Cooperating Over Losses and Competing Over Gains: a Social Dilemma Experiment

Alessandro Ispano (THEMA - Universite de Cergy-Pontoise)
Peter Schwardmann (University of Munich)

Discussion Paper No. 23

March 23, 2017

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Dr. Klaus M. Schmidt, University of Munich, 80539 Munich, Germany
+49 (89) 2180 3405 | info@rationality-and-competition.de

Cooperating over losses and competing over gains: a social dilemma experiment[☆]

Alessandro Ispano[a,*], Peter Schwardmann[b,*]

[a] *THEMA - Université de Cergy-Pontoise, 33 boulevard du Port, 95011 Cergy-Pontoise, France*
[b] *Department of Economics, University of Munich, Ludwigstr. 28, D-80539 Munich, Germany*

Abstract

Evidence from studies in international relations, the politics of reform, collective action and price competition suggests that economic agents in social dilemma situations cooperate more to avoid losses than in the pursuit of gains. To test whether the prospect of losses can induce cooperation, we let experimental subjects play the traveler's dilemma in the gain and loss domain. Subjects cooperate substantially more over losses. Furthermore, our results suggest that this treatment effect is best explained by reference-dependent risk preferences and reference-dependent strategic sophistication. We discuss the implications of our results and relate our findings to other experimental games played in the loss domain.

Keywords: traveler's dilemma, loss domain, diminishing sensitivity, strategic sophistication
JEL classification: C90, D01, D03, D81

March 2017

1. Introduction

In many contexts it is efficient for two or more agents to cooperate, yet the cooperative outcome leaves each agent with an incentive to deviate at the expense of the others. How cooperation can be sustained in such social dilemmas is the focus of several literatures across the social sciences. But one empirical observation has thus far evaded a convincing unifying explanation: cooperation appears to be more likely when agents face losses. For example, there is evidence that governments reach international agreements more readily when losses are at stake, that reforms, which require the cooperation of several stakeholders, are often spurred by crises, and that price competition between firms is on average less severe during economic downturns (see section 2.1 for a more detailed discussion of our motivating examples).

[☆]We thank Giuseppe Attanasi, Astrid Hopfensitz, Hannah Braun, Antonio Cabrales, Pascal Lavergne, Sébastien Pouget, Paul Seabright, Frances Spies, Johannes Spinnewijn, Klaus Schmidt and Jean Tirole as well as seminar participants at Toulouse School of Economics, Paris School of Economics, CREST, ECORE, LMU Munich, the ESA World Meeting 2013, Journees de Microeconomie Appliquee 2014, the Annual Meeting of the AFSE 2014 and the European Meeting of the Econometric Society 2014 for useful comments. We received funding from the grant ANR: 2010 JCJC 1803 01 TIES. Financial support by Deutsche Forschungsgemeinschaft through CRC TRR 190 is also gratefully acknowledged.
[*]*Email addresses*: alessandro.ispano@gmail.com and peter.schwardmann@econ.lmu.de

It is difficult to establish the causal effect of losses on cooperative behavior in the field, where confounds cannot be ruled out and where it is hard to know for certain whether a given outcome truly places an agent in the loss domain. We therefore conduct a controlled laboratory experiment in which we let subjects play a once-off, anonymous social dilemma and vary exogenously whether payoffs are framed as gains or as losses. The experiment is also designed to investigate the mechanism by which the prospect of losses impacts on cooperation.[1]

Our experimental game is the traveler's dilemma (henceforth TD),[2] in which two players simultaneously submit claims that may take any value between a lower bound and an upper bound (Basu, 1994). Both players then receive the lower of the two submitted claims and a reward of size R is paid to the player making the lower claim, while a penalty of size R is deducted from the payoff of the player making the higher claim. Each player in the TD has an incentive to minimally undercut the other and in the unique Nash equilibrium, both players are thus maximally uncooperative and claim the lower bound of the action space.

We implement two main treatments. In the *gain treatment* admissible claims lie between 3 and 8 euros. In the *loss treatment* subjects are given 11 euros in cash for their participation at the beginning of the experiment and then stand to lose money, with admissible claims ranging from -8 to -3 euros. In both treatments the reward/penalty parameter R takes a value of 3 euros and subjects play the game repeatedly, being anonymously matched with a new partner each time. The loss and gain treatments differ only in how payoffs are framed[3] and would be equivalent if subjects were expected utility maximizers or if our framing did not have an impact on subjects' reference points or strategic sophistication.

We find that behavior is substantially more cooperative in the loss treatment. In particular, in early periods of the experiment average claims in the loss treatment are up to 1.42 euros or 42 percent higher than in the gain treatment. As learning sets in and strategic uncertainty is reduced, claims eventually converge to Nash play in both treatments. This suggests that higher cooperation in the loss domain is likely to arise mostly in surprising, complicated or uncertain settings, like those that characterize our motivating examples.

We consider three plausible drivers of our treatment effect: gain-loss differences in risk preferences, in social preferences and in strategic sophistication, i.e. subjects propensity to make mistakes in strategic reasoning. If any such gain-loss difference in preferences or sophistication is driving our treatment effect, then the variable in question should exhibit a treatment effect when elicited in an individual decision-making task and it should correlate with behavior in the traveler's dilemma.

We find that risk tolerance is higher and that strategic sophistication is lower in the loss domain. The former effect is implied by prospect theory's diminishing sensitivity (Kahneman and Tversky,

[1] We define "cooperation" as the act of working together toward social efficiency. According to our definition, this behavior need not reflect deeper cooperative or prosocial motives.

[2] In section 3, we provide several arguments for why we believe the TD, rather than other commonly studied social dilemmas, to be the ideal research vehicle for the question at hand.

[3] Suppose, for example, player 1 chooses the fully cooperative action, which is 8 in the gain treatment and -3 in the loss treatment, and player 2 undercuts her by 0.1. Then, in both the loss and the gain treatment, player 1 leaves the experiment with 4.9 euros and player 2 with 10.9 euros.

1979). In the first test of gain-loss asymmetries in social preferences we are aware of, we find that dictator game giving is *not* responsive to the gain-loss frame. Furthermore, a higher risk tolerance and less strategic sophistication, but not dictator game giving, are associated with higher claims in the TD for the average subject.

This paper makes two main contributions. First, it corroborates suggestive evidence from the field that the prospect of losses causes economic agents to cooperate, while ruling out explanations that do not rely on the gain-loss frame. Second, it uncovers that this treatment effect is driven by greater risk tolerance and lower strategic sophistication when losses are at stake. In section 5 we discuss these drivers in more detail and point to additional evidence for their relevance. In particular, our proposed mechanism is consistent with a stylized model of game play in the traveler's dilemma under strategic uncertainty, estimates from a structural cognitive hierarchy model, further experimental treatments and suggestive evidence from a questionnaire after the experiment.

In the following section we describe motivating examples from the field and relate our paper to the broader experimental literature. In section 3, we describe our experimental design and in section 4 our results. Section 5 discusses the drivers of our treatment effect before section 6 concludes and highlights some implications of our results.

2. Related evidence from the field and lab

2.1. Motivating examples

A number of studies in international relations find that the threat of losses, more than the promise of gains, induces national governments to cooperate on multilateral economic surveillance and military strategy (Stein and Pauly, 1992; Mercer, 2005). For example, it has been argued that the threat of losses motivated the international cooperation of setting up the Bretton Woods agreement (Pauly, 1992); lead to the Structural Impediments Initiative of 1989-1990, which saw Japan and the United States agree to costly domestic reforms in order to facilitate a better trading relationship (Mastanduno, 1992); and was the foundation of Israel's cooperation with the US during the Gulf war (Welch, 1992).

Within countries, reforms are often spurred by crises (Weyland, 2002; Vis, 2009). Where inefficient policies are the non-cooperative outcome of a game between political stakeholders, crises can bring about a switch to more cooperative policymaking (Velasco, 1998; Tommasi, 2004). Being put in the loss domain by a crisis appears to equip policy makers, the electorate and other stakeholders with the risk tolerance required to pursue risky reforms and cooperation (Weyland, 1996; Tommasi, 2004; Vis and Van Kersbergen, 2007).

Collective action to revolt seems puzzling in light of individuals' incentives to freeride or defect (Coleman, 1994; Moore, 1995). According to Berejikian (1992) and Fanis (2004), it is the threat of losses that helps individuals overcome the freeriding problem and engage in a revolt.

In the context of organized labor, the credible threat of a strike, which requires a workforce to cooperate, in response to wage cuts may help explain the downward stickiness of wages. Kahn (1997) documents workers' and firms' resistance to nominal pay cuts and interprets it as evidence that current wage levels constitute an important reference point. Bewley (2002) provides additional evidence and various different explanations for the downward rigidity of wages.

Price competition between firms is less fierce during economic downturns (Rotemberg and Saloner, 1986). Rojas (2012) and Ruffle (2013) show this effect in experimental games. Rotemberg and Saloner (1986) explain it as the outcome of a repeated game, in which low demand decreases the temptation to cheat. But since Bertrand competition is a social dilemma, our results suggest an alternative mechanism. Bad demand conditions may lead to more cooperative or collusive outcomes because they put managers in the loss domain.

2.2. Related experimental literature

Our social dilemma game is maximally simply. It does not have an equilibrium in dominant strategies. And we take great care to credibly embed our subjects in the loss domain. Related experimental papers differ from our paper along one or more of these dimensions.

Some studies in experimental economics compare subjects' propensity to contribute to public goods to their propensity to take from common resources (see Fosgaard et al. (2014) or Cox (2015) for a list of papers in this literature).[4] In these experiments, strategically equivalent public good games are presented to subjects under either a giving frame (public good) or a taking frame (common resource). The latter may conceivably also evoke a loss frame if subjects integrate the common resource into their wealth.[5] Evidence is mixed, with several studies finding higher contributions under the giving frame and several other studies, including a large sample experiment by Fosgaard et al. (2014), finding the opposite.

Our paper differs along two crucial dimensions from these experiments and we do not take a strong stand on their findings. First, in line with our motivating examples and contrary to these experiments, we attempt to truly put the subjects in our loss treatment in the loss domain, by paying out physical currency before the experiment that has to be returned to the experimenter in case of losses. Second, the linear public good games typically featured in the above experiments have a dominant strategy. To capture the inherent riskiness of our motivating examples, our game does not. If agents are self-interested, a dominant strategy implies that risk preferences should not impact on behavior.

