point earnings be (= point earnings from private account + point earnings from project)?

Assume you contribute 20 [*withdraw 0*] tokens to [*from*] the project. Each of the other three members of the group also contributes 20 [*withdraws 0*] tokens to [*from*] the project. What will your total earnings be (= point earnings from private account + point earnings from project)?

Assume you contribute 0 [*withdraw 20*] tokens to [*from*] the project and the other group members contribute [*leave*] in total 60 tokens to [*in*] the project. What will your total earnings be (= point earnings from private account + point earnings from project)?

Assume you contribute 20 [*withdraw 0*] tokens to [*from*] the project and the other group members contribute in total 0 tokens to the project [*leave 0 tokens in the project*]. What will your total earnings be (= point earnings from private account + point earnings from project)?

You are now ready to make your decisions. Your task is based on the decision problem described above**.**

As you know, you will have 20 tokens at your disposal. You can put them into a private account or into a project. [*As you know, there will be 80 tokens in a project. You can withdraw up to 20 tokens and put them into a private account or leave them in the project*.]

All group members have **two** tasks, which we will refer to below as the **"unconditional contribution [*withdrawal*]"** and the **"contribution [*withdrawal*] table"**.

In the **unconditional** contribution [*withdrawal*] task you simply decide how many tokens (up to 20) you want to contribute to [*withdraw from*] the **project**.

Your second task is to fill in a "contribution [*withdrawal*] table" where you indicate how many tokens **you want to contribute** to **[*withdraw from*]** the project **for each possible average contribution [*withdrawal*] of the other group members** (rounded to the next integer). Here, you can condition your contribution [*withdrawal*] on that of the other group members. You can see the table that you will have to fill in if you scroll down.

This is a one-off decision problem that is finished once you have made both decisions.

How your bonus will be determined

When all participants in your group have made their decisions, we will randomly select three group members for whom the unconditional contribution [*withdrawal*] will be relevant for their earnings. For the non-selected group member, the contribution [*withdrawal*] table will be relevant for his/her earnings. This means that you should *Maintenance* both the unconditional contribution [*withdrawal*] and the contribution [*withdrawal*] table equally seriously because you don't know yet which one will be relevant for calculating your bonus.

Example:

- Imagine that the unconditional contributions [*withdrawals*] of group members 1, 2, 3, and 4 are 20, 15, 10 and 0, respectively.
- Assume that for group members 1, 3 and 4 the unconditional contributions [*withdrawals*] are relevant for their earnings and for group member 2 the contribution [*withdrawal*] table will be used to calculate earnings.
- Then we calculate the average of the three unconditional contributions [*withdrawal*] -- in our example: (20 + 10 + 0)/3 = 10.
- To determine the contribution [*withdrawal*] of group member 2 we will *Maintenance* the contribution [*withdrawal*] this group member indicates in his/her contribution [*withdrawal*] table if others contribute [*withdraw*] on average 10.
- Imagine that this group member contributes [*withdraws*] 12 if others contribute [*withdraw*] 10 on average. Then the total sum of contributions to [*withdrawals from*] the project is 20 + 12 + 10 + 0 = 42 and earnings are calculated as explained above.

We now ask you to make the unconditional contribution [*withdrawal*] decision, followed by filling in the contribution [*withdrawal*] table.

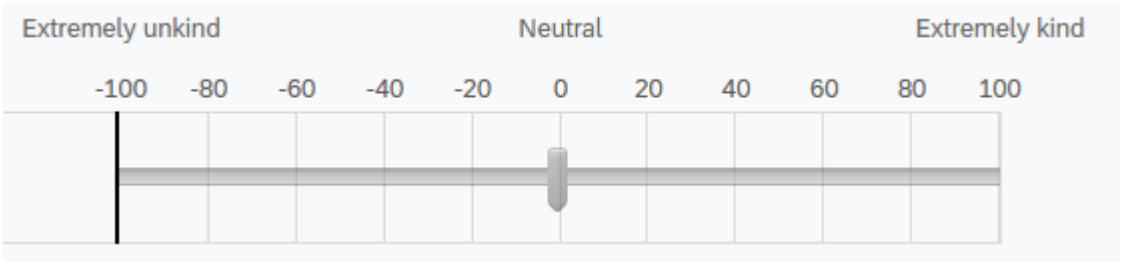
A.3 – Kindness Survey

After having read the general decision situation of either Maintenance or Provision (see above), participants were asked the following (results are in Sections 6 and 7):

In the following we ask you in various scenarios to evaluate the kindness of the other three group members on a scale from -100 to +100 where -100 corresponds to extremely unkind and +100 corresponds to extremely kind.

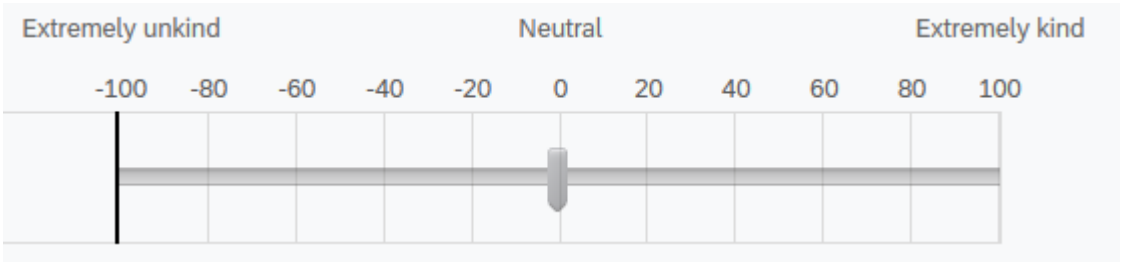
1. Assume that the other three group members contribute [*withdraw*] on average £0 [£20] to [*from*] the project.

How kind do you think they are?



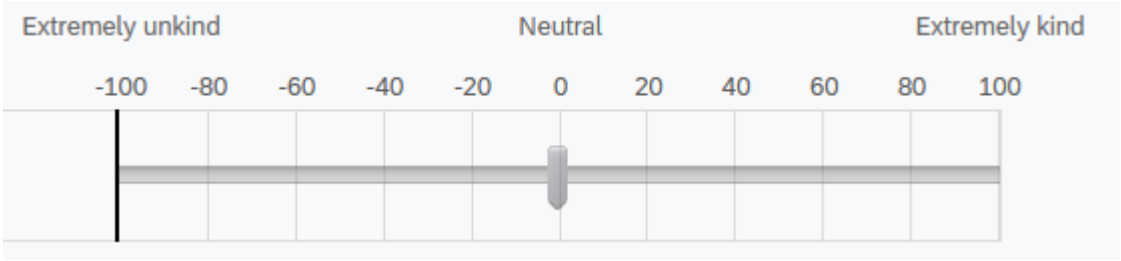
2. Assume that the other three group members contribute [*withdraw*] on average £10 [£10] to [*from*] the project.

How kind do you think they are?



3. Assume that the other three group members contribute [*withdraw*] on average £20 [£0] to [*from*] the project.

How kind do you think they are?



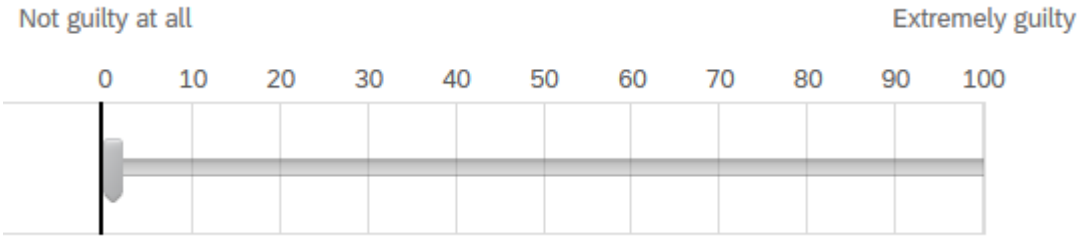
A.4 – Guilt Survey

After having read the general decision situation of either Maintenance or Provision (see above), participants were asked the following (results are in Section 8):

In the following we ask you in various scenarios to evaluate how guilty you would feel by contributing £0 [*withdrawing £20*] to [*from*] the project given various average contributions [*withdrawals*] of the other three group members. Please indicate your answer on a scale from 0 to 100, where 0 means "not guilty at all" and 100 means "extremely guilty". Please click on the slider to submit your answer.

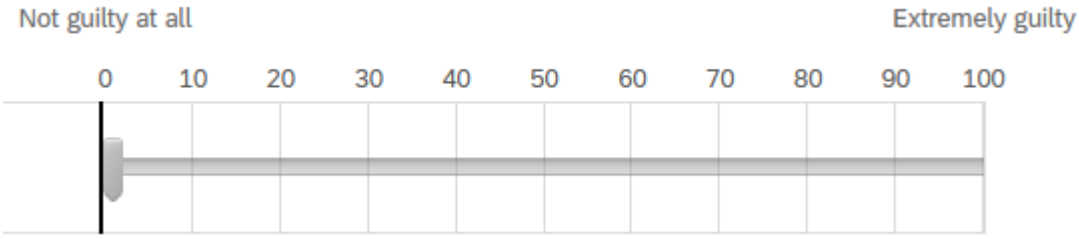
1. Assume that the other three group members move first, and you observe that they have contributed [*withdrawn*] on average **£0 [*£20*]** to [*from*] the project.

How guilty would you feel if you contribute **£0** [***withdraw £20***] in response?



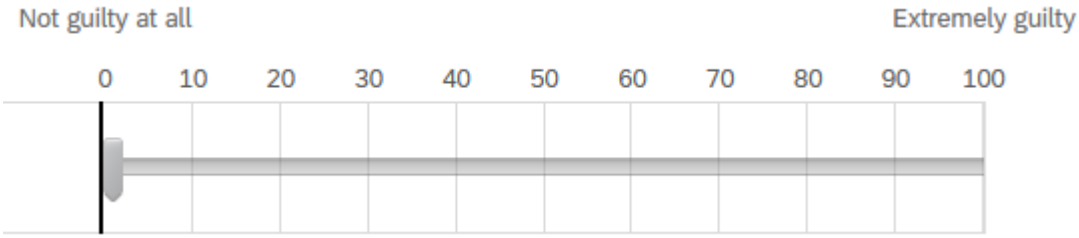
2. Assume that the other three group members move first, and you observe that they have contributed [*withdrawn*] on average **£10 [*£10*]** to [*from*] the project.

How guilty would you feel if you contribute **£0** [***withdraw £20***] in response?



3. Assume that the other three group members move first, and you observe that they have contributed [*withdrawn*] on average **£20 [*£0*]** to [*from*] the project.

How guilty would you feel if you contribute **£0** [***withdraw £20***] in response?



ONLINE APPENDIX B – Supplementary Analyses

B.1 – Procedural details and further supporting evidence for Section 3

Disaggregating the type category 'others'

In our main classification, described in Section 2 of the main text, we have only used three categories as our main focus is on conditional cooperators and free riders. The category "others", however, can be broken down into different patterns of behavior. In particular, in addition to free-riders and conditional cooperators, we classify a subject as (*i*) *unconditional cooperator* if she contributes a constant positive amount irrespective of the others' contributions, (*ii*) *triangle cooperator* if her contribution schedule is monotonically increasing up to a maximum of κ and thereafter monotonically decreasing, (*iii*) *anti-conditional cooperator* if either her contribution (withdrawal) schedule exhibits a (weakly) monotonically decreasing pattern, or if the Spearman correlation coefficient between her schedule and the others' average contribution (withdrawal) is negative and significant at $p < 0.01$, (*iv*) *other* if none of the criteria above apply.

Table B1 below reports the classification of cooperation types according to these criteria. Panel A reports the data from Gächter et al. (2017), and Panel B from our MTurk replication. In both samples the distribution of types is significantly different across treatments, in line with the results reported in the main text of the paper. Moreover, in both samples we find significantly more anti-conditional cooperators in *Maintenance* compared to *Provision*, and significantly fewer unconditional cooperators in *Maintenance* than in *Provision* in our MTurk sample.

Table B1. Classification of types – disaggregating the category 'others'

Panel A - Gächter et al. (2017, n = 704)

	Provision	*Maintenance*	χ^2 test
Conditional cooperators	63%	43%	$p < 0.001$
Free riders	17%	28%	$p = 0.001$
Unconditional cooperators	3%	3%	$p = 0.473$
Triangle cooperators	9%	10%	$p = 0.387$
Anti-conditional cooperators	3%	7%	$p = 0.006$
Others	5%	9%	$p = 0.065$
Test overall distribution		$\chi^2(5) = 34.15, p < 0.001$	

Panel B - MTurk experiment (n = 703)

	Provision	*Maintenance*	χ^2 test
Conditional cooperators	80%	67%	$p < 0.001$
Free riders	9%	17%	$p = 0.001$
Unconditional cooperators	7%	3%	$p = 0.020$
Triangle cooperators	2%	4%	$p = 0.244$
Anti-conditional cooperators	1%	7%	$p < 0.001$
Others	1%	2%	$p = 0.440$
Test overall distribution		$\chi^2(5) = 38.61, p < 0.001$	

Testing whether differences in types across differs across samples

As described in the main text, the distribution of types differs between our UK student and our US MTurk sample, with more conditional cooperators and less free riders and others in the latter. We observe these shifts for both *Maintenance* and *Provision* (compare Table B1). To test whether these shifts influence the differences across the two dilemmas, we run logistic regressions in which we use the different types as dependent variable, a treatment dummy, a sample dummy, and an interaction between the latter two as independent variables. The results are reported in Table B2 below. They confirm that there are significantly more conditional cooperators and significantly less free riders and others in *Provision* than *Maintenance* and on MTurk, but that there is no significant interaction effect (Provision × MTurk). These results indicate that the treatment differences across the two dilemmas are similar across both samples.

Table B2: Logistic regressions on the treatment differences in types across the Gächter et al. (2017) UK student and the US MTurk sample

	(1) CC	(2) FR	(3) OT
Provision *(1 if Provision, 0 otherwise)*	0.854*** (0.155)	-0.591*** (0.184)	-0.591*** (0.179)
MTurk *(1 if MTurk, 0 otherwise)*	1.003*** (0.158)	-0.630*** (0.187)	-0.785*** (0.187)
Provision × MTurk	-0.190 (0.233)	-0.172 (0.300)	0.207 (0.283)
Constant	-0.690*** (0.108)	-0.965*** (0.120)	-0.853*** (0.117)
N	1407	1407	1407

Notes: The dependent variable takes value 1 if a participant is classified as a conditional cooperator (Model (1)), free rider (Model (2)), or other (Model (3)) and 0 otherwise. Robust standard errors are in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Temporal Stability

To test the temporal stability of our revealed preference measure, four months after their first participation we re-invited a subset of $n = 288$ participants who had participated in our first experiment reported in Gächter et al. (2017) and in the left panel of Figure 1.[1] Without knowing in advance, participants took part in sessions that were identical to the ones in which they participated before. This allows us to observe a participant's cooperation type at two different points in time and to assess its temporal stability. We report results from $n = 119$ participants; $n = 65$ in *Provision* and $n = 54$ in *Maintenance* who showed-up in both waves.[2] Table B3 below reports the frequencies of types in the two waves.

Table B3: Distribution of types in Maintenance (M) and Provision (P) across waves.

	Wave 1			**Wave 2**		
	M	P	χ^2 - test	M	P	χ^2 - test
Conditional Cooperators	43%	66%	*p = 0.010*	39%	60%	*p = 0.022*
Free Riders	18%	20%	*p = 0.839*	24%	29%	*p = 0.528*
Others	39%	14%	*p = 0.002*	37%	11%	*p = 0.001*
χ^2-test M vs. P	*p = 0.006*			*p = 0.003*		

Notes: M corresponds to *Maintenance* and P for *Provision*.

In addition to the analyses reported in the main text, here we provide further information on the stability of types at the individual level. Our results are summarized in Table B4, displaying the relative frequency of all possible type combination across the two waves. As can be seen, the largest number of observations lie on the main diagonal, corresponding to participants who display the same type in both waves. The numbers on the off-diagonal, in contrast, reveal changes in types across the two waves. As can be seen, we observe changes of types in all directions. For example, while some participants classified as conditional cooperators in Wave 1 are classified as free riders in Wave 2, some others display an opposite change.

[1] Our experiment was conducted in several waves as the third part differed across waves (see Gächter et al. (2017)). The participants we re-invited were the ones who participated in the first wave. The data we report here as Wave 2 were not used in Gächter et al. (2017).

[2] Due to attrition, we also invited additional participants to be able to form groups of four. For the ease of exposition, the data from these $n = 48$ additional participants are not reported here but are available upon request. The results from these participants, however, are very similar to the ones reported above. More importantly, we can rule out any selection effects with respect to our main variable of interest, as the attrition we found between the two waves was not related to cooperation attitudes. The distribution of types in Wave 1 does not significantly differ between participants who did or did not show up for the second experiment (*Maintenance:* $\chi^2(2) = 0.35$, $p = 0.839$; *Provision* $\chi^2(2) = 2.77$, $p = 0.251$). Participants also did not differ with respect to important socio-demographic characteristics such as age, gender, nationality, or field of studies.

Table B4: Relative frequency of all possible combination of types in the two waves.

		Maintenance			*Provision*		
Wave 2		CC	FR	OT	CC	FR	OT
	CC	**24.1%**	9.3%	9.3%	**49.2%**	10.8%	6.2%
Wave 1	FR	7.4%	**9.3%**	1.8%	4.6%	**13.9%**	1.5%
	OT	7.4%	5.5%	**25.9%**	6.1%	4.6%	**3.1%**

To test whether some types are more likely to be persistent over time, we can compare the fraction of 'stable' participants across the different types. In *Maintenance*, we find that 57% of all participants classified as conditional cooperator in Wave 1 are classified as the same type in Wave 2. Similar numbers are observed for free riders and others, for which we observe consistency rates of 50% and 67%, respectively. We find no significant differences in the consistency across types ($\chi^2(2) = 0.90$; $p = 0.636$). In *Provision*, consistency rates for conditional cooperators and free riders amount to 74% and 69%, respectively. These numbers are neither significantly different from each other ($\chi^2(1) = 0.14$; $p = 0.711$), nor are they different from the consistency rates of the same type in *Maintenance* (conditional cooperators: $\chi^2(1) = 2.21$; $p = 0.137$; free riders: $\chi^2(1) = 0.88$; $p = 0.349$). The only difference we observe is about *others*, who in *Provision* display a consistency rate of only 22%, which is lower than the one observed in *Maintenance* ($\chi^2(1) = 5.00$; $p = 0.025$), and lower than the one observed for the other types in *Provision* ($\chi^2(1) = 9.00$; $p = 0.003$).

To assess if individual-level stability occurs more frequently than would be expected if types would change randomly, we follow the approach of Volk et al. (2012) and simulate a distribution of types assuming that each participant randomly picks a type in Wave 2 with a probability equal to the observed frequency in our data. In 100 runs of this simulation, we find that in *Provision* and *Maintenance* participants are predicted to be of the same type in both waves in on average 47% and 36% of the cases, respectively. Testing the simulated distribution of stable types against the observed proportion in the experiment reveals that the hypothesis of random types can be rejected for both *Provision* and *Maintenance* (two-sided t-tests, both $p < 0.001$).

To further evaluate overall individual-level stability, we compare our results to the ones reported in Volk et al. (2012) who use a provision game in a similar design to ours. In line with our results, Volk et al. find that 64% of participants are classified as the same type between two waves that are 2.5 months apart. No such comparison is possible for *Maintenance* as (to the best of our knowledge) no previous study has investigated the stability of cooperation preferences using a maintenance game.

As a final step, we check whether the differences in the distribution of cooperation types is also robust when only considering the subset of participants who are classified as the same type in both waves. A significant difference in this subsample would indicate that the

stability of the treatment effect across waves is systematic as it is due to differences in cooperation attitudes that are stable at the individual level. Strikingly, the effect among this subsample is even stronger in terms of percentage point differences across types compared to the whole sample. We find 74% (41%) conditional cooperators, 21% (15%) free riders and 5% (44%) others in *Provision* (*Maintenance*) ($\chi^2(2) = 16.92$, $p < 0.001$).

Predictive Power

To test the predictive power of our revealed preference measure we follow Fischbacher et al. (2012), combining the data from our strategy-method elicitation with the data from a one-shot direct-response game that followed immediately after. After participating in the strategy method experiment, subjects were re-matched in a perfect stranger protocol to play a one-shot simultaneous game where we also elicited beliefs about the average effective contribution of the other three members of their group.[3] Using cooperation preferences and stated beliefs allows to make a point prediction about the contribution decision in the direct-response game, $\hat{c}_i$. For each individual, we then compare the predicted contribution with their actual contribution in the direct-response game, c_i, delivering an individual-level measure of consistency. For this exercise, we can use all the experiments where the strategy method experiment was followed by a one-shot direct-response game with belief elicitation. For this we have (1) n = 288 observations from participants from our first experiment reported in Gächter et al. (2017) and in the left panel of Figure 1, and (2) n = 703 observations from participants in our MTurk experiment. We further report data from (3) a set of n = 116 participants for which the elicitation of cooperation preferences and the direct-response game took place in two separate sessions that lay five months apart. [4]

Further to the analyses reported in the main text, we follow Fischbacher et al. (2012) and define a subject as consistent if their actual cooperation decision does not deviate by more than ± 2 tokens (10 percent of their endowment) from their predicted contribution. We find that 63 and 62 percent of lab participants are consistent in *Maintenance* and *Provision*, respectively. The one percent difference is not statistically significant ($\chi^2(1) = 0.15$; p = 0.903). We observe similar numbers in the MTurk sample with no differences across treatments (*Maintenance*: 63%, *Provision* 65%, $\chi^2(1) = 0.18$; $p = 0.672$). In our experiment with a delay of five months, consistency is still remarkably high: 50% and 55% of participants are classified as consistent in *Maintenance* and *Provision*, respectively.

[3] Participants were paid for the accuracy of their beliefs. If their beliefs matched the other's average contribution exactly, participants earned 3 points. When their belief deviated by 1 (2) point(s) from the correct estimate they earned 2 (1) points. If their estimation was off by more than two points, they received no additional money.

[4] In this new set of experiments, we had a total of n = 696 participants. Out of these, five months after the first experiment we randomly re-invited n = 312 participants to participate in a one-shot direct-response experiment using the same dilemma, *Maintenance* or *Provision*. None of these participants had played a direct-response game in the first experiment. We report the elicited cooperation attitudes of these n = 696 participants in Section 5. As before, we can rule out that attrition was related to cooperation attitudes, as the distribution of types in the sample of participants who did not show up five months later is not statistically different from the distribution for the ones who did show up ($\chi^2(2) = 0.09$, $p = 0.953$ and $\chi^2(2) = 0.23$, $p = 0.894$ in *Maintenance* and *Provision*, respectively). Participants also did not differ with respect to important socio-demographic characteristics such as age, gender, nationality, or field of studies.

Compared to the case without delay, consistency is lower, but this difference is only marginally significant ($\chi^2(1) = 3.38$; $p = 0.066$). As before, we find no difference in the distribution of deviations across treatments ($\chi^2(1) = 0.33$; $p = 0.564$). We conclude that our proxy for cooperation preferences is an equally good predictor for actual gameplay in *Maintenance* and *Provision.*

B.2 – Supporting evidence for Section 4

Differences in kindness perceptions across Maintenance and Provision

Table B5: OLS regressions on the evaluation of kindness of others' contributions

	(1) Lab	(2) MTurk
Other's average contribution	3.752*** (0.400)	4.704*** (0.268)
Other's average contribution × Provision	2.410*** (0.483)	1.703*** (0.342)
Provision *(1 if Provision, 0 otherwise)*	-26.317*** (6.310)	-19.694*** (4.390)
Constant	-21.690*** (5.229)	-25.875*** (3.597)
N	555	1203
R^2	0.541	0.562

Notes: Dependent variable: Kindness evaluations on a scale from -100 to +100 (where -100 corresponds to 'very unkind' and +100 corresponds to 'very kind'). Robust standard errors clustered on the individual level are in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Model (1) includes survey responses from $n = 185$ students and model (2) includes survey responses from $n = 401$ MTurkers.

B.3 – Supporting evidence for Section 5

Testing whether differences in types across differs across samples

As described in the main text, the distribution of types differs between our initial student sample and the one in which we administered additional incentivized control questions prior to eliciting cooperation attitudes (compare footnote 19). Specifically, we find a significant change in the distribution of types in both *Provision* ($\chi^2(2) = 14.26$, $p = 0.001$) and *Maintenance* ($\chi^2(2) = 11.01$, $p = 0.004$), with more free riders (*Provision*: 27% vs. 17%, $\chi^2(1) = 10.43$, $p = 0.001$, *Maintenance* 39% vs. 28%, $\chi^2(1) = 10.44$, $p = 0.001$) and fewer conditional cooperators (*Provision*: 63% vs. 51%, $\chi^2(1) = 12.50$, $p < 0.001$, *Maintenance* 33% vs. 43%, $\chi^2(1) = 5.84$, $p = 0.016$) in the new experiment. To test whether these shifts have an effect on the differences across the two dilemmas, we run logistic regressions in which we use the different types as dependent variable, a treatment dummy, a sample dummy, and an interaction between the latter two as independent variables. The results are reported in Table B6 below. As indicated by the insignificant interaction effect (Provision × Misperception sample), despite the differences in the distribution of types across both samples, we find no evidence that this had any systematic effect on our treatment comparison.

Table B6: Logistic regressions on the treatment differences in types across the Gächter et al. (2017) and the student sample reported in Section 5

	(1) CC	(2) FR	(3) OT
Provision *(1 if Provision, 0 otherwise)*	0.854*** (0.155)	-0.591*** (0.184)	-0.591*** (0.179)
Misperception sample *(1 if Misperception sample, 0 otherwise)*	-0.387** (0.161)	0.534*** (0.166)	-0.132 (0.172)
Provision × Misperception sample	-0.144 (0.220)	0.048 (0.246)	0.314 (0.251)
Constant	-0.301*** (0.108)	-0.965*** (0.120)	-0.853*** (0.117)
N	1400	1400	1400

Notes: The dependent variable takes value 1 if a participant is classified as a conditional cooperator (Model (1)), free rider (Model (2)), or other (Model (3)) and 0 otherwise. Robust standard errors are in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Disaggregating the type category 'others'

As described in Section B1 above, we can break down the category "others" into the following subcategories: (*i*) *unconditional cooperator*, (*ii*) *triangle cooperator*, (*iii*) *anti-conditional cooperator*, and (*iv*) *other*. Table B7 below reports the classification of cooperation types according to these criteria. Panel A reports the data from the full sample, while Panels B, C, and D report the results after dropping participants who are classified as misperceiving according to Measures 1, 2, and 3, respectively.

Table B7. Classification of types disaggregating the category 'others' and controlling for misperceptions

Panel A – All (n = 696)

	Provision	*Maintenance*	χ^2 test
Conditional cooperators	51%	33%	$p < 0.001$
Free riders	27%	39%	$p = 0.001$
Unconditional cooperators	2%	4%	$p = 0.409$
Triangle cooperators	13%	14%	$p = 0.770$
Anti-conditional cooperators	3%	5%	$p = 0.158$
Others	4%	5%	$p = 0.409$
Test overall distribution	$\chi^2(5) = 22.69, p < 0.001$		

Panel B – No mistake in payoff questions (n = 572)

	Provision	*Maintenance*	χ^2 test
Conditional cooperators	49%	34%	$p < 0.001$
Free riders	29%	41%	$p = 0.005$
Unconditional cooperators	2%	2%	$p = 0.325$
Triangle cooperators	14%	14%	$p = 0.933$
Anti-conditional cooperators	3%	4%	$p = 0.744$
Others	3%	5%	$p = 0.332$
Test overall distribution	$\chi^2(2) = 14.53, p = 0.013$		

Table B7 continued

Panel C – No mistake in goal questions (n = 463)

	Provision	*Maintenance*	χ^2 test
Conditional cooperators	45%	35%	$p = 0.042$
Free riders	32%	43%	$p = 0.019$
Unconditional cooperators	2%	2%	$p = 0.564$
Triangle cooperators	14%	14%	$p = 0.919$
Anti-conditional cooperators	3%	2%	$p = 0.605$
Others	4%	4%	$p = 0.748$
Test overall distribution		$\chi^2(2) = 6.60, p = 0.252$	

Panel D – No mistaken conditional cooperation (n = 632)

	Provision	*Maintenance*	χ^2 test
Conditional cooperators	46%	34%	$p = 0.002$
Free riders	31%	41%	$p = 0.007$
Unconditional cooperators	2%	3%	$p = 0.344$
Triangle cooperators	14%	14%	$p = 0.878$
Anti-conditional cooperators	3%	4%	$p = 0.679$
Others	4%	4%	$p = 0.830$
Test overall distribution		$\chi^2(2) = 11.25, p = 0.047$	

Differences in kindness perceptions across Maintenance and Provision

Table B8: OLS regressions on the evaluation of kindness of others' contributions controlling for misperceptions

	(1)	(2)	(3)	(4)
	All subjects	No mistake in payoff questions	No mistake in goal questions	No mistaken conditional cooperation
Other's average contribution	4.684***	4.589***	4.791***	4.839***
	(0.444)	(0.471)	(0.592)	(0.480)
Other's average contribution × Provision	1.813***	1.824***	1.897***	1.631***
	(0.523)	(0.563)	(0.681)	(0.565)
Provision	-23.623***	-23.484***	-22.325**	-20.019**
(1 if Provision, 0 otherwise)	(7.342)	(8.150)	(9.708)	(7.981)
Constant	-16.244***	-12.779**	-18.370**	-18.509***
	(5.928)	(6.335)	(7.760)	(6.418)
Observations	600	471	381	525
R^2	0.583	0.568	0.564	0.574

Notes: Dependent variable: Kindness evaluations on a scale from -100 to +100 (where -100 corresponds to 'very unkind' and +100 corresponds to 'very kind'). Robust standard errors clustered on the individual level are in parentheses, * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Model (1) includes responses from $n = 200$ students for whom we elicited cooperation attitudes before the kindness questionnaire. Models (2) - (4) use only those subjects who are not classified as misperceiving according to our first, second and third measure, respectively.

ONLINE APPENDIX C – Theoretical Considerations

This appendix complements Section 6 in the main text. We first show why models of distributional social preferences cannot predict differences conditional cooperation across *Maintenance* and *Provision* unless one assumes that parameters are context-dependent. Then, we analyze models of reciprocity and guilt aversion. Finally, we derive propositions applying the theory of revealed altruism (Cox et al. (2008)) to the two social dilemmas.

Models of distributional social preferences

Theories of distributional social preferences cannot explain the difference in conditional cooperation between *Maintenance* and *Provision* unless one assumes dilemma-specific parameters. To illustrate this, we use as a workhorse the model of inequity aversion by Fehr and Schmidt (1999). All conclusions derived here apply similarly to other models of distributional preferences. In the Fehr and Schmidt (1999) model, subjects are assumed to care about their material payoff and to bear a "psychological cost" for inequality in payoffs between themselves and each of the other players. This cost is higher if the inequality is to their disadvantage rather than to their advantage. The utility function for an inequity averse player is given by:

$$u_i(\pi_i, \pi_k) = \pi_i - \frac{\alpha_i}{n-1}\sum_{k\neq i}\max(0, \pi_k - \pi_i) - \frac{\beta_i}{n-1}\sum_{k\neq i}\max(0, \pi_i - \pi_k)$$

where α_i is player *i's* aversion to disadvantageous inequality and β_i is subject *i's* aversion to advantageous inequality, with $\alpha_i \geq \beta_i$ and $0 \leq \beta_i < 1$.

This model can reconcile conditional cooperation if β_i is high enough. The player will bear a psychological cost in free-riding when the average contribution of the others is positive. This cost can offset the material gain from free-riding if β_i is high enough. In that case, the decision maker will prefer matching the contribution of others rather than free-riding to avoid the (advantageous) inequality.

Note, however, that since for a given average effective contribution of the other group members the inequalities vis-à-vis the other players will be the same across *Maintenance* and *Provision*, this model cannot predict any difference in conditional cooperation across the two dilemmas, unless one would be willing to assume β_i, i.e., the distaste for advantageous inequity to be context-dependent. In particular, if one would assume β_i to be larger in *Provision* than in *Maintenance*, then this could rationalize our finding of more conditional cooperation in *Provision.* We note, however, that there is no psychological mechanism in the formal assumptions of the theory that postulates such a shift in β_i.

Theories of reciprocity

In theories of reciprocity (Rabin (1993); Dufwenberg and Kirchsteiger (2004); Falk and Fischbacher (2006)), agents' motivations derive from their material payoff as well as a psychological payoff that depends on their first- or second-order beliefs about others' actions. These theories are natural candidates to explain conditional cooperation because they postulate that agents want to reward kind actions (or intentions) with kindness and to punish hostile actions (or intentions) with unkindness. Kindness is evaluated relative to a reference point that is a statistic derived from the set of material consequences (typically an equitable payoff calculated as an average between the maximum and the minimum payoff a player can get).

Here, we follow Dufwenberg et al. (2011) and apply Dufwenberg and Kirchsteiger (2004) theory to the public goods game as follows: let a'_{ik} denote *i*'s 'first-order' belief about *k*'s cooperation (where $i \neq k$). We denote k_{ik} as the kindness of individual *i* towards individual *k.* Kindness is defined as the difference between what *i* actually gives to *k* and the average of the maximum (= 20) and minimum (= 0) that *i* could give to *k*. Applying this definition, we get $k_{ik} = a_i - 10$. We further define λ_{iki} as the belief of player *i* about the kindness of player *k*, i.e., the belief that *i* has about $k_{ki} = a_k - 10$. Hence, we plug-in the first order belief about *k*'s contribution to obtain $\lambda_{iki} = a'_{ik} - 10$.

The utility function of a reciprocal agent can then be described as follows:

$$u(a_i, a_k, a''_{iki}) = \pi_i + Y_i \sum_{k \neq i} k_{ik} \cdot \lambda_{iki}$$

where Y_i is an individual-level degree of reciprocity. Using the kindness terms defined above, we can rewrite the utility function as:

$$u(a_i, a_k, a''_{iki}) = \pi_i + Y_i \sum_{k \neq i} [(a_i - 10) \cdot (a'_{ik} - 10)]$$

The idea behind the model is that if the first order belief is greater than 10, individual *i* would maximize her utility by a cooperation level higher than 10. The opposite is true if her belief is below 10 and λ_{iki} is negative. Then, individual *i* would maximize her utility by a negative k_{ik}, which means she will cooperate less than 10 tokens.

Regarding the potential differences across dilemmas, the reference point which is only based on material payoffs must be the same across *Maintenance* and *Provision* because of equivalent game forms. Therefore, the only channel through which simultaneous game play may differ across *Maintenance* and *Provision* is via differences in beliefs. As argued in Dufwenberg et al. (2011), in a linear public goods game the evaluation of others' kindness only depends on first-order but not on second-order beliefs. Therefore, since in our strategy-method experiment first-order beliefs are fixed because participants condition their contributions on all possible average contributions of others, these models do not predict differences in cooperation preferences across *Maintenance* and *Provision*.

As for the theories of distributional preferences, this prediction hinges on the assumption that the concern for reciprocity as measured by a reciprocity parameter, Y_i, does not vary across dilemmas. If one would be willing to relax this assumption and instead assume that the reciprocity parameter is stronger under *Provision* than *Maintenance*, then reciprocity theory could also rationalize our finding that conditional cooperation is more frequent in *Provision* than *Maintenance*. However, also in this theory there is no psychological mechanism in the formal assumptions that postulates such a shift in Y_i.

Guilt Aversion

In modelling *guilt aversion*, we rely on a model of simple guilt by Battigalli and Dufwenberg (2007) and applied to public goods games in Dufwenberg et al. (2011). An agent's utility depends on her material payoff as well as her second-order beliefs, that is, what she believes the other players believe she will do. Guilt aversion predicts that player *i* will suffer guilt if *i* contributes less than what *i* thinks the other three group members expect *i* to contribute (on average). If the disutility from guilt becomes large enough, player *i* has an incentive to contribute whatever she thinks others expect her to contribute. Adopting the same specification as in Dufwenberg et al. (2011), *i*'s utility in the *Provision* dilemma is defined by:

$$u_i\left(c_i, c_j, c_k, c_l, b_{iji}, b_{iki}, b_{ili}\right) = 20 - c_i + 0.4\sum_{j=1}^{4} c_j - \gamma_i max\left(\frac{b_{iji}+b_{iki}+b_{ili}}{3} - c_i, 0\right),$$

where j, k and l denote the other players; $b_{iji}, b_{iki}, b_{ili}$ denote the second-order beliefs of player *i* and γ_i measures player *i*'s game-independent degree of guilt aversion. An analogous specification describes utilities in the *Maintenance* dilemma.

Since the strategy-method experiment fixes only first-order beliefs, differences in second-order beliefs can predict differences in cooperation preferences: If more participants in *Provision* than in *Maintenance* have second-order beliefs that others expect them to reciprocate their contributions, the perceived guilt from not matching others' contributions in *Provision* may be higher than in *Maintenance*, which, in turn, can lead to a higher fraction of conditional cooperators in *Provision* than *Maintenance.* Similarly, if more participants in *Maintenance* than in *Provision* believe that others expect them to free ride, the perceived guilt from actual free riding may be lower in *Maintenance* than *Provision*, which would predict more free riders in *Maintenance* than *Provision*. Hence, in contrast to the theories of inequity aversion and reciprocity, guilt aversion can reconcile our results without making ad-hoc assumptions on the preference parameters.

Revealed Altruism

Finally, we consider the theory of revealed altruism by Cox et al. (2008) which explicitly postulates a mechanism for the differences between *Maintenance* and *Provision*. Consider a generic player i and define an opportunity set at a given node of the game as a subset of $\mathbb{R}^2_+$ that contains all feasible payoffs for player i. We start by stating a similar definition of Definition 2 from Cox et al. (2008), which allows an ordering of opportunity sets contained in $\mathcal{C}$ (the set containing all the opportunity sets for player *i*):

Definition 2 (compare Cox et al. (2008), p. 36): Opportunity set G is more generous than (MGT) opportunity set F if (a) $\pi^*_{iG} - \pi^*_{iF} \geq 0$ and (b) $\pi^*_{iG} - \pi^*_{iF} \geq \bar{\pi}^*_{-iG} - \bar{\pi}^*_{-iF}$. In this case, we say $G\ MGT\ F$.

where π^*_{iF} stands for the maximum feasible payoff of player i in opportunity set F and $\bar{\pi}^*_{-iF}$ stands for the maximum feasible *average* payoff of the other three group members in opportunity set F. Notice that we assume that player i will compare her earnings with the average earnings of the other three members, which is a slightly modified version of Definition 2 in Cox et al. (2008) that accounts for the fact that subjects in our experiment were confronted with possible averages contributions of the other group members and not with the entire vector of contributions.

Applying Definition 2 to *Maintenance* and *Provision* leads to our first proposition:

PROPOSITION 1. **(a)** *In Provision, an opportunity set generated by an average contribution* $\bar{c}^A_{-i}$ *of the other three group members is* more generous than *(MGT) an opportunity set generated by another average contribution* $\bar{c}^B_{-i}$ *if and only if* $\bar{c}^A_{-i} > \bar{c}^B_{-i}$. **(b)** *In Maintenance, an opportunity set generated by an average withdrawal* $\bar{w}^A_{-i}$ *is more generous than an opportunity set generated by another average withdrawal* $\bar{w}^B_{-i}$ *if and only if* $\bar{w}^A_{-i} < \bar{w}^B_{-i}$.

Proof. Applying Definition 2 to the *Provision* problem, we show that an opportunity set generated by an average contribution $\bar{c}^A_{-i}$ of the other three group members is more generous than (MGT) an opportunity set generated by another average contribution $\bar{c}^B_{-i}$ if $\bar{c}^A_{-i} > \bar{c}^B_{-i}$. Consider the payoff function of the *Provision* game:

$$\pi_i = 20 - c_i + 0.4 \sum_{j=1}^{4} c_j$$

In the strategy method experiment, as the average contribution of the other group members goes from $\bar{c}^B_{-i}$ to $\bar{c}^A_{-i}$, a subject gains $0.4 \times 3 \times (\bar{c}^A_{-i} - \bar{c}^B_{-i})$ on her maximum feasible payoff which satisfies condition (a) of Definition 2. Regarding condition (b), her gain on the maximum feasible payoff $\pi^*_{iG} - \pi^*_{iF}$ is always greater than $\bar{\pi}^*_{-iG} - \bar{\pi}^*_{-iF}$, as the first term is positive and the second term is negative if $\bar{c}^A_{-i} > \bar{c}^B_{-i}$.

Now consider the payoff function of the *Maintenance* problem:

$$\pi_i = w_i + 0.4\left(80 - \sum_{j=1}^{4} w_j\right)$$

Using $\bar{c}_{-i}^{A} = 20 - \bar{w}_{-i}^{A}$ and $\bar{c}_{-i}^{B} = 20 - \bar{w}_{-i}^{B}$, it follows that if $\bar{c}_{-i}^{A} > \bar{c}_{-i}^{B}$ then $\bar{w}_{-i}^{A} < \bar{w}_{-i}^{B}$. Given that the payoff functions of *Provision* and *Maintenance* are isomorphic, similar to above we can show that an opportunity set generated by an average withdrawal of $\bar{w}_{-i}^{A}$ is more generous than an opportunity set generated by another average withdrawal $\bar{w}_{-i}^{B}$ if $\bar{w}_{-i}^{A} < \bar{w}_{-i}^{B}$. Q.E.D.

Next, we apply Axiom R of Cox et al. (2008), p. 40, which specifies how differences in generosity across opportunity sets translate into differences in preferences.

Formally, **Axiom R** states that if $G, F \in \mathcal{C}$ and $G\ MGT\ F$, then $A_G\ MAT\ A_F$, where A_G indicates the preferences induced by opportunity set G and MAT stands for "more altruistic than".

Hence, for any two opportunity sets, G and F, if $G\ MGT\ F$, then the preferences induced by G are more altruistic than (*MAT*) the preferences induced by F. Applying Axiom R to our context leads to our second proposition:

PROPOSITION 2. **(a)** *In Provision, the preferences induced by an average contribution $\bar{c}_{-i}^{A}$ are more altruistic (MAT) than the preferences induced by $\bar{c}_{-i}^{B}$ if and only if $\bar{c}_{-i}^{A} > \bar{c}_{-i}^{B}$.* **(b)** *In Maintenance, the preferences induced by an average withdrawal $\bar{w}_{-i}^{A}$ are more altruistic (MAT) than the preferences induced by $\bar{w}_{-i}^{B}$ if and only if $\bar{w}_{-i}^{A} < \bar{w}_{-i}^{B}$.*

Proof. As shown before, Proposition 2 establishes that opportunity set $G_A\ MGT\ G_B$if and only if $\bar{c}_{-i}^{A} > \bar{c}_{-i}^{B}$ ($\bar{w}_{-i}^{A} < \bar{w}_{-i}^{B}$). Axiom R states that if $G_A\ MGT\ G_B$, then $A_A\ MAT\ A_B$, where A_A are the preferences induced by opportunity set G_A. It follows that $A_A\ MAT\ A_B$ if and only if $\bar{c}_{-i}^{A} > \bar{c}_{-i}^{B}$ ($\bar{w}_{-i}^{A} < \bar{w}_{-i}^{B}$). Q.E.D.

Finally, we derive the implications of Axiom S in Cox et al. (2008), p. 41. Assume $\mathcal{C}$ is composed by at least two opportunity sets, one of which is the status quo. Denote as A'_C the preferences induced by opportunity set C when C is the status quo and A_C when C is not the status quo. If $\mathcal{C}$ is singleton, preferences induced by the only feasible opportunity set C are indicated as A°_C. This is a case where the action is forced, i.e., there is no alternative than that particular opportunity set.

The first part of **Axiom S** (Cox et al. (2008)**,** p. 41) states that if $G\ MGT\ F$ and either G or F is the status quo, then $A_G\ MAT\ A'_G, A^{\circ}_G$ and $A'_F, A^{\circ}_F\ MAT\ A_F$.[5]

Hence, if there are two identical opportunity sets G and G' and the latter is the status quo, then the preferences induced by G are *more altruistic than* (*MAT*) the ones induced by G'; and that if there are two identical opportunity sets F and F' and the latter is the status quo,

[5] We only describe the first part of Axiom S for the ease of exposition. See Cox et al. (2008) for the complete Axiom. All our predictions depend on the first part of the Axiom only.

then the preferences induced by F' are more altruistic than the ones induced by F. Intuitively, Axiom S strengthens or weakens Axiom R, depending on whether the status quo opportunity set is more or less generous than the opportunity set under consideration. Applied to our context, we can derive the following proposition:

PROPOSITION 3. *Consider an average level of cooperation by the other group members of* $c_{-i} = 20 - w_{-i}$ *and the implied opportunity sets* G_c *for Provision and* G_w *for Maintenance. The preferences induced by* G_c *are (weakly) more altruistic than the ones induced by* G_w.

Proof. Consider an average contribution of $\bar{c}_{-i}^{A} > 0$ and an average withdrawal of $\bar{w}_{-i}^{A} < 20$, where $\bar{c}_{-i}^{A} = 20 - \bar{w}_{-i}^{A}$. We prove that the best response to an average $\bar{c}_{-i}^{A}$ is higher than the best response to an average of $\bar{w}_{-i}^{A}$ (see Cox et al. (2013) and Frackenpohl et al. (2016) for similar proofs).

In the Provision problem, consider $\bar{c}_{-i}^{A} > 0$ and another average contribution $\bar{c}_{-i}^{A*} = \bar{c}_{-i}^{A}$ that is generated by nature ($\mathcal{C}$ is singleton). Assume that c_i is the best reply to the latter average contribution. Axiom S implies the following ranking of the best replies (br):

$$br^{\bar{c}_{-i}^{A}} \geq c_i$$

We turn now to the Maintenance problem. Consider the average withdrawal $\bar{w}_{-i}^{A}$, where $\bar{c}_{-i}^{A} = 20 - \bar{w}_{-i}^{A}$. Consider further another average withdrawal $\bar{w}_{-i}^{A*} = \bar{w}_{-i}^{A}$ that is generated by nature ($\mathcal{C}$ is singleton) and w_i is the best reply to that average withdrawal. From Axiom S it follows that:

$$br^{\bar{w}_{-i}^{A}} \geq w_i$$

From Axiom R, it follows that the two best replies to the contribution and withdrawal generated by nature are isomorphic. Hence, we can express $w_i = 20 - c_i$ and combine the two inequalities above as follows:

$$br^{\bar{c}_{-i}^{A}} \geq c_i \geq 20 - br^{\bar{w}_{-i}^{A}}$$

The inequality proves that according to revealed altruism, best response for a given effective average contribution is (weakly) higher in *Provision* compared to *Maintenance*. Q.E.D.

The intuition for Proposition 3 is the following: in *Provision*, the status quo opportunity set (in which no tokens are yet contributed to the public good) is the least generous possible. Hence, any other opportunity set compared with the status quo will increase the effect of Axiom R. Conversely, in *Maintenance* the status quo opportunity set (in which no token is yet withdrawn from the public good) is the most generous possible. Hence, any other opportunity set compared with the status quo will decrease the effect of Axiom R. This implies that for the same effective average contribution of the other three group members, preferences will be (weakly) more altruistic in *Provision* than in *Maintenance*.

To further illustrate the effects of Axiom R and Axiom S, we introduce a utility function that represents preferences as in Cox et al. (2008). Consider the following utility function:

$$u_i = \pi_i - \gamma_i^F \max(\bar{c}_{-i} - c_i, 0),$$

where π_i is the material payoff for player i and γ_i^F represents player i's degree of reciprocity under $F \in (M, P)$ *Maintenance* or *Provision*. The effective contribution of player i is c_i and the average effective contribution of the other group members is denoted by $\bar{c}_{-i}$.

Taking the average contribution $\bar{c}_{-i}$ of the others as given (as it is the case in our strategy-method experiment), the best response of individual i depends on i's degree of reciprocity, γ_i^F. Given the parameters of our experiment, any token contributed to (not withdrawn from) the public good generates a material cost of $0.4 - 1 = -0.6$. Hence if $\gamma_i^F > 0.6$ the best reply of individual i will be to match others' contributions, while if $\gamma_i^F < 0.6$ her best reply is to free ride and contribute nothing for every possible average contribution of the others. In terms of two axioms described above, people with $\gamma_i^F > 0.6$ satisfy Axiom R. Hence, under this parametrization, depending on the distribution of the reciprocity parameter, γ_i^F, some people will be conditional cooperators, and some will be free riders.

Applying Axiom S to our parametrization implies that the reciprocity parameter γ_i^F is not the same across *Provision* and *Maintenance*, but that (on average) $\gamma_i^P \geq \gamma_i^M$. While for participants satisfying either $0.6 > \gamma_i^P \geq \gamma_i^M$ or $\gamma_i^M \geq \gamma_i^P > 0.6$, behavior is predicted to be identical across dilemma types, participants with $\gamma_i^P > 0.6 > \gamma_i^M$ are predicted to be conditional cooperators in *Provision* and free riders in *Maintenance*. Hence, this model is consistent with more conditional cooperators in *Provision* compared to *Maintenance*.

ONLINE APPENDIX D – Guilt Survey

As reported in Section 6 in the main text, to investigate whether guilt aversion can explain our findings of more conditional cooperators and fewer free riders in *Provision* than in *Maintenance*, we conducted an online survey. As discussed in Chang et al. (2011) and Bellemare et al. (2019), there are different approaches to test guilt aversion. Here, we rely on eliciting *ex post* feelings of guilt.[6] In particular, using an online questionnaire, we test whether being a free rider generates stronger feelings of guilt in *Provision* compared to *Maintenance.* Similar to the elicitation of kindness perceptions, we conducted two studies using two different subject pools. In the first study, we recruited $n = 347$ students from the University of Nottingham, while in the second study we recruited $n = 402$ participants via MTurk (none had participated in any of our experimental sessions before). In the questionnaire, we first explained participants either a *Maintenance* or a *Provision* dilemma. We then asked them on a scale from 0 to 100 (where 0 corresponds to "not guilty at all" and 100 corresponds to "very guilty") to assess how guilty they would feel if as a response to others' contributions (withdrawals) of 0, 10, or 20, they would contribute 0 (withdraw 20) tokens (for the exact wording, see Online Appendix A4).[7]

In Table D1 below we report estimates from two regression models, one for each sample, in which we regress the guilt score on a treatment dummy for *Maintenance,* others' average contributions, as well as an interaction term of the last two variables. As indicated by the significant coefficient for others' contributions, the results show that people feel more guilty about free riding the more others contribute. Furthermore, even when others contribute nothing, people feel somewhat guilty about free riding as indicated by the constant, which is significantly larger than zero.

[6] A potential alternative would have been to elicit second-order beliefs. However, given that we would need second-order beliefs for each cell of the strategy method, an incentive-compatible elicitation mechanism would require the elicitation of 21 contribution decisions, 21 first-order beliefs, and 21 second order beliefs. We believe that this procedure would have been too lengthy and too hard to understand for our participants. For this reason, we rely on the elicitation of *ex post* feelings of guilt. Research on guilt aversion in games also supports our approach. Chang et al. (2011) show that participants who match actions with second-order beliefs are in fact more likely to experience *ex post* feelings of guilt and conclude that the two measures point to the same psychological construct. Beranek et al. (2015) find that survey-based measures of guilt are positively correlated with advantageous inequality aversion, which can explain conditional cooperation. Bellemare et al. (2019) also find that a survey measure of guilt is correlated with game behavior.

[7] Like with the elicitation of kindness perceptions, answers were not incentivized because we elicited personal judgments. We did, however, incentivize participation. As before, student participants were offered three randomly drawn prizes of £50 each, and MTurk participants received a flat payment of $2.

Table D1: OLS regressions on feelings of guilt as a consequence of free riding

	(1) Lab	(2) MTurk
Maintenance *(1 if Maintenance, 0 otherwise)*	-4.961** (2.483)	-2.064 (2.698)
Others' average contribution	2.609*** (0.135)	2.403*** (0.149)
Maintenance × Others' average contribution	0.593*** (0.185)	-0.017 (0.215)
Constant	17.138*** (1.889)	20.390*** (1.999)
N	1041	1206
R^2	0.422	0.268

Note: Dependent variable: Guilt score. Others' average contribution = 0 is omitted category. Robust standard errors clustered on the individual level are in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Model (1) includes survey responses from 347 students and model (2) includes survey responses from 402 MTurkers.

Regarding potential differences across treatments, we find somewhat mixed results. For the MTurk sample, we find that the *Maintenance* dummy as well as the interaction term are both not significantly different from zero, indicating no difference in feelings of guilt between the two treatments. For the student sample, in contrast, we find a significantly lower intercept and a significantly steeper slope in *Maintenance* compared to *Provision*. This suggests that free-riding on others' effective contribution of zero triggers lower feelings of guilt in *Maintenance* than *Provision*, but that it generates more guilt in *Maintenance* than *Provision* when others contribute large amounts. Based on this, we conclude that if feelings of guilt would be the only driver behind the differences in cooperation preferences across the two dilemmas, we should observe *more* conditional cooperation in *Maintenance* than in *Provision*, which is the opposite of what we find. Overall, the results from our online survey suggest that ex post feelings of guilt are not a good predictor of the observed differences in cooperation preferences

Supplementary References

Battigalli, P., Dufwenberg, M., 2007. Guilt in games. The American Economic Review 97, 170-176.

Bellemare, C., Sebald, A., Suetens, S., 2019. Guilt aversion in economics and psychology. Journal of Economic Psychology 73, 52-59.

Beranek, B., Cubitt, R., Gächter, S., 2015. Stated and revealed inequality aversion in three subject pools. Journal of the Economic Science Association 1, 43-58.

Chang, Luke J., Smith, A., Dufwenberg, M., Sanfey, Alan G., 2011. Triangulating the Neural, Psychological, and Economic Bases of Guilt Aversion. Neuron 70, 560-572.

Cox, J. C., Friedman, D., Sadiraj, V., 2008. Revealed altruism. Econometrica 76, 31-69.

Cox, J. C., Ostrom, E., Sadiraj, V., Walker, J. M., 2013. Provision versus Appropriation in Symmetric and Asymmetric Social Dilemmas. Southern Economic Journal 79, 496-512.

Dufwenberg, M., Gächter, S., Hennig-Schmidt, H., 2011. The framing of games and the psychology of play. Games and Economic Behavior 73, 459-478.

Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games and Economic Behavior 47, 268-298.

Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior 54, 293-315.

Fehr, E., Schmidt, K. M., 1999. A theory of fairness, competition, and cooperation. Quarterly Journal of Economics 114, 817-868.

Fischbacher, U., Gächter, S., Quercia, S., 2012. The behavioral validity of the strategy method in public good experiments. Journal of Economic Psychology 33, 897-913.

Frackenpohl, G., Hillenbrand, A., Kube, S., 2016. Leadership effectiveness and institutional frames. Experimental Economics 19, 842-863.

Gächter, S., Kölle, F., Quercia, S., 2017. Reciprocity and the tragedies of maintaining and providing the commons. Nature human behaviour 1, 650.

Rabin, M., 1993. Incorporating fairness into game-theory and economics. American Economic Review 83, 1281-1302.

Volk, S., Thöni, C., Ruigrok, W., 2012. Temporal stability and psychological foundations of cooperation preferences. Journal of Economic Behavior & Organization 81, 664-676.