Preferences for Fair Competition[§]

Nadine Chlaß[*] and Gerhard Riener[‡]

April 28, 2025

Abstract

We study in how far people's aversion against lying, spying, and sabotage can be explained by an intrinsic concern for the opponent's freedom to choose. In three constant sum games with identical payoffs, one party either fabricates, spies, or replaces her opponent's decision unbeknownst to that opponent, or opts into fair competition. Thereby, lying and sabotage affect the opponent's decision rights; spying does not. We observe substantial aversion to lying and sabotage to the extent that people forego *all* payoff, but no aversion to spying. In a variation, we change the setup such that lying and sabotage leave the opponent's decision rights intact. No aversion to lying, spying, or sabotage occurs. Individuals' beliefs about how undesirable their opponent deems lying, spying, or sabotage show that individuals are particularly selfish where they expect to let down the opponent the most. In a third variation, we grant the opponent exogenous decision rights through punishment and reward options to make rights more equal and remove concerns about the opponent's freedom to choose. Indeed, lying, spying, and sabotage aversion are now uniquely driven by people's punishment and reward expectations. A (redundant) instrumental variable approach with independent data confirms our results.

JEL: D02,D03,D63,D64

Keywords: lying, spying, sabotage, decision rights, autonomy, equality of opportunity, moral judgement

---

[§]This is the completed version of University of Mannheim Working Paper ECON #15-17. Since both the theory and the instrumental variable approach on which this paper builds took a decade to be published, this work has not yet been submitted for publication.

[*]Friedrich-Schiller-University Jena, Carl-Zeiss-Str. 3, 07743 Jena, Germany. Corresponding author. Email: nadine.chlass@uni-jena.de, Phone: 0049 (0)3641 943243. The manuscript was first drafted under Leverhulme Visiting Fellowship DKA 7200 at the Economics Department, University of Essex, UK.

[‡]University of Southampton, University Road, Southampton SO17 1BJ, United Kingdom. Email: g.riener@soton.ac.uk. We thank audiences at the EEA/ESEM, the Royal Economic Society, research seminars at the Universities of Cologne, Düsseldorf, Jena, Mannheim, Maastricht, and Turku for discussions. Markus Prasser and Martin Schneider assisted in the data collection. We thank Marie-Claire Villeval, Kaisa Kotakorpi, Topi Miettinen, and Paul Heidhues for commenting the manuscript. This research was funded by the German Research Foundation under grant RTG 1411. All experimental treatments have been reported.

1 Introduction

In 2013, Edward Snowden's leaks of classified information about global surveillance activities by the U.S. secret service led to an international diplomatic crisis. The leaks documented that – in pursuit of preventing terrorist attacks – the U.S. secret service had systematically and preemptively intercepted and stored private communications and information on U.S. citizens, foreign governments, heads of friendly nations and had sabotaged internet encryption as a means to this end.[1] Similarly, insiders broke practices of 'parallel construction' in the U.S. Drug Enforcement Administration to Reuter's journalists Shiffman and Cooke (2013): the 'fabrication' of investigative trails to cover up that actual trails were based on inadmissible evidence from NSA warrantless surveillance. In his interviews with the Guardian, Snowden stated that *'he was willing to sacrifice all [...] because he could not in good conscience allow the destruction of privacy and basic liberties [...]'* (Greenwald et al., 2013).

In this paper, we explore the idea that people's aversion to lying, sabotage, and spying is linked to such an intrinsic concern for others' basic liberties and rights. Some of the agents we observe precommit to fabrication-free, sabotage-free, and espionage-free interactions when fabrication, sabotage, and spying earns them full control over all payoff. Others who either by chance, or by their own choice, end up being forced to fabricate, sabotage, or spy, give *all* payoff away to their opponent. Both courses of action relate to the same ethical criterion, and must therefore be seen as different reactions to the same moral concern. Agents who consider basic liberties and rights, have preferences purely over the rules which make up an interaction: how many times a party is called upon to decide and between how many options, how much information the situation accords her, and how these rights compare to those awarded to other parties. We find that agents pursue their own rational self-interest *least*, where they expect it to be *most* acceptable from their opponent's viewpoint. A simple explanation that agents' altruism issues from social norms and expectations therefore, does not do.

That the rules of a game may violate basic liberties and rights and this way, give rise to people's aversion against some actions explicitly prescribed by these rules, is perhaps most intuitive for the case of *sabotage*. A person who sabotages attaches new material consequences to their opponent's choice, redefining by how much the opponent prefers their choice over the other choices they might have had, that is, how much the opponent's choice assists them in the pursuit of their ends.[2] Sabotage therefore directly interferes with the opponent's freedom to choose, overriding the opponent's will and right to look after her own self-interest. *Lying* or *fabrication* in turn is often studied in sender-receiver, or so-called *deception* games (Gneezy, 2005). Therein, a sender with private information about her opponent's payoffs, informs the opponent which course of

[1] NSA officials did not deny these activities, stating they are 'hardly surprising' (Larson et al., 2013).

[2] In (Lazear, 1989; Konrad, 2000; Chen, 2003; Münster, 2007; Harbring et al., 2007; Carpenter et al., 2010; Harbring and Irlenbusch, 2011), agents choose an effort linked to some output, performance, or chance of winning or choose the latter directly. Sabotage reduces (or increases, and is then called *help*) either effort or output unbeknownst to the agent, relinking the agent's original choice to a new payoff.

action yields them the higher payoff. The sender therefore defines by how much their opponent can prefer their choice over another, and, quite similarly to the case of sabotage, directly interferes with the opponent's freedom to choose and their decision rights.[3] Each truthful statement leaves the receiver's decision rights intact; each lie inevitably takes some away, provided the sender expects to be believed, and the receiver does indeed believe the message. A sender who intrinsically values the opponent's decision rights, seeks to keep the opponent's freedom to choose intact by correctly informing the receiver which option they prefer, and by how much. The sender becomes averse to lying. *Spying* collects information about the opponent's choice but leaves the degree to which the opponent prefers it over other options, intact. A person who spies, increases her own information rights at the expense of the opponent's.

That a party values her *own* decision rights *intrinsically* beyond the mere instrumental value which these rights afford, has first been documented in (Fehr et al., 2013; Bartling et al., 2014). Increasing an opponent's or an agent's decision rights therefore quite rightly has a positive welfare effect beyond what reciprocity can explain (Charness et al., 2012; Kala, 2024). Chlaß et al. (2019) embed the intrinsic value of decision rights into an inequity aversion framework resulting in a preference for the equality of decision rights, or put differently, an intrinsic concern that agents should hold equal rights to pursue their own self interest. Individuals pay some amount for such an equality when, given their pure strategy beliefs and actions, no preference model predicts they should do so; or accept lesser payoff under more equal rights; or refuse to increase their own rights to increase their payoff. These departures from rational self-interest all link to the same ethical criterion: social contract reasoning and the equality of rights stipulated therein. Sugden and Wang (2020) intend much the same with their concept of strategy fairness, concluding much the opposite – that inequality in decision rights does not lead to greater altruism but to destructive acts on all sides. Note, however, that this is due

[3]Thereby, a truthful statement need not classify as lying aversion (Sutter, 2009) and need not reflect any concern about the opponent, unless the sender does not also expect her opponent to believe her message. Recent work (Kajackaite and Gneezy, 2017; Gneezy and Kajackaite, 2020) builds novel features into the deception game to resolve the issue. A second strand of literature studies lies in individual decision making. Here, individuals report the outcome of a random mechanism to the experimenter such as a die roll (Fischbacher and Föllmi-Heusi, 2013) which, purportedly, remains their private information. Subject's payoffs depend on their report, not the actual die roll. In this setup, there is no opponent other than the experimenter whose research funds the lie depletes. Fischbacher and Föllmi-Heusi (2013) compare actual and experimenter opponents but find no difference in lying aversion. However, the two strands of literature disagree on whether lying responds to incentives, or not. In deception games, it does, in the individual decision task, it does not. Recent efforts therefore translate one setup into the other, first, by letting agents simply *imagine* rather than actually perform the die roll (Kajackaite and Gneezy, 2017). Agents now know for sure that the die roll is hidden from the experimenter. A significant share of lies responds to incentives, suggesting that the original individual decision making task did indeed suffer from experimenter demand. It therefore seems that the experimenter effect in the individual decision task provides as strong a motive for lying aversion as do the opponent's decision rights – if indeed active – in deception games. Gneezy and Kajackaite (2020) further introduce a payoff externality of the die roll lie on a passive opponent. Senders lie less than in absence of such an externality, and do not lie more when honesty leaves them worse off than the opponent which makes a bad case for inequity aversion in deception games. Note, however, that lying with a mere payoff externality on a passive opponent does not yet impair the latter's decision rights since she has no freedom of choice in the first place. Quite rightly, the effect of the payoff externality is not robust to increased stakes – the latter increase the value of the sender's options to lie, i.e. her own freedom of choice.

to an error which has gone unnoticed[4]. Cappelen et al. (2024) find that, in a large scale field experiment in Scandinavia and the U.S., a significant majority of respondents is tolerant of unequal opportunities, since/once these have been earned or merited. This finding poses no challenge, however, to the idea that purely procedural preferences lead to a compensation of lesser decision rights. An option inherits its diversity from the degree to which it can be preferred over other options (Sugden, 1998). The cost of effort in earning the option obviously causes a disutility which reduces the degree to which the option increases an individual's freedom to choose, reduces the advantage in opportunities which then warrants lesser or no compensation toward the fellow human being with lesser opportunities because she has incurred lesser effort.[5] Here, we continue this small (Trautmann, 2023) strand of research by asking whether people's aversion to lying, spying, and sabotage – all departures from narrow rational self-interest – are caused by an intrinsic valuation for the opponent's decision rights.

Lying, spying and sabotage are fundamentally relevant to Economics. Economics' core idea that *competition fosters welfare* – that market agents who seek innovation in competition for revenue and gain benefit society in pursuit of their own narrow rational self-interest (Smith, 1904) –, rests on the assumption that agents may not outcompete others by simply manipulating the latters' cost, fabricating information, or spying business secrets. Where such externalities are at work, competition no longer selects the most innovative or cost-effective agent, and the market no longer regulates itself by Adam Smith's invisible hand. Externalities appear to be particularly problematic where asymmetric: when some agents, by propensity, chance, ability, or construction, use them to advance their place in a competition, putting more able or innovative competitors who lack such opportunity or willingness, behind.

Examples for the externalities at hand are the recent explosion of sabotaged communication devices in Lebanon, or spoofing by high frequency traders who either slow down other traders' algorithms by inflating the amount of data to be processed through fake orders or place their orders in between more slowly timed orders changing the price at which the latter are executed (German Central Bank, 2016). Similarly, Kirch media's downfall was brought about by Deutsche Bank's CEO Breuer fabricating rumours about the group's insolvency in an interview with Bloomberg (Reuters, 2012). In fact,

[4]The authors use a game of cards, each with a number, in which one party may ask for *more* new deals (i.e. 'replacement cards') than her opponent. The highest numbered card wins the game. Subsequently, players' emotions are elicited by means of a vendetta game in which players may alternately reduce each others' payoffs. The sequential vendetta game, however, solely intended to measure the emotions about the unequal number of new deals in the first game, accidentally grants unequal opportunities to both parties: along *25 of 31* possible courses of play through this game, the loser of the card game (player 1) *always* has more opportunities to choose from than her opponent. *In the remaining 6 courses of play*, the opponent also has lesser options, except for the final period. Even here therefore, player 1 can for sure prevent this from happening; she can make a decision such that the opponent terminates the interaction at a decision rights with lesser options. Quite rightly, if the loser of the card game now holds greater opportunities in the vendetta game, and the winner of the card game now holds lesser opportunities in the vendetta game, both players have good reason to be upset and destructive.

[5]Note that the same does not hold for Sugden and Wang (2020)'s strategy fairness since the concept is explicitly not based on Sugden's (1998) principle of eligibility when constructing the set of opportunities: a player need not actually hold a preference such that she prefers an option over all others she has.

lying, spying, and sabotage are part of many people's work lives (Abratt and Penman, 2002): Online shops collect, analyze, and complete information on client behaviour to obtain comprehensive customer profiles and induce customers to buy; personnel managers screen social media to obtain information about the social life and character of job candidates (Brown and Vaughn, 2011), credit reference agencies such as the German Schufa collect and analyze information on financial incidents in people's lives, rate their creditworthiness and sell these ratings to customers.

Experimentally, individuals' attitudes toward lying, spying, and sabotage differ substantially. In the original deception game, some 52% of all senders lie when their own gain is large and their lie reduces the opponent's payoff by an equally large extent; 48% consequently tell the truth (Gneezy, 2005).[6] Sabotage in tournaments rests on a level of, on average, roughly 10% and rises to 20% with maximal incentives (Harbring and Irlenbusch, 2011), leaving considerable room for maneuver unused. Not so with spying; the 31 potential spies in (Lee, 2023) spy their opponent 494 of 496 times. This is in line with Beresford et al. (2012) who find that spying is seen as a largely legitimate action in pursuit of one's self-interest. These differences suggest that, indeed, different liberties and rights may be at risk in each case.

We design three basic setups LIE, SPY, and SAB with *identical payoffs*. In each setup, one party chooses the interaction: either a matching pennies game where, by construction, opponents have equal decision and information rights, or a matching pennies game where the party who chooses the interaction, fabricates (LIE), looks up (SPY), or overrides (SAB) the opponent's choice to dictate the allocation.[7] In LIE and SAB, the party who chooses the interaction takes away the opponent's decision rights when dictating the allocation; in SPY, she always leaves these rights intact. In three further treatments, we remove the opponent's decision rights from all matching pennies games such that the party who chooses the interaction *always* dictates the allocation, and can therefore lie, spy, and sabotage without impairing the opponent's freedom to choose. If concerns about the opponent's decision rights are at work in the basic setup, they should disappear here and decisions be close to rational self-interest. Instructions in the basic setups and these further three treatments are *identically* worded, except for two strategy combinations exchanging payoffs. If no differences across LIE, SPY and SAB occur anymore, treatment differences in the basic setup could not be due to the necessarily different descriptions of lying, spying, and sabotage.

In each session, we repeat the task once, allowing the opponent to punish or reward the party's choice of interaction by a symbolic amount. Lying, spying, and sabotage remain in the party's rational self-interest. We elicit the party's punishment and reward

[6] In simple dictator games which allow the dictator (sender) to implement the same allocations without lying, senders take the higher payoff two times more often.

[7] In order to uniquely vary decision rights (and not information rights at the same time), a party in LIE and SAB never knows the true state of the world, i.e. her opponent's choice. The party's message is then dishonest by Sobel's (2020) definition that she knows true states (opponent choices) exist such that her message *is* a lie. The Deutsche Bank example in the main text illustrates this feature; sabotage requires no knowledge of the actual state of the world either: Iran's uranium centrifuges were sabotaged without knowledge whether they were enriching uranium for energy production, or for an atomic bomb.

expectations to see how undesirable she believes her choice of interaction to be from the opponent's point of view, and study how her pursuit of rational self-interest varies along with these expectations. In a methodologically redundant effort, we confirm that where lying, spying, and sabotage impair the opponent's basic liberties and rights, *all* departures from rational self-interest – parties actively foregoing lying, spying, and sabotage or parties lying, spying, and sabotaging to give all payoff away – are significantly linked to a statistical instrument for purely procedural preferences developed in (Chlaß et al., 2019) – ethical preferences for equal basic liberties and rights which are purely deontological by nature. The instrument builds upon an early spillover of moral philosophy to developmental psychology which, in a century's work, has developed and validated robust tools to measure and quantify moral judgement.[8] These measures are identically distributed in all samples, whether we observe lying, spying, and sabotage aversion, or pure rational self-interest, and capture the respective elements in social or moral preference models which motivate departures from rational self-interest – the respective ethical criteria which people deem binding. Thereby, we consider a comprehensive *list* of such criteria – reward or punishment, social image, expectations, social norms, status, the equality of rights, and universal principles of conscience – or put differently, *Kantian* morality (Alger and Weibull, 2013).

The paper is organized as follows. Section 2 briefly reviews the workings of our basic setup, section 3 describes the experimental design. All treatments are reported. Section 4 presents our experimental results on parties' choice of the interaction, the allocation they impose, and parties' beliefs about how the opponent punishes or rewards their choice of the interaction. Section 5 shows that, from a comprehensive list of ethical criteria, parties' choices exclusively link to social contract reasoning – *Kohlberg class 5* – the instrument for purely procedural preferences, controlling for critical latent variables which might intercept that link; section 6 presents the theory underneath our experimental design, and section 7 concludes.

[8] There are two psychometric tools to characterize individuals' moral judgement in a comprehensive way, Georg Lind's *moral competence/moral judgement test M-J/C-T* (Lind, 1978; Lind, 2008) and Rest's Defining Issues test (Rest, 1979; Bebeau, 2002), the first freely available for research purposes at https://moralcompetence.net, the second to be ordered from https://ethicaldevelopment.ua.edu. These tests are standard in other disciplines, the D-I-T having several methodological disadvantages such as straightforward fakeability (Barnett et al., 1995) and hence, from an Economist's point of view, sensitivity to ex-post rationalization and experimenter demand effects. (Chlaß et al., 2019) introduce the M-J/C-T to Economics to confirm the ethical foundation of the purely procedural fairness preferences they suggest. Contrary to the D-I-T, the M-J/C-T assesses moral competence and actual preferences for ethical criteria *separately* which is a *positive* approach and very much in line with Economics' methodological individualism. Cultural validation studies exist. Next to these two comprehensive inventories, tools such as Tanner et al.'s (2009) sacred values exclusively target the deontological dimension of individuals' moral judgement, leaving other – including deontological – criteria, uncontrolled; divide ethical criteria into coarser categories such that the distinction between outcome-based and outcome-independent criteria is less precise than required by the Economic preferences we study; new concepts such as moral identity operate on a frivilously narrow empirical basis, operationalize moral behaviour as a compound of preferences, attitudes and behaviour and most seriously, perhaps, use measurement tools such as the dictator game which, in themselves, depend on the the taxonomy and ethical criteria we use.

2 Lying, spying, sabotaging: rules and payoffs

Table 1 summarizes our setup. In the constant sum game on the left, two parties A and B each choose between L and R. Each option may yield the only nonzero payoff of the game; each option may therefore be preferred over the other in some contingency, each by the same degree, and each by the same degree by each party.[9] A and B therefore have equal decision rights. Neither knows the opponent's choice, and A and B are therefore also equally well off in terms of information.

Suppose now that, in this situation, B may set A's choice to a given option, may fabricate A's choice in a payoff-relevant way, or may look up which option A has chosen. By doing so, she transforms the game into a dictator game. To see this, suppose B chooses L herself, and sets A's choice to L, yielding strategy combination LL. Thereby, she ensures that A and B end up in the upper left cell where B earns Zero, and A earns all payoff. Looking at the transformed matrix on the right, we see that for LL^A, B quite rightly earns Zero, and A all payoff, whatever A's actual choice. A can no longer prefer L over R or vice versa in any contingency; she only has a single option she must necessarily adopt, and therefore, zero decision rights.

Table 1: How can B profit from sabotaging or fabricating A's decisions? Left: S_1 – equal decision rights; Right: S_2 – unequal decision rights

B \ A	L	R
L	100 0	0 100
R	0 100	100 0

$\xrightarrow[\text{Fabrication}]{\text{Sabotage}}$

B \ A	L	R
LL^A	100 0	100 0
RL^A	0 100	0 100
LR^A	0 100	0 100
RR^A	100 0	100 0

Notes. L: Left; R: Right; LL^A: B chooses L herself, and replaces (or fabricates) A's choice by L^A.

Spying, in contrast, affects B's information. Fig. A1 shows S_1 in its extensive form, as if A chose first, and B second. A does not know B's choice since it lies in the future; B does not know whether A 'has chosen' L or R, since the two possible 'histories' of what may have happened thus far, are tied to the same information set. If B spies A's choice, she moves each history to a separate, singleton, information set.

Next, we discuss how these concepts of lying, spying, and sabotage build upon previous work. Sabotage typically increases the opponent's cost of effort (Harbring et al., 2007), reduces (Harbring and Irlenbusch, 2011; Falk et al., 2008) or fabricates (Carpenter

[9] A player's decision rights are the number of her *effective* opportunities; an opportunity is effective if diverse in that a player can prefer it over all other options she has.

et al., 2010)[10] her output, in order to lessen the opponent's chance of winning a contest or earning a high wage. In each definition, sabotage relinks the sabotaged party's strategy to some other outcome, that is, to the expected payoff which lower effort and/or output would have yielded, see Appendix B. Note that, in this line of work, parties do not know which strategy they sabotage, and do not know if they are being sabotaged themselves. In our setup, B therefore overrides A's unknown choice; and A may, at the payoff stage, encounter the consequences of a strategy she did not choose.

Lying is often studied in deception games (Gneezy, 2005).[11] A sender with private information about the game's payoffs, falsely informs the receiver which of two options give the receiver more payoff. Thereby, the sender decrees which of the receiver's options the receiver actually prefers[12], and thus fully controls the receiver's decision rights. A truthful statement falls into the same category as a lie if the sender expects the truth to be mistaken for a lie (Sutter, 2009). In our setup, B therefore reports A's *unknown*[13] choice to C who invariably implements the choice. The resulting message is unmistakeably a lie by which B fully controls A's decision rights. B's lie is black if she takes all payoff for sure; white, if she gives all payoff to A, and neither white nor black if she tosses a fair coin to decide wether to give or take all payoff. Occasionally, senders in deception

[10]In this instance, a party who sabotages must necessarily lie; the authors therefore also refer to the literature on lying in the context of sabotage and use the notion 'sabotage' in a generic way.

[11]Another strand of research studies lies as individual decision making. Subjects report the outcome of a random mechanism such as a die roll which is their private information. Lies do not harm any opponent and do not affect an opponent's decision rights, unless the experimenter herself is considered whose research funds the lie depletes. There are two implications for our setup: first, the ethical criterion we study is not at play – and one may well ask of what economic interest a lie can be which is made privately and without any effect other than for the self – if, as soon as there are effects on others, new ethical criteria apply –; second, the setup puts the experimenter center stage as the only potentially harmed party. Indeed, one main reason not to lie in such setups, are reputational and social image concerns (Abeler et al., 2019) which *require* consideration of and for an audience. It is worrisome that precisely in this setup, lying aversion does not seem to depend on payoffs (as if subjects still expected to be observed by the experimenter) whereas lying in deception games *does* depend on senders' incentives to do so. Kajackaite and Gneezy, 2017 therefore move the random mechanism (die roll) into players' heads where it literally becomes unobservable, and indeed, a significant share of lies in the individual decision task starts responding to incentives. Gneezy and Kajackaite, 2020 further modify the game such that the individual decision task has a payoff externality on a passive second player. Subjects lie less, but not once the stakes are increased. There is underreporting (similar to the white lies in our setup), and subjects do not lie more when being truthful makes them worse off than their passive counterpart (which makes a bad case for disadvantageous inequality aversion). In our setup, C is B's audience and C, similar to the experimenter in the individual decision task mentioned, does not see whether B lied intentionally, i.e. paid for S_2 and fabricated A's choice, or paid for S_1, but by chance, found herself bereft of an option to tell the truth. Gibson et al., 2013 embed an individual decision task (whether ot not to accurately report a company's earnings) into a sender (CEO)-receiver (market stakeholders) frame: senders are asked to imagine a hypothetical setting in which their message governs whether stakeholders prefer one of their options over another, i.e. holding the stake vs. selling. Indeed, senders' cost of lying unexpectedly depends on the potential harm caused to (the purely hypothetical) stakeholders *and at the same time*, also on senders' *sacred values*, i.e. outcomes and outcome-invariant values are linked (which, conceptually, they should not be), further complicated by the observation that the cost of lying varies *within* individuals across situations.

[12]Contrary to sabotage, lying does not actually relink the strategy to another outcome, but makes the receiver believe her strategy has an outcome it does not actually have.

[13]This 'subtle form of lying' (Gill et al. 2013, pp. 121) fabricates a fact; the exact message B sends says *'A has chosen:'*, followed by either *'L'* or *'R'*, see the screen in Appendix E.

games are given a third option to avoid the decision to lie, either by not sending any message (Sánchez-Pagés and Vorsatz, 2009), or by delegating the task to another party (Erat, 2013). In our setup, B can do so by opting into S_1.

Spying collects information about an opponent's moves (Lee, 2023) and to disentangle it from mere information acquisition, must take on a form of covert surveillance (Klebe Treviño and Weaver, 1997) in that the opponent is at least uncertain of being observed. As a result, the sequentality of moves becomes uncertain (Penta and Zuazo-Garin, 2022). In Lee, 2023, the opportunity to spy in a constant sum game is costless and exogenously given with a commonly known probability. In our setup, B herself sets the likelihood of observability which is costly: she *actively* decides to spy A's choice by opting into S_2. The likelihood of observability remains B's private information.[14]

Throughout, we study fabrication, spying, and sabotage as *clandestine* activities. A does not know whether B spies, fabricates, or sabotages. She cannot distinguish between S_1 and S_2, and therefore ignores which of her options she actually prefers. B sets the likelihood of S_2 at a small cost of maximally 5 ECU (25 Euro Cents), comparable to a nudge.[15] The default is the toss of a fair coin. Two additional setups sever or weaken the link between fabrication, sabotage, spying, and B's concern for A's lack of decision rights: Appendix C redefines S_1 in a way which allows B to take all payoff while A always has zero decision rights. B can now fabricate, spy, and sabotage without affecting the latter. Appendix D adds symbolic punishment and reward to the setup. A submits a costly punishment and reward schedule for B's choice of S_2 in which A reduces or increases B's payoff by up to 30 units. A's decision rights increase, but fabrication, spying, and sabotage are still in B's rational self-interest.