Iturbe-Ormaetxe et al. (2011) run an experiment that features a public good game without a dominant strategy. In a control condition, a subject gains a prize g if at least k subjects (possibly including herself) contribute c from their endowment. In the treatment, a subject can avoid a loss of g if at least k subjects contribute and she gains c if she does not contribute. Endowments are identical across treatments. Our experiment differs in two crucial ways. First, our loss treatment is deeply embedded in the loss domain and we therefore can not interpret our results through the lens of loss aversion. Second, unlike in our experiment, the two treatments in Iturbe-Ormaetxe et al. (2011) would *not* be identical if subjects were expected utility maximizers. Nonetheless, if we interpret their treatment as inducing a loss frame, their finding that contributions are higher in the treatment than in the control group when k is equal to the group size is consistent with our results.[6]

[4] In the field, Ostrom (1990) argues that groups frequently overcome individual incentives to exploit open access common pool resources such as forests and fisheries, especially when appropriators are very dependent on the resource and its risk of depletion is high (Ostrom, 2000).

[5] See Cox (2015) for a discussion of how take/give and gain/loss frames may be confounded in some of these studies.

[6] When the whole group needs to contribute for the public good to be created, contributing is risky in that it yields

Some experiments in psychology compare prisoner's dilemma play in the gain and loss domain, but no clear treatment effect emerges (see De Dreu and McCusker (1997) and De Heus et al. (2010) for reviews of the evidence). Similar to the linear public good game, the prisoner's dilemma has a dominant strategy and risk preferences are irrelevant if a player is self-interested. Interestingly, however, De Dreu and McCusker (1997) find that individuals they characterize as collaborators do in fact cooperate more in the loss domain. This is in line with our findings. An individual with sufficiently strong social preferences like inequity aversion, for example, obtains the highest utility from both parties cooperating, while cooperating and being defected upon constitutes the worst-case scenario. When there is strategic uncertainty, defecting is therefore not a dominant strategy for collaborators and risk preferences matter in determining their favored action much like they do in the TD. This is also consistent with evidence in Mengel (2014), who finds that the downside risk of cooperating is the most important driver of cooperation in prisoner's dilemmas with random matching.

In a two-player game of chicken, player 1's preference ordering over outcomes is given by (defect, cooperate) $\succ$ (cooperate, cooperate) $\succ$ (cooperate, defect) $\succ$ (defect, defect). De Heus et al. (2010) find that individuals cooperate less when they play the game of chicken in the loss domain. They argue that the payoff structure of the game implies that defecting is the high-potential-reward and high-variance action and that subjects defect more frequently in the loss domain because of diminishing sensitivity. Therefore, their findings lend support to our risk-preference based explanation and highlight the importance of a game's exact payoff structure in shaping gain-loss differences in strategic behavior. Their version of the chicken game is not a social dilemma, but even if it were we find that the TD's assumption that (defect, defect) $\succ$ (cooperate, defect) better fits our motivating examples. Moreover, their study ignores strategic sophistication and does not elicit preferences in individual decision making tasks.

The TD can also be understood as capturing a setting of imperfect price competition between two firms with differentiated products or capacity constraints.[7] The effect of the loss treatment then captures the anticompetitive effects of sunk costs or bad demand conditions. Offerman and Potters (2006) find that auctioning off entry fees or imposing fixed sunk costs in a Bertrand oligopoly increases collusion among experimental entrants, Kachelmeier (1996) finds that sunk costs have no effect in a double auction and Buchheit and Feltovich (2011) find that experimental market prices are first increasing and then decreasing in sunk costs. Our experiment differs in that it is simpler and explicitly designed to induce losses and uncover the mechanism behind our treatment effect.

Previous experiments featuring the TD show that for low values of R relative to the upper bound, claims are clustered around the highest possible claim (Capra et al., 1999; Goeree and Holt, 2001; Becker et al., 2005; Rubinstein, 2007), a result we replicate in online appendix E. As the reward/penalty parameter grows larger, however, claims converge to the Nash equilibrium play (Capra et al., 1999; Goeree and Holt, 2001). This comparative static in R is well explained by models of noisy decision

$g - c$ if everybody else is contributing and $-c$ if at least one person fails to contribute. Not contributing, on the other hand, is safe in that it yields zero payoffs with certainty.

[7] See Capra et al. (2002) for an experiment in which firms engage in Bertrand competition and the firm setting the higher price has a non-vanishing market share. In this case, the size of the residual market is the counterpart of the reward/penalty parameter in the TD.

making (Capra et al., 1999) and models of strategic uncertainty (Baghestanian, 2014), suggesting that, as we assume, uncertainty plays a key role in driving behavior in the TD.

We explain our treatment effect by asserting that subjects are less risk averse and less sophisticated in the loss domain. Diminishing sensitivity (Kahneman and Tversky, 1979), has been documented in the decision making of experimental subjects and the general population (Booij et al., 2010; Tymula et al., 2012). In a review of studies that estimate the curvature of individuals' utility functions in the gain and loss domain 9 out of 11 studies find risk loving preferences in the loss domain *and* risk aversion in the gain domain (Booij et al., 2010). Camerer (2003) provides some examples from the field. However, our explanation is based on a weaker condition than diminishing sensitivity: we merely require that individuals are less risk averse in the loss domain.

In line with our result on subjects' sophistication, Tymula et al. (2012) find that in a sample of over 8000 individual choices, stochastic dominance violations in choices between simple lotteries and certain payoffs are more likely in the loss domain. In line with our result that dictator game giving is not correlated with actions in the TD, Brañas-Garza et al. (2011) document that pro-social considerations are absent from a subjects' ex-post explanation of the action they chose in a TD experiment.

3. Design

We let subjects in a computerized experiment play the traveler's dilemma, a two-player game in which a player's payoff, if she claims x_i and her opponent claims x_j, is given by

$$\pi_i = \begin{cases} x_i + R & \text{if} \quad x_i < x_j \\ x_j - R & \text{if} \quad x_i > x_j \\ x_i & \text{if} \quad x_i = x_j \end{cases} .$$

The TD captures the basic trade-off between what is privately and socially optimal that characterizes our motivating examples. Yet its simple payoff structure makes it easy to explain to experimental subjects. Compared to its close relative, the prisoner's dilemma, it delivers a rich distribution of actions.[8] The TD is preferable to linear public good games because these lack the uncertainty about the marginal private returns to cooperation that characterizes the real-world social dilemmas we seek to capture. At the same time, it is analytically simpler than step-level public good games that contain said uncertainty, but often feature multiple equilibria.

Subjects were randomly assigned to sessions belonging to either a *gain* or a *loss* treatment. In the gain treatment, subjects received no participation fee and admissible claims lay between 3 and 8 euros. In the loss treatment, they received a participation fee of 11 euros before the experiment and admissible claims lay between -8 and -3 euros. Subjects then had to reimburse the experimenter for any losses they incurred. R was set equal to 3 euros in both treatments.

In a first set of sessions, which we call *experiment 1*, subjects played the TD five times[9] and the action set was finely grained: any multiple of 0.1 euros between the lower and upper bound of the

[8] This facilitates estimating the structural model we use to get at the mechanism that drives our treatment effect in appendix B.

[9] They then played five more surprise periods that serve as a consistency check that we discuss in online appendix E.

claims range was admissible. In a second set of sessions, which we call *experiment 2*, we let subjects play the TD ten times and elicited risk preferences, social preferences and propensity to make mistakes in strategic reasoning. In experiment 2, we restricted claims to be multiples of 0.5.[10] Experiment 1 was designed to investigate the treatment effect in a clean fashion, without contamination or excessive cognitive load from other tasks. Experiment 2 serves to investigate the drivers of the treatment effect we observe and provides a view of how play evolves in later periods.

In both experiments, we paired each subject with a different person in each period to avoid dynamic strategic considerations and we used a randomly selected period to determine final earnings to avoid wealth effects. In experiment 2, periods that featured preference elicitation tasks were also eligible to be selected for payment. The TD was introduced to subjects in its abstract form (see the instructions in online appendix F).

The experiments were programmed in z-Tree (Fischbacher, 2007). We conducted a total of twelve experimental sessions at the Toulouse School of Economics experimental laboratory with student subjects. Experiment 1 featured two sessions under the gain treatment (with 34 subjects), two under the loss treatment (32 subjects), and two under the gain-loss treatment (34 subjects) that we describe in appendix D. Sessions lasted less than 30 minutes and subjects earned on average 6.30 euros. Experiment 2 featured three sessions under the gain treatment (36 subjects) and three under the loss treatment (42 subjects). Sessions lasted 45 minutes and subjects earned on average 9.80 euros.

Table 1 depicts the design of experiment 2. The dashed arrows indicate that we switched the timing of risk and social preference elicitation in half of the sessions, in order to detect potential order effects. Payoffs from preference elicitation periods were only revealed at the end of part A. Part B of the experiment came as a surprise and subjects were paid for it separately. Part C consisted of a questionnaire on the motives behind subjects' choices in the TD (see tables C.4 and C.5 in appendix C).

Risk preferences. We elicited risk preferences by letting subjects choose between different lotteries, as in Holt and Laury (2002) (see table C.1 in appendix C). Outcomes of lotteries in the loss treatment were 11 euros less than those in the gain treatment and thereby fully embedded in the loss domain. The *HL switching point* refers to the least favorable pair of lotteries for which a subject prefers the riskier one. It is increasing in risk aversion. As a second non-strategic risk preferences elicitation task, we let subjects play the TD of the corresponding treatment against the computer, which was programmed to choose each claim with equal probability. The more risk averse a subject is, the lower her *risk TD* claim should be.

Social preferences. To elicit social preferences, we paired each subject with a randomly selected anonymous partner.[11] In a dictator game (see table C.2 in appendix C), dictators in the gain treatment had

[10] Experiment 2 therefore also allowed us to check whether our treatment effect replicates in a setting with a coarser action set.

[11] In order to provide proper incentives without losing observations, each subject performed all social preference elicitation tasks. Then, with equal probability, either her choice or that of her partner determined earnings for the period.

Gain treatment	**Loss treatment**
Part A:	Part A:
• no fee	• 11 € fee
• elicitation of risk preferences over gains	• elicitation of risk preferences over losses
• 10 periods of the TD over gains	• 10 periods of the TD over losses
• elicitation of social preferences over gains	• elicitation of social preferences over losses
Part B:	Part B:
• 3 € fee	• 9 € fee
• quiz on the TD with a reward for correct answers	• quiz on the TD with a penalty for wrong answers
Part C: questionnaire	Part C: questionnaire

Table 1 The sequencing of tasks in experiment 2

to choose how to split 8 euros with their partners, while dictators in the loss treatment had to choose how to allocate a loss of 8 euros. The more prosocial a subject is, the higher her *dictator giving* should be, i.e. the higher the dictator game payoff she allocates to her partner. As in the risk preference elicitation, each subject also played the TD of her corresponding treatment against a computer that randomly chose claims. This time, however, the computer's payoff went to the subject's partner. The difference between choices in this task, which we denote by *social TD* claim, and the risk TD claim provides another measure of prosociality.

Mistakes and strategic sophistication. In part B of experiment 2 we elicited subjects' strategic sophistication by presenting them with a quiz featuring 8 questions of varying degrees of difficulty, ranging from simple questions on the payoff structure to more advanced questions on optimal actions conditional on an opponent's behavior. Hypothetical payoffs and actions in the quiz were framed according to the treatment the subject was in. Furthermore, subjects in the gain treatment earned 0.5 euros for a correct answer, while subjects in the loss treatment lost 0.5 euros for a wrong answer. The quiz thus captured the effect of incentivizing subjects through losses as well as the difficulty subjects may experience in thinking in terms of losses. We also elicited subjects' beliefs about their own and others' performance in the quiz, while providing incentives for accuracy as high as 2 euros. The initial fees for part B were 3 and 9 euros in the gain and loss treatment respectively, such that a given number of correct answers yields identical earnings in the two treatments.

4. Results

4.1. The treatment effect on strategic behavior

Figure 1 depicts the gain and loss treatments' distributions of claims, pooling all 5 periods of experiment 1. For the sake of comparability, we present results in terms of *net claims*, which obtain by adding the participation fee of the corresponding treatment to actual claims.[12] The distribution

[12]That is, we translate claims in the loss treatment to claims into the action space of the gain treatment by adding 11 euros.

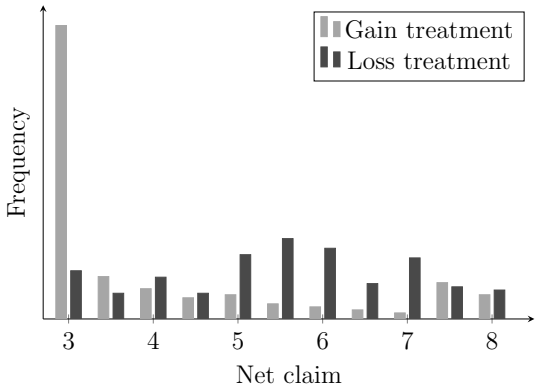


Figure 1 Distribution of claims in experiment 1

in the gain treatment has most of its mass at the Nash equilibrium of 3. The distribution in the loss treatment is more dispersed, with its modal claim in the center of the action set.

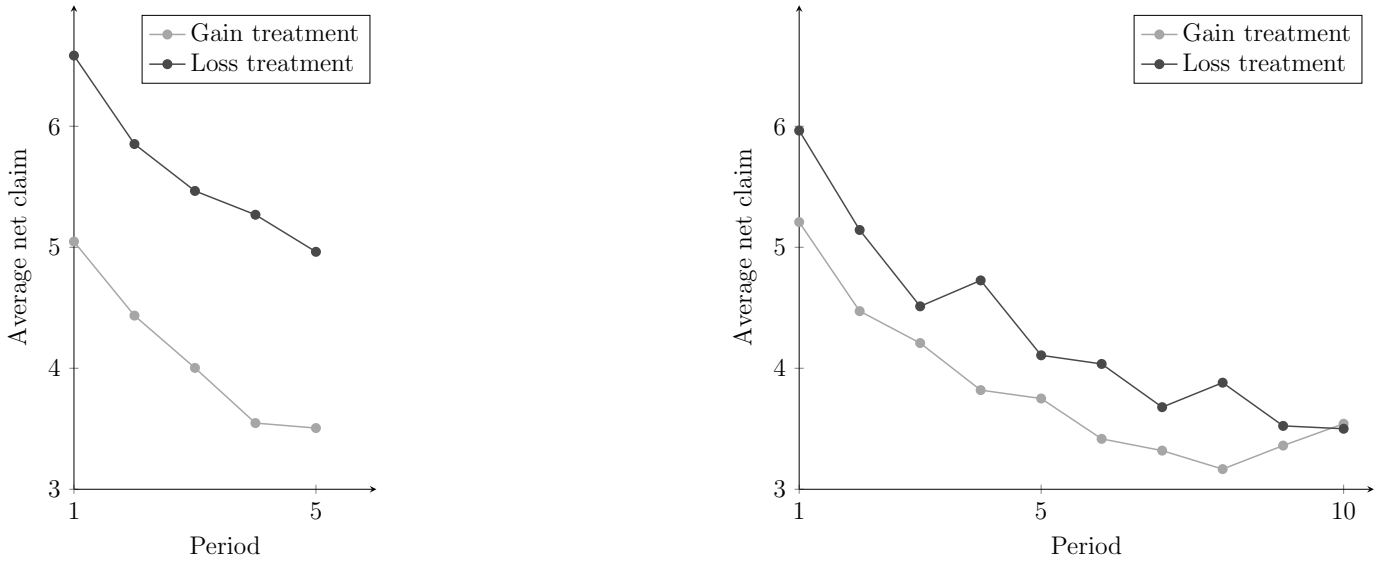
Figure 2a shows average net claims across periods in experiment 1. In each period, average claims are higher in the loss treatment than in the gain treatment. In period 5, the average claim is 3.51 euros in the gain treatment and 4.96 euros in the loss treatment. Framing payoffs as losses therefore increases cooperation by up to 41 percent. As depicted in figure 2b, average net claims in experiment 2 are also higher in the loss treatment, except for in the last period. Statistically, we can test for differences in claims between treatments by comparing randomly drawn claims from either treatment (using a Wilcoxon rank-sum test), by comparing means (using a two-sided t-test), or by comparing the frequency of Nash play (using a two-sided Fischer's exact test). In both experiments these three comparisons tell us that claims in the loss treatment are significantly higher.[13]

Note that the treatment effect on strategic behavior is smaller in experiment 2 than in experiment 1. This may be due to experiment 2's coarser action set.[14] Alternatively, it may be due to the preference elicitation tasks that preceded the 10 periods of TD play and may have influenced a subject's expectation about her opponent's behavior.

Experiment 2 allows us to observe TD play after 5 periods. In the final periods of the experiment we observe convergence to Nash equilibrium play in both treatments. This may be the result of a learning

[13]Pooling data across periods of experiment 1, we find that a randomly selected claim from the loss treatment is more likely than not to exceed a randomly selected claim from the gain treatment ($p = 0.00$, Wilcoxon rank-sum test), that average claims are higher in the loss treatment ($p = 0.00$, t-test), and that the frequency of Nash play is higher in the gain treatment ($p = 0.00$, Fischer's exact test). Significance at the 1 or 5 percent level also obtains for these three tests when looking at every period individually. Pooling data across periods for experiment 2, we again find that a randomly selected claim from the loss treatment is more likely than not to exceed a randomly selected claim from the gain treatment ($p = 0.00$, Wilcoxon rank-sum test), that average claims are higher in the loss treatment ($p = 0.00$, t-test), and that the frequency of Nash play is higher in the gain treatment ($p = 0.00$, Fischer's exact test). Per period comparisons for experiment 2 show that differences between the treatments are mostly significant, albeit only marginally in some and not at all in the last two periods. Consider the following per period tests (Wilcoxon rank-sum test, t-test, Fischer's exact test): period 1 ($p = 0.041, p = 0.031, p = 0.010$); period 2 ($p = 0.012, p = 0.027, p = 0.005$); period 3 ($p = 0.079, p = 0.313, p = 0.069$); period 4 ($p = 0.000, p = 0.001, p = 0.001$); period 5 ($p = 0.027, p = 0.141, p = 0.020$); period 6 ($p = 0.002, p = 0.013, p = 0.012$); period 7 ($p = 0.028, p = 0.078, p = 0.064$); period 8 ($p = 0.000, p = 0.003, p = 0.000$); period 9 ($p = 0.426, p = 0.459, p = 0.449$); period 10 ($p = 0.743, p = 0.874, p = 0.801$).

[14]Consider the logic of the cognitive hierarchy model in appendix B, in which subjects with higher levels of sophistication undercut less sophisticated types. Of course, a subject who plans to minimally undercut her opponent in experiment 2, mechanically has to choose a lower claim than a subject in experiment 1.



(a) Experiment 1 **(b)** Experiment 2

Figure 2 Average claims per period

process by which subjects become better at predicting their opponent's action or simply modify their behavior on the basis of negative and positive reinforcement. Comparing earlier and later periods demonstrates that our treatment effect and explanation are likely to be more pertinent in noisy and uncertain settings. Arguably, most real world social dilemmas take place in precisely such environments.

4.2. Treatment effects on preferences and mistakes

The individual decision making tasks in experiment 2 can help us identify a plausible mechanism underlying our treatment effect. For either risk preferences, social preferences or sophistication to constitute a valid explanation of our treatment effect, the variable in question, when measured directly, should respond to the loss treatment in the expected way.

Table 2 tells us whether preferences and the propensity to make mistakes exhibit treatment effects. The first four rows of the table feature the switching points in a Holt and Laury (2002) lottery choice list. Over the full sample, a t-test indicates that individuals are significantly more risk averse in the gain domain. The pattern persists in row 2, where we only consider subjects for whom risk preferences were elicited before the TD was played. Rows 3 and 4 exclude those subjects that did not switch within the ten options of the HL lottery, thereby revealing themselves to irrationally preferring a sure 6 euros to a sure 10.5 euros. Subjects again exhibit more risk aversion in the gain domain, although this difference falls just short of being significant at the 10 percent level when we only consider risk preferences elicited in the first period. But since HL switching points do not exhibit order effects (see table C.3 in appendix C), we can safely consider the whole sample.

Rows 5 and 6 of table 2 provide the first test of the impact of gain/loss frames on social preferences we are aware of: subjects playing a dictator game do not give significantly more in the loss domain compared to the gain domain.

Row 7 indicates that subjects do significantly better in a quiz about optimal play in the TD, when they are paid with gains rather than penalized by losses and when payoffs in the quiz are framed as

Table 2 Elicited preferences and propensity to make mistakes

		Gains		Losses		
	Sample	Mean	N	Mean	N	Difference
HL switching point	full	6.83	36	5.95	42	0.88**
	first round	6.72	18	5.62	21	1.10*
HL switching point (<11)	full	6.31	32	5.56	39	0.75**
	first round	6.19	16	5.35	20	0.84
Dictator giving	full	3.14	36	3.21	42	-0.07
	first round	3.17	18	3.62	21	-0.45
Quiz score	full	5.89	36	4.21	42	1.67***
Risk TD claim	full	4.94	36	5.57	42	-0.62**
	first round	5.00	18	6.05	21	-1.05***
Social TD claim	full	4.72	36	5.37	42	-0.65**
	first round	4.81	18	6.00	21	-1.19***
Social TD − Risk TD	full	-0.22	36	-0.20	42	-0.02
	first round	-0.19	18	-0.05	21	-0.18

Note: The last column features t-tests on the difference in means between the gain and the loss treatment: * ($p < 0.1$), ** ($p < 0.05$), *** ($p < 0.01$).

gains rather than losses.

When we transform the TD into an individual decision making task in which subjects play against a computer who picks a claim at random, subjects make considerably more risk averse choices in the gain domain. This is reflected in the gain treatment's lower Risk TD claim. Adding a social component by giving the computer player's payoffs to another subject does not significantly increase the difference in behavior between the gain and loss domain. This is made precise in the last two rows of table 2.

To summarize, there is a treatment effect on risk preferences and the propensity to make mistakes, as captured by the quiz, but not on social preferences. In section 5 we discuss how greater risk tolerance and a greater propensity to make mistakes could plausibly map into higher claims.

4.3. Do elicited preferences and propensity to make mistakes correlate with behavior in the traveler's dilemma?

Further suggestive evidence for the mechanism can be gleaned from the correlations between individually elicited subject characteristics and their strategic behavior. In particular, we expect any driver of our treatment effect to actually correlate with strategic behavior.

Table 3 lists the results of eight OLS regressions of net claims in the TD on various independent variables. Regressions 1 and 2 imply that the correlation between net claims in the TD and choices in the simple non-strategic TD is higher than the correlation between strategic play and the non-strategic TD that features a social motive.[15] The absolute value of the coefficient in regression 1 may not be informative, because, by design, the non-strategic TD and the strategic TD are very similar and hence,

[15]Because Risk TD claim and Social TD claim both exhibit order effects (see table C.3 in appendix C), regressions 1 and 2 are run over those subjects whose respective risk and social preferences were elicited before the strategic TD. Since the intersection of these two samples is zero, a valid regression featuring both Risk TD claim and Social TD claim cannot be run.

Table 3 Determinants of net claims

	1	2	3	4	5	6	7	8
Risk TD claim	0.794***							
	(0.10)							
Social TD claim		0.588***						
		(0.13)						
HL switching point			-0.183*			-0.145	-0.073*	-0.008
			(0.10)			(0.10)	(0.04)	(0.03)
Dictator giving				-0.067		-0.109	0.008	0.012
				(0.10)		(0.09)	(0.04)	(0.03)
Quiz score					-0.165**	-0.165**	-0.095***	-0.094***
					(0.07)	(0.07)	(0.03)	(0.02)
Period(s) of TD play	1	1	1	1	1	1	1-5	6-10
N	39	39	71	71	71	71	355	355
R-squared	0.473	0.321	0.048	0.007	0.070	0.119	0.042	0.052

Note: OLS regression with robust standard errors in parenthesis; significance levels of coefficients: * ($p < 0.1$), ** ($p < 0.05$), *** ($p < 0.01$); Observations with HL switching point = 11 are not included in the regressions.

drivers of behavior unrelated to risk preferences, like focal points in the action space or the description of the game, could be at work in both. But the relative difference between the coefficients in regression 1 and regression 2 tells us that adding a social motive to the independent variable makes it less predictive of strategic play. This speaks against an important social motive in the strategic behavior.

Regressions 3 through 5 indicate that a subject's HL switching point and her performance in the quiz, but not dictator giving, are correlated with claims in the first period of TD play. Since these independent variables do not exhibit significant order effects, we can run a regression over the entire sample of subjects and include measures of risk preferences, social preferences and sophistication. In regression 6, the quiz score is the only significant independent variable. The HL switching points falls just short of the 10 percent level.

Regression 7 pools observations from the first 5 periods of TD play: risk preferences and sophistication exhibit the expected correlation with strategic behavior. Regression 8 pools observations form the last 5 periods, after some learning would have taken place. In these periods, it is mainly subjects that did not fully understand the game (as captured by a low quiz score) that still deviate from Nash play, while the median subject now plays the lowest action.

5. Discussion

Our experimental results suggest that greater risk tolerance and less strategic sophistication in the loss domain are at the core of our treatment effect on strategic behavior. In this section, we provide further context for this assertion and discuss several appendices that speak to our proposed mechanism.

To demonstrate that greater risk tolerance and less strategic sophistication are plausible drivers of higher claims in the traveler's dilemma, it is useful to study analytically how best to play the TD in the presence of strategic uncertainty. In appendix A, we provide a simple theoretical example in which we

model strategic uncertainty as stemming from an unsophisticated opponent who sometimes deviates from Nash play.[16] We show that our treatment effect on strategic behavior is consistent with subjects exhibiting either more risk tolerance, more altruism or a higher propensity to make mistakes in the loss domain.

To see why risk preferences matter, note that strategic uncertainty implies that moving away from the least cooperative action of our experimental game may yield a higher expected payoff because it allows a player to reap the benefits of a higher claim, while still undercutting her unsophisticated opponent some of the time. But cooperation also exposes a player to the downside risk of being undercut. More risk tolerance enables a subject to embrace this gamble that cooperation entails.

A lack of strategic sophistication, or a greater propensity to make mistakes, mechanically increases the claims of a "mistaken" player, given that the counterfactual is Nash play at the least cooperative action. Sophisticated payers will respond to this by also raising their claims, since higher claims are now associated with a lower risk of being undercut than before.

In appendix B, we enrich our simple analytical example and consider the interaction between players of several different levels of sophistication in a cognitive hierarchy model (Camerer et al., 2004). We augment this model to allow for, respectively, curvature in players' utility function and other-regarding preferences.

The added complexity of a cognitive hierarchy model entails a loss of analytical tractability. Nonetheless, we can estimate the model's parameters structurally from the experimental data. This allows us to ascertain whether our suggested mechanism organizes the data well. We therefore use the cognitive hierarchy not as an intuition pump but as an empirical test of the plausibility of our proposed mechanism.[17] The structural estimates indicate that two parametrizations of the cognitive hierarchy model best rationalize the experimental data: 1) more risk tolerance and 2) less sophistication in the loss domain or $\hat{1}$) more altruism and $\hat{2}$) less sophistication in the loss domain. Both of these models provide a better fit to the data than a model predicated solely on differences in strategic sophistication.

Together, the analytical example in appendix A and the cognitive hierarchy model in B demonstrate that a combination of greater risk tolerance and lower strategic sophistication provide a plausible mechanism for our treatment effect, but neither can rule out an explanation based on social preferences. However, a social preference based explanation is forcefully rejected by the fact that social preferences, elicited in individual decision-making tasks, do not exhibit a treatment effect (section 4.2) and do not correlate with strategic behavior (section 4.3).

When we elicit the self-proclaimed motives behind subjects' chosen claims in a questionnaire after experiment 2 (see table C.4 and C.5 in appendix C), subjects' answers are again more consistent with a risk-preference based than a social-preference based explanation of the treatment effect. In particular,

[16] In the absence of uncertainty, neither risk preferences nor plausible social preferences can explain cooperation in the TD. Noisy behavior in the TD is also empirically plausible: Oppenheimer et al. (2011) find that from period to period, behavior of many individuals in social dilemmas appears to be almost random. Furthermore, models premised on noisy decision making do well at explaining the effect of changing the reward/penalty parameter on subjects' behavior in the TD (Capra et al., 1999).

[17] For example, the structural parameter estimates could refute our proposed mechanism by reflecting greater risk aversion or greater strategic sophistication in the loss domain.

subjects in the loss domain consider "trying to gain a lot" to be a more important reason to choose a high claim and "avoiding the risk of being undercut" to be a less important reason for choosing a low claim than subjects in the gain domain.

We conducted two additional sets of treatments to test the scope of our risk-preference and sophistication based explanation. In the gain-loss treatment, in online appendix D, both gains and losses are possible. This treatment therefore lies between our two main conditions. We find that subjects' behavior exhibits a form of loss avoidance (Cachon and Camerer, 1996): they predominantly choose the action that minimizes their risk of loosing money. We estimate a cognitive hierarchy model to find out what kind of preference function can rationalize this behavior. In line with our explanation for the behavior in the two main conditions, the utility function we estimate in the gain-loss treatment has the S-shape implied by diminishing sensitivity and has its steepest gradient, i.e. its implied reference point, at the experimentally induced reference point.

In a further set of treatments, found in online appendix E, we set the reward/punishment parameter to 0.5. Compared to the case of $R = 3$, cooperation increases in both the loss and the gain domain. Moreover, when $R = 0.5$, subjects cooperate slightly more in the gain than in the loss domain. This is consistent with our main results. Out-of-sample predictions of the cognitive hierarchy model that is parametrized using the data from our main treatments with $R = 3$ match the empirical means of this low-risk treatment. The intuition for this as follows. When the payoff loss associated with being undercut is low, high cooperativeness is privately optimal for sophisticated players. Also, risk preferences naturally matter less. In this setting, a higher propensity to make mistakes in the loss domain then implies less cooperation.

Comparing the treatments with $R = 0.5$ with our main treatments suggests that our treatment effect and proposed mechanism pertain mostly to high stakes situation. Moreover, looking at the later periods of experiment 2 suggests that our treatment effect is likely to be more pronounced in noisy and uncertain settings (see figure 2b). This is again consistent with a risk preference based explanation. Risk preferences only matter when there is strategic uncertainty. In its absence, raising one's claim is definitely a bad idea. As a result, risk preferences do not govern behavior once sufficient learning has taken place (see also column 8 in table 3).

6. Conclusion

We find that individuals are more likely to take a chance on one another when losses are at stake. The prospect of losses induces cooperation in the traveler's dilemma because subjects' greater risk tolerance and lower sophistication in the loss domain drive them to choose the risky action of deviating from Nash play.[18]

An appreciation of not just the treatment effect on strategic behavior, but also its drivers, is helpful in organizing existing experimental evidence. The key role that risk preferences play in driving our results suggests that losses will lead to more cooperative behavior primarily in situations in which

[18]We find these effects in a once-off, anonymous interaction. Doing away with anonymity may open up further channels through which losses impact on cooperation. For instance, the joint experience of losses my allow individuals to "bond".

selfish behavior is not a dominant strategy. We therefore conjecture that our proposed mechanism could shed light on when and how a loss frame is likely to impact on behavior in non-linear public good games, coordination games and Bertrand competition. Our proposed mechanism also elucidates the scope of applicability of our experimental results. The treatment effect is more likely to play a role when strategic uncertainty is high, as in the early periods of our experiment, and when the payoff variance is large, as in our main treatment.

Our results yield tentative implications for scholars of management and political science. A manager may want to keep subordinates from cooperating to provide low effort or organizing to demand higher wages. As a result, she may try to evoke the gain frame whenever possible, for example, by attempting to hide necessary cuts in real wages behind nominal wage increases. Abeler et al. (2011) and Fryer et al. (2012) find that by embedding payment schemes in the loss domain, loss aversion can be leveraged as a cheap way to increase effort in the lab and in the field respectively. Furthermore, De Quidt (2014) finds that doing so has no negative impact on employee participation. In light of this evidence, it is puzzling that the loss frame is not more widely used as an incentive device. Our results speak to this puzzle. Where the organizational structure involves complicated games between employers and several employees, employee cooperation fueled by risk tolerance may undermine the incentives to provide effort generated by loss aversion.

Politicians who seek to foster cooperation among the different stakeholders to a reform or policy initiative may find it useful to evoke a loss frame. The loss frame may not only increase the perceived stakes through loss aversion, but also foster cooperation through the mechanism we uncover in this paper. Indeed, leaders that wish to garner support for war, will often exaggeratedly point to the "threat to our way of life" the enemy poses. Similarly, economic reforms are often advertised as the only way to "remain competitive" or "avoid sliding deeper into recession". However, our results also imply that cooperation will primarily emerge in uncertain settings and that it will decline as players gain experience. Therefore, a loss frame is probably most potently applied under exceptional circumstances and cannot be exploited repeatedly to foster cooperation.

References

Abeler, J., A. Falk, L. Goette, and D. Huffman (2011): "Reference points and effort provision," *American Economic Review*, 101, 470–92.

Arad, A. and A. Rubinstein (2012): "The 11-20 money request game: A level-k reasoning study," *American Economic Review*, 102, 3561–3573.

Baghestanian, S. (2014): "Unraveling the traveler' s dilemma puzzle. A level-k approach," *Working paper*.

Basu, K. (1994): "The traveler's dilemma: Paradoxes of rationality in game theory," *American Economic Review*, 84, 391–395.

Becker, T., M. Carter, and J. Naeve (2005): "Experts playing the traveler's dilemma," Hohenheimer Diskussionsbeiträge 252/2005, University of Hohenheim, Germany.

Berejikian, J. (1992): "Revolutionary collective action and the agent-structure problem." *American Political Science Review*, 86, 647–657.

Bewley, T. F. (2002): *Why Wages Don't Fall During A Recession*, Harvard University Press.

Booij, A. S., B. M. V. Praag, and G. van de Kuilen (2010): "A parametric analysis of prospect theory's functionals for the general population," *Theory and Decision*, 68, 115–148.

Brañas-Garza, P., M. P. Espinosa, and P. Rey-Biel (2011): "Travelers' types," *Journal of Economic Behavior & Organization*, 78, 25–36.

Buchheit, S. and N. Feltovich (2011): "Experimental evidence of a sunk-cost paradox: A study of pricing behavior in bertrand-edgeworth duopoly," *International Economic Review*, 52, 317–347.

Cachon, G. P. and C. F. Camerer (1996): "Loss-avoidance and forward induction in experimental coordination games," *The Quarterly Journal of Economics*, 111, 165–194.

Camerer, C. F. (2003): "Prospect theory in the wild: Evidence from the field," *Colin F. Camerer, George Loewenstein, and Matthew. Rabin, eds., Advances in Behavioral Economics*, 148–161.

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2004): "A cognitive hierarchy model of games," *The Quarterly Journal of Economics*, 119, 861–898.

Capra, C. M., J. K. Goeree, R. Gomez, and C. A. Holt (1999): "Anomalous behavior in a traveler's dilemma?" *The American Economic Review*, 89, 678–690.

Capra, C. M., J. K. Goeree, R. Gomez, and C. A. Holt (2002): "Learning and noisy equilibrium behavior in an experimental study of imperfect price competition," *International Economic Review*, 43, 613–636.

Coleman, J. S. (1994): *Foundations of Social Theory*, Harvard University Press.

Cox, C. A. (2015): "Decomposing the effects of negative framing in linear public goods games," *Economics Letters*, 126, 63–65.

De Dreu, C. K. and C. McCusker (1997): "Gain–loss frames and cooperation in two-person social dilemmas: A transformational analysis." *Journal of Personality and Social Psychology*, 72, 1093–1106.

De Heus, P., N. Hoogervorst, and E. Van Dijk (2010): "Framing prisoners and chickens: Valence effects in the prisoner's dilemma and the chicken game," *Journal of Experimental Social Psychology*, 46, 736–742.

De Quidt, J. (2014): "Your loss is my gain: a recruitment experiment with framed incentives," Technical report, SSRN Working Paper 2418218.

Fanis, M. (2004): "Collective action meets prospect theory: An application to coalition building in chile, 1973–75," *Political Psychology*, 25, 363–388.

Fehr, E. and K. M. Schmidt (1999): "A theory of fairness, competition, and cooperation," *Quarterly Journal of Economics*, 114, 817–868.

Fischbacher, U. (2007): "z-tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 10, 171–178.

Fosgaard, T. R., L. G. Hansen, and E. Wengström (2014): "Understanding the nature of cooperation variability," *Journal of Public Economics*, 120, 134–143.

Fryer, R. G., S. D. Levitt, J. List, and S. Sadoff (2012): "Enhancing the efficacy of teacher incentives through loss aversion: A field experiment," Technical report, National Bureau of Economic Research Working Paper.

Goeree, J. K. and C. A. Holt (2001): "Ten little treasures of game theory and ten intuitive contradictions," *American Economic Review*, 91, 1402–1422.

Goeree, J. K., C. A. Holt, and T. R. Palfrey (2002): "Quantal response equilibrium and overbidding in private-value auctions," *Journal of Economic Theory*, 104, 247–272.

Goeree, J. K., C. A. Holt, and T. R. Palfrey (2003): "Risk averse behavior in generalized matching pennies games," *Games and Economic Behavior*, 45, 97–113.

Holt, C. A. and S. K. Laury (2002): "Risk aversion and incentive effects," *American Economic Review*, 92, 1644–1655.

Iturbe-Ormaetxe, I., G. Ponti, J. Tomás, and L. Ubeda (2011): "Framing effects in public goods: Prospect theory and experimental evidence," *Games and Economic Behavior*, 72, 439–447.

Kachelmeier, S. J. (1996): "Do cosmetic reporting variations affect market behavior? a laboratory study of the accounting emphasis on unavoidable costs," *Review of Accounting Studies*, 1, 115–140.

Kahn, S. (1997): "Evidence of nominal wage stickiness from microdata," *American Economic Review*, 87, 993–1008.

Kahneman, D. and A. Tversky (1979): "Prospect theory: An analysis of decision under risk," *Econometrica*, 47, 263–292.

Mastanduno, M. (1992): "Framing the japan problem: The bush administration and the structural impediments initiative," *International Journal*, 47, 235–264.

Mengel, F. (2014): "Risk and temptation: A meta-study on social dilemma games," *Available at SSRN 2519997*.

Mercer, J. (2005): "Prospect theory and political science," *Annual Review of Political Science*, 8, 1–21.

Moore, W. H. (1995): "Rational rebels: Overcoming the free-rider problem," *Political Research Quarterly*, 48, 417–454.

Offerman, T. and J. Potters (2006): "Does auctioning of entry licences induce collusion? an experimental study," *Review of Economic Studies*, 73, 769–791.

Oppenheimer, J., S. Wendel, and N. Frohlich (2011): "Paradox lost: Explaining and modeling seemingly random individual behavior in social dilemmas," *Journal of Theoretical Politics*, 23, 165–187.

Ostrom, E. (1990): *Governing the Commons: The Evolution of Institutions for Collective Action*, Cambridge University Press.

Ostrom, E. (2000): "Reformulating the commons," *Swiss Political Science Review*, 6, 29–52.

Pauly, L. W. (1992): "The political foundations of multilateral economic surveillance," *International Journal*, 47, 293–327.

Rojas, C. (2012): "The role of demand information and monitoring in tacit collusion," *The RAND Journal of Economics*, 43, 78–109.

Rotemberg, J. J. and G. Saloner (1986): "A supergame-theoretic model of price wars during booms," *The American Economic Review*, 76, 390–407.

Rubinstein, A. (2007): "Instinctive and cognitive reasoning: A study of response times," *The Economic Journal*, 117, 1243–1259.

Ruffle, B. J. (2013): "When do large buyers pay less? experimental evidence," *The Journal of Industrial Economics*, 61, 108–137.

Stein, J. G. and L. W. Pauly (1992): "Choosing to co-operate: How states avoid loss," *International Journal*, 47, 199–201.

Tommasi, M. (2004): "Crisis, political institutions, and policy reform the good, the bad, and the ugly," *Toward Pro-Poor Policies: Aid, Institutions and Globalization, Washington, DC: World Bank*, 135–64.

Tymula, A., P. W. Glimcher, I. Levy, and L. A. R. Belmaker (2012): "Separating risk and ambiguity preferences across the life span: Novel findings and implications for policy," *Unpublished manuscript*.

Velasco, A. (1998): "The common property approach to the political economy of fiscal policy," in F. Sturzenegger and M. Tommasi, eds., *The Political Economy of Reform*, MIT, chapter The Common Property Approach to the Political Economy of Fiscal Policy, 165–184.

Vis, B. (2009): "The importance of socio-economic and political losses and gains in welfare state reform," *Journal of European Social Policy*, 19, 395–407.

Vis, B. and K. Van Kersbergen (2007): "Why and how do political actors pursue risky reforms?" *Journal of Theoretical Politics*, 19, 153–172.

Welch, D. A. (1992): "The politics and psychology of restraint: Israeli decision-making in the gulf war," *International Journal*, 47, 328–369.

Weyland, K. (1996): "Risk taking in latin american economic restructuring: Lessons from prospect theory," *International Studies Quarterly*, 40, 185–207.

Weyland, K. (2002): *The Politics of Market Reform in Fragile Democracies: Argentina, Brazil, Peru, and Venezuela*, Princeton University Press.

Appendix

A. A stylized model of strategic uncertainty

To gain an intuition for how bounded rationality, risk preferences and social preferences may shape behavior in the TD, it is useful to transform the game into a non-strategic decision problem.[19] Suppose that player 1, a rational agent, is playing the TD with player 2, a non-strategic opponent, who plays the Nash equilibrium strategy $x_2 = 3$ with probability $(1-p) \in [0,1]$ and $x_2 = m+0.1$ with complementary probability, where $m > 3$.[20] We may interpret p as the probability of a mistake and m as its size, so that both p and m are measures of player 2's lack of sophistication.

Assume first that player 1 is self-interested and that her utility is given by $U(z_1)$, where z_1 represents her material payoff and $U'(z_1) > 0$. Faced with player 2, player 1 then only has two sensible choices: she may either play her Nash strategy $x_1 = 3$, or she may try to capitalize on player 2's mistake by minimally undercutting, i.e. by playing $x_1 = m$. Given the payoff structure of the TD, player 1 finds it optimal to play $x_1 = m$ if and only if[21]

$$(1-p)U(3-R) + pU(m+R) > (1-p)U(3) + pU(3+R). \tag{A.1}$$

When player 1 plays $x_1 = m$ instead of $x_1 = 3$, she risks being undercut in situations in which player 2 does not make a mistake. This downside risk associated with moving away from the Nash outcome is a key feature of the social dilemmas in the uncertain environments we outline in the introduction.[22] Crucially, the spread of payoffs is higher on the left-hand side of equation (A.1) than on the right-hand side: the minimum and maximum payoffs from playing $x_1 = m$ are respectively lower and higher than the minimum and maximum payoffs from playing $x_1 = 3$. Player 1 therefore has to decide between two gambles with equal probabilities but different variances. A risk loving player 1 is more comfortable with a higher variance in payoffs and thus more comfortable with playing $x_1 = m$ than a risk averse individual.

We define a risk averse, a risk neutral and a risk loving individual as an individual whose utility function is characterized by $U''(z_1) < 0$, $U''(z_1) = 0$ and $U''(z_1) > 0$ respectively and we index them by a, n and l. The following proposition then establishes more formally the above intuition, as well as that cooperating is more attractive when mistakes on behalf of player 2 are large and frequent.

Proposition A.1. *There exists a unique threshold $p^* \in (0,1)$ such that*
- player 1 claims $x_1 = m$ if and only if $p > p^*$ and $x_1 = 3$ otherwise;
- p^* is decreasing in m;

[19] The more elaborate model we adopt in the next section cannot be solved for analytically, which makes getting at the intuition behind comparative statics harder.

[20] Without loss of generality we are implicitly assuming here that the TD is played with the fine action set.

[21] For ease of exposition, we assume that player 1 chooses the Nash strategy whenever indifferent.

[22] The downside risk would also be present if player 1 was choosing whether to defect or cooperate in a prisoners' dilemma. However, contrary to the TD, in a prisoner's dilemma, cooperating never yields higher utility for a *self-interested* individual. For a sufficiently altruistic individual, on the other hand, (cooperate, cooperate) may be the favored prisoners' dilemma outcome.

- p^* is lower for a risk loving than for a risk neutral individual and higher for a risk averse than for a risk neutral individual, i.e. $p_l^* < p_n^* < p_a^*$.

Proof. The difference between the lhs and rhs of equation (A.1) is increasing in p, positive when $p = 1$ and negative when $p = 0$. There thus exists a unique threshold for which player 1 is indifferent between playing 3 and m. This threshold is given by

$$p^* = \frac{U(3) - U(3-R)}{U(3) - U(3-R) + U(m+R) - U(3+R)},$$

which is decreasing in m.

To show that $p_i^* < p_j^*$ is to show that $\frac{1}{p_i^*} > \frac{1}{p_j^*}$ for $i, j \in \{a, n, l\}$ and $i \neq j$. We have that

$$\frac{1}{p^*} = 1 + \frac{U(m+R) - U(3+R)}{U(3) - U(3-R)}$$

By the mean value theorem, we can write

$$\frac{1}{p^*} = 1 + \frac{\big((m+R) - (3+R)\big)U'(c)}{\big(3 - (3-R)\big)U'(b)}$$

with $m + R > c > 3 + R$ and $3 > b > 3 - R$. From the definitions of risk preferences and the fact that $c > b$ it follows that $\frac{U'(c)}{U'(b)} > 1$ for a risk loving individual, $\frac{U'(c)}{U'(b)} = 1$ under risk neutrality, and $\frac{U'(c)}{U'(b)} < 1$ if the individual is risk averse. Consequently, $p_l^* < p_n^* < p_a^*$. □

Let us now impose risk neutrality and allow for social preferences on behalf of player 1. In particular, suppose that the utility of player 1 takes the form of $U(z_1; z_2) = z_1 + \beta z_2$, where z_2 is the material payoff of player 2 and $\beta \in (0, 1)$ a measure of player 1's altruism.[23] The following proposition shows that altruism, like risk lovingness, provides an incentive for player 1 to deviate from Nash play and that this incentive is stronger the less sophisticated player 2 is:

Proposition A.2. *There exists a unique threshold $p_\beta^* \in (0, 1)$ such that*
- player 1 deviates from $x_1 = 3$ if and only if $p > p_\beta^*$;
- p_β^* is decreasing in m and β.

Proof. When $U(z_1; z_2) = z_1 + \beta z_2$, the best deviation from $x_1 = 3$ is either $x_1 = m$ or $x_1 = m + 0.1$, depending on whether

$$m + R + \beta(m - R) \geq m + 0.1 + \beta(m + 0.1),$$

that is, on whether $\beta \leq \frac{R-0.1}{R+0.1} \equiv \hat{\beta}$. If $\beta \leq \hat{\beta}$, the best deviation is $x_1 = m$ and player 1 finds it optimal to deviate if and only if

$$p\big(m + R + \beta(m - R)\big) + (1 - p)\big(\beta 3 + (3 - R)\big) > p\big(3 + R + \beta(3 - R)\big) + (1 - p)\big(3 + 3\beta\big). \quad \text{(A.2)}$$

[23] We model prosocial considerations in the simplest possible way, but results generalize to more elaborate other-regarding preferences. For instance, if player 1's utility is as in Fehr and Schmidt (1999), similar predictions obtain based on her aversion to advantageous inequality.

The difference between the lhs and rhs of equation (A.2) is increasing in p, positive at $p = 1$, negative at $p = 0$ and equal to zero at

$$p^*_{\beta\leq\hat{\beta}} = \frac{R}{m + R - 3 + (m-3)\beta},$$

which is decreasing in β and m.

When $\beta > \hat{\beta}$, the best deviation is $x_1 = m + 0.1$ and player 1 finds it optimal to deviate if and only if

$$p\big(m + 0.1 + \alpha(m + 0.1)\big) + (1-p)\big(\beta 3 + (3 - R)\big) > p\big(3 + R + \beta(3 - R)\big) + (1-p)\big(3 + 3\beta\big).$$

The difference between the lhs and rhs of this inequality is again increasing in p, positive at $p = 1$, negative at $p = 0$ and equal to zero at

$$p^*_{\beta>\hat{\beta}} = \frac{10R}{10(m + (m + R)\beta - 29(1 + \beta))},$$

which is also decreasing in β and m. To conclude the proof, let p^*_β be equal to $p^*_{\beta\leq\hat{\beta}}$ for $\beta \leq \hat{\beta}$ and to $p^*_{\beta>\hat{\beta}}$ for $\beta > \hat{\beta}$ and note that p^*_β is continuous in β because $p^*_{\beta\leq\hat{\beta}} = p^*_{\beta>\hat{\beta}}$ when $\beta = \hat{\beta}$. □

Propositions A.1 and A.2 tell us that changes in risk preferences and altruism can systematically impact on subjects' tendency to cooperate in a TD with underlying uncertainty. Since we observe that individuals cooperate more in the loss domain, the propositions imply that the treatment effect could be accounted for by higher levels of risk lovingness or altruism in the loss domain. Of course, a higher propensity to make mistakes, as reflected in a higher p, or larger mistakes, as reflected in a larger m, would also lead to more frequent deviations from Nash.[24] In this stylized example, player 2 does not react strategically to player 1's actions. The next section demonstrates that the simple intuitions developed here have bite when we allow for a larger set of strategic types.

B. Estimates from a cognitive hierarchy model

The structural estimations of this section can be viewed as a comparative statics exercise that maps the experimental data from our two treatments into implied differences in risk preferences, social preferences and sophistication. Results indicate that an explanation based on less sophistication in the loss domain alone does a bad job at explaining the treatment effect, but they do not allow us to distinguish between explanations that feature either reference-dependent risk preferences and reference-dependent sophistication or reference-dependent social preferences and reference-dependent sophistication.

We adopt an augmented version of the cognitive hierarchy model in Camerer et al. (2004). Subjects belong to different steps of sophistication, indexed by an integer k. Steps in the population follow a Poisson distribution parametrized by t, both the distribution's mean and variance. Each step $k > 0$ player myopically believes to be the most sophisticated player and, in particular, that the distribution

[24] Note that high claims are only a "mistake" if player 2 makes them. High claims by player 1 are perfectly rational.

of other players follows a Poisson distribution parametrized by t, but right-truncated at $k-1$. A step $k>0$ player's belief about the frequency of step $h<k$ players is thus given by

$$g_k(h) = \frac{\frac{e^{-t}t^h}{h!}}{\sum_{l=0}^{k-1}\frac{e^{-t}t^l}{l!}}. \tag{B.1}$$

Each step $k>0$ player maximizes her expected utility U^e given her beliefs specified in (B.1) and anticipating that each step $h<k$ behaves in a similar fashion. The model solves recursively after specifying the behavior of step 0 players. We assume that a fraction $1-w$ of step 0 players randomize over the action set, while the remaining fraction w play the highest action.[25]

With the exception of Goeree et al. (2002) and Goeree et al. (2003), experimental papers that make use of models of noisy strategic interactions impose that players are risk neutral and selfish. In order to capture risk attitudes, we adopt the more general functional form $U(z_1)=z_1^\alpha$, where z_1 represents a player's material payoff.[26] Then, $\alpha<1$, $\alpha=1$ and $\alpha>1$ respectively capture risk aversion, neutrality and lovingness. We also consider an alternative model with altruism, in which utility takes the form $U(z_1;z_2)=z_1+\beta z_2$, where z_2 is the material payoff of the other player.

Both models have three parameters, respectively (t,w,α) and (t,w,β), and solve numerically for any parameter values. Using maximum likelihood techniques, we determine which combination of parameters best replicates the observed distribution of our subjects' actions in each period. The trends in figure 2 might be interpreted as evidence of learning as play progresses, which in the models would translate into higher values of t and lower values of w. Conversely, there is no a priori reason to expect α or β to change across periods. We therefore also estimate each model jointly over the five periods under the respective restriction that α or β remain constant.

Table B.1 reports estimates of the model that allows for different risk preferences. Estimates of α confirm the link between risk preferences on cooperation established in section A.[27] For instance, in the joint model we obtain $\hat{\alpha}=0.95$ in the gain treatment and $\hat{\alpha}=3.25$ in the loss treatment. We use a likelihood ratio test that compares these models with a model that imposes the risk-neutrality restriction ($\alpha=1$) to check whether estimates of α are significantly different from 1. In the loss treatment we are able to reject risk-neutrality in favor of risk-lovingness at the 1 percent significance level. In the gain treatment we cannot reject risk neutrality. These results are confirmed in the models that are estimated over individual periods. Thus, the model that best matches the observed distribution of actions is one that supposes individuals are more risk tolerant in the loss domain.

Estimates of t suggest that the average subject is also less sophisticated in the loss domain.[28] But

[25] Arad and Rubinstein (2012) adopt a similar specification for the behavior of level 0 players in a game which bears some similarities with the TD. Qualitative results are robust to imposing that $w=0$.

[26] As we run all estimations using net claims, there is no need to define $U(z_1)$ piecewise for the gain and the loss domain.

[27] One can show numerically that the optimal action of a step 1 player (and hence of all more sophisticated players) is weakly increasing in α and coincides with Nash play when α is lower than some cutoff $\bar{\alpha}(w)$, where $\bar{\alpha}(w)$ is decreasing in w and $\bar{\alpha}(0)=\frac{6}{5}$. When the likelihood attains the maximum in the $\alpha<\bar{\alpha}$ region, in case of multiple maximizers we adopt the convention to select the largest α.

[28] Overall, values of t are a slightly smaller than typical estimates from the literature. This is due to the fact that the action set is large, play is fairly dispersed and $k>0$ players overall use only a small set of actions. Thus, the model

Table B.1 Estimates of risk preferences in a cognitive hierarchy model

Treatment	Period	Restriction	α	t	w	LogL
Gain	1	None	0.95	0.31	0.20	−110.04
			(0.120)	(0.102)	(0.090)	
	2	None	1.10	0.51	0.06	−99.64
			(0.062)	(0.155)	(0.059)	
	3	None	1.15	0.56	0.01	−94.17
			(0.034)	(0.156)	0.034	
	4	None	1.15	0.90	0.01	−77.96
			(0.034)	(0.222)	(0.034)	
	5	None	1.05	1.30	0.11	−53.94
			(0.123)	(0.268)	(0.101)	
	all 5	α const	0.95			−435.75
Loss	1	None	6.00***	0.11	0.11	−117.16
			(1.720)	(0.065)	(0.062)	
	2	None	2.30***	0.16	0.06	−118.67
			(0.678)	(0.082)	(0.054)	
	3	None	4.00***	0.26	0.06	−119.08
			(0.560)	(0.113)	(0.053)	
	4	None	3.25***	0.21	0.01	−116.87
			(0.434)	(0.094)	(0.027)	
	5	None	1.50	0.16	0.06	−121.53
			(0.880)	(0.084)	(0.039)	
	all 5	α const	3.25***			−609.10

Note: Bootstrapped standard errors in parentheses; significance levels from a likelihood-ratio test against the restriction $\alpha = 1$: * ($p < 0.1$), ** ($p < 0.05$), *** ($p < 0.01$).

the likelihood ratio test on the restriction of risk neutrality implies that risk-preferences in combination with sophistication fares better at explaining our treatment effect than an explanation based on less sophistication in the loss domain alone. Note also that estimates of t show a clear increasing pattern over periods in the gain treatment but less so in the loss treatment, while estimates of w tend to decrease in both treatments.

Table B.2 presents estimates of the model that allows for altruism. Estimates confirm the prediction of section A that more altruism in the loss domain can also explain our treatment effect. Moreover, the log-likelihoods of the model featuring risk preferences and the model featuring altruism are very similar, suggesting that a social-preference based explanation fits the aggregate data just as well as a risk-based explanation.

Figure B.1 displays the distribution of subjects' net claims in each period and treatment as well as the distributions obtained from the two cognitive hierarchy models using the correspondent estimates from table B.1 and B.2. It demonstrates how both models are able to replicate actual play remarkably well. Numerical simulations provide additional evidence that the insights from the simple example of section A still hold in the cognitive hierarchy models and, hence, likely also in practice. As an example,

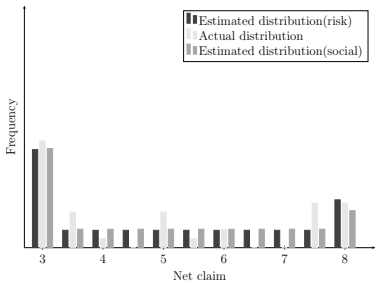
attributes actions outside this set to step 0 players.

Table B.2 Estimates of altruism in a cognitive hierarchy model

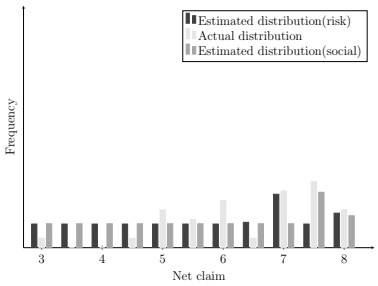
Treatment	Period	Restriction	β	t	w	LogL
Gain	1	None	0	0.31	0.15	−110.09
			(0.020)	(0.103)	(0.040)	
	2	None	0.05	0.51	0.075	−99.59
			(0.024)	(0.150)	0.052)	
	3	None	0.05	0.61	0	−93.93
			(0.007)	(0.172)	(0.025)	
	4	None	0.05	0.86	0	−77.80
			(0.029)	(0.233)	(0.031)	
	5	None	0	1.31	0.10	−53.93
			(0.025)	(0.340)	(0.063)	
	all 5	β const	0			−435.34
Loss	1	None	0.90***	0.11	0.10	−117.02
			(0.273)	(0.061)	(0.064)	
	2	None	0.30***	0.15	0.025	−118.86
			(0.121)	(0.080)	(0.030)	
	3	None	0.35***	0.26	0.15	−120.00
			(0.059)	(0.109)	(0.052)	
	4	None	0.40***	0.21	0	−117.02
			(0.068)	(0.096)	(0.011)	
	5	None	0.05	0.11	0	−121.60
			(0.239)	(0.052)	(0.022)	
	all 5	β const	0.40***			−605.91

Note: Bootstrapped standard errors in parentheses; significance levels from a likelihood-ratio test against the restriction $\beta = 0$: * ($p < 0.1$), ** ($p < 0.05$), *** ($p < 0.01$).

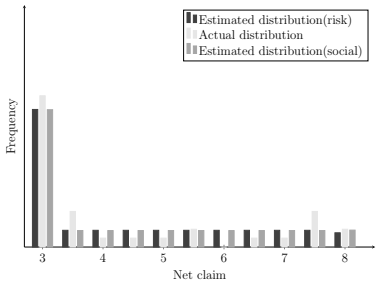
figure B.2 illustrates how average claims increase with risk tolerance and altruism and decrease with strategic sophistication. Figure B.3 disaggregate these effects by considering the whole distribution of net claims, who shifts to the right or left accordingly.



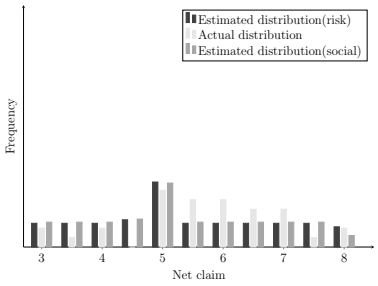
Period 1: gain treatment



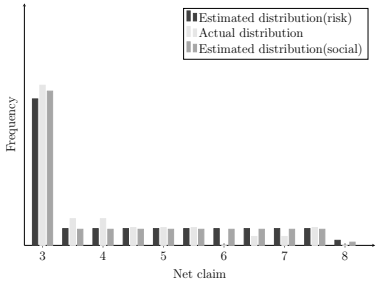
Period 1: loss treatment



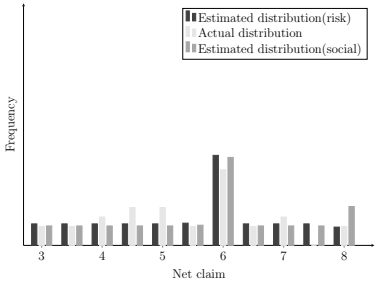
Period 2: gain treatment



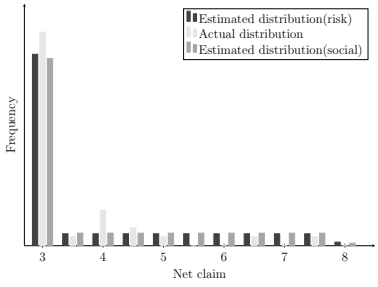
Period 2: loss treatment



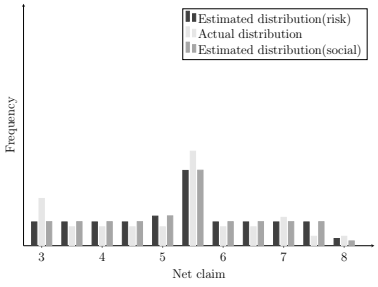
Period 3: gain treatment



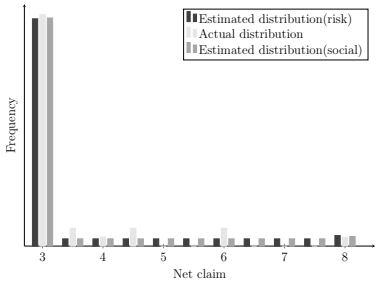
Period 3: loss treatment



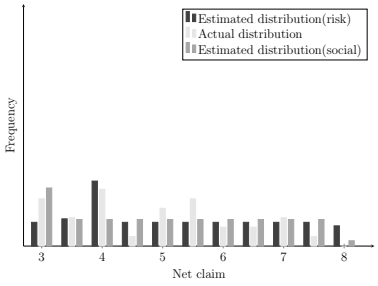
Period 4: gain treatment



Period 4: loss treatment

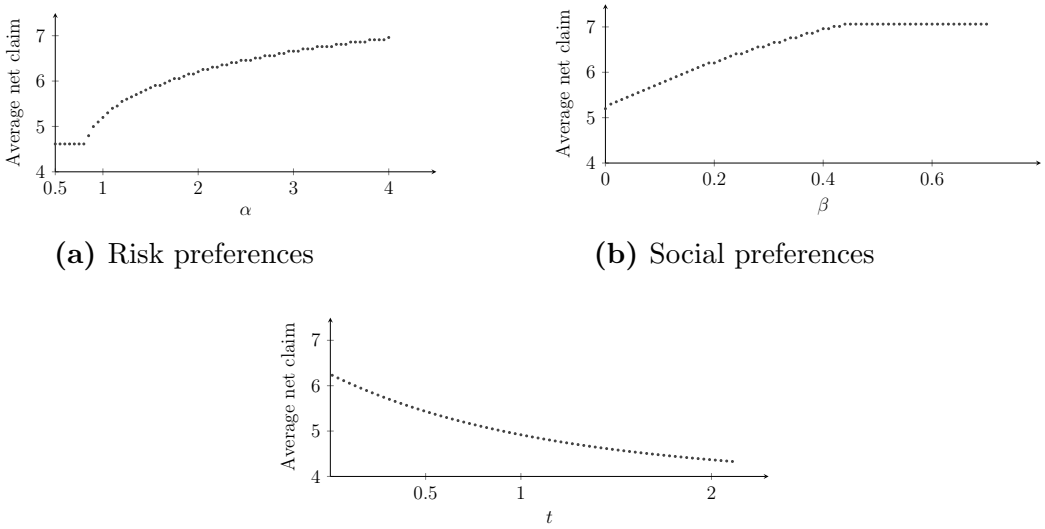


Period 5: gain treatment



Period 5: loss treatment

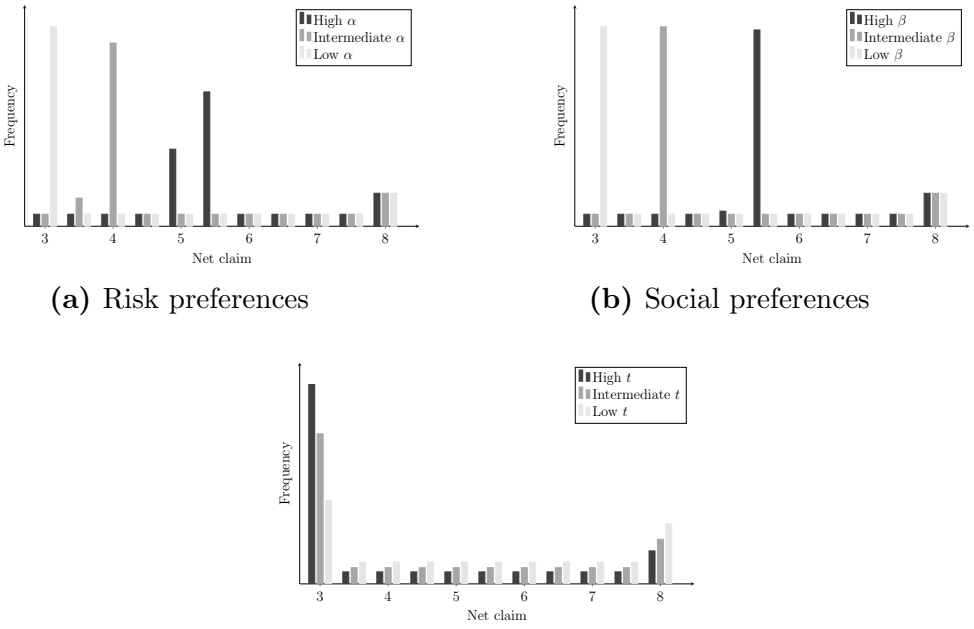
Figure B.1 Actual and estimated distribution from the cognitive hierarchy models



(a) Risk preferences **(b)** Social preferences

(c) Strategic sophistication

Figure B.2 Comparative statics on average claims from the cognitive hierarchy models



(a) Risk preferences **(b)** Social preferences

(c) Strategic sophistication

Figure B.3 Comparative statics on the distribution from the cognitive hierarchy models

C. Additional material

Table C.1 The HL risk preferences elicitation task in the gain treatment

Row	Option L	Option R
1	6 with $p = \frac{1}{10}$; 5 with $p = \frac{9}{10}$	10.5 with $p = \frac{1}{10}$; 0.5 with $p = \frac{9}{10}$
2	6 with $p = \frac{2}{10}$; 5 with $p = \frac{8}{10}$	10.5 with $p = \frac{2}{10}$; 0.5 with $p = \frac{8}{10}$
3	6 with $p = \frac{3}{10}$; 5 with $p = \frac{7}{10}$	10.5 with $p = \frac{3}{10}$; 0.5 with $p = \frac{7}{10}$
4	6 with $p = \frac{4}{10}$; 5 with $p = \frac{6}{10}$	10.5 with $p = \frac{4}{10}$; 0.5 with $p = \frac{6}{10}$
5	6 with $p = \frac{5}{10}$; 5 with $p = \frac{5}{10}$	10.5 with $p = \frac{5}{10}$; 0.5 with $p = \frac{5}{10}$
6	6 with $p = \frac{6}{10}$; 5 with $p = \frac{4}{10}$	10.5 with $p = \frac{6}{10}$; 0.5 with $p = \frac{4}{10}$
7	6 with $p = \frac{7}{10}$; 5 with $p = \frac{3}{10}$	10.5 with $p = \frac{7}{10}$; 0.5 with $p = \frac{3}{10}$
8	6 with $p = \frac{8}{10}$; 5 with $p = \frac{2}{10}$	10.5 with $p = \frac{8}{10}$; 0.5 with $p = \frac{2}{10}$
9	6 with $p = \frac{9}{10}$; 5 with $p = \frac{1}{10}$	10.5 with $p = \frac{9}{10}$; 0.5 with $p = \frac{1}{10}$
10	6 with $p = \frac{10}{10}$; 5 with $p = \frac{0}{10}$	10.5 with $p = \frac{10}{10}$; 0.5 with $p = \frac{0}{10}$

Note: The task in the loss treatment is identical, except that 11 is subtracted from each outcome.

Table C.2 The dictator task in the gain treatment

Choice	Consequences
1	you obtain 11; the other person obtains 3
2	you obtain 10; the other person obtains 4
3	you obtain 9; the other person obtains 5
4	you obtain 8; the other person obtains 6
5	you obtain 7; the other person obtains 7
6	you obtain 6; the other person obtains 8
7	you obtain 5; the other person obtains 9
8	you obtain 4; the other person obtains 10
9	you obtain 3; the other person obtains 11

Note: The task in the loss treatment is identical except that 11 is subtracted from each outcome.

Table C.3 Order effects in elicited preferences

		First round		Last round		
	Treatment	Mean	N	Mean	N	Difference
HL switching point	Gain	6.72	18	6.94	18	-0.22
	Loss	5.62	21	6.28	21	-0.66
HL switching point (<11)	Gain	6.19	16	6.44	16	-0.25
	Loss	5.35	20	5.79	19	-0.44
Dictator giving	Gain	3.17	18	3.11	18	0.05
	Loss	3.62	21	2.81	21	0.81
Risk TD claim	Gain	5.00	18	4.89	18	0.11
	Loss	6.05	21	5.10	21	0.95**
Social TD claim	Gain	4.81	18	4.64	18	0.16
	Loss	6.00	21	4.74	21	1.26***

Note: The last column features t-tests on the difference in means between subjects who performed the task in the first and the last round: * ($p < 0.1$), ** ($p < 0.05$), *** ($p < 0.01$).

Table C.4 Self-reported explanations of play (importance on a 1 to 4 scale)

	Gains		Losses		
	Mean	N	Mean	N	Difference
Reasons to choose a high claim					
• try to gain a lot	2.38	36	2.71	42	-.32*
• let the other gain a lot	2.3	36	2.16	42	.13
• reward who chooses a high claim	2.44	36	2.4	42	.03
• it is fair	2.3	36	2.4	42	-.09
Reasons to choose a low claim					
• avoid the risk of being undercut	3.69	36	3.26	42	.43***
• let the other gain little	2	36	2.3	42	-.309
• punish who plays a low claim	2	36	2.42	42	-.42***
• it is rational	3.08	36	3.04	42	-.03

Note: The last column features t-tests on the difference in means between the gain and the loss treatment: * ($p < 0.1$), ** ($p < 0.05$), *** ($p < 0.01$).

Table C.5 Self-reported opinions on social norms (% of "yes")

	Gains		Losses		
	Mean	N	Mean	N	Difference
A plays a high claim and B a low one					
• is B unfair?	25%	36	23.8%	42	1.2%
• would you incur a cost to punish B ?	41.6%	36	38.1%	42	3.5%

Note: The last column features t-tests on the difference in means between the gain and the loss treatment: * ($p < 0.1$), ** ($p < 0.05$), *** ($p < 0.01$).