[14]There are forms of spying which affect information *and* decision rights as in (Solan and Yariv, 2004). This occurs when a party first chooses an option and then spies her opponent's move whereupon she may revise her own choice. In this case, the party who spies obtains additional decision rights by spying: she may either stay with her initial choice at some future stage, or change it.

[15]Nudging oneself into S_1 or S_2 could be a party's choice to walk to her own desk without passing her colleague's (or deliberately passing the latter, respectively) in order to forego (or obtain) the chance to spy or manipulate that colleague's progress. Similarly, it could be avoiding the coffee corner to prevent being part in creating or spreading rumours about others.

3 Experiment

The experiment proceeded in three parts as shown in table 2. At the outset, subjects were seated at visually isolated computer terminals, and handed a hard copy of the German instructions for our baseline treatment.[16] Instructions for two upcoming parts 2 and 3 were shown on screen, once the experiment had proceeded this far; at no point in time did subjects have information about any upcoming parts. Once participants had confirmed on screen they had read the instructions, the experiment started automatically by a set of control questions which all participants answered successfully. Subsequently, participants were randomly assigned either role A, B, or C. A and B participants were randomly matched into pairs and two C participants assigned to each session of treatment LIE. Next, Bs chose between two situations S_1 and S_2 at their own discretion, S_1 offering symmetric decision and information rights, and S_2 affording B the opportunity to either LIE, SPY, or SABOTAGE. A and B next submitted their choices for the situation determined by B – A's options being identical across S_1 and S_2 such that the situation remained B's private information always. No feedback was given after part 1. Part 2 proceeded the same way, except that A was announced to have an option to punish or reward B's choice of the situation. Again, B chose between S_1 and S_2 after which both A and B made their decisions for the situation selected by B. No feedback was given after part 2. Part 3 elicited risk and envy preferences, demographics, administered a pen-and-paper moral judgement test, and the pen-and-paper ranking scales for materialist and postmaterialist values. Only one of the first two parts was paid out, part 3 was always paid; average payments included a show-up fee of €2.50 and amounted to €7.94 (min: €3.60, max: €12.10) where €1 $\widehat{=}$ \$1.28 at the time. 630 subjects participated, 49% of them female.[17]

3.1 Part 1: Baseline Treatments LIE, SPY, and SABOTAGE

Figure 1 formalizes part 1. A and B have an initial endowment of 50 ECU, the show-up fee of €2.50. B moves first and chooses the probability Prob (S_2) for situation S_2 which is initialized at 50%. This default has two purposes: first, it portrays an unintentional choice and second, does not point subjects toward either S_1 or S_2. Each one percent change to this default costs B 0.1 ECU where 1 ECU = €0.05. B may therefore select one situation for sure at the relatively small cost of 5 ECU or 25 Euro Cents. Next, situations S_1 and S_2 are drawn according to B's choice of Prob (S_2). A neither knows Prob (S_2), nor the situation which is drawn. She chooses between L (left), R (right), and the toss of a fair coin between the two. B's choices in turn depend on the situation which is drawn. If S_1 is drawn, B's choices are the same as A's, and neither A nor B know the opponent's choice. If S_2 is drawn in treatment SPY, B's choices are the same as A's, but B *sees* A's choice. If S_2 is drawn in treatment SABOTAGE, B overrides

[16]See appendix F for translations of these instructions into English for our three baseline treatments LIE, SPY, and SABOTAGE.

[17]A session lasted approximately 50 minutes including payment. Subjects were undergraduate students and native German speakers at Friedrich-Schiller-University Jena, randomly recruited from all fields of study via ORSEE (Greiner, 2004). At the time of the experiment, the subject pool counted around 3000 students. The experiment was programmed in z-Tree (Fischbacher, 2007). Payouts were distributed in sealed envelopes; receipts did not match subjects' names with their client numbers.

Treatment	*Spy*		*Sabotage*		*Lie*	
Payoff regime	Neutral	Competitive	Neutral	Competitive	Neutral	Competitive
B-participants	# 52	# 54	# 53	# 53	# 47	# 44
Part 1	**Baseline** B chooses probability Prob(S_2) of situation S_2 A chooses L or R In S_2, B learns A's choice	In S_2, B overrules A's choice	In S_2, B transmits A's choice to C B chooses L or R					
Part 2	**Reward and Punishment** B chooses probability Prob(S_2) of situation S_2 A chooses L or R In S_2, B learns A's choice	In S_2, B overrides A's choice	In S_2, B transmits A's choice to C B chooses L or R A chooses punishment/reward schedule without knowing Prob(S_2), the situation, or B's choice. B submits 1st order beliefs about A's punishment and reward schedule.					
Part 3	**Covariates** Risk Preferences Envy Moral Judgement Test (pen and paper) Materialist and Postmaterialist values (pen and paper) Demographics					

Table 2: Experimental design

A's unknown choice by L (left) or R (right), and chooses for herself between L (left), R (right), and the toss of a fair coin betweeen the two. If S_2 is drawn in treatment LIE, B transmits some choice for A and her own choice to participant C who implements the choices transmitted. Throughout SPY, SABOTAGE and LIE, A and B have an equal ex-ante chance to obtain the payout of a constant sum game if B selects S_1; in S_2, B has all allocation power and can secure this payout. Fabrication, spying, and sabotage therefore turn the constant sum game S_1 shown in table 3a into dictator game S_2. Thereby, SABOTAGE and LIE allow B to take decision rights from A, whereas SPY allows B to increase her own information rights. In a second variant *payoff neutrality*, S_1, too, is a dictator game such that B's power to take A's decision rights is removed from all treatments and B may fabricate and sabotage without affecting A's decision rights in any way. Table 3b shows S_1 in variant *payoff neutrality*: since A can no longer prefer either L over R, or vice versa, she has zero decision rights and B dictates the allocation also in S_1, without, however, resorting to fabrication, sabotage, or spying.

competitive payoffs

		A: L	A: R
B	L	$u_B^* = 0$, $u_A = 100$	$x_B^* = 100$, $x_A = 0$
	R	$v_B^* = 100$, $v_A = 0$	$y_B^* = 0$, $y_A = 100$

payoff neutrality

		A: L	A: R
B	L	$u_B^* = 0$, $u_A = 100$	$x_B^* = 0$, $x_A = 100$
	R	$v_B^* = 100$, $v_A = 0$	$y_B^* = 100$, $y_A = 0$

Table 3: Payoffs in S_1.

Note: Table 3a on the left reviews A's and B's payoffs in S_1 for treatment *competitive payoffs*, table 3b on the right reviews A's and B's payoffs in S_1 for treatment *payoff neutrality.* Thereby, u_B^* disregards the cost B has incurred from choosing Prob(S_2), that is, $u_B^* - 0.1 \cdot |50\% - \text{Prob}(S_2)| = u_B$ where u_B denotes B's actual payout.

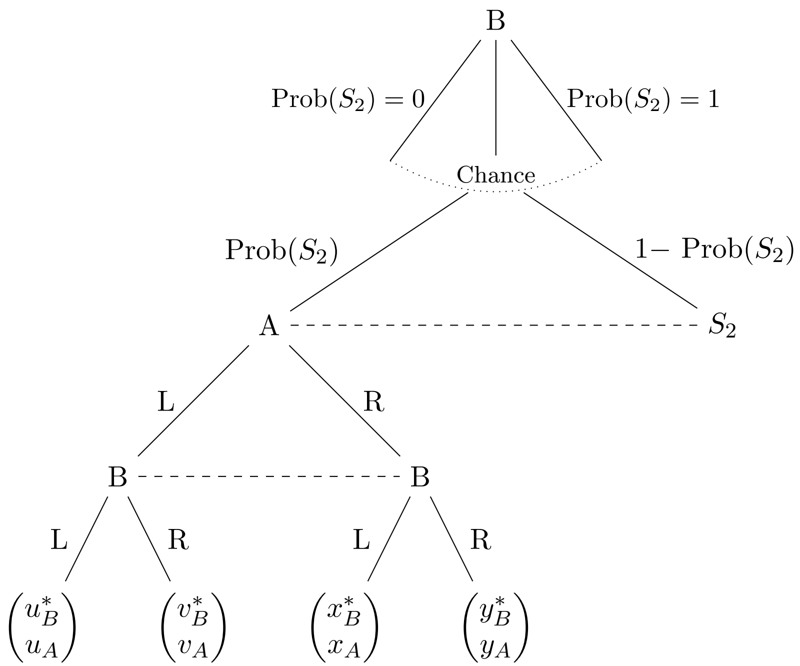


Figure 1: Basic Game Structure

Note: This tree illustrates our baseline treatments from table 2. S_2 is a place holder for Figure 2 in treatment SPY, for Figure 3 in treatment SABOTAGE and for Figure 4 in treatment LIE.

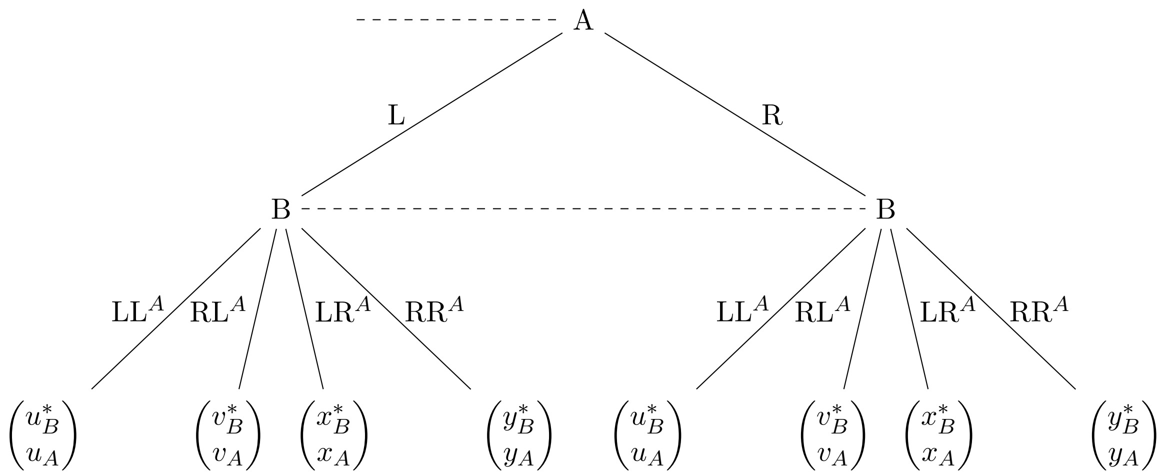


Figure 3: S_2 in treatment SABOTAGE.

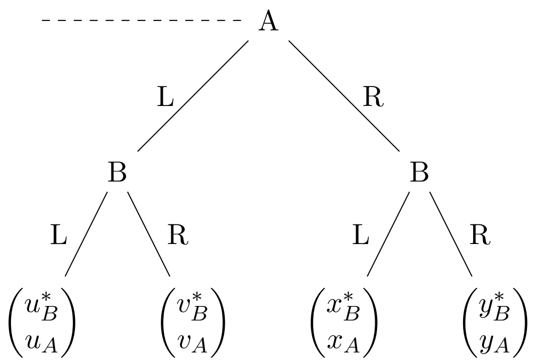


Figure 2: S_2 in treatment SPY.

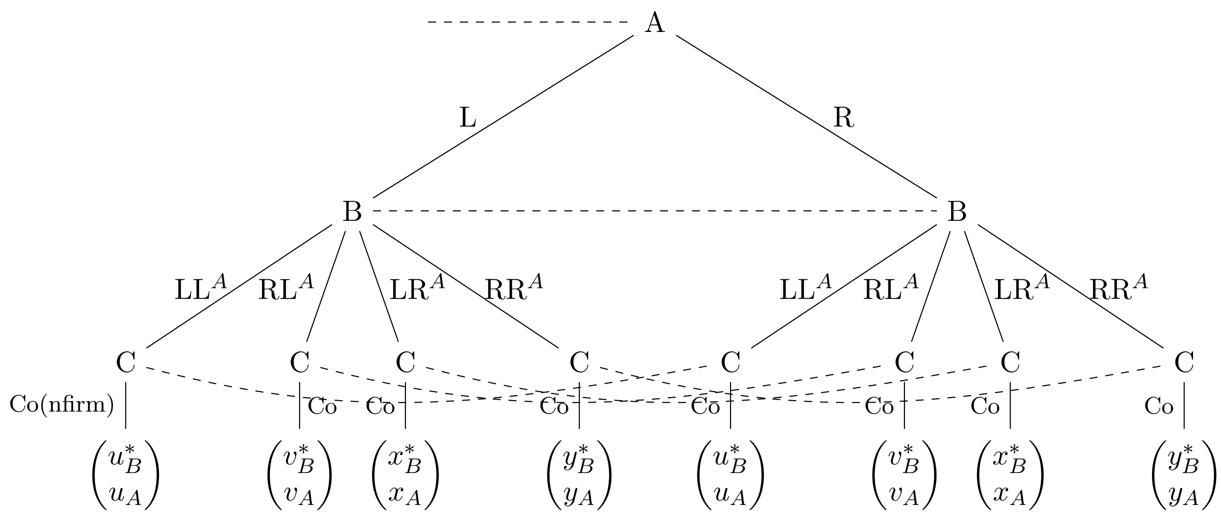


Figure 4: S_2 in treatment LIE.

If B's concern for A's decision rights makes her averse to fabrication and sabotage, this aversion will disappear in *payoff neutrality* and LIE, SPY, and SABOTAGE yield similar results. *Payoff neutrality* is worded identical to *competitive payoffs* such that any difference in wording between LIE, SPY, and SABOTAGE is preserved along with its potential effects. Appendix E shows screen shots for B's choice of Prob(S_2) in Fig. A4, for situation S_1 in Figure A5, situation S_2 SPY in Figure A6, S_2 SABOTAGE in Figure A7, and S_2 LIE in Figure A8. Note that throughout S_1 and S_2, B, in addition to her choices L (left) and R (right), is given the explicit option to toss a fair coin. This way, B can always equalize A's and B's chances of obtaining all payoff which, coincidentally, is also a feature of the equilibrium solution for S_1 as we discuss in theory section 6. Having already had the opportunity to randomize between S_1 and S_2, B participants, contrary to our concerns, never use this option in the experiment.

3.2 Part 2: Giving A a symbolic punishment or reward option

In part 2 of each session, A and B repeat part 1 with a new opponent, knowing that A can punish or reward B's choice of Prob(S_2). This affords A new decision rights which grant A some control over B's freedom of choice in that she can magnify or reduce the degree by which B prefers S_1 over S_2. Again, we expect B's concern about A's lack of decision rights to crowd out. In particular, A submits a punishment and reward schedule in which she may subtract up to 30 ECU, or may add up to 30 ECU to B's payoff, depending on whether B chooses S_1 (1) for sure, (2) with Prob(S_1) $\in [75\%, 99\%]$, (3) with Prob(S_1) $\in]50\%, 75\%[$, (4) with Prob(S_1) $= 50\%$, or chooses S_2 with (5) Prob(S_2) $\in]50\%, 75\%[$, with (6) Prob(S_2) $\in [75\%, 99\%]$, or (7) for sure. Any 1 ECU change to B's payoff costs A 1 ECU. B participants submit their beliefs about A's punishment and reward schedule. The correct guess of A's entire schedule earned B 35 ECU, the correct guess for any of the seven cases above, earned B 5 ECU. For each ECU by which B misguessed A's actual plan, B earned 0.08 ECU less. Figure A9 in appendix A shows the corresponding screen shot. Appendix D shows S_1 and S_2 from B's point of view: In S_1, B has suddenly lesser decision rights than A whereas in S_2, B still retains greater decision rights but cannot reduce A's to zero.

3.3 Part 3/ Controls and Instrumental Variable

Part 3 began by eliciting *envy* (Kirchsteiger, 1994) to see how much B participants dislike being materially worse off than others. To this end, subjects were randomly rematched with a new opponent and submitted their choice between "10 ECU for themselves and 10 ECU for the other" or "10 ECU for themselves and 20 ECU for the other". A fair coin determined whether their own, or their opponent's decision would be payoff-relevant (Bartling et al., 2009). Part 3 also elicited *risk preferences* in a Holt-Laury price list format (Holt and Laury, 2002) with subjects choosing ten times between a lottery and a sure payoff of 25 ECU. Each lottery paid either 10 or 35 ECU whereby lotteries systematically increased the chance of paying 10 ECU by 10%.

Next, an on-screen announcement pointed to a copy of Lind's (1978, 2008) standardized moral judgement test (M-J-T) placed upside down at the side of each desk. All information pertaining to the name or purpose of this test[18] had been removed. The test draws upon an inventory by Jean Piaget and Lawrence Kohlberg (Piaget, 1948; Kohlberg, 1969; Kohlberg, 1984) who, in the 20th century, conducted extensive field research to observe and classify which criteria individuals use to make moral judgements. The test elicits Bs' preferences over these criteria: if, and by how much she uses a given criterion to judge whether a course of action is ethically right.

As by *Kohlberg class 1* and *2*, individuals deem those actions ethically right which are either not punished in material terms, or are rewarded instead. By *Kohlberg class 3*, individuals judge actions ethically right if the latter comply with a social norm, with others' expectations, were done with a good intention, or assist their social image with their peers. By *Kohlberg class 4*, individuals resort to the law, and to the idea of maintaing the status quo and the social order to judge whether an action is ethically right. By *Kohlberg class 5*, an action is deemed right if it respects parties' equality rights granted by a democratic social contract, and by *Kohlberg class 6*, if it satisfies some universal principle of conscience such as parties' human rights, parties' right to state their own will, or their human dignity. Chlaß et al. (2019) show in particular, that *purely procedural preferences* link to subjects' *Kohlberg class 5* scores and point out which demographic data might intercept this link.

The test introduces two vignettes, a first portraying workers who break into a factory in order to find and steal evidence that management was listening in on them, and a second, portraying a woman who is fatally ill and asks a doctor to medically assist her suicide. After each vignette, subjects are asked for their opinion whether or not the respective protagonists' behaviour was right or wrong. Next, the test lists 24 arguments (12 arguments after each vignette, six to judge the behaviour in question was wrong, each pertaining to one *Kohlberg class*; another six to judge it was right) and asks subjects how much they would agree or disagree on a nine-point Likert scale to judge the protagonists' course of action by each argument. In sum, we obtain four ratings per subject for each of the six *Kohlberg classes*, and a set of six preferences. Thereby, the test is constructed such that subjects who do not give their actual opinion in the test, answering, for instance, in what they deem a socially acceptable way, do not succeed in biasing the sample distribution of scores but add noise to the latter.

The experiment resumed with a payoff screen after which subjects submitted their age, gender, field of study, semester, and the type of degree they were studying for. Thereof, relevant controls for *Kohlberg class 5* scores are *field of study: Law*, and *gender*; relevant controls for *Kohlberg class 6* are *age*, *gender*, and *fields of study: Law*, *IT*, *Education*, and *Medicine*.

Finally, subjects filled in a questionnaire to elicit their *materialism* and *postmaterialism* values (Inglehart, 1977; Baker and Inglehart, 2000; Klages and Gensicke, 2006)

[18]Freely available for research purposes from Georg Lind's webpage at http://moralcompetence.net. Appendix K reproduces a standardized English version. See also:

where materalists appreciate power, order, obedience, and hierarchy, whereas postmaterialists value individualism, autonomy, and self-fulfillment. Some people may seek and condone power to put to rights what they see as ethically wrong, trading off monetary value against power, whereas others may deem that some individual rights are inalienable and must be reinstated; if such attitudes exist, they may explain why some subjects amend their opponents' rights whereas others seek additional rights to compensate the opponent materially. In (Chlaß et al., 2019), both behaviours were observed and linked to *Kohlberg class 5* and would, in our setup, imply postmaterialist B participants to opt into S_1, and materialist B participants to opt into S_2 and give all payoff away.[19]

3.4 Summary of Treatments

purely procedural aspects:	competitive payoffs[11]			payoff neutrality			competitive pun/rew[12]			payoff neutral pun/rew		
decision rights ↓	LIE	SPY	SAB	LIE	SPY	SAB	LIE	SPY	SAB	LIE	SPY	SAB
A has decision rights	+	+	+	-	-	-	+	+	+	+	+	+
B can take some of A′s decision rights	+	-	+	-	-	-	+	-	+	-	-	-
B can take all of A′s decision rights	+	-	+	-	-	-	-	-	-	-	-	-
wording of instructions is identical between treatments:												

[11] LIE + SAB competitive. S_1 – A and B have equal decision rights. S_2 – B has *greater* decision rights.
[12] LIE + SAB competitive pun/rew: S_1 – B has *lesser* decision rights than A. S_2 – B has *greater* decision rights than A.

Hypothesis 1 – B's concern for A's decision rights causes high levels of altruism. In LIE and SABOTAGE with competitive payoffs, many Bs therefore depart from rational self-interest, but not in SPY where B exerts no influence over A's decision rights.

Hypothesis 2 – These results by the Rubin causal model are confirmed by an instrumental variable: B's altruism links to B's Kohlberg class 5 scores *after controlling for latent correlates of the latter which might intercept the link.*

Hypothesis 3 – As B's influence over A's decision rights declines, so does her altruism, dropping significantly in LIE/SABOTAGE payoff neutrality, and in LIE/SABOTAGE with punishment/reward. Residual altruism does not link to B's Kohlberg class 5 scores.

[19]We elicit these value groups by the 'Speyerer value inventory' (Klages and Gensicke, 2006) which consists of 12 items to be rated on a seven point Likert scale (1 – not important at all, to 7 – very important). Three items load on a first scale *'duty and acceptance values'*, four on a second *'hedonistic and materialist values'*, and three on a third, *'idealistic values and political participation'*. Typically, five value groups (clusters) emerge; amongst them *'conventionalists'* – Inglehart's original materialists, and so-called *'idealists'* – Inglehart's original postmaterialists. We use individuals' absolute ratings of all three scales for our analysis. Klages and Gensicke's measurement instrument has three main advantages over Inglehart's in our setup: first, the items being directly validated on German samples, second, the use of separate scales for materialism and postmaterialism values (Inglehart obtains these as opposite ends of the same scale; they are therefore by construction consistent and cannot be used to check the other) and third, the possibility of hybrid value groups which, in Inglehart's measurement, need to be post-assigned to the only two value groups allowed. For details, see appendix L; a concise review of Klages' research in English is found in (Borg et al., 2019) who also show that Klages' three scales emerge as the first three principal components of the popular Schwartz' portrait value scales.

4 Results

4.1 Descriptives: B's choice of situation and allocation

4a)[20] LIE, SPY, SABOTAGE – competitive payoffs

treatment (obs.) →	LIE (44)		SPY (53)		SAB (54)	
situation →	S_1	S_2	S_1	S_2	S_1	S_2
Bs who pay for S_1 or S_2, respectively...	9 20%	5 11%	5 9%	36 68%	2 4%	37 69%
...set Prob(S_1) or Prob(S_2) to median	0.6	0.7	1	0.8	0.7	0.8
Bs in each situation	19	25	13	40	26	28
Bs who give A all payoff in S_2		17 68%		0 0%		20 71%

4b) LIE, SPY, SABOTAGE – payoff neutrality

treatment (obs.) →	LIE (47)		SPY (53)		SAB (52)	
situation →	S_1	S_2	S_1	S_2	S_1	S_2
Bs who pay for S_1 or S_2, respectively...	8 17%	3 6%	2 4%	19 36%	4 8%	18 35%
...set Prob(S_1) or Prob(S_2) to median	0.6	0.8	0.8	0.7	0.7	0.7
Bs in each situation	25	22	20	33	22	30
Bs who give A all payoff in S_1 or S_2	5 20%	5 23%	2 10%	2 6%	4 18%	4 13%

4c) LIE, SPY, SABOTAGE – competitive payoffs punishment/reward

treatment (obs.) →	LIE (44)		SPY (53)		SAB (54)	
situation →	S_1	S_2	S_1	S_2	S_1	S_2
Bs who pay for S_1 or S_2, respectively...	7 16%	13 30%	2 4%	37 70%	4 7%	36 67%
...set Prob(S_1) or Prob(S_2) to median	0.6	0.7	0.7	0.9	0.6	0.7
Bs in each situation	20	24	16	37	14	40
Bs who give A all payoff in S_2		12 50%		0 0%		19 48%

4d) LIE, SPY, SABOTAGE – payoff neutrality punishment/reward

treatment (obs.)→	LIE (47)		SPY (53)		SAB (52)	
situation →	S_1	S_2	S_1	S_2	S_1	S_2
Bs who pay for S_1 or S_2, respectively...	6 13%	6 13%	4 8%	15 28%	10 19%	11 21%
...set Prob(S_1) or Prob(S_2) to median	0.7	0.6	0.6	0.7	0.7	0.8
Bs in each situation	30	17	25	28	29	23
Bs who give A all payoff in S_1 or S_2	5 17%	1 6%	1 4%	4 14%	7 24%	0 0%

Table 4: B's choice of situation and her choice of situation by allocation

Tables 4 list, how many B participants pay for situation S_1, how many for S_2, which probability they set for their preferred situation, and which allocation B participants impose if given the opportunity. Table 4a summarizes our baseline treatments with *competitive payoffs*. In LIE, 20% (9 of 44) B participants pay for S_1, compared with 9% (5 of 53) in SPY and 4% (2 of 54) in SABOTAGE. 11% (5 of 44) B participants pay for S_2, compared with seven times as many, i.e. 68% (36 of 53), in SPY and 69% (37 of 54) in SABOTAGE. In sum, significantly *fewer* B participants fabricate than spy or sabotage by Fisher's exact tests, all *p-values* < 0.02. Turning to altruism, 68% (17 of 25) B participants in S_2 give all payoff to A in LIE, *none* of the 40 B participants in S_2 does so in SPY, and 71% (20 of 28) do so in SABOTAGE. Most altruism – most departures from rational self-interest – does therefore occur, when rational self interest requires B to impair A's decision rights.

Result 1. In LIE and SABOTAGE with competitive payoffs, significantly more B participants give all payoff to A than in SPY with competitive payoffs where B's only source of power is her advantage in information (Fisher's Exact tests, p-value < 0.01).

[20] Reading example: In treatment LIE, there are 44 B participants, 5 of which (11%) pay for S_2 and set Prob(S_2) to median 0.7. 25 B participants arrive in S_2, 17 of which (68%) give all payoff to A.

Table 4B summarizes LIE, SPY, and SABOTAGE under *payoff neutrality.* Roughly as many B participants as before pay for S_1, but only half as many for S_2. 21% (S_1: 5 + S_2: 5 = 10 of 47) give all payoff to A in LIE and 15% (S_1: 4 + S_2: 4 = 8 of 52) do so in SABOTAGE which are significiantly fewer than before by Fisher's exact tests, all *p-value* < 0.001. Treatment SPY remains unchanged by Fisher's exact test, *p-value* $= 0.136$ with 8% (S_1: 2 + S_2: 2 = 4 out of 53) giving all payoff to A. Again, altruism decreases where self-interest does not impair A's decision rights.

Turning to tables 4C and D, symbolic punishment and reward sustains a considerable level of altruism, maintained, however, by a largely different set of individuals as shown by the contingency tables in appendix G. Roughly 40% of Bs opt for a different situation in LIE and SABOTAGE, some 30% do so under *payoff neutrality.* Altruism among altruists from part 1 drops by one third in LIE, by two thirds in SABOTAGE, and more strongly so under *payoff neutrality*, i.e. by 80% in LIE and 88% in SABOTAGE. Throughout, behaviour in SPY is least affected. Symbolic punishment and reward might therefore indeed crowd out B's concern for A's decision rights, if, in addition, a new ethical criterion – preferably referring to *reward and punishment* – were at play.[21] Figure 5 illustrates Bs' choice of the situation as by the allocation they impose, for all treatments.

Result 3. As B's influence over A's decision rights decreases, so does her altruism: in LIE/SABOTAGE with payoff neutrality and LIE/SABOTAGE with punishment/reward.

4.2 Descriptives: Bs' beliefs

Figure 6A) illustrates that Bs believe to be rewarded for opting into S_1, and less so as this choice tends toward the toss of a fair coin. At this point, they expect neither reward nor punishment. Bs believe As to punish S_2, and increasingly so as S_2 becomes certain. Expected average punishment is 9.77 ECU for Prob(S_2) = 1, 7.60 ECU for Prob(S_2) $\in [75\%, 99\%]$, and 5.27 ECU for Prob(S_2) $\in\,]50\%, 75\%[$, each category significantly greater than the next.[22] Bs therefore believe that S_2 – fabrication, spying, and sabotage – is undesirable in As' eyes. Appendix I shows that this pattern is strongest in SPY where Bs opt most frequently into S_2, less strong in SABOTAGE, and least so in LIE where Bs hardly opt into S_2. Bs' choice of Prob(S_2) and their beliefs about what As wish them to do therefore vary at odds with each other across LIE, SPY, and SABOTAGE. This paradox is apparent from individuals' beliefs pertaining to their

[21] If no ethical criterion at all were at play, Bs might simply have adopted new behaviours to keep the task interesting. If previously selfish Bs felt guilt, *others' expectations*, i.e. *Kohlberg class 3*, would explain Bs' choices. In section 5, we show that Bs' choices link to *Kohlberg class 1* which derives the right course of action from material punishment and reward. Note that if guilt were at play, at least *some* selfish Bs in SPY where Bs expect to let down A particularly strongly, should turn altruistic in part 2. Nobody does.

[22] Bs expect punishment to be highest for Prob(S_2)=1 (SPY: Wilcoxon Signed Rank *p-value* < 0.001, SABOTAGE: *p-value* < 0.032, LIE: *p-value* < 0.043), second highest for Prob(S_2) $\in [75\%, 99\%[$ (SPY: *p-value* < 0.001, SABOTAGE: *p-value* < 0.005, LIE: *p-value* $= 0.23$), third highest for Prob(S_2) $\in [50\%, 75\%[$, and least for Prob(S_2) = 50% (SPY: *p-value* < 0.001, SABOTAGE: *p-value* < 0.001, LIE: *p-value* < 0.09). In LIE where the order is least pronounced, average aggregate punishment for S_2 is weakly significantly larger than for the toss of a fair coin, *p-value* < 0.059.

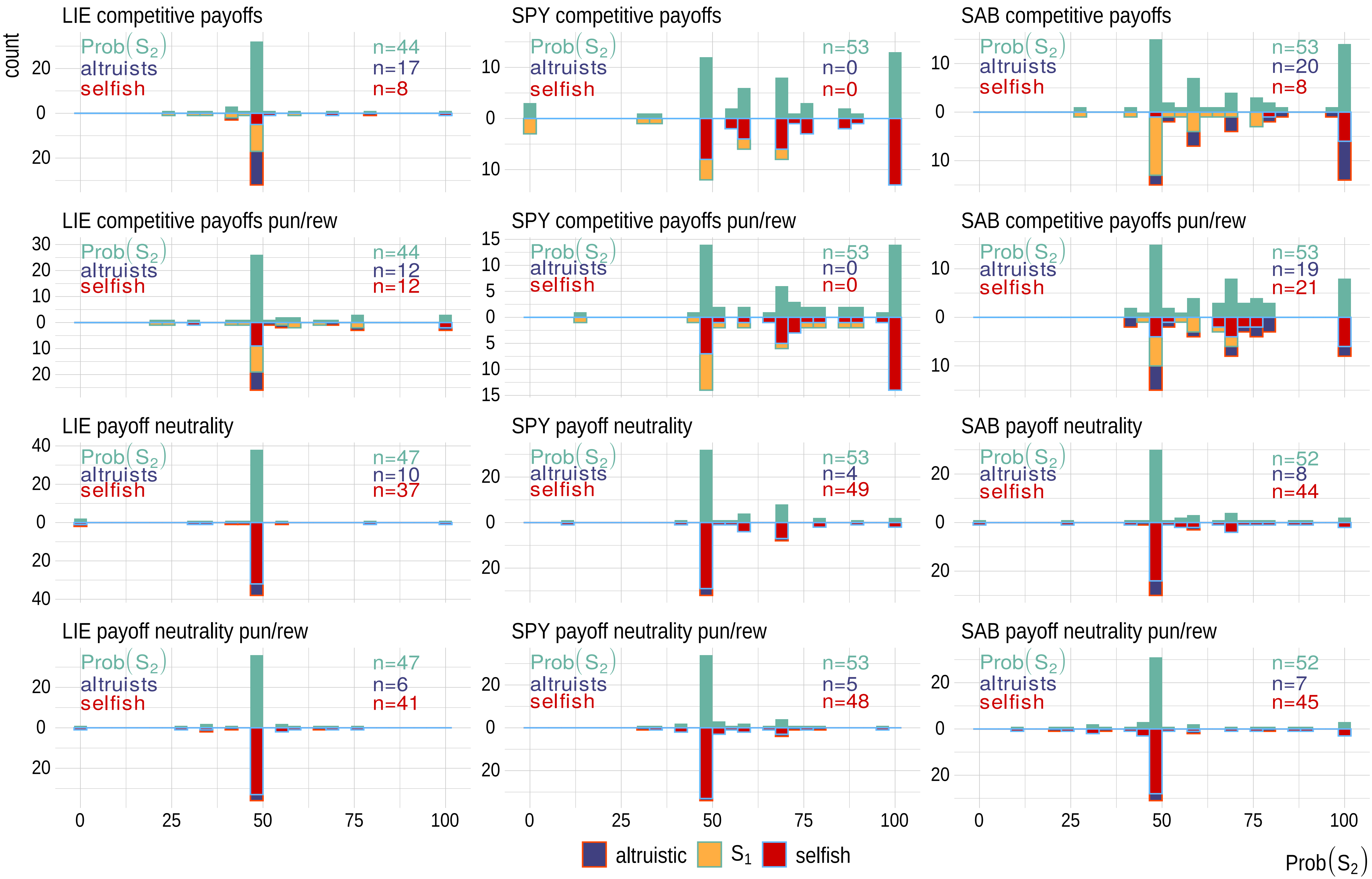
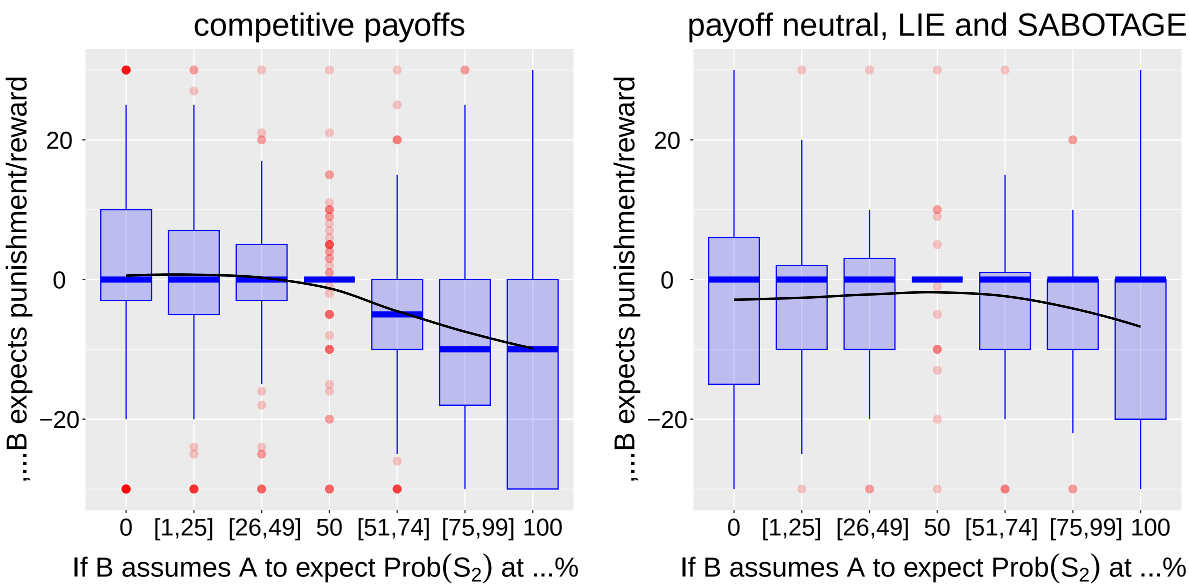


Figure 5: B's choice of situation, mirrored against her choice of situation by allocation imposed.

actual choice of Prob(S_2), as well as from their beliefs about the entire (hypothetical) choice set.[23] Arguably, punishment beliefs also provide access, however imperfect, to social norms which might regulate lying, spying, and sabotage differently. If social norms guide Bs' beliefs about As' punishment, the social norm against spying turns out strongest, followed by the norm against sabotage, and then lying. Beliefs and choices are logically linked: Bs who opt into S_2 and take all payoff, make S_2 as likely as possible while keeping punishment at a reasonable level; Bs who opt into S_1 or toss a fair coin, expect As to punish S_2 significantly more than their actual choice. Bs who give all payoff to A show no such belief patterns.[24] Thereby, beliefs and moral judgement are not linked in our data, and seem to describe what Bs believe As *actually* do, rather than *should* do.

Figure 6B) shows that when A has zero decision rights always, Bs expect punishment for every intentional choice, increasing in its intentionality. In these cases, exerting one's rights to choose the procedure when one dictates the allocation always, increases the asymmetry in decision rights even further. Bs expect to be punished for Prob(S_2) $= 0$ (p-*value* < 0.034), Prob(S_2) $\in\ ]0\%, 25\%]$ (p-*value* < 0.02), Prob(S_2) $\in\ ]25\%, 50\%[$ (p-*value* < 0.008), not for tossing a fair coin (p-*value* $= 0.350$), and again for Prob(S_2) $\in\ ]50, 75[$ (p-*value* < 0.007), Prob(S_2) $\in\ [75\%, 99\%[$ (p-*value* < 0.002) and Prob(S_2) $= 1$ (p-*value* < 0.001). In SPY payoff neutrality, the original punishment belief pattern remains intact.

Figure 6: B's beliefs about A's decision to punish or reward B's choice of $Prob(S_2)$. Left: 6a) competitive payoffs; Right: 6b) payoff neutrality.



[23] Bs expect more punishment for Prob(S_2) $> 50\%$ in SPY than in LIE (Wilcoxon Rank Sum, p-*value* < 0.001) or SABOTAGE (p-*value* < 0.034). In LIE where Bs rarely opt into S_2, Bs expect *less* punishment for S_2 than in SABOTAGE (p-*value* < 0.036) where 69% opt into S_2. Similarly, Bs expect more severe punishment for their actual choice in SPY than in LIE (p-*value* < 0.016) or SABOTAGE (p-*value* < 0.026).

[24] In SABOTAGE, Bs who set Prob(S_2) $\leq 50\%$, opt for an average of 47.5% (SPY: 47.5%), expecting greater punishment for Prob(S_2) $= 1$ (Wilcoxon Signed Rank tests, p-*value*$= 0.002$, SPY: p-*value*$= 0.002$), for Prob(S_2) $\in\ [75\%, 99\%]$ (p-*value*< 0.001, SPY: p-*value*$= 0.006$), and for Prob(S_2) $\in]50\%, 75\%[$ (p-*value*< 0.001, SPY: p-*value*$= 0.060$) than for their own actual choice. Bs who take all payoff, set Prob(S_2) to an average 79.94% (SPY: 85.77%), expecting greater punishment for Prob(S_2) $= 1$ (p-*value* < 0.021, SPY: p-*value*$= 0.001$) and for Prob(S_2) $\in]75\%, 99\%]$ (p-*value* < 0.312, SPY: p-*value* < 0.011) than for their actual choice, but lesser punishment for Prob(S_2) $\in]50\%, 75\%]$ (p-*value*< 0.063, SPY: p-*value*< 0.177) and all categories Prob(S_2) $< 50\%$ (p-*values*< 0.010). Bs who give all payoff to A, set Prob(S_2) $> 50\%$ to an average Prob(S_2) $= 76.08\%$ and do not expect greater or lesser punishment for other choices.

5 Ethical criteria at work

Next, we study which ethical criteria – if any – underlie B's decision not to opt into S_2 and secure all payoff. B might, for instance, avoid the option out of concern for her social image, in order not to disappoint A's expectations (Battigalli and Dufwenberg, 2007), not to violate some, or several, social norms[25], or in order to signal her own generous intentions (Falk and Fischbacher, 2006).[26] In the previous section, we saw that Bs are particularly selfish where they expect A to punish this selfishness most: a desire to avoid letting A down or to comply with a social norm would have implied a different pattern of altruism across LIE, SPY, and SABOTAGE. Finally, B might deem that fabrication, sabotage and spying violate the opponent's civil rights granted by the social contract (Chlaß et al., 2019), or that stripping the opponent of all freedom to choose violates her human rights and dignity (Chlaß and Moffatt, 2017).

If indeed, Bs' altruism arose from a concern purely about As' decision rights, Bs' decision to opt into S_2 and give all payoff to A must link to *Kohlberg class 5* which regroups criteria around the equality of rights as stipulated by a democratic social contract. Chlaß et al. (2019) identify a link between the latter and individuals' willingness to pay for changes in the information and decision structure of a formally defined game when these changes are either of no, or against individuals' material self-interest. The link occured in particular where individuals paid to improve the *opponent's* position of rights. The link at hand was intercepted by two demographic variables, i.e. *field of study: Law*, and *gender*. Any link between Bs' altruism and *Kohlberg class 5* must therefore be robust to including these as well as the complete set of six Kohlbergian classes.[27]

In a series of Logit models, we contrast each variant of altruism: I) paying for S_1, II) paying for S_2 and giving away all payoff, and III) tossing a fair coin, against IV) opting into S_2 and taking all payoff. To account for the entirety of the data, we assign altruists who arrive in S_2 by dint of a fair coin or by paying for S_1, to II. B participants who pay for S_2 and end up in S_1 are also assigned to this group such that they may operate most effectively *against* a potential effect of *Kohlberg class 5*.[28] Appendix M shows the actual count of Bs' behaviours per treatment. We regress the resulting pairs of behaviour on B's average rankings over all six Kohlberg classes[29], and a treatment Dummy. To avoid omitted variable bias and, at the same time, preserve the estimator's efficiency, models are tested downward, removing insignificant variables which do not affect the goodness-of-fit.

[25] More precisely, if B were guilt averse, she would seek to avoid feeling guilt. She would feel guilty, if she opted into S_2 and took all payoff while expecting A to expect her not to do so (Battigalli and Dufwenberg, 2007) or knowing that a social norm (Miettinen, 2013) bans the actions in question.

[26] Note that if B simply tried to allocate outcomes, or their ex-ante expectations in a fair or kind way, she would also refer to social norms (Fehr and Schmidt 1999, p. 820-821, Bolton and Ockenfels 2000, p. 172, Bolton et al. 2005, p. 1068), or intentions (Dufwenberg and Kirchsteiger, 2004; Sebald, 2010).

[27] The current paper uses the same subject pool with a gap of five years. Since *all six Kohlberg scores* are very similarly distributed across both studies in Figure A16, we do not expect new variables to intercept the link between *Kohlberg class 5* and individuals' purely procedural concerns in the context of our study.

[28] Suppose these B participants opted for S_2 to take all payoff. In this case, their *Kohlberg class 5 scores* – if the latter do explain altruism – would be smaller than those of the actually observed altruists, weakening the effect. If they intended to give all payoff, the effect simply remains intact.

[29] B's *Kohlberg class 1 (2,3,4,5,6)* score averages her (four) ratings of the (four) arguments pertaining to *Kohlberg class 1 (2,3,4,5,6)* in the moral judgement test, adjusted for B's personal use of the 9-point Likert scale (the difference between the maximal and minimal rating a subject ever ticks). Average ratings are standardized by subtracting the sample mean, then dividing by the sample standard deviation. All moral judgement variables are computed the same way as in (Chlaß et al., 2019) and (Chlaß et al., 2023).

Dependent Variable: Variant of Altruism vs. Rational Self-Interest			
	S_1 (1) vs. Selfish (0)	S_2 + Give All (1) vs. Selfish (0)	Fair Coin (1) vs. Selfish (0)
B participants	19 (10 vs. 9)	83 (73 vs. 10)	16 (6 vs. 10)
Kohlberg class 1	-0.051 (0.088)	-0.140^a (0.055)	-0.315^b (0.142)
Kohlberg class 3	-0.126 (0.093)	0.050 (0.042)	-0.063^c (0.034)
Kohlberg class 5	0.451^a (0.087)	0.081^b (0.032)	0.567^a (0.147)
Kohlberg class 6	-0.253^a (0.086)	0.027 (0.038)	-0.068 (0.097)
Dummy LIE		0.086^b (0.039)	0.123 (0.136)
postmaterialism	0.160^c (0.094)		
materialism		-0.054^b (0.021)	-0.062^b (0.028)
Count R^2	0.90	0.90	0.88

Table 5: Ethical determinants of B participants' departures from rational self-interest (marginal effects).

Note: Significance levels of z-tests are indicated by $a : p < .01$, $b : p < .05$, $c : p < .10$

In order to clearly see whereto likelihood is shifted away from rational self-interest, we first specify independent binary Logits with robust errors and return a trifle too conservative *p-values* (Agresti, 2002) in Table 5. Estimated Logits yield a Count R^2 beyond 88%. Results are robust and easily reach a 1% significance level under a multinomial specification, an increase in sample size, a more balanced ratio of successes and failures after adding treatment SPY, an inclusion of all six *Kohlberg classes*, as well as critical (in fact, all collected) demographics in appendix O.

→ Kohlberg class 1. The more strongly B deems that an action which is not punished, cannot be wrong, the more likely she opts into S_2 and takes all payoff. Per one-unit increase in the strength of this conviction, she is 14%, *p-value* = 0.01, less likely to give all payoff to A and 31.5%, *p-value* = 0.026, less likely to toss a fair coin.

→ Kohlberg classes 2,3, and 4. *Kohlberg classes 2* and *4* are not significant in any binary comparison, neither on the reduced, nor on the full model – see appendix O – and, for the sake of efficiency and fit, left out from Table 5. Note that if our results were caused by this omission, both variables would either separately, or jointly have needed to turn out significant themselves. *Kohlberg class 3*, the extent to which B refers to her social image, others' expectations, social norms, or intentions to derive the right course of action, does not make her less inclined to behave selfishly either.

→ Kohlberg class 5. The more strongly B resorts to the social contract and the (equality of) civil rights granted therein to derive the right course of action, the more likely she opts into S_1, i.e. 45.1%, *p-value* = 0.000, the more likely she opts into S_2 and gives all payoff to A, i.e. 8.1%, *p-value*[30] = 0.011, and the more likely she tosses a fair

[30] In the reduced model with variables significant at least at the 10% level only, the effect becomes 10.7%, *p-value* = 0.005; the effect also reaches a 1% significance level on the full model in appendix O.

coin, i.e. 56.7%, *p-value* = 0.000, rather than being selfish.

→ Kohlberg class 6. The more strongly B resorts to universal principles of conscience such as human rights, individuals' freedom of choice, will and dignity, the more likely she gives all payoff to A in S_2 rather than opt into S_1, i.e. 13.9%, *p-value* = 0.01 in Table A10. The effect from Table 5 on S_1, however, is not robust.

→ Materialism and Postmaterialism. The more B values power and authority ('materialism'), the more likely she seeks S_2 and exerts her allocation power in S_2 to take all payoff rather than giving all payoff to A, i.e. 5.4%, *p-value*=0.011, or tossing a fair coin, i.e. 6.2%, *p-value* = 0.028. The more she values autonomy ('postmaterialism'), the more likely she opts into S_1[31] and reinstates A's decision rights.

→ Sample size, False Positives, Omitted Variable Bias & Randomization. Table A9 in Appendix O increases the sample size to 151 by adding treatment SPY, adds *all six* Kohlberg classes and critical demographic data. The effect of *Kohlberg class 5* reaches a 1% significance level. The multinomial analysis in Table A10 reports the effect at the same significance level for a reduced model. Figure A15 shows very similar distributions of Bs' six *Kohlberg scores* across LIE, SPY, and SABOTAGE, suggesting effective randomization *and* an absence of effective ex-post rationalization.

→ Punishment/Reward, Payoff Neutrality. Table A11 shows that once A's position of rights improved, and improved even beyond B's own position in S_1, *Kohlberg class 5* makes B behave more selfishly. B's symbolic punishment and to a weaker extent, also her reward beliefs regarding her own actual choice, dominate. The more B expects A to actually exert her rights – be it through punishment or reward[32] –, the more likely B looks after her own rational self-interest. The ethical criteria which underlie *all* departures from rational self-interest in this setting, are of *Kohlberg class 4*, which is also the case for payoff neutrality with punishment/reward. In payoff neutrality, the only active ethical criterion is *Kohlberg class 2*. Note, however, that, except for *Kohlberg class 5* and *6*[33], no actual experimental validation of these moral judgement variables has been undertaken.

Result 2. Departures from rational self-interest in LIE and SABOTAGE with competitive payoffs are indeed motivated by Kohlberg class 5, also after controlling for critical demographics.

Result 3. Symbolic punishment and reward crowds out Kohlberg class 5. Other ethical criteria are at work which is also the case under payoff neutrality.

6 Underlying Preferences & Discussion

In this section, we discuss theoretically which preferences can explain the variation of B participants' behaviour across treatments, and whether or not our empirical results are consistent with each preference type. We restrict our attention to the competitive payoff

[31] Removing insignificant variables *Kohlberg class 1* and *3*, *postmaterialism* has a marginal effect on opting into S_1 of 20.8%, *p-value* = 0.001. Similarly, removing *Kohlberg class 6* in the last column, *materialism* has a marginal effect on tossing a fair coin of −7.1%, *p-value* = 0.010. Reductions of the models in each case increases the goodness-of-fit; *Kohlberg class 3* never turns significant.

[32] The analysis looks at B's expectations about how much A punishes or rewards B's *actual* choice. These options are those which eventually yield genuinely different outcomes, and hence, *effectively* increase A's freedom to choose/decision rights.

[33] *Kohlberg class 6* statistically explains giving in (anonymous) dictator games – in 'Give' and 'Take' games, with earned income, with windfall profits, and with repetition (Chlaß and Moffatt, 2017).

setting where B must lie, spy, or sabotage to secure all payoff for sure.

Self-interested opportunism. If B only cares about her own material payoff, she spies, lies, or sabotages for sure to take all payoff. She pays 5 ECU to set $\text{Prob}\,(S_2) = \alpha = 1$ and in S_2, opts for strategy combination $\{B : RL^A, A : \{\cdot\}\}$, or $\{B : LR^A, A : \{\cdot\}\}$ to achieve allocation (B: 100, A: 0). Altogether, B receives $100 - 5 = 95$ ECU, and A receives 0 ECU in treatments LIE, SPY, and SABOTAGE[34]. Self-interested opportunism can therefore neither explain the variation in B participants' procedural choices across treatments LIE, SPY, and SABOTAGE, nor the empirical link between B's behaviour and her ethical preferences over *Kohlberg class 5* documented in section 5 and appendix O.

Altruism. If B cares more about A's material payoff than about her own – in Charness and Rabin's (2002) notation, for instance, B weights A's payoff by σ and her own payoff by $1-\rho$ where $\sigma > 1-\rho$, – she prefers allocation (B: 0, A: 100) to (B: 100, A: 0). To achieve this allocation, she pays 5 ECU for setting $\text{Prob}\,(S_2) = \alpha = 1$ and arrives for sure in S_2 where she imposes allocation (B: 0, A: 100) either via strategy combination $\{B : LL^A, A : \{\cdot\}\}$ or $\{B : RR^A, A : \{\cdot\}\}$. B receives -5 ECU and A 100 ECU in LIE, SPY, and SABOTAGE. Altruistic preferences therefore do not explain the variation in B participants' procedural and allocation choices across treatments LIE, SPY, and SABOTAGE, or the empirical link between B's behaviour and her ethical preferences over *Kohlberg class 5*.

Preferences for equal expected payoffs. B may be willing to forego some of her payoff in order to ex-ante grant A more equal chances on the one ex-post nonzero payoff. Formally, if B is inequity-averse over expected payoffs (Bolton et al., 2005), she has utility $u_B = a_B \cdot E(y_B) - 0.5 b_B \left(E(y_B) \cdot 100^{-1} - 0.5\right)^2$. Thereby, y_B denotes her own expected payoff, $a_B \geq 0$ her aversion against disadvantageous inequality, and $b_B \geq 0$ her aversion against advantageous inequality, both forms of aversion being driven by a social norm of payoff equality. In S_1, two perfectly selfish players would each choose to toss the fair coin between L and R which, coincidentally, also guarantees ex-ante equality in payoffs. B's corresponding utility is $a_B \cdot 50$ with no disutility from advantageous inequality. In S_2, B can implement any distribution of chances she prefers with the explicit option of tossing a fair coin. If B has a_B, b_B such that she cannot reach her preferred distribution of chances in S_1, she prefers S_2. This decision is identical in LIE, SPY and SABOTAGE. Preferences for equal expected payoffs do therefore not explain the variation in B participants' procedural and allocation choices across said treatments. Similarly, we could not confirm that B participants predominantly resort to social norms, a criterion located in *Kohlberg class 3*.

Preferences for kind procedures (Sebald 2010). A and B may care for the *kindness* of a procedural choice (whereby the kindness of a person who chooses a procedure is equal to the kindness of the distribution of outcomes which this procedure is expected to induce) and, upon observing a kind (unkind) procedural choice, be kind (unkind) in return. In our setting, it is commonly known that A never observes B's procedural choice. However,

[34] 95 ECU is the maximal payout as can be seen from comparing the following cases: If B opts into S_1 for sure, she pays 5 ECU to set $\alpha = 0$ and receives an expected equilibrium payout of 50 ECU in S_1, overall $50 - 5 = 45$ ECU. If B leaves the default $\alpha = 0.5$, she receives an equilibrium payout of 50 ECU from S_1 which occurs with 50% probability, and a payoff of 100 ECU from S_2 which also occurs with 50% probability. Hence, her overall expected payoff from not influencing the set of rules is $0.5 \cdot 50$ ECU $+0.5 \cdot 100 = 75$ ECU. Making S_2 one per cent more likely costs 0.1 ECU, but yields an expected payoff increase of $0.01 \cdot (95 - 75) = 0.2$ ECU. Hence, the 95 ECU which B earns from making S_2 sure are her maximal payoff.

A may hold expectations about B's procedural choice, and B may expect A to have such expectations. *a) suppose B expects A to expect S_2*. In this case, A expects to have no opportunity to reciprocate and she is always neutral toward B. This implies that B's payoff from reciprocity is zero and her preferences in S_2 coincide with self-interest: B chooses either $\{B : RL^A, A : \{\cdot\}\}$, or $\{B : LR^A, A : \{\cdot\}\}$ which earn her $100 - 5 = 95$ ECU. *b) suppose instead that B expects A to expect S_1*. When B is called upon to choose in S_1, she only considers her efficient strategies: yet, all are efficient since neither L nor R destroy the pie. If B believes A plays L with probability q_L and R with $1-q_L$, B's kindness in choosing L equals $q_L \cdot 100 + (1-q_L) \cdot 0 - (q_L \cdot 100 + (1-q_L) \cdot 0 + q_L \cdot 0 + (1-q_L) \cdot 100)/2$,[35] and her kindness in choosing R equals $q_L \cdot 0 + (1-q_L) \cdot 100 - (q_L \cdot 100 + (1-q_L) \cdot 0 + q_L \cdot 0 + (1-q_L) \cdot 100)/2$. If B believes that A tosses the fair coin, i.e. $q_L = 0.5$ which is the only equilibrium in S_1, then B's choice of L and R is exactly neutral toward A. Since B is not unkind in equilibrium, A need not reciprocate, and the payoffs from reciprocity in S_1 are zero. Hence, A and B implement the selfish solution and each tosses a fair coin which yields both players 50 ECU. Therefore, B participants who prefer kind over unkind procedures opt into S_2 which earns them $100 - 5$ ECU. Even if B held off-equilibrium beliefs in S_1, any reciprocation she expects in S_1 would be identical across SPY, LIE, and SABOTAGE. No variation in B's procedural or allocation choice should occur. In terms of ethical criteria, A and B assess their own and each others' choices in terms of *intentions*, and the degree to which the intended outcomes comply with a social norm of payoff equality. We could not confirm that B participants strongly invoke social norms or intentions which are both located in *Kohlberg class 3*.

Guilt aversion. If B is guilt-averse, she seeks to avoid disappointing A's payoff expectation and seeks to avoid being blamed by A for doing so (Battigalli and Dufwenberg, 2007). In part two – see section 3.2 – B submits her expectations about A's symbolic punishment and reward plan[36], a plan which lists by how much A increases or decreases B's payoff for any given choice of procedure Prob (S_2). This plan fuses information about how much A disapproves of a given procedural choice along with the allocation A expects this choice to entail. B participants expect more symbolic punishment for choosing S_2 in SPY than in LIE (one-sided Wilcoxon Rank Sum tests, *p-value* < 0.01 for $\alpha \in]0.5, 0.75[$, for $\alpha \in]0.75, 0.99[$, and for $\alpha = 1$), expect similar punishment for S_2 across LIE and SABOTAGE and also across SPY and SABOTAGE. Since B participants choose S_2 often in SPY and rarely in LIE, their procedural choices run contrary to their beliefs about what A approves them to do. Similarly, guilt aversion does not explain why, given that Bs expect the same punishment for S_2 in SPY and SABOTAGE, we observe substantial altruism in SABOTAGE, but none in SPY. In terms of ethical criteria, we could not confirm that B

[35] $q_L \cdot 100 + (1 - q_L) \cdot 0$ is A's payoff from B choosing L when B believes A plays L with probability q_L. This payoff is compared to the average payoff for A over all pure strategies which are still available to B at a given node: since B can still choose between L and R, this average payoff for A over B's pure strategies L and R is: $(q_L \cdot 100 + (1 - q_L) \cdot 0 + q_L \cdot 0 + (1 - q_L) \cdot 100)/2$. A payoff for A equal to this average payoff is neutral, payoffs for A greater than this average are kind (Dufwenberg and Kirchsteiger, 2004).

[36] A's expectations about B's choice of the interaction structure, and B's choice of the allocation may differ across LYING, SPYING, and SABOTAGING, for instance, because there are different social norms regarding lying, spying, or sabotaging which may in turn imply that the shares of individuals in the population who lie, spy, and sabotage differ, or because individuals also hold expectations whether or not others lie, spy, or sabotage, and expect others to have such expectations, too.

participants predominantly resort to others' expectations which are located in *Kohlberg class 3*.

Purely Procedural Preferences. B participants may have ethical reservations against being favored by the rules of the game, noteably in terms of decision or information rights (Chlaß et al., 2019). Suppose B's utility function includes some element similar to: $-\beta_B \max\{\#S_B - \#S_A, 0\} - \alpha_B \max\{\#S_A - \#S_B, 0\}$ where $\#S_B - \#S_A$ and $\#S_A - \#S_B$ count the difference between A's and B's number of effective pure strategies: strategies which induce genuinely different outcomes and therefore add to their freedom of choice – see section 2; where β_B denotes B's dislike of having greater, and α_B her dislike of having lesser rights. For LIE and SABOTAGE, B has two such pure strategies in S_1, and two in S_2 whereas A has two in S_1 but none in S_2. In S_1, therefore, B has no disutility from the rules of the game themselves whereas in S_2, her disutility is $\beta_B \cdot 2$. If this disutility is larger than the utility from her payoff advantage in S_2, then $S_1 \underset{B}{\succ} S_2$. In SPY, on the other hand, A and B always have equal decision rights: two effective pure strategies in S_1, and two in S_2. B's allocation power in S_2 arises from an advantage in information.[37] Summing up, LIE, SABOTAGE and SPY put different rights at stake and also differ in how B brings her advantage in S_2 about: in LIE and SABOTAGE, B takes decision rights *away from A* whereas in SPY, she assigns *herself* more information rights. We conclude that, if B has preferences over A's decision rights, she may prefer S_1 over S_2 in LIE and SABOTAGE, but not in SPY. In terms of ethical criteria, we can statistically confirm that B participants predominantly resort to the notion of civic rights as granted by a democratic social contract located in *Kohlberg class 5*, the very criterion underlying Chlaß et al. (2019)'s *purely procedural preferences*, after controlling for all known potential confounds for this link. Looking at B's giving all payoff to A in S_2, note that already Chlaß et al. (2019) find individuals who value decision rights, and yet reduce the opponent's rights while paying that opponent off, thus trading off monetary payoff and rights. B participants who opt into S_2, give all payoff to A and are motivated by *Kohlberg class 5* belong to this group. If B cares for A's decision rights, we therefore expect altruism in LIE and SABOTAGE but not in SPY.

Preferences for power & control. If B prefers to maintain power and control (Bartling et al., 2014), she opts for interaction structure S_2, thereby avoids any interference from A and implements whatever allocation she prefers. Preferences for power and control therefore do not predict variation in B participants' procedural or allocation choices across LIE, SPY, and SABOTAGE. Similarly, B participants would not resort to any ethical (fairness) criterion; yet, we observe such a link with *Kohlberg class 5*.[38] Preferences for

[37] We can express this advantage by the cardinalities (the fineness) of A's and B's information partitions over all possible terminal histories $z \in Z$. Again, B's utility function might include some element similar to $-b_B \max\{\#\mathcal{I}_B^z - \#\mathcal{I}_A^z, 0\} - a_B \max\{\#\mathcal{I}_A^z - \#\mathcal{I}_B^z, 0\}$ where $\#\mathcal{I}_B^z - \#\mathcal{I}_A^z$ and $\#\mathcal{I}_A^z - \#\mathcal{I}_B^z$ measure the difference between the cardinalities of A's and B's information partitions over all possible terminal histories, and a_B and b_B express B's aversion against having greater, or lesser, information rights. In S_1, B knows her own, but not A's choice and B's partition over the four terminal histories of S_1 has cardinality two. In S_2, B's partition over the four terminal histories has cardinality four: at the time of her decision, she knows which terminal history she will reach. A's information partition over the terminal histories in turn has cardinality one always, since she does not know whether she operates in S_1 or S_2. B's choice of S_2 therefore increases her own information rights, but does not reduce A's. In LIE and SABOTAGE, information rights are distributed the same way in S_1 and S_2: B's information partition has cardinality two always, A's partition always cardinality one.

[38] A preference for power would be a preference for maximizing one's own rights. The *purely procedural preferences* above build this idea into a framework of inequity aversion over decision rights (Chlaß et al.,

power might, however, explain why some B participants *within* the same treatment, opt for S_1 whereas others opt into S_2 and give all payoff away, both motivated by the same ethical criterion *Kohlberg class 5.* If B dislikes power, she might prefer to reinstate A's decision rights by opting into S_1; if she values power, she might seek and exert her power to compensate A for her lack of rights. Indeed, we find such a correspondence between B's choice and her materialism and postmaterialism values who, amongst other aspects, measure B's attitudes toward power, hierarchy, and autonomy.

Risk attitudes. In S_1 and S_2, B can achieve the same payoffs ex-post: 100 ECU, and 0 ECU. In S_2, however, B can obtain the 100 ECU for sure which is why a risk-averse B prefers S_2 where she takes all payoff.[39] Indeed, B's risk aversion slightly correlates with B's choice of S_2 and her taking all payoff, see table O. Risk attitudes do not predict varying degrees of altruism across LIE, SPY, or SABOTAGE.

Experimenter demand effects. Other than having addressed any of these preferences, we might— despite a neutral framing — have induced a cognitive or a social experimenter demand effect (Zizzo, 2010) in that the existence of an experimenter, or the awareness of participating in an experiment affected B participants' behaviour. If so, B participants should be strongly motivated by a desire to satisfy our expectations and to behave in a socially acceptable way. We could not confirm that B participants strongly resort to ethical criteria such as others' expectations, social norms, social image concerns or a desire to be taken as a nice person – all located in *Kohlberg class 3* – when choosing either procedure or allocation.

7 Conclusion

We show, for the first time, that individuals value fair rules of competition for their own sake and prefer to compete with opponents who are in a position to look after their own self-interest. In particular, individuals prefer to forego *all* payoff rather than fabricate or sabotage the opponent's decision, when doing so would win a constant sum game but at the same time, also take away the opponent's decision rights.

We begin with classic treatment interventions, and design three different ways to compete unfairly within the same setup. Two of them affect the opponent's decision rights – fabrication and sabotage –, and one does not – spying. Substantial amounts of altruism occur in the first two treatments, and little to none in the third. We formally discuss at length that the only preference to produce this difference must be one *purely* over the rules of the game, and that only an intrinsic concern for the opponent's rights does so. In particular social norms or guilt aversion should, according to individuals' actual beliefs, produce either no, or the exact opposite difference in altruism.

In a next step, we supplement the intervention study by an independent instrumental variable approach. Chlaß et al. (2019) built individuals' purely procedural preferences for

2019) [one feels the infringement of one's own rights more immediately than one feels the infringement of another individual's rights], a preference for power would imply a disutility from losing control over the payoff distribution to other individuals, but no disutility at all from taking decision rights from others.

[39] Since B cannot obtain a higher ex-post payoff than these 100 ECU through incurring additional risk, also risk-loving or risk-neutral Bs prefer S_2 and take all payoff, but they prefer S_2 to a lesser extent than a risk-averse B.

equal decision and information rights on the ethical criterion of equal rights stipulated by the social contract – equal freedom of choice, opportunity, and participation – and elicit subjects' preferences over these along with a comprehensive list of miscellaneous ethical criteria. We repeat this psychometric analysis and show that the same ethical criterion explains all departures from rational self-interest, after controlling for the critical latent variables we found in our earlier work. In a second set of interventions, worded identical to the first, we remove all influence individuals hold over their opponent's decision rights in all treatments, such that fabrication, sabotage, and spying become merely different frames. All treatments show similarly low occurences of altruism. The instrument no longer shows any effect on individuals' decisions.

A third set of interventions reinforces the opponent's decision rights by a set of options to reward or punish fabrication, sabotage, and spying, first, to neutralize concerns about the opponent's position of rights and, second, to provide the connecting dots with the literature which reports that altruism depends on our beliefs about how desirable we believe our actions are from others' point of view. Indeed, altruism does now indeed aim at avoiding expected punishment and earning expected reward. This change in the nature of altruism is signalled by a significant drop in its amount which remains, however, substantial. The instrument now shows a negative effect on individuals' departures from rational self-interest. Our results also indicate that altruism can be caused *purely* by the rules of the game and in this case, can be understood as a behavioural strategy to compensate aspects of these rules which individuals deem unethical. These unethical aspects out of the way, and no others being present, we can observe *exact* rational self-interest.

References

Abeler, J., D. Nosenzo, and Collin B. Raymond (2019). "Preferences for truth-telling". In: *Econometrica* 87 (4), pp. 1115–1153.

Abratt, R. and N. Penman (2002). "Understanding Factors Affecting Salespeople's Perceptions of Ethical Behavior in South Africa". In: *Journal of Business Ethics* 35 (4). 10.1023/A:1013872805967, pp. 269–280. ISSN: 0167–4544.

Agresti, Alan (2002). *Categorical Data Analysis.* 2nd. Wiley Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons, Inc.

Alger, Ingela and Jörgen W. Weibull (2013). "Homo Moralis—Preference Evolution Under Incomplete Information and Assortative Matching". In: *Econometrica* 81 (6), pp. 2269–2302.

Baker, D. E. and R. Inglehart (2000). "Modernization, Cultural Change, And the Persistence of Traditional Values". In: *American Sociological Review* 65 (1), pp. 19–51.

Barnett, R., J. Evens, and J. Rest (1995). "Faking moral judgement on the Defining Issues Test". In: *British Journal of Social Psychology* 34, pp. 267–278.

Bartling, B., E. Fehr, and H. Herz (2014). "The Intrinsic Value of Decision Rights". In: *Econometrica* 82.6, pp. 2005–2039.

Bartling, B. et al. (2009). "Egalitarianism and Competitiveness". In: *The American Economic Review* 99.2, pp. 93–98. DOI: 10.1257/aer.99.2.93. URL: http://www.ingentaconnect.com/content/aea/aer/2009/00000099/00000002/art00016.

Battigalli, P. and M. Dufwenberg (2007). "Guilt in Games". In: *The American Economic Review* 2.97, pp. 170–176.

Bebeau, Muriel J. (2002). "The Defining Issues Test and Four Component Model: Contributions to Professional Education". In: *Journal of Moral Education* 31 (3), pp. 271–295.

Beresford, A., D. Kübler, and S. Preibusch (2012). "Unwillingness to Pay for Privacy: a Field Experiment". In: *Economics Letters* 112, pp. 25–27.

Bolton, G. E., J. Brandts, and A. Ockenfels (2005). "Fair Procedures: Evidence from Games Involving Lotteries*". In: *The Economic Journal* 115.506, pp. 1054–1076. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1468--0297.2005.01032.x/full;%20http://digital.csic.es/bitstream/10261/1924/1/48301.pdf.

Borg, I. et al. (2019). "Do the PVQ and the IRVS scales for personal values support Schwartz's value circle model or Klages' value dimensions model?*". In: *Measurement Instruments for the Social Sciences* 2/2019.3, pp. 2–14. URL: https://doi.org/10.1186/s42409-018-0004-2.

Brown, V. R. and E. D. Vaughn (2011). "The Writing On the (Facebook) Wall: The Use of Social Networking Sites in Hiring Decisions". In: *Journal of Business Psychology* 26, pp. 219–225.

Busch, W. (1906). *Max und Moritz: Eine Bubengeschichte in sieben Streichen.* 53rd ed. Braun und Schneider.

Cappelen, Alexander W. et al. (2024). *Fairness in A Society of Unequal Opportunities.* Tech. rep. 2024-17. Norwegian School of Economics, Department of Economics.

Carpenter, J., P.H. Matthews, and J. Schirm (2010). "Tournaments and Office Politics: Evidence from a real effort experiment". In: *The American Economic Review* 100.1, pp. 504–517.

Charness, G. and M. Rabin (2002). "Understanding Social Preferences with Simple Tests". In: *The Quarterly Journal of Economics* 117.3, pp. 817–869.

Charness, Gary et al. (2012). "The Hidden Advantage of Delegation: Pareto Improvements in a Gift Exchange Game". In: *American Economic Review* 102.5, pp. 2358–79.

Chen, K.-P. (2003). "Sabotage in Promotion Tournaments". In: *Journal of Law, Economics, and Organization* 19.1, pp. 119–140.

Chlaß, N., L. Gangadharan, and K. Jones (2023). "Charitable Giving and Intermediation: a principal-agent problem with hidden prices." In: *Oxford Economic Papers* 75, pp. 941–961.

Chlaß, N., W. Güth, and T. Miettinen (2019). "Purely procedural preferences – beyond procedural equity and reciprocity." In: *European Journal of Political Economy* 59, pp. 108–128.

Chlaß, N. and P. Moffatt (2017). *Giving in Dictator Games, Experimenter Demand Effect, or Preference over the Rules of the Game?* Tech. rep. 2017-05. CBESS, University of East Anglia.

Dufwenberg, M. and G. Kirchsteiger (2004). "A Theory of Sequential Reciprocity". In: *Games and Economic Behavior* 47.3, pp. 268–98.

Erat, Sanjiv (2013). "Avoiding lying: the case of delegated deception". In: *Journal of Economic Behavior and Organization* 93, pp. 273–278.

Falk, A., E. Fehr, and D. Huffman (2008). *The Power and Limits of Tournament Incentives.* Tech. rep. 2008. University of Zurich.

Falk, A. and U. Fischbacher (2006). "A Theory of Reciprocity". In: *Games and Economic Behavior* 54, pp. 293–315.

Fehr, Ernst, Holger Herz, and Tom Wilkening (2013). "The Lure of Authority: Motivation and Incentive Effects of Power". In: *American Economic Review* 103.4, pp. 1325–59.

Fischbacher, U. (2007). "z–Tree: Zurich Toolbox for Ready–Made Economic Experiments". In: *Experimental Economics* 10.2, pp. 171–178. URL: `http://link.springer.com/article/10.1007/s10683--006--9159--4;%20http://link.springer.com/content/pdf/10.1007_252Fs10683--006--9159--4.pdf`.

Fischbacher, U. and F. Föllmi-Heusi (2013). "Lies in Disguise: An Experimental Study on Cheating". In: *Journal of the European Economic Association* 11.3, pp. 525–547.

German Central Bank (Oct. 2016). "Significance and impact of high-frequency trading in the German capital market". In: *Monthly Report.* URL: `https://www.bundesbank.de/en/publications/reports/monthly-reports/monthly-report-october-2016-707572`.

Gibson, R., C. Tanner, and A. F. Wagner (2013). "Preferences for Truthfulness: Heterogeneity Among and Within Individuals". In: *American Economic Review* 103.1, pp. 532–548.

Gneezy, U. (2005). “Deception: The Role of Consequences”. In: *The American Economic Review* 95.1, pp. 384–394. DOI: 10.1257/0002828053828662.

Gneezy, U. and A. Kajackaite (2020). “Externalities, stakes, and lying”. In: *Journal of Economic Behavior and Organization* 178, pp. 629–643.

Greenwald, G., E. MacAskill, and L. Poitras (June 2013). “Edward Snowden: the Whistleblower Behind the NSA Surveillance Revelations”. In: *The Guardian* 11. URL: http://www.theguardian.com/world/2013/jun/09/edward-snowden-nsa-whistleblower-surveillance?guni=Podcast:in%20body%20link.

Greiner, B. (2004). “An online recruitment system for economic experiments”. In: *Forschung und wissenschaftliches Rechnen* 63. Ed. by K. Kremer and V. Macho. URL: http://mpra.ub.uni--muenchen.de/13513/;%20http://mpra.ub.uni--muenchen.de/13513/1/MPRA_paper_13513.pdf.

Harbring, C. and B. Irlenbusch (2011). “Sabotage in Tournaments: Evidence from a Laboratory Experiment.” In: *Management Science* 57.4, pp. 611–627.

Harbring, C. et al. (2007). “Sabotage in Corporate Contests–An Experimental Analysis”. In: *International Journal of the Economics and Business* 14, pp. 201–223.

Holt, C. A. and S. K. Laury (2002). “Risk Aversion and Incentive Effects”. In: *American Economic Review* 92.5, pp. 1644–1655.

Inglehart, R. (1977). *The Silent Revolution: Changing Values and Political Styles Among Western Publics.* Princeton University Press.

Ishida (2006). “How do Scores of the DIT and the MJT differ? A Critical Assessment of the Use of Alternative Moral Development Scales in Studies of Business Ethics.” In: *Journal of Business Ethics* 67, pp. 63–74.

Kajackaite, A. and U. Gneezy (2017). “Incentives and Cheating”. In: *Games and Economic Behavior* 102, pp. 433–444.

Kala, Namrata (2024). “The Impacts of Managerial Autonomy on Firm Outcomes”. In: *Econometrica* 92.6, pp. 1777–1800.

Kirchsteiger, G. (1994). “The Role of Envy in Ultimatum Games”. In: *Journal of Economic Behavior & Organization* 25.3, pp. 373–389. URL: http://www.sciencedirect.com/science/article/pii/0167268194901066;%20http://www.ecares.org/ecare/personal/kirchsteiger/publications/1994jebo--envy.pdf.

Klages, H. and Th. Gensicke (2006). “Wertesynthese - Funktional oder Dysfunktional?” In: *Kölner Zeitschrift für Soziologie and Sozialpsychologie* 58.2, pp. 332–351.

Klebe Treviño, Linda and Gary R. Weaver (1997). “Ethical Issues in Competitive Intelligence Practice: Consensus, Conflicts, and Challenges”. In: *Competitive Intelligence Review* 8, pp. 61–72.

Kohlberg, L. (1969). *Stage and Sequence: the Cognitive–Developmental Approach to Socialization.* Ed. by D.A. Goslin. Handbook of Socialization and Endash; Theory and research. Chicago: McNally.

— (1984). *The Psychology of Moral Development.* San Francisco: Harper & Row.

Konrad, K.A. (2000). “Sabotage in Rent-Seeking Contests”. In: *Journal of Law, Economics, and Organization* 16 (1), pp. 155–165.

Larson, J., N. Perlroth, and S. Shane (Sept. 2013). "Revealed: The NSA's Secret Campaign to Crack, Undermine Internet Security". In: *New York Times* 5. URL: http://www.propublica.org/article/the-nsas-secret-campaign-to-crack-undermine-internet-encryption.

Lazear, E.P. (1989). "Pay Equality and Industrial Politics". In: *Journal of Political Economy* 97 (3), pp. 561–580.

Lee, Natalie (2023). "Feigning Ignorance for Long-Term Gains". In: *Games and Economic Behavior* 138, pp. 42–71.

Lind, G. (1978). "Wie misst man moralisches Urteil? Probleme und alternative Möglichkeiten der Messung eines komplexen Konstrukts". In: *Sozialisation und Moral*. Ed. by G. Portele. Weinheim: Beltz, pp. 1215–1259.

— (2008). "The Meaning and Measurement of Moral Judgment Competence Revisited – A Dual–Aspect Model". In: *Contemporary Philosophical and Psychological Perspectives on Moral Development and Education*. Ed. by D. Fasko and W. Willis. Cresskill. NJ: Hampton Press, pp. 185–220.

Miettinen, T. (2013). "Promises and Conventions – An Approach to Pre-Play Agreements". In: *Games and Economic Behaviour* 80.80, pp. 68–84.

Münster, J. (2007). "Selection Tournaments, Sabotage, and Participation". In: *Journal of Economics & Management Strategy* 16.4, pp. 943–970.

Penta, Antonio and Peio Zuazo-Garin (2022). "Rationalizability, Observability and Common Knowledge". In: *Review of Economic Studies* 89, pp. 948–975.

Piaget, J. (1948). *The Moral Judgment of the Child*. Glencoe, Illinois: Free Press.

Rest, James (1979). *Development in judging moral issues*. Minneapolis: University of Minnesota Press.

Reuters (Dec. 2012). "Court says Deutsche Bank must pay damages in Kirch case". In: URL: https://www.reuters.com/article/business/court-says-deutsche-bank-must-pay-damages-in-kirch-case-idUSBRE8BD0RK/.

Sánchez-Pagés, S. and M. Vorsatz (2009). "Enjoy the silence: an experiment on truth-telling". In: *Experimental Economics* 12, pp. 220–241.

Sebald, A. (2010). "Attribution and Reciprocity." In: *Games and Economic Behavior* 68.1, pp. 339–352.

Smith, A. (1904). *An Inquiry into the Nature and Causes of the Wealth of Nations*. 5th ed. London: Methuen & Co., Ltd.

Solan, E. and L. Yariv (2004). "Games with Espionage". In: *Games and Economic Behavior* 47.1, pp. 172–199.

Sugden, R. (1998). "The Metric of Opportunity". In: *Economics and Philosophy* 14, pp. 307–337.

Sugden, R. and M. Wang (2021). "Equality of opportunity and the acceptability of outcome inequality". In: *European Economic Review* 130, p. 103597.

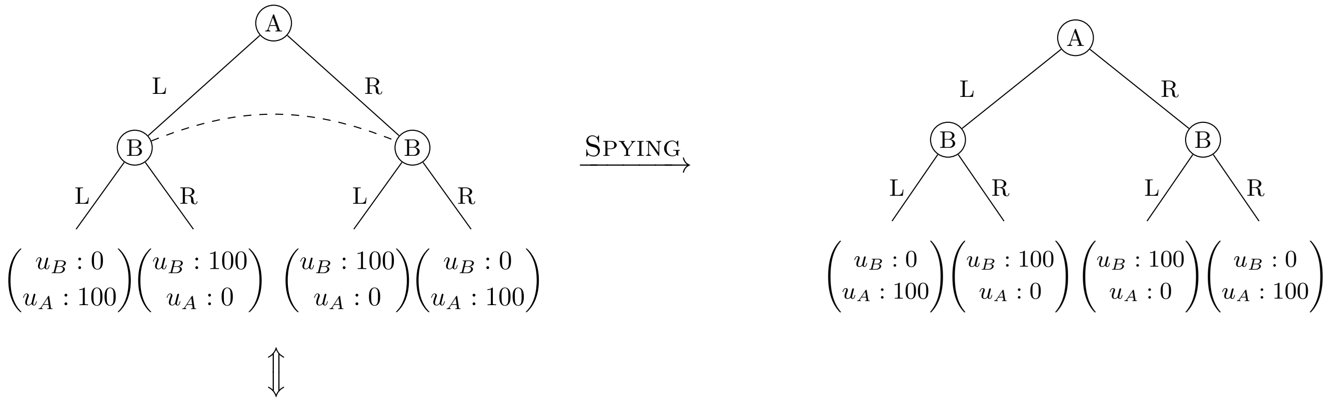
Sutter, M. (2009). "Deception Through Telling the Truth?! Experimental Evidence From Individuals and Teams". In: *The Economic Journal* 119.534, pp. 47–60. DOI: 10.1111/j.1468--0297.2008.02205.x.

Tanner, C., B. Ryf, and M. Hanselmann (2009). "Protected Values Measure: construction and first validation of an instrument to assess protected values". In: *Diagnostica* 3.55, pp. 174–183.

Trautmann, Stefan T. (2023). "Procedural fairness and equality of opportunity". In: *Journal of Economic Surveys* 37.5, pp. 1697–1714.

Zizzo, D. J. (2010). "Experimenter Demand Effects in Economic Experiments". In: *Experimental Economics* 13.1, pp. 75–98.

A Lying, Spying, Sabotaging: conceptualization



		A: L	A: R
B	L	100 / 0	0 / 100
	R	0 / 100	100 / 0

Notes. The normal form game from the left of Table 1 in its extensive form, before (left) and after (right) spying. Player indexes are kept the same across normal and extensive form such that player 2, i.e. A, moves first, and player 1, i.e. B, moves last in the extensive form; payoffs at the terminal nodes are listed by increasing player index – first B, then A.

Figure A1: How can B profit from Spying? Left: equal information rights, right: unequal information rights.

B Defining sabotage: Max and Moritz (Busch, 1906)

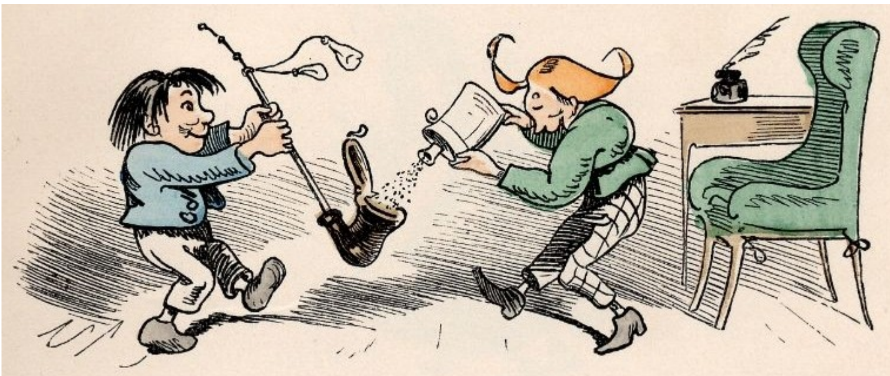


Figure A2: Max and Moritz fill their teacher's smoking pipe with black powder.

Figure A3: Lighting the smoking pipe has a new consequence for the teacher.

C Rules and payoffs: payoff neutrality

Table A1: Payoff neutrality: B sabotages and fabricates without removing A's decision rights. Left: S_1 – unequal decision rights, Right: S_2 – unequal decision rights.

B \ A	L	R
L	100 0	100 0
R	0 100	0 100

$\xrightarrow[\text{Fabrication}]{\text{Sabotage}}$

B \ A	L	R
LL^A	100 0	100 0
RL^A	0 100	0 100
LR^A	100 0	100 0
RR^A	0 100	0 100

D Rules and payoffs: symbolic reward and punishment

Table A2: A obtains additional decision rights in S_1 (Above) and S_2 (Below): next to L and R, she may choose between L and R <u>and</u> reduce (or increase) B's payoff within $d_A \in [-30, 30] \setminus 0$ ECU.

B \ A	L	R	$L^{Pu/Rew}$	$R^{Pu/Rew}$
L	100 0	0 100	$100 - \|d_A\|$ $0 - d_A$	$0 - \|d_A\|$ $100 - d_A$
R	0 100	100 0	$0 - \|d_A\|$ $100 - d_A$	$100 - \|d_A\|$ $0 - d_A$

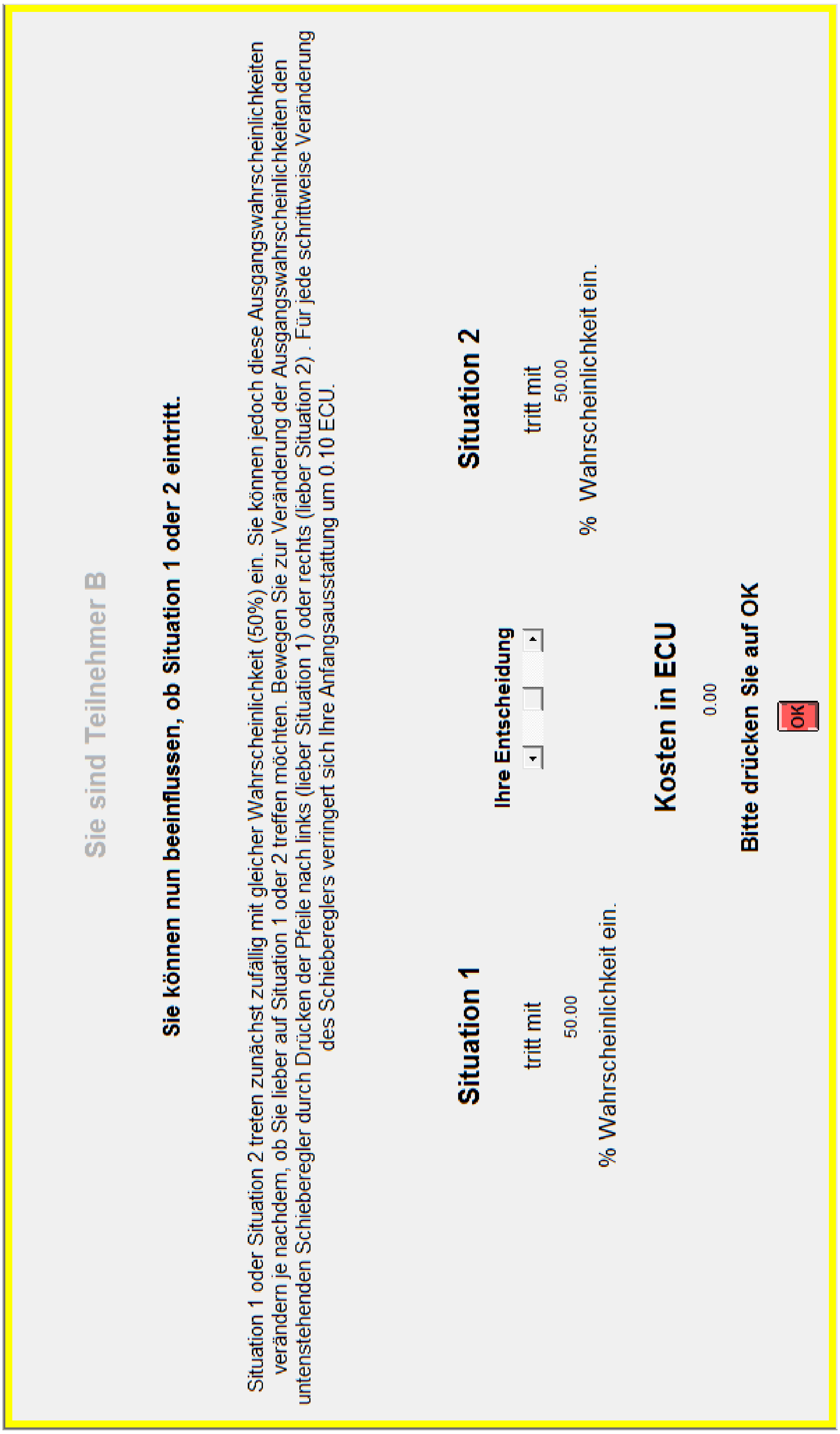
Sabotage ↓ Fabrication

B \ A	L	R	$L^{Pu/Rew}$	$R^{Pu/Rew}$
LL^A	100 0	100 0	$100 - \|d_A\|$ $0 - d_A$	$100 - \|d_A\|$ $0 - d_A$
RL^A	0 100	0 100	$0 - \|d_A\|$ $100 - d_A$	$0 - \|d_A\|$ $100 - d_A$
LR^A	100 0	100 0	$100 - \|d_A\|$ $0 - d_A$	$100 - d_A$ $0 - d_A$
RR^A	0 100	0 100	$0 - \|d_A\|$ $100 - d_A$	$0 - \|d_A\|$ $100 - d_A$

Notes. Every B may have her own d_A, depending on the A she encounters. $d_A \in [-30, 30] \setminus 0$ is the amount by which A punishes or rewards B's choice of Prob(S_2). Suppose the latter actually is 75%, and A decided to reduce B's payoff by 30 for all Prob(S_2) $\in$ [75%, 99%]. Then, $d_A = -30$. Strategies expand the individual's freedom of choice, if they yield genuinely different outcomes. To see that the new options can be preferred, think of an A participant who is spiteful. She enjoys more utility from reducing B's payoff than disutility from reducing her own and can therefore prefer the new punishment options over the options she already had. The first two columns L and R show A's old options L and R which amount to A choosing $d_A = 0$.

E Screenshots

E.1 B's choice Prob(S_2) of the situation



[You are participant B.

You may now influence whether situation 1 or situation 2 occurs.

For the time being, situation 1 or situation 2 occur randomly with equal probability (50%). You may, however, change these initial probabilities, depending on whether you prefer to encounter situation 1 or situation 2.

To change the initial probabilities, move the slider below by clicking on the arrows on its left (rather situation 1) or on its right (rather situation 2). For every stepwise change of the slider, your initial endowment decreases by 0.10 ECU.

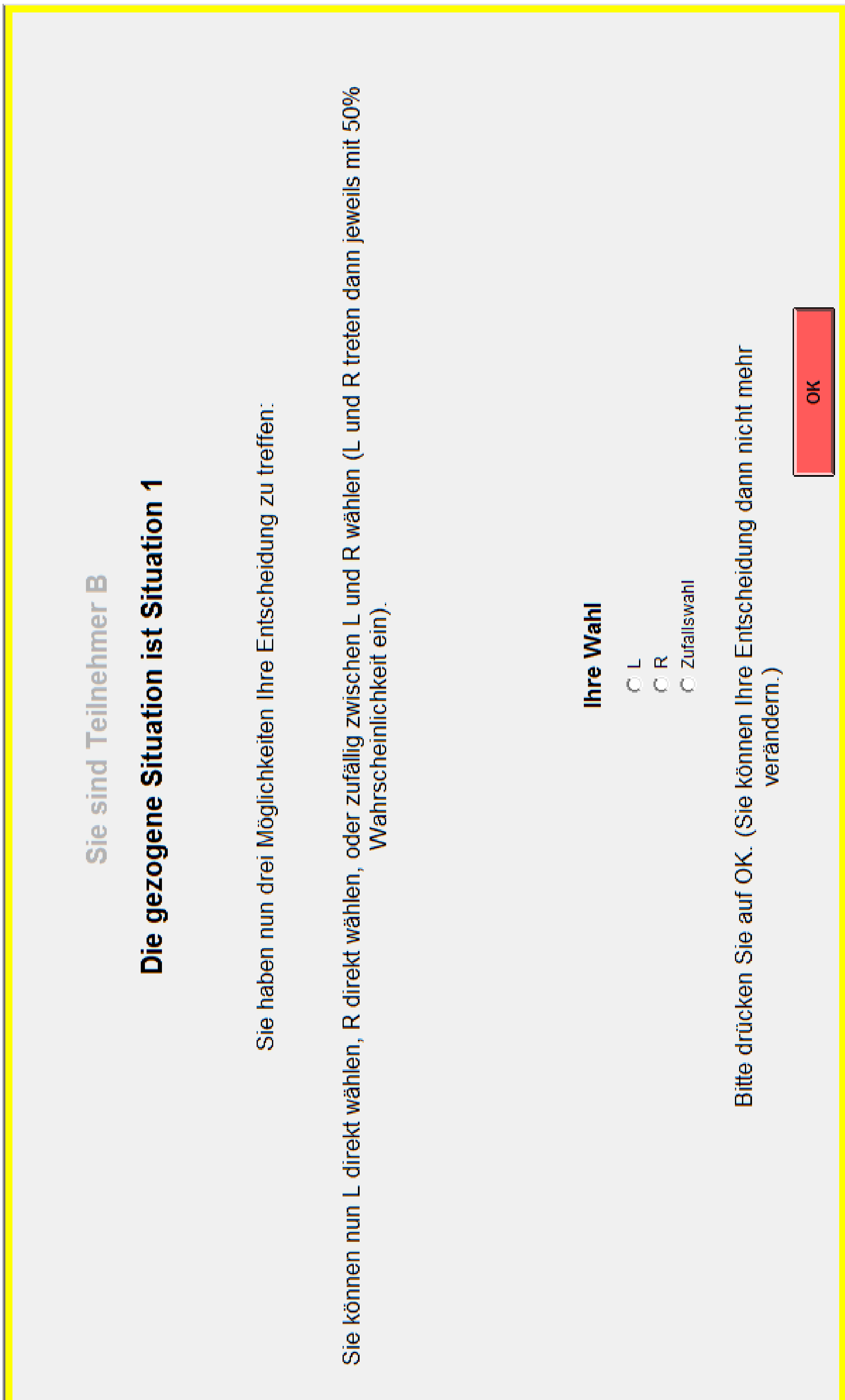
Your decision:

Cost in ECU

Please click OK

Figure A4: *B*'s PROBABILITY CHOICE PROB(S_2) OF THE SITUATION.

E.2 B's choices in situation S_1



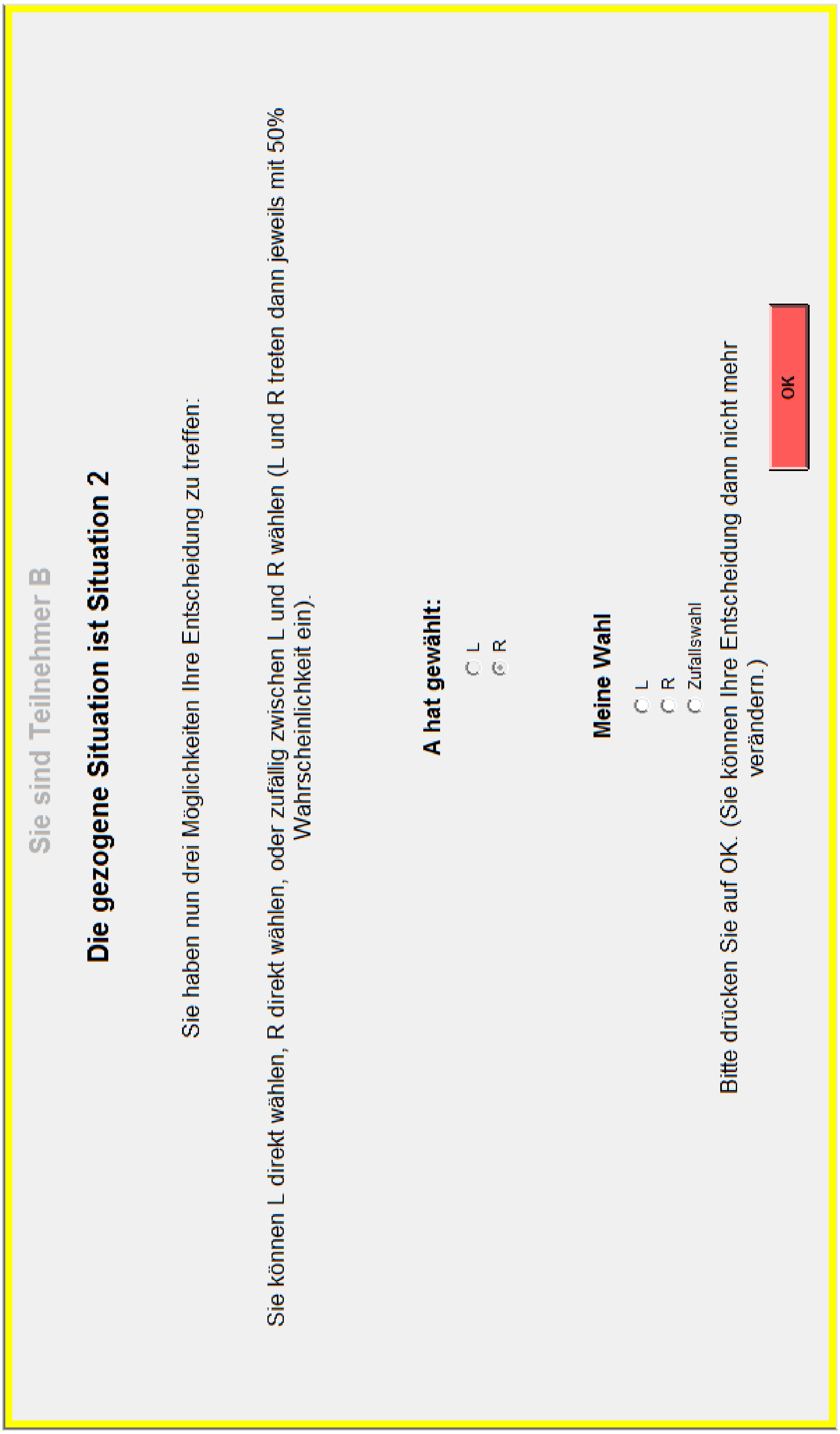
[You are participant B.
The situation drawn is situation 1.
You now have three possibilities to make a decision.
You can choose L directly, R directly, or choose randomly between L and R (L and R then occur with 50% probability each.)

Your choice:
○ L
○ R
○ random choice

To confirm, please click OK. (You can no longer change your decision afterwards).]

Figure A5: *B*'S CHOICES IN SITUATION S_1.

E.3 B's choices in situation S_2 SPY



[You are participant B.

The situation drawn is situation 2.

You now have three possibilities to make a decision. You can choose L directly, R directly, or choose randomly between L and R (L and R then occur with 50% probability each.)

A has chosen:

○ L

○ R

○

My choice:

○ L

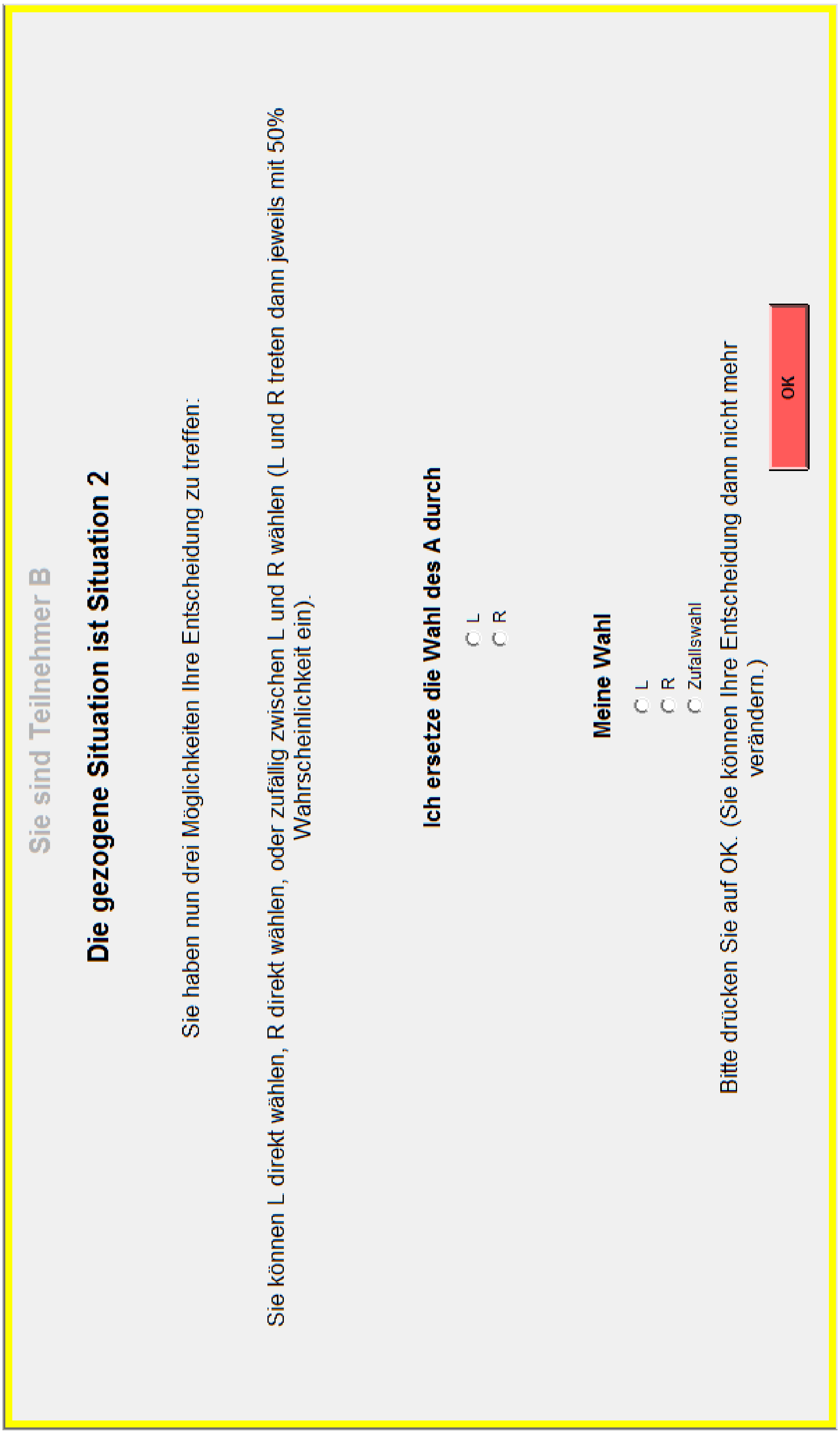
○ R

○ random choice

To confirm, please click OK. (You can no longer change your decision afterwards).]

Figure A6: B's decision screen in S_2, treatment SPY.

E.4 B's choices in situation S_2 SABOTAGE



[You are participant B.
The situation drawn is situation 2.
You now have three possibilities to make a decision. You can choose L directly, R directly, or choose randomly between L and R (L and R then occur with 50% probability each.)

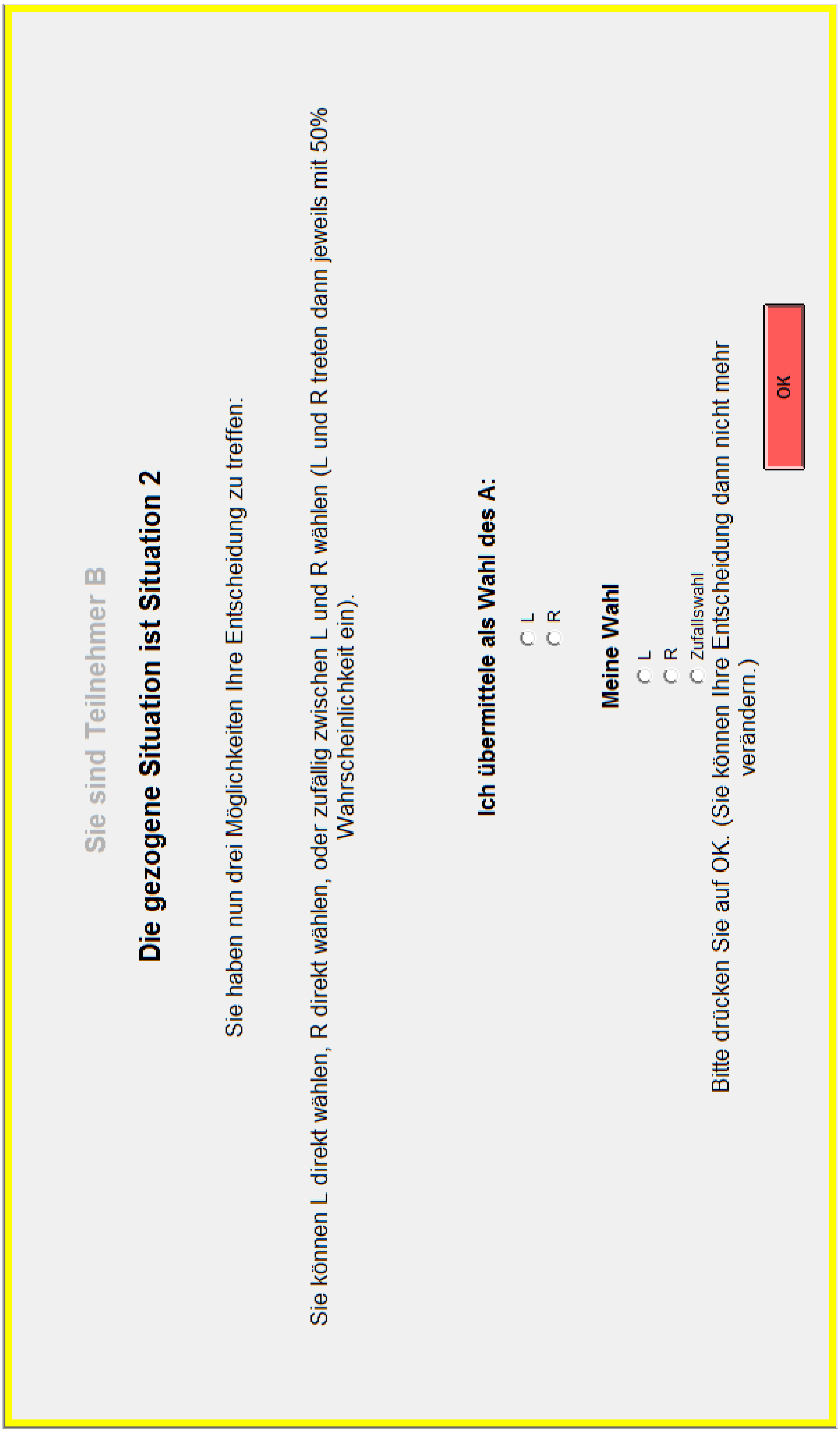
I replace A's choice by:
○ L
○ R

My choice:
○ L
○ R
○ random choice

To confirm, please click OK. (You can no longer change your decision afterwards).]

Figure A7: B's decision screen in S_2, treatment SABOTAGE.

E.5 B's choices in situation S_2 LIE



[You are participant B.
The situation drawn is situation 2.
You now have three possibilities to make a decision. You can choose L directly, R directly, or choose randomly between L and R (L and R then occur with 50% probability each.)

I transmit as A's choice:
○ L
○ R

My choice:
○ L
○ R
○ random choice

To confirm, please click OK. (You can no longer change your decision afterwards).]

Figure A8: *B*'s decision screen in S_2, treatment LIE.

E.6 B's punishment beliefs in the punishment/reward stage

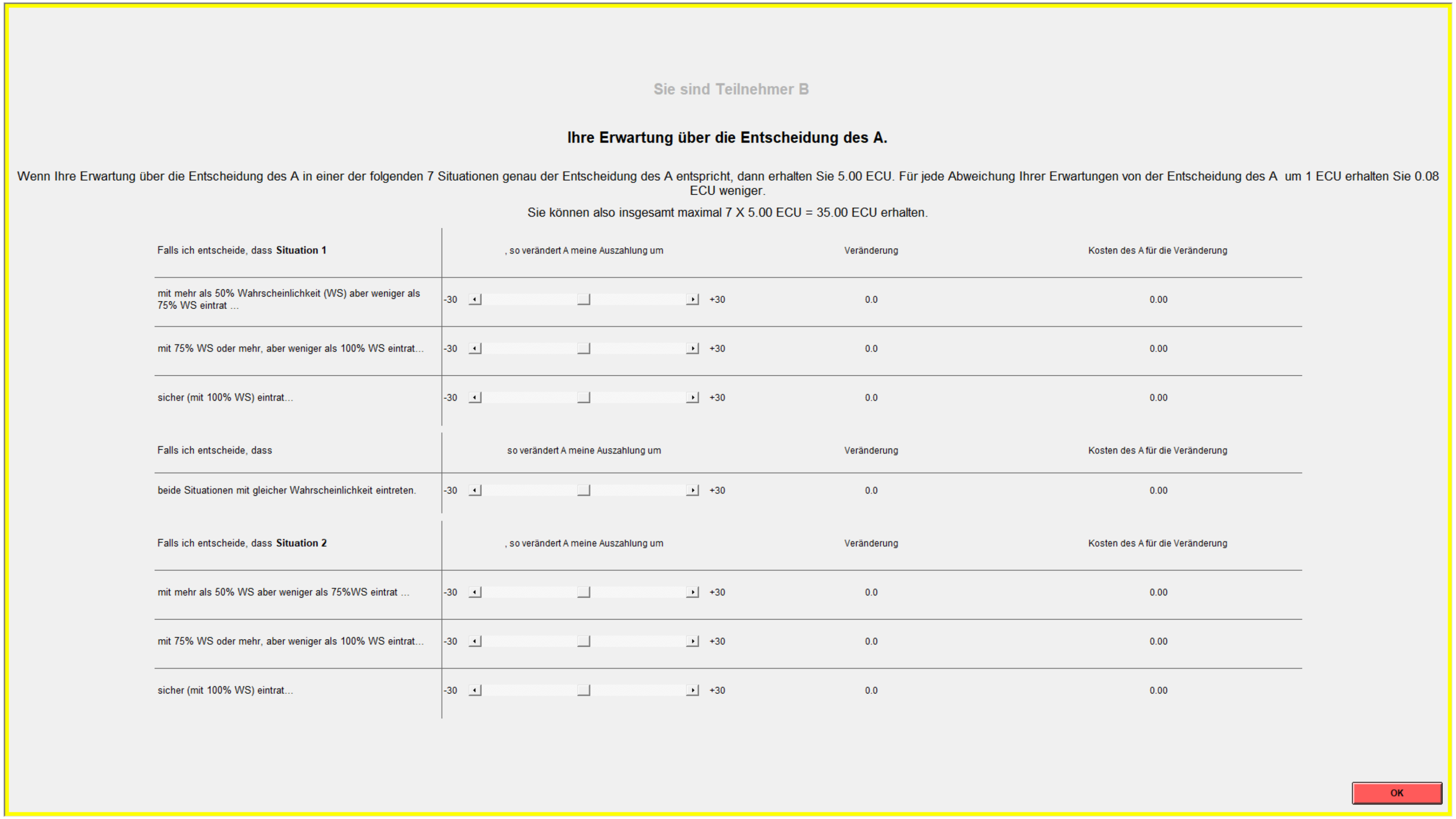


Figure A9: *B*'S DECISION SCREEN TO SUBMIT HER 1ST ORDER BELIEFS ABOUT *A*'S PUNISHMENT AND REWARD OF PROB(S_2).

[You are participant B.

Your expectation regarding *A*'s choice.

If your expectation about *A*'s decision is exactly equal to *A*'s decision in one of the following 7 situations, you receive 5.00 ECU. For every ECU by which your expectations differ from *A*'s actual decision, you receive 0.08 ECU less. That is, you can altogether maximally earn 7x5=35 ECU.

If I decide that **situation 1**	, *A* changes my payoff by	change	*A*'s cost for the change
occured with more than 50% but less than 75% probability (Pr)...	-30 $\longleftrightarrow$ +30	0.0	0.00
occured with 75% Pr or more but less than 100% Pr...	-30 $\longleftrightarrow$ +30	0.0	0.00
occurred for sure (with 100% Pr)...	-30 $\longleftrightarrow$ +30	0.0	0.00
If I decide that	, *A* changes my payoff by	change	*A*'s cost for the change
both situations occured with equal probability...	-30 $\longleftrightarrow$ +30	0.0	0.00
If I decide that **situation 2**	, *A* changes my payoff by	change	*A*'s cost for the change
occured with more than 50% Pr but less than 75% Pr...	-30 $\longleftrightarrow$ +30	0.0	0.000
occured with 75% Pr or more, but less than 100% Pr...	-30 $\longleftrightarrow$ +30	0.0	0.00
occured for sure (with 100% Pr)...	-30 $\longleftrightarrow$ +30	0.0	0.00

F Experimental Instructions

F.1 Instructions[40]

Instructions

Welcome and thank you for participating in this experiment. Please read the following instructions carefully. The instructions are identical for all participants. Communication with other participants must cease from now on. Please turn off your mobile phone. If you have any questions, please raise your hand - we will answer them individually at your seat. Do not ask your questions aloud.

During the experiment, monetary amounts are denoted in ECU (Experimental Currency Units). The sum of your payoffs from all rounds will be disbursed to you in cash at the end of the experiment (exchange rate 1 ECU=0.05 Euro). Your initial endowment is 50 ECU.

Information about the experiment

In this experiment, you interact with other anonymous participants. Participants take on different roles **A** and **B** [TREATMENT LIE: and **C**]. Roles are randomly determined at the beginning and remain the same throughout the experiment. The experiment consists of several rounds. In each round, you are matched with a new participant. In each round, you encounter two situations. These situations are initialized to occur with probability 50%. At the beginning of each round, B can decide which situation actually occurs, and can make one situation more likely than the other. Making one situation 10 percent more likely costs 1 ECU. The two situations are characterized as follows.

Situation 1. Participant A chooses between two options L and R. Participant B does not see which option A has chosen. B then also chooses between options L and R. Both participants can also choose options L and R with equal probability.

Situation 2. Participant A chooses between two options L and R. Participant B does not see which option A has chosen. B sets *A*'s choice to either L or R. B then also chooses between options L and R. Both participants can also choose options L and R with equal probability.

[IN TREATMENT LIE, SITUATION 2 READ AS FOLLOWS:

Situation 2. Participant A chooses between two options L and R. Participant B does

[40] Instructions of the experiment were written in German. This appendix produces a translation into English for treatment SABOTAGE with competitive payoffs. Instructions for treatments SPY and LIE differed by the text in square brackets. TEXT IN CAPITAL LETTERS WAS NOT PART OF THE ORIGINAL INSTRUCTIONS. Emphases in bold or italic font are taken from the original text. Instructions for the payoff neutral treatment were worded identically, the only difference being the respective numbers in the payoff table: If *B* chose R and *A* chose L, *A* received 0, and *B* 100 ECU. If *B* chose R and *A* chose R, *A* received 0 ECU and *B* 100 ECU. If *B* chose L and *A* chose L, *A* received 100 ECU and *B* 0 ECU. If *B* chose L and *A* chose R, *A* received 100 ECU and *B* 100 ECU.

not see which option A has chosen. B transmits A's choice to participant **C**. B then also chooses between options L and R. Both participants can also choose options L and R with equal probability.]

[IN TREATMENT SPY, SITUATION 2 READ AS FOLLOWS:
Situation 2. Participant A chooses between two alternatives L and R. Participant B sees which option A has chosen. B then also chooses between options L and R. Both participants can also choose options L and R with equal probability.]

Your Payoff

The table below shows which payoffs A and B receive for their choices in a given round. At the end of the experiment, one round will be paid out (exchange rate 1 ECU=0.05 Euro). The computer selects this round randomly and with equal probability. [TREATMENT LIE: Participant C receives a fixed payoff of 125 ECU.]

Decisions	Payoffs for these decisions
B chooses L, **A** chooses L	**A** receives 100 ECU, **B** receives 0 ECU
B chooses L, **A** chooses R	**A** receives 0 ECU, **B** receives 100 ECU
B chooses R, **A** chooses L	**A** receives 0 ECU, **B** receives 100 ECU
B chooses R, **A** chooses R	**A** receives 100 ECU, **B** receives 0 ECU
B chooses 'randomly' and/or **A** chooses 'randomly'	Chance decides with equal probability whether **A** receives 100 ECU and**B** receives 0 ECU, or whether **A** receives 0 ECU and **B** receives 100 ECU.

If B chooses L, B receives 0 ECU (and A 100 ECU) if A also chooses L. If B chooses R, B receives 0 ECU (and A 100 ECU) if A also chooses R. If B chooses L, B receives 100 ECU (and A 0 ECU) if A chooses R. If B chooses R, B receives 100 ECU (and A 0 ECU) if A chooses L. If B chooses 'randomly' and/or A chooses 'randomly', B receives with 50% probability 100 ECU (and A 0 ECU), and with 50% probability 0 ECU (and A 100 ECU.).

Please be patient until all participants have read the instructions. Before the experiment starts, please answer the following comprehension questions.

F.2 Comprehension Questions

Comprehension Questions

Question 1 Assume B chooses 'L'. What are A's and B's payoffs in situation 2?

Participant A's payoff is:

Participant B's payoff is:

Question 2 What are A's and B's payoffs if B chooses 'L' in situation 1?

Participant A's payoff is:

Participant B's payoff is:

Question 3 Assume B chooses 'R'. What are A's and B's payoffs?

Participant A's payoff is:

Participant B's payoff is:

Question 4 If B chooses 'random choice',...

...both participants receive 100 ECU: ○ false ○ true

...both participants receive with equal probability either 0 or 100 ECU: ○ false ○ true

Question 5 Please answer the following true/false statements.

In situation 2, participant B can determine A's choice, irrespective of what A has chosen: ○ false ○ true

In situation 1, participant B cannot influence A's decision: ○ false ○ true

[In treatment LIE, Question 5 read as follows:

Question 5 Please answer the following true/false statements.

In situation 2, participant B transmits A's and B's decisions to participant C without learning A's actual decision: ○ false ○ true

In situation 1, participant C does not learn either A's or B's decision: ○ false ○ true

]

[In treatment SPY, Question 5 read as follows:

Question 5 Please answer the following true/false statements.

In situation 2, participant B learns participant A's decision ○ false ○ true

In situation 1, no participant learns the other participant's decision: ○ false ○ true

]

G Results: Bs' behaviour across part 1 and part 2.

G.1 Competitive payoffs

LIE competitive ($n = 44$)

		Prob(S_2) part 2 $< 50\%$	50%	$> 50\%$
Prob(S_2) part 1	$< 50\%$	3	1	5
	50%	4	20	6
	$> 50\%$	0	3	2

Notes: 43% within]24%, 63%] (19 of 44) Bs opt for a different situation in part 2.

SPY competitive ($n = 53$)

		Prob(S_2) part 2 $< 50\%$	50%	$> 50\%$
Prob(S_2) part 1	$< 50\%$	2	0	3
	50%	0	8	4
	$> 50\%$	0	6	30

Notes: 25% within]11%, 42%] (13 of 53) Bs opt for a different situation in part 2.

SAB competitive ($n = 54$)

		Prob(S_2) part 2 $< 50\%$	50%	$> 50\%$
Prob(S_2) part 1	$< 50\%$	0	1	1
	50%	3	5	7
	$> 50\%$	1	8	28

Notes: 39% within]22%, 57%] (21 of 54) Bs opt for a different situation in part 2.

LIE competitive

		part 2 S_1	alt	self
part 1	S_1	10	4	5
	alt	5	8	4
	self	5	0	3

Notes: 33%]6%, 73%] (4 of 12) altruists who arrive in S_2 again, are selfish in part 2.

SPY competitive

		part 2 S_1	alt	self
part 1	S_1	6	0	7
	alt	0	0	0
	self	10	0	30

Notes: No altruism occurs either in part 1 or part 2.

SAB competitive

		part 2 S_1	alt	self
part 1	S_1	6	11	9
	alt	6	5	9
	self	2	3	3

Notes: 64%]27%, 91%] (9 of 14) altruists who arrive in S_2 again, are selfish in part 2.

G.2 Payoff neutrality

LIE neut ($n = 47$)

		Prob(S_2) part 2 $< 50\%$	50%	$> 50\%$
Prob(S_2) part 1	$< 50\%$	2	4	2
	50%	4	29	3
	$> 50\%$	0	2	1

Notes: 32% within]15%, 51%] (15 of 47) Bs opt for a different situation in part 2.

SPY neut ($n = 53$)

		Prob(S_2) part 2 $< 50\%$	50%	$> 50\%$
Prob(S_2) part 1	$< 50\%$	1	1	0
	50%	2	25	5
	$> 50\%$	1	8	10

Notes: 32% within]16%, 50%] (17 of 53) Bs opt for a different situation in part 2.

SAB neut ($n = 52$)

		Prob(S_2) part 2 $< 50\%$	50%	$> 50\%$
Prob(S_2) part 1	$< 50\%$	3	0	1
	50%	2	26	2
	$> 50\%$	5	5	8

Notes: 29% within]14%, 47%] (15 of 52) Bs opt for a different situation in part 2.

LIE neut

		part 2 alt	self
part 1	alt	2	8
	self	4	33

Notes: 80%]35%, 98%] (8 of 10) altruists who arrive in S_2 again, are selfish in part 2.

SPY neut

		part 2 alt	self
part 1	alt	3	1
	self	4	45

Notes: 25%]0%, 89%] (1 of 4) altruists who arrive in S_2 again, are selfish in part 2.

SAB neut

		part 2 alt	self
part 1	alt	1	7
	self	6	38

Notes: 88%]35%, 99%] (7 of 8) altruists who arrive in S_2 again, are selfish in part 2.

H How does B expect A to punish or reward B's procedural choice?

H.1 Average positive and negative expected changes to B's payoff

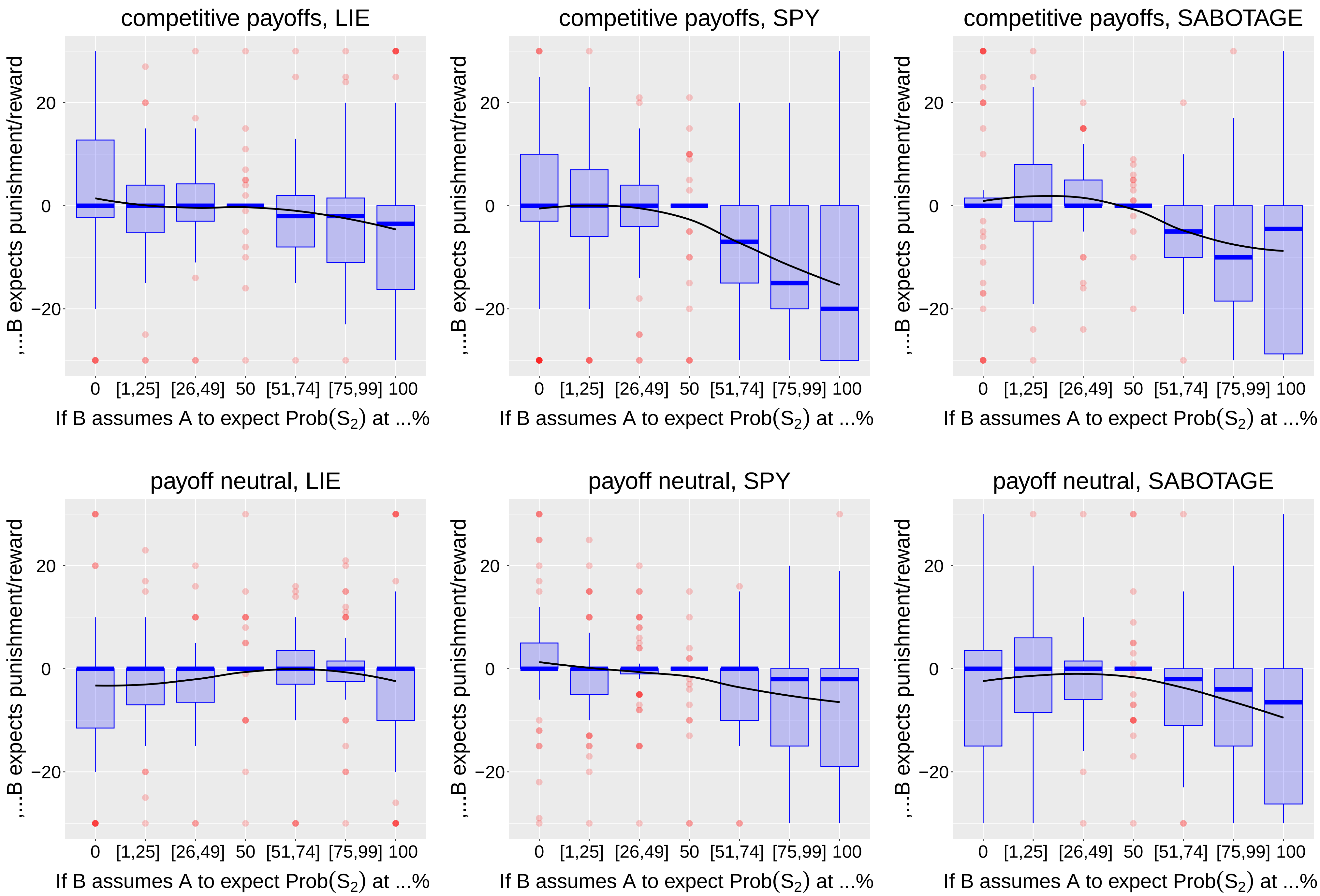
COMPETITIVE PAYOFFS ↓			B'S CHOICE OF Prob(S_2): 0	]0%, 25%]	]25%, 49%[	0.5	]50%, 75%]	]75%, 99%]	1
LIE	expected change to B's payoff	> 0	18 16.06	16 10.31	14 8.86	8 9.88	14 9.64	12 15.42	9 23.67
		0	14	12	13	30	7	8	12
		< 0	12 -17.92	16 -11.75	17 -8.65	6 -11.67	23 -8.57	24 -11.75	23 -18.17
SPY	expected change to B's payoff	> 0	19 15.0	20 12.35	16 9.25	8 10.38	6 11.5	6 12.33	5 17
		0	19	16	19	36	11	8	10
		< 0	15 -20.4	17 -15.29	18 -12	9 -17.22	36 -13.86	39 -18.21	38 -23.37
SAB	expected change to B's payoff	> 0	14 20.57	18 13.17	22 8.59	9 4.67	8 7.13	8 14.12	10 15.40
		0	27	19	19	41	10	10	14
		< 0	13 -17.08	17 -10.59	13 -8.00	4 -9.25	36 -10	36 -14.67	30 -20.70

PAYOFF NEUTRALITY ↓			B'S CHOICE OF Prob(S_2): 0	]0%, 25%]	]25%, 49%[	0.5	]50%, 75%]	]75%, 99%]	1
LIE	expected change to B's payoff	> 0	10 17.40	7 11.29	9 9.22	8 11.62	15 7.60	12 11.33	11 15.91
		0	18	18	18	33	17	20	19
		< 0	19 -16.68	22 -11.14	20 -10	6 -13.50	15 -10.40	15 -9.27	17 -17.47
SPY	expected change to B's payoff	> 0	17 14.82	13 10.85	13 8.92	5 6.60	7 8.29	7 9.14	8 14.50
		0	23	23	26	39	20	19	18
		< 0	13 -12.62	17 -10.94	14 -9.00	9 -12.11	26 -9.89	27 -12.89	27 -16.89
SAB	expected change to B's payoff	> 0	15 13.80	19 9.74	17 6.65	8 12.25	11 9.27	10 7.60	6 16.50
		0	18	14	13	33	11	13	15
		< 0	19 -17.53	19 -12.79	22 -9.77	11 -10.91	30 -11.27	29 -14.24	31 -19.00

Table A3: B's expectations: average expected reward, average expected punishment per choice of Prob(S_2), and number of B participants expecting A to reward, punish, or not to change their own payoff at all.

Notes. As reading example, take the first panel which displays treatment LIE with competitive payoffs. There are 18 B participants who expect A will increase their payoff if they set Prob(S_2) = 0. On average, these 18 B participants expect A to reward (increase B's payoff) by 16.06 ECU. 14 B participants do not expect A will change their payoff if they set Prob(S_2) = 0. 12 B participants expect A will decrease their payoff of they set Prob(S_2) = 0. On average, these 12 B participants expect A will punish (decrease B's payoff) by 17.92 ECU.

Figure A10: B's beliefs about A's decision to punish or reward B's choice of Prob(S_2). Upper three graphs: Competitive Payoffs – B impairs A's decision rights in S_2 in LIE and SABOTAGE; Lower three graphs: Payoff Neutrality – B cannot impair A's decision rights.



I How does B expect A to punish or reward B's procedural choice?

Figure A11: B's BELIEFS ABOUT A's DECISION TO PUNISH OR REWARD B's CHOICE OF $Prob(S_2)$. LEFT: PAYOFF NEUTRALITY – B CANNOT IMPAIR A's DECISION RIGHTS; RIGHT: COMPETITIVE PAYOFFS – B IMPAIRS A's DECISION RIGHTS IN S_2 IN TREATMENTS LIE AND SABOTAGE.

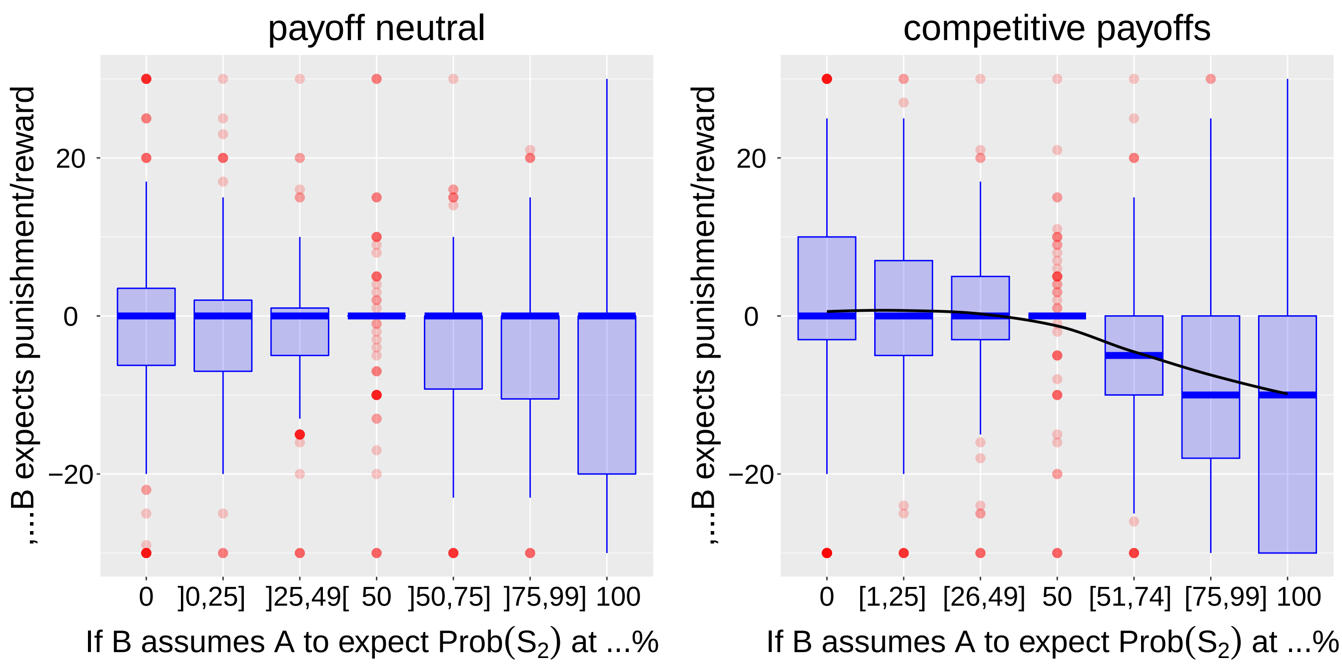


Figure A12: TREATMENT LIE: B's BELIEFS ABOUT A's DECISION TO PUNISH OR REWARD B's CHOICE OF $Prob(S_2)$. LEFT: PAYOFF NEUTRALITY – B CANNOT IMPAIR A's DECISION RIGHTS; RIGHT: COMPETITIVE PAYOFFS – B IMPAIRS A's DECISION RIGHTS IN S_2.

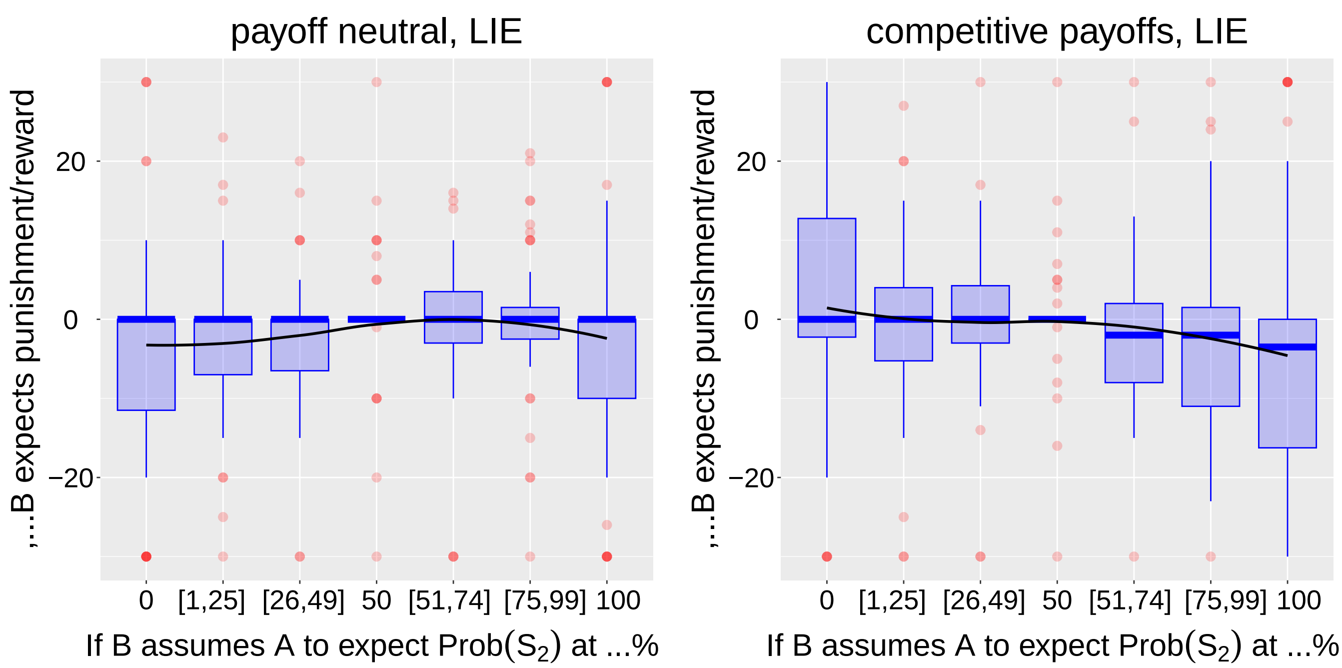


Figure A13: Treatment SPY: B's beliefs about A's decision to punish or reward B's choice of $Prob(S_2)$. Left: payoff neutrality – B cannot impair A's information or decision rights; Right: competitive payoffs – B impairs A information (but not her decision) rights.

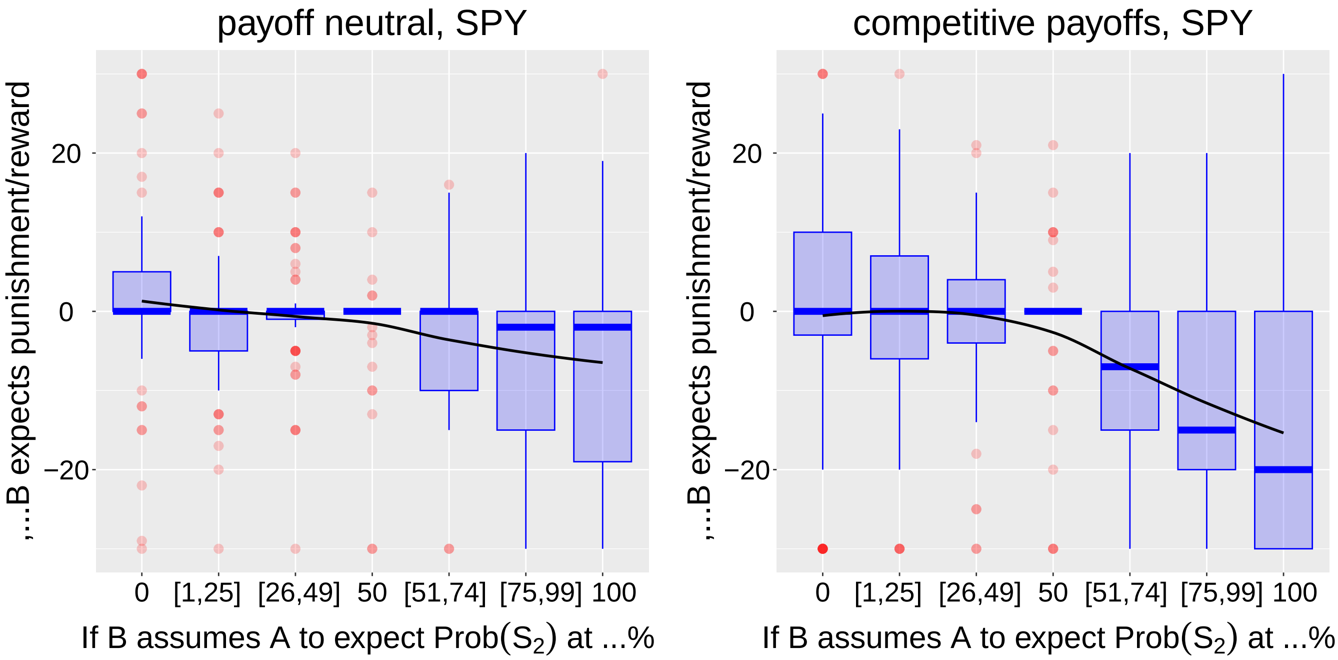
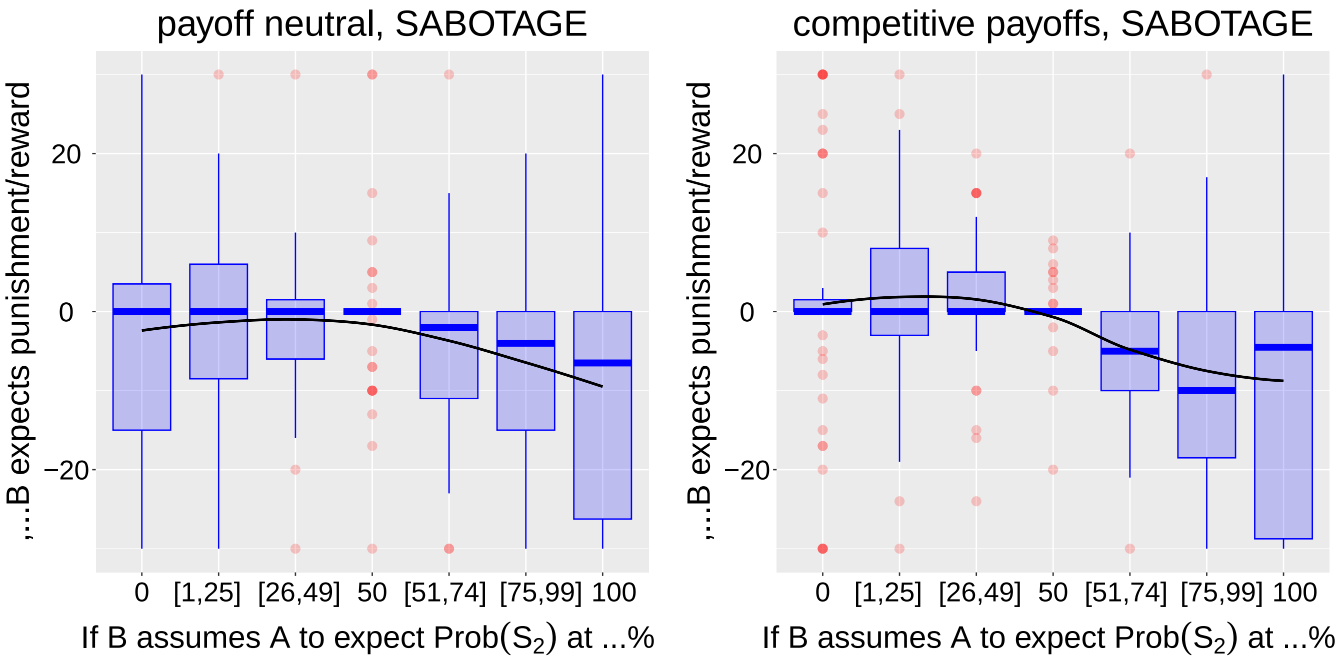


Figure A14: Treatment SABOTAGE: B's beliefs about A's decision to punish or reward B's choice of $Prob(S_2)$. Left: payoff neutrality – B cannot impair A's decision rights; RIGHT: competitive payoffs – B impairs A's decision rights in S_2..



J Taxonomy of moral argumentation by Kohlberg and Piaget

preconventional argumentation	preference models
Kohlberg 1. Orientation toward punishment and obedience, physical and material power. Rules are obeyed to avoid punishment. ***Kohlberg 2.*** Naïve hedonistic orientation. The individual conforms to obtain rewards.	(...)
conventional argumentation	**preference models**
Kohlberg 3. Orientation toward interindividual mutual relations. "Good boy/girl" orientation to win the approval and maintain the expectations of one's immediate group. The individual conforms with norms and expectations and shows good intentions to avoid disapproval. One earns approval by being "nice". ***Kohlberg 4.*** Orientation toward law and order, and in particular societal expectations and moral rules from outside one's immediate peer group since these maintain and ensure the continuity of the social order.	guilt aversion (Battigalli and Dufwenberg 2007), inequity aversion (Fehr and Schmidt 1999, Bolton and Ockenfels 2000), reciprocal preferences (Falk and Fischbacher 2006, Dufwenberg and Kirchsteiger 2007), preferences for equal expected payoffs (Bolton et al. 2005) and preferences for kind procedures (Sebald 2010)
postconventional argumentation	**preference models**
Kohlberg 5. Orientation toward the social contract. Duties are defined by the social contract and the equality of rights resulting from the social contract. Emphasis of the mutual commitment and obligation in a liberal democratic basic order..	Chlaß et al.'s (2019) purely procedural preferences: equality of decision rights, information rights, transparency
Kohlberg 6. Orientation toward the universal ethical principle of conscience such as Kant's categorical imperative. Rightness of an act is derived from abstract, consistent such as the inalienability of human rights, the free will, and individuals' freedom to choose. Ethical principles are a priori truths inherent in rational beings as laid down by Kant's categorical imperative.	Alger and Weibull's (2013) Homo Moralis

Table A4: The six categories of Lawrence Kohlberg's taxonomy of moral argumentation and a list of economic preference models built on the respective criteria listed in each category.

Notes. Sources: Kohlberg, 1984; Ishida, 2006. Economic preference models listed are *preferences for kind procedures.* Sebald, A. (2010), Attribution and Reciprocity, Games and Economic Behavior, 68, pp. 339-352; *preferences for equal expected payoffs.* Bolton, G., Brandts, J.,Ockenfels A. (2005), Fair Procedures: Evidence From Games Involving Lotteries, Economic Journal, 115, pp. 1054-1076; *guilt aversion.* Battigalli, P. and M. Dufwenberg (2007), Guilt in Games, American Economic Review, Papers and Proceedings, 97, pp. 170-176; *reciprocal preferences* Falk, A., Fischbacher, U. (2006), A Theory of Reciprocity, Games and Economic Behavior, 54, pp. 293-315, Dufwenberg, M., Kirchsteiger, G. (2004), A Theory of Sequential Reciprocity, Games and Economic Behavior, 47, pp. 268-98. *inequity aversion* Fehr and Schmidt (1999), Bolton, G., Ockenfels A. (2000) ERC - A Theory of Equity, Reciprocity and Competition, American Economic Review, 90, pp. 166-193, Fehr, E., Schmidt, G. (1999), A Theory of Fairness, Competition and Cooperation, Quarterly Journal of Economics, 114, pp. 817-868, *purely procedural preferences.* Chlaß N., Güth, W., Miettinen, T. (2019), Purely procedural preferences – Beyond procedural equity and reciprocity, European Journal of Political Economy, 59, pp. 108-128; *Homo Moralis.* Alger, I. and Weibull, J.W. (2013), Homo Moralis - Preference Evolution Under Incomplete Information and Assortative Matching, Econometrica 81(6), pp. 2269-2302.

K An Excerpt of the Moral Judgement Test by Georg Lind (1976, 2008)

Doctor

A woman had cancer and she had no hope of being saved. She was in terrible pain and so weak that a large dose of a pain killer such as morphine would have caused her death. During a temporary period of improvement, she begged the doctor to give her enough morphine to kill her. She said she could no longer stand the pain and would be dead in a few weeks anyway. The doctor decided to give her a overdose of morphine.

Do you agree or disagree with the doctor's action ...

I strongly disagree						I strongly agree
-3	-2	-1	0	1	2	3

How acceptable do you find the following arguments *in favor* of the doctor's actions?
Suppose someone argued he acted *rightly*...

...because the doctor had to act according to his conscience. The woman's condition justified an exception to the moral obligation to preserve life

I strongly reject								I strongly accept
-4	-3	-2	-1	0	1	2	3	4

. . .

...because the doctor was the only one who could fulfill the woman's wish; respect for her wish made him act as he did.

I strongly reject								I strongly accept
-4	-3	-2	-1	0	1	2	3	4

How acceptable do you find the following arguments *against* the doctor's actions?
Suppose someone argued he acted *wrongly*

. . .

...because he acted contrary to his colleagues' convictions. If they are against mercy-killing the doctor shouldn't do it.

I strongly reject								I strongly accept
-4	-3	-2	-1	0	1	2	3	4

. . .

...because one should be able to have complete faith in a doctor's devotion to preserving life even if someone with great pain would rather die

I strongly reject								I strongly accept
-4	-3	-2	-1	0	1	2	3	4

NOTE: This excerpt of the moral judgement test MJT is reprinted with kind permission by Georg Lind. It does not faithfully reproduce the formatting of the original test. For ease of readability, the original test numbers each item, and the alignment slightly differs from this excerpt. Dots represent items which have been left out. The full test cannot be republished due to copyright protection, but is freely available from https://moralcompetence.net – the password for the download area has been permanently published on the website after Georg Lind's demise. Upon submission to a journal, the test can be sent along for refereeing and editorial purposes.

L Klages's and Gensicke's (2006) materialism - postmaterialism scales[41]

Table A5: QUESTIONNAIRE ITEMS FOR EACH OF KLAGES'S AND GENSICKE'S THREE VALUE DIMENSIONS (CATEGORIES) TO IDENTIFY MATERIALISTS, POSTMATERIALISTS, AND MIXED VALUE TYPES IN THE GERMAN POPULATION (KLAGES AND GENSICKE, 2006).

value category I **duty and acceptance values**	**value category II** **hedonistic and materialistic values**	**value category III** **idealistic values and public participation**[42]
✓ respect law and order	✓ have a high living standard	✓ develop one's fantasy and creativity
✓ need and quest for security	✓ hold power and influence	✓ help socially disadvantaged and socially marginal groups
✓ be hard-working and ambitious	✓ enjoy life to the fullest	✓ also tolerate opinions with which one actually cannot really agree
	✓ assert oneself, and one's needs against others	✓ be politically active

conventionalists	high scores on value category I (Inglehart's classic materialist values). Intermediate scores for value categories II and III. Clear hierarchy between value category I and II/III → approximate Inglehart's 'materialists' but Inglehart classifies value category II as 'materialist' values (with the exception of item 3) and not as a separate dimension.
idealists	high scores on value category III. Intermediate scores for value category II. Clear hierarchy between both value categories. Lower scores on value category I than conventionalists → approximation of Inglehart's postmaterialists.
hedonic materialists	score lower than conventionalists in value category I and lower than idealists in value category III. No hierarchy between value categories (all similarly important).
resigned without perspective	lower scores on category I than conventionalists and lower scores on value category III than idealists. Lowest scores in value category II. One of Inglehart's 'mixed types'.
realists	second lowest value hierarchy after hedonists, high scores on category I and relatively high scores on category II; 'synthesis' of values. One of Inglehart's 'mixed types'.

[41] Klages and Gensicke (2006) use these value categories to obtain the clusters (types) below: conventionalists, resigned people, realists, hedo-materialists, and idealists. In this paper, we do not cluster people into these groups; we use each individuals' average rating for all three value categories to model B participants' choice of the fair rules (type i)), or their altruism (type ii) under the unfair rules as opposed to the selfish type (type iv). The average rating is the mean rating over all questionnaire items pertaining to the same value category. Individuals rate each item from 1 to 7.

[42] Category III corresponds to Ingelhart's postmaterialism value scale. Higher mean ratings on value category III make the procedural type i) in section 5 more likely. Category II mostly belongs to Inglehart's materialist values. Higher mean ratings of this value category makes the altruistic type ii) in section 5 more likely. Value category I does not significantly influence B participants' choices in the experiment.

M Raw data: B participants' choices of situation and allocation

M.1 Competitive payoffs

altruistic allocation

	Prob(S_2) < 50	Prob(S_2) ≥ 50
LIE	2	15
SPY	0	0
SAB	0	20

Notes. S_2 + GIVE ALL: 2+15+20+12+8+24=81.

situation 1

	Prob(S_2) < 50%	Prob(S_2) ≥ 50%
LIE	7	12
SPY	5	8
SAB	2	24

Notes. S_1: 7+5+2=14

selfish allocation

	Prob(S_2) < 50%	50%	> 50%
LIE	0	5	3
SPY	0	8	32
SAB	0	1	7

Notes. SELFISH: 3+32+7=42. FAIR COIN: 5+8+1=14.

M.2 Competitive payoffs: punishment and reward

altruistic allocation

	Prob(S_2) < 50	Prob(S_2) ≥ 50
LIE	1	11
SPY	0	0
SAB	2	17

Notes. S_2 + GIVE ALL: 1+2+11+17+16+14+12=73.

situation 1

	Prob(S_2) < 50%	Prob(S_2) ≥ 50%
LIE	4	16
SPY	2	14
SAB	2	12

Notes. S_1: 4+2+2=8

selfish allocation

	Prob(S_2) < 50%	50%	> 50%
LIE	2	8	2
SPY	0	7	30
SAB	0	4	17

Notes. SELFISH: 2+30+17=49. FAIR COIN: 2+8+7+4=21.

M.3 Payoff Neutrality

altruistic allocation

	Prob(S_2) < 50	50	> 50
LIE	3	6	1
SPY	0	3	1
SAB	1	6	1

$\sum 22$

selfish allocation

	Prob(S_2) < 50%	50%	> 50%
LIE	5	30	2
SPY	2	29	18
SAB	3	24	17

$\sum 130$

M.4 Payoff Neutrality: punishment and reward

altruistic allocation

	Prob(S_2) < 50	50	> 50
LIE	2	3	1
SPY	1	1	3
SAB	2	3	2

$\sum 18$

selfish allocation

	Prob(S_2) < 50%	50%	> 50%
LIE	4	32	5
SPY	3	33	12
SAB	8	28	9

$\sum 134$

N B participants' demographics and their ethical preferences

dependent variable →	*Kohlberg class 1*	*Kohlberg class 2*	*Kohlberg class 3*	*Kohlberg class 4*	*Kohlberg class 5*	*Kohlberg class 6*
B participants →	151	151	151	150	151	151
Intercept	2.412^a (0.925)	2.462^a (0.831)	2.220^b (1.098)	2.497^a (0.939)	1.650 (0.999)	2.529^a (0.843)
Age	−0.040 (0.031)	−0.029 (0.029)	−0.026 (0.033)	−0.027 (0.029)	−0.018 (0.033)	−0.048 (0.030)
Gender: female	−0.090 (0.170)	0.187 (0.180)	0.087 (0.165)	0.130 (0.174)	0.074 (0.185)	0.018 (0.152)
Envy	−0.189 (0.163)	-0.327^c (0.167)	-0.313^c (0.159)	-0.455^a (0.160)	−0.278 (0.169)	−0.154 (0.154)
Risk aversion	0.032 (0.049)	−0.051 (0.050)	−0.029 (0.050)	−0.027 (0.047)	−0.020 (0.051)	-0.091^b (0.043)
Education	-1.086^a (0.370)	−0.571 (0.353)	−0.619 (0.696)	−0.694 (0.519)	−0.439 (0.431)	−0.337 (0.213)
Law	-1.796^a (0.403)	-1.504^a (0.383)	-1.713^b (0.715)	-1.730^a (0.511)	-1.327^a (0.482)	-1.352^a (0.283)
IT	−1.269 (0.853)	-1.686^b (0.761)	−1.637 (1.001)	-1.984^b (0.774)	−1.294 (0.859)	-1.919^c (1.022)
Philosophy	-0.635^c (0.364)	−0.078 (0.335)	0.020 (0.694)	−0.196 (0.511)	-1.207^a (0.421)	0.305 (0.218)
Social and Behavioral Sciences	-1.503^a (0.361)	-1.167^a (0.363)	−1.021 (0.689)	-1.099^b (0.521)	-0.833^c (0.435)	-0.656^a (0.228)
Medicine	-1.684^a (0.484)	-1.246^a (0.420)	−1.221 (0.775)	−0.933 (0.566)	−0.735 (0.488)	-0.730^c (0.392)
Business and Economics	-1.433^a (0.400)	-1.172^a (0.432)	-1.235^c (0.732)	-1.137^b (0.558)	-1.072^b (0.477)	-0.965^a (0.308)
Engeneering	-1.317^a (0.500)	-1.026^b (0.450)	−0.880 (0.767)	−0.959 (0.584)	−0.627 (0.528)	−0.260 (0.339)
Languages	-1.167^a (0.441)	-1.020^a (0.405)	−0.727 (0.724)	-0.978^c (0.569)	−0.252 (0.446)	−0.031 (0.288)
Sciences	-1.381^a (0.440)	-0.808^c (0.475)	−0.836 (0.733)	-1.014^c (0.605)	−0.406 (0.524)	−0.326 (0.285)

Table A6: Correlation of all six *Kohlberg classes* and *B* participants' demographics, current study, treatment competitive payoffs.

Notes. Linear regression with robust standard errors of *Kohlberg classes 1-6* from table 5 on variables listed. **Age** – *B* participants' age in whole years, ranges from 18 to 35 with a median of 23; **Envy** – Dummy taking on a value of One if *B* allocated 10 ECU to herself and 10 ECU to *A*, rather than 10 ECU to herself and 20 ECU to *A*; **Risk aversion** – ordinal variable ranging from 1 to 10; indicates at which lottery on a ten-item Holt-Laury lottery list *B* starts to prefer the sure payoff of 25 ECU to a binary lottery of 10 ECU and 35 ECU, see section 3.3; **Fields of study** – miscellaneous category is Field of Study: University of Applied Sciences. Coefficients will depend on changes in the base category (for instance, all positive if *Law* is the base category), significance levels of particular fields, however, are reassuringly robust to changes in the base category. **Philosophy**: single participant, individual effect. **Materalism and postmaterialism scores**. *Kohlberg class 5* does not significantly correlate with the former (−0.017, *p-value* = 0.82) or the latter (−0.072, *p-value* = 0.48).

O Robustness

O.1 Bs' altruism: multinomial models for table 5

DEPENDENT VARIABLE: THREE DEPARTURES FROM RATIONAL SELF-INTEREST (BASE CATEGORY)				
$x_j \downarrow$	S_1	S_2 + GIVE ALL	FAIR COIN	$\partial F/\partial x_j$
nr. of obs.	9	37	6	62
Intercept	4.935 (4.756)	10.755^a (3.977)	5.743 (6.529)	
Kohlberg class 1	-2.052^c (1.133)	-3.122^a (1.057)	-3.022^b (1.196)	-0.259^a (0.066)
Kohlberg class 3	−0.487 (1.077)	0.991 (0.858)	0.356 (1.077)	0.066 (0.066)
Kohlberg class 5	4.635^a (1.570)	1.430^c (0.758)	2.487^a (0.869)	0.164^a (0.057)
Kohlberg class 6	-2.029^c (1.151)	0.554 (0.654)	0.866 (1.030)	0.022 (0.051)
Dummy LIE	3.688^b (1.877)	2.321 (1.640)	4.256^b (2.027)	0.227^b (0.095)
postmaterialism	0.054 (0.644)	-0.919^c 0.528	−0.427 (0.707)	-0.067^c (0.034)
materialism	-1.466^a (0.436)	-0.886^a (0.308)	-1.136^b ((0.484))	-0.084^a (0.026)
Nagelkerke's R^2		0.56		

Table A7: The independent Logits from table 5 in one multinomial model, S_2+ Give all <u>*without*</u> Bs who pay for S_2 and end up in S_1.

DEPENDENT VARIABLE: THREE DEPARTURES FROM RATIONAL SELF-INTEREST (BASE CATEGORY)				
$x_j \downarrow$	S_1	S_2 + GIVE ALL	FAIR COIN	$\partial F/\partial x_j$
nr. of obs.	9	73	6	98
Intercept	4.464 (4.000)	10.034^a (3.230)	4.640 (5.472)	
Kohlberg class 1	-1.715^b (0.804)	-2.432^a (0.710)	-2.575^a (0.910)	-0.162^a (0.047)
Kohlberg class 3	0.037 (0.844)	1.109 (0.703)	0.665 (0.863)	0.068 (0.045)
Kohlberg class 5	2.476^a (0.808)	1.002^b (0.466)	1.623^a (0.612)	0.080^b (0.034)
Kohlberg class 6	-0.594^c (0.660)	0.485 (0.486)	0.856 (0.632)	0.027 (0.030)
Dummy LIE	3.134^b (1.461)	1.537 (1.174)	3.698^b (1.566)	0.121^c (0.066)
postmaterialism	−0.090 (0.503)	-0.617^c 0.404	−0.295 (0.576)	−0.038 (0.025)
materialism	-1.099^a (0.340)	-1.031^b (0.262)	-1.136^b ((0.417))	-0.064^a (0.021)
Nagelkerke's R^2		0.40		

Table A8: The independent Logits from table 5 in one multinomial model. S_2+ Give all <u>*with*</u> Bs who pay for S_2 and end up in S_1.

Notes. Multinomial Logit regression of 4 nominal categories l where $l = 0 \leftrightarrow$ rational self-interest with B participants opting into S_2 and taking all payoff; $l = 1 \leftrightarrow B$ participants who toss a fair coin, arrive in S_2 and take all payoff; $l = 2 \leftrightarrow B$ participants who opt into S_1, and $l = 3 \leftrightarrow B$ participants who opt into S_2 and give all payoff, B participants who toss a fair coin, arrive in S_2 and give all payoff and, for table A8 only: also B participants who opt into S_2 but end up in S_1. Marginal effects represent the average over all individual marginal effects. Marginal effects are the <u>*negative*</u> average marginal effect of each variable x_j on the base category $l = 0$, revealing by how much a one unit change in the variable increases the likelihood of a <u>departure</u> from rational self-interest. Standard errors of marginal effects are obtained by the Delta method. **Moral judgement variables** – *Kohlberg class 1* ranges from -1.7845 to 2.8872 with median 0.2177; *Kohlberg class 3* from -1.6258 to 3.1242 with median 0.3534; *Kohlberg class 5* from -1.9995 to 3.2368 with median 0.2914. **Miscellaneous variables** – *Dummy LIE* takes on a value of One if treatment was LIE and Zero otherwise, *postmaterialism* ranges from 2 to 7 with a median of 6, *materialism* from 1 to 7 with a median of 5.

O.2 *Bs*' altruism: sample size augmented by SPY and demographic controls

Dependent Dummy variable →	S_1 (1) vs. Selfish (0)	S_2 + Give All (1) vs. Selfish (0)	S_1 (1) vs. S_2 + Give All (0)
B participants	56 (14 vs. 42)	123 (81 vs. 42)	94 (14 vs. 81)
Kohlberg class 1	-0.132^b (0.063)	-0.101^a (0.035)	0.084 (0.063)
Kohlberg class 2	0.289^c (0.156)	−0.015 (0.045)	0.004 (0.056)
Kohlberg class 3	0.105^c (0.063)	0.067^b (0.034)	−0.093 (0.060)
Kohlberg class 4	−0.118 (0.072)	0.055 (0.056)	0.060 (0.063)
Kohlberg class 5	0.228^a (0.071)	0.101^a (0.035)	0.028 (0.057)
Kohlberg class 6	-0.297^a (0.073)	0.000 (0.033)	-0.139^a (0.052)
Dummy LIE	0.340^a (0.093)	0.519^a (0.040)	
Dummy SABOTAGE	-0.135^c (0.077)	0.467^a (0.050)	
Risk aversion	-0.058^b (0.028)	0.017 (0.018)	-0.030^c (0.015)
Envy	0.148^c (0.080)	0.018 (0.065)	−0.043 (0.078)
Age	0.002 (0.011)	0.000 (0.010)	−0.002 (0.015)
Gender:female	-0.280^a (0.073)	−0.050 (0.077)	−0.030 (0.066)
Business and Economics	0.094	0.129	0.620^a
Medicine	NA	0.119	NA
Law	0.389^b	0.128	0.599^a
Social and Behavioral Sciences	0.310^a	0.015	0.548^a
Sciences	NA	0.192^c	0.641^a
Philosophy	NA	NA	NA
IT	NA	0.139	NA
Engeneering	0.470^a	0.160	0.643^a
Languages	NA	−0.053	NA
Education	0.322^a	0.026	0.635^a
Count R^2	0.89	0.86	0.86

Table A9: Ethical determinants of *B* participants' departures from rational self-interest, and the type of altruistic behavior they adopt (marginal effects).

Note: Significance levels of z-tests are indicated by $a : p < .01$, $b : p < .05$, $c : p < .10$.

Notes. Logit regressions with robust standard errors of – column 1: *B* participants opting into S_1 vs. *B* participants opting into S_2 and taking all payoff; – column 2: *B* participants who arrive in S_2 and give all payoff and *B* participants who opt into S_2 but end up in S_1 vs. *B* participants opting into S_2 and taking all payoff; and – column 3: *B* participants who opt into S_1 vs. *B* participants who arrive in S_2 and give all payoff and *B* participants who opt into S_2 but end up in S_1. Materialism and postmaterialism scores were only collected in LIE and SABOTAGE where we expected altruism and are not available elsewhere. Demographic controls include those intercepting the link between *Kohlberg class 5* and *B* participants' *purely procedural concerns* (Chlaß et al., 2019; Chlaß et al., 2023). Marginal effects take the mean over all individual marginal effects; their standard errors are obtained by the Delta method. **Age** – *B* participants' age in whole years, ranges from 18 to 35 with a median of 23; **Envy** – Dummy taking on a value of One if *B* allocated 10 ECU to herself and 10 ECU to *A*, rather than 10 ECU to herself and 20 ECU to *A*; **Risk aversion** – ordinal variable ranging from 1 and 10; indicates at which lottery in a ten-item Holt-Laury lottery list *B* starts to prefer the sure payoff of 25 ECU to a binary lottery of 10 ECU to 35 ECU, see section 3.3.; **Fields of study** – miscellaneous categories are Arts, Philosophy, and University of Applied Sciences.

Dependent Variable: three Departures from Rational Self-Interest (Base Category)			
$x_j \downarrow$ specification $\rightarrow$	(1)	(2)	(3)
Kohlberg class 1	-0.071^c (0.039)	-0.079^b (0.039)	
Kohlberg class 2	-0.036 (0.045)		
Kohlberg class 3	0.086^b (0.040)	0.088^b (0.037)	
Kohlberg class 4	0.041 (0.053)		
Kohlberg class 5	0.105^b (0.042)	0.101^a (0.037)	
Kohlberg class 6	-0.031 (0.041)		
Dummy LIE	0.444^a (0.084)	0.438^a (0.086)	
Dummy SABOTAGE	0.240^a (0.062)	0.221^a (0.055)	
Risk aversion	-0.004 (0.019)	-0.002 (0.018)	-0.021 (0.022)
Envy	0.054 (0.063)	0.055 (0.061)	-0.015 (0.073)
Age	-0.003 (0.011)	-0.002 (0.011)	0.007 (0.014)
Gender:female	-0.093 (0.066)	-0.090 (0.066)	-0.085 (0.081)
Business/Economics	0.036	0.057	-0.044
Medicine	-0.374^b	-0.374^b	-0.435^a
Law	0.139	0.167	0.208
Social and Behavioral Sciences	-0.073	-0.059	0.058
IT	-0.740^a	-0.713^a	-0.754^a
Languages	-0.412^a	-0.420^a	-0.340^b
Education	-0.071	0.075	-0.026
Nagelkerke's R^2	0.60	0.58	0.16
B participants	150	151	151

Table A10: The independent Logits from Table A9 in one multinomial model (Marginal Effects), three different specifications.

Notes. Multinomial Logit regression of 4 nominal categories l where $l = 0 \leftrightarrow$ rational self-interest with B participants opting into S_2 and taking all payoff; $l = 1 \leftrightarrow B$ participants who toss a fair coin, arrive in S_2 and take all payoff; $l = 2 \leftrightarrow B$ participants who opt into S_1, and $l = 3 \leftrightarrow B$ participants who opt into S_2 and give all payoff, B participants who opt into S_2 but end up in S_1, and B participants who toss a fair coin, arrive in S_2 and give all payoff. Marginal effects represent the average over all individual marginal effects. We report the *negative* average marginal effect of each variable x_j on the base category $l = 0$, revealing by how much a one unit change in the variable increases the likelihood of a departure from rational self-interest. Standard errors of marginal effects are obtained by the Delta method. For information on variables, see notes to table A9. Since all categories are regressed simultaneously and the same variables must be available for each category, fields of study are reduced to *Law*, *Social and Behavioral Sciences*, and *Business and Economics* which showed a latent correlation with *Kohlberg class 5* in table A6, omitting *Philosophy* with a single participant.

P Improving $A's$ position of decision rights: treatment punishment/reward

P.1 LIE and SABOTAGE competitive payoffs only

Dependent variable → x_j ↓	S_1 (1) vs. Selfish (0)	S_2 + give all (1) vs. Selfish (0)		Fair Coin (1) vs. Selfish (0)	All Four Categories	
B participants	15 (6 vs. 19)	49 (30 vs. 19)	78 (59 vs. 19)	33 (14 vs. 19)	69	97
Kohlberg class 1	-0.169^a (0.064)	-0.098^b (0.049)	-0.073 (0.062)	-0.219^a (0.078)	-0.080 (0.052)	-0.063 (0.054)
Kohlberg class 2	-0.199^a (0.064)	-0.082 (0.052)	0.024 (0.068)	0.091 (0.137)	-0.037 (0.065)	0.044 (0.058)
Kohlberg class 3	0.241^a (0.059)	-0.049 (0.050)	-0.056 (0.069)	-0.241^a (0.094)	-0.051 (0.060)	-0.054 (0.056)
Kohlberg class 4	0.278^a (0.090)	0.302^a (0.066)	0.118^c (0.066)	-0.053 (0.112)	0.218^a (0.065)	0.106^c (0.058)
Kohlberg class 5	-0.275^a (0.094)	-0.228^a (0.075)	-0.125^b (0.057)	0.266^b (0.112)	-0.206^a (0.063)	-0.123^a (0.047)
Kohlberg class 6	0.161^a (0.061)	0.096^a (0.057)	0.069 (0.066)	0.165^a (0.052)	0.083 (0.053)	0.047 (0.053)
Dummy LIE	0.445^a (0.070)	0.340^a (0.056)	0.289^a (0.050)	0.313^b (0.126)	0.332^a (0.085)	0.283^a (0.075)
postmaterialism		0.099^a (0.026)	(0.057) (0.037)		0.053 (0.038)	0.036 (0.037)
materialism	-0.027^c (0.015)	0.035 (0.028)	0.025 (0.039)	-0.115^a (0.036)	0.029 (0.037)	0.013 (0.035)
expected punishment, ranging from 0 to 30 ECU	-0.023^a (0.007)	-0.017^a (0.003)	-0.014^a (0.003)	-0.120^a (0.019)	-0.028^a (0.005)	-0.016^a (0.003)
expected reward, ranging from 0 to 30 ECU		-0.005 (0.003)	-0.008 (0.005)		-0.011^b (0.005)	-0.009^c (0.005)
Law		-0.425^a	-0.117		-0.235	-0.061
Social/Behavioral Sciences		-0.242^a	-0.041		-0.192^c	-0.037^c
Languages					-1.528^a	1.463^a
Count R^2 or Nagelkerke's R^2	0.96	0.84	0.83	0.88	0.74	0.60

Table A11: Independent Logits analogous to table 5 and a multinomial model, S_2+ Give all *without* (column 2 & 5) and *with* (column 3 & 6) Bs who pay for S_2 and end up in S_1 (Marginal Effects).

Notes. Variables listed which are not fitted, cannot be fitted for technical reasons (multicollinearity). **Columns 2-5**: Binary Logit models with robust errors analogous to Table 5 in the main text. S_2 + Give All vs. Selfish presents a first column with B participants who opt into S_2, arrive there, and give all payoff and B participants who toss a fair coin, arrive in S_2 and also give all payoff; and a second column which, in addition, includes B participants who opt into S_2 but end up in S_1. Marginal effects represent the average over all individual marginal effects. **Columns 6-7**: All Four Categories presents multinomial models (robust errors) of all four categories with a first column in which category S_2 vs. Give All excludes B participants who opt into S_2 but end up in S_1 and a second column where the category at hand includes the latter. Marginal effects of both multinomial models are the *negative* average marginal effect of each variable x_j on the base category Selfish, revealing by how much a one unit change in the variable increases the likelihood of a departure from rational self-interest. Standard errors of marginal effects are obtained by the Delta method. **Expected punishment** – B participants' beliefs by how much A will punish their choice of $Prob(S_2)$, ranges from 0 to 30 ECU with mean 4.633. **Expected reward** – B participants' beliefs by how much A will reward their choice of $Prob(S_2)$, ranges from 0 to 30 ECU with mean 3.214. For binary Logits, we continue to report the Count R^2 as goodness-of-fit measure, for multinomial Logits, we continue to report Nagelkerke's R^2.

P.2 adding treatment SPY, omitting materialism/postmaterialism scores

Dependent Dummy variable →	S_1 (1) vs. Selfish (0)	S_2 + give all (1) vs. Selfish (0)	Fair Coin (1) vs. Selfish (0)
B participants → x_j ↓	57 (8 vs. 49)	122 (73 vs. 49)	70 (21 vs. 49)
Kohlberg class 1	-0.108^c (0.057)	-0.094 (0.058)	-0.149^a (0.050)
Kohlberg class 2	-0.133^b (0.056)	0.016 (0.063)	-0.001 (0.064)
Kohlberg class 3	0.200^a (0.054)	0.064 (0.063)	-0.049^c (0.054)
Kohlberg class 4	0.098^b (0.050)	0.070 (0.069)	0.121^b (0.051)
Kohlberg class 5	-0.172^a (0.060)	-0.134^a (0.051)	-0.029 (0.046)
Kohlberg class 6	0.112^b (0.050)	0.069 (0.050)	0.056 (0.061)
Dummy LIE	0.387 (0.267)	0.460^a (0.040)	0.465^a (0.078)
Dummy SABOTAGE	-0.047 (0.053)	0.149^c (0.079)	0.029 (0.086)
expected punishment, ranging from 0 to 30 ECU	-0.012^b (0.005)	-0.010^a (0.004)	-0.037^a (0.007)
expected reward ranging from 0 to 30 ECU	-0.005 (0.005)	-0.008 (0.006)	-0.019^a (0.004)
Risk aversion	-0.004 (0.067)	-0.004 (0.018)	
Envy	-0.224^b (0.065)	-0.024 (0.078)	
Age	0.003 (0.011)	-0.019 (0.013)	
Gender: female	0.007 (0.122)	-0.038 (0.078)	
Law	-0.014	0.171	
Languages	0.211	0.217^a	
Count R^2	0.95	0.79	0.87

Table A12: Ethical determinants of B participants' departures from rational self-interest, and the type of altruistic behavior they adopt (marginal effects).

Note: Significance levels of z-tests are indicated by $a : p < .01$, $b : p < .05$, $c : p < .10$.

Notes. Variables listed which are not fitted, cannot be fitted for technical reasons (multicollinearity). Binary Logit regressions with robust errors analogous to table A9. S_1 vs. Selfish: B participants opting into S_1 vs. B participants opting into S_2 and taking all payoff; S_2 + Give all vs. Selfish: B participants who opt into S_2 and give all payoff, B participants who opt into S_2 but end up in S_1, and B participants who toss a fair coin (including 2 who actually pay for S_1), arrive in S_2, and give all payoff vs. B participants opting into S_2 and taking all payoff; and Fair coin vs. Selfish: B participants who toss a fair coin, arrive in S_2, and take all payoff vs. B participants opting into S_2 and taking all payoff. Materialism and postmaterialism scores were not collected in SPY since we did not expect any altruism. **Expected punishment**, see notes to table A11, ranges from 0 to 30 ECU with mean 6.44; **Expected reward**, see notes to table A11, ranges from 0 to 30 with mean 2.73. Independent Logits allow for individual specifications but run more easily into estimation difficulties. Multinomial specifications in table A13 specifically control for a wide range of demographics.

Dependent Variable: three Departures from Rational Self-Interest (Base Category)				
$x_j \downarrow$ *specification* $\rightarrow$	(1)	(2)	(3)	(4)
Kohlberg class 1	-0.102^a (0.039)	-0.050 (0.040)		
Kohlberg class 2	0.016 (0.064)			
Kohlberg class 3	0.054 (0.051)			
Kohlberg class 4	0.058 (0.066)			
Kohlberg class 5	-0.103^b (0.043)	-0.070^c (0.042)		
Kohlberg class 6	0.073^c (0.042)	0.117^a (0.043)		
Dummy LIE	0.417^a (0.067)	0.422^a (0.072)		
Dummy SABOTAGE	0.078 (0.069)	0.069^a (0.060)		
expected punishment, ranging from 0 to 30 ECU	-0.012^a (0.003)	-0.016^a (0.003)	-0.019^a (0.004)	
expected reward, ranging from 0 to 30 ECU	-0.009^b (0.005)	-0.009^b (0.004)	-0.007 (0.005)	
Risk aversion	0.002 (0.016)			
Envy	-0.050 (0.061)			
Age	-0.018 (0.012)			
Gender:female	-0.040 (0.074)			
Business/Economics	0.133	0.138	0.071	0.058
Medicine	-0.320^c	0.315	-0.444^b	-0.776^a
Law	0.258^c	0.323^b	0.373^b	0.437^a
Social and Behavioral Sciences	0.205^b	0.216^b	0.262	0.315^a
Languages	-0.003	-0.009	-0.056^b	-0.188
Sciences	-0.129	-0.144	-0.211^b	-0.160
Education	0.011	0.018	-0.070	0.029
Nagelkerke's R^2	0.64	0.58	0.43	0.26
B participants	150	151	151	151

Table A13: The independent Logits from Table A12 in one multinomial model (Marginal Effects), four different specifications.

Notes. Multinomial Logit regression of 4 nominal categories l where $l = 0 \leftrightarrow$ rational self-interest with B participants opting into S_2 and taking all payoff; $l = 1 \leftrightarrow B$ participants who toss a fair coin (including 2 who actually pay for S_1), arrive in S_2 and take all payoff; $l = 2 \leftrightarrow B$ participants who opt into S_1, and $l = 3 \leftrightarrow B$ participants who opt into S_2 and give all payoff, B participants who opt into S_2 but end up in S_1, and B participants who toss a fair coin (including 2 who actually pay for S_1), arrive in S_2 and give all payoff. Marginal effects represent the average over all individual marginal effects. We report the *negative* average marginal effect of each variable x_j on the base category $l = 0$ (rational self-interest), revealing by how much a one unit change in the variable increases the likelihood of a departure from rational self-interest. Standard errors of marginal effects are obtained by the Delta method. For information on variables, see notes to table A7 to A12. Fields of study are reduced to those available for each category; the remaining set includes all categories showing a latent correlation with *Kohlberg class 5* in table A6.

P.3 Taking A's decision rights out of B's hands: treatment payoff neutrality

Note: In treatment payoff neutrality, B's rational self-interested choice is to toss a fair coin, i.e. leave Prob(S_2) = 50%, and to take all payoff. A has no decision rights in either S_1 or S_2.

Dependent Dummy variable →	S_1 (1) vs. Fair Coin (0)		S_2 (1) vs. Fair Coin (0)		All Three Categories	
B participants → x_j ↓	97 (14 vs. 83)	111 (14 vs. 98)	123 (40 vs. 83)	138 (40 vs. 98)	134	151
Kohlberg class 1	-0.088^c (0.047)	-0.091^b (0.045)	-0.074 (0.052)	-0.083^c (0.048)	-0.107^c (0.058)	-0.121^b (0.054)
Kohlberg class 2	0.129^a (0.044)	0.127^a (0.041)	0.072 (0.051)	0.074 (0.047)	0.143^b (0.057)	0.140^a (0.052)
Kohlberg class 3	0.021 (0.047)	0.016 (0.044)	0.050 (0.046)	0.051 (0.045)	0.049 (0.052)	-0.048 (0.049)
Kohlberg class 4	0.049 (0.043)	0.050 (0.039)	0.027 (0.050)	0.053 (0.045)	0.049 (0.056)	0.077 (0.052)
Kohlberg class 5	-0.046 (0.034)	-0.048 (0.033)	-0.055 (0.048)	-0.061 (0.046)	-0.077 (0.052)	-0.087^c (0.049)
Kohlberg class 6	-0.052 (0.042)	-0.041 (0.036)	-0.036 (0.051)	-0.021 (0.052)	-0.051 (0.054)	-0.035 (0.051)
Dummy LIE	0.205 (0.130)	0.186 (0.127)	-0.318^a (0.046)	-0.291^a (0.041)	-0.305^c (0.159)	-0.259^b (0.125)
Dummy SABOTAGE	0.080 (0.103)	0.045 (0.096)	0.028 (0.084)	-0.014 (0.077)	0.063 (0.096)	0.034 (0.090)
Risk aversion	-0.021 (0.024)	-0.014 (0.022)	-0.038 (0.025)	-0.029 (0.024)	-0.042 (0.028)	-0.028 (0.025)
Envy	-0.123^b (0.061)	-0.131^b (0.058)	-0.048 (0.077)	-0.037 (0.075)	-0.090 (0.086)	-0.088 (0.080)
Age	-0.016 (0.019)	-0.014 (0.015)	-0.006 (0.011)	-0.006 (0.011)	-0.007 (0.014)	-0.008 (0.014)
Gender:female	-0.017 (0.065)	-0.014 (0.061)	-0.130 (0.081)	-0.111 (0.073)	-0.117^c (0.081)	-0.107^c (0.074)
Education	0.148	0.136	-0.108	-0.048	0.030	0.021^c
Social and Behavioral Sciences	0.217	0.197	-0.054	0.006	0.080	0.118
Business and Economics	0.009	0.027	0.109	0.159	0.128	0.161
Sciences	0.172	0.180	-0.107	-0.031	0.026	0.093
Engeneering	0.040	0.046	-0.203^b	-0.162^c	-0.160	-0.114
Medicine					-0.816^a	-0.743
Law					-0.450	-0.526^b
IT					-0.676^a	-0.733^a
Count R^2/Nagelkerke's R^2	0.85	0.87	0.76	0.75	0.43	0.41

Table A14: Ethical determinants of B participants' departures from rational self-interest, and the type of altruistic behavior they adopt (marginal effects).

Note: Significance levels of z-tests are indicated by $a : p < .01$, $b : p < .05$, $c : p < .10$.

Notes. See joint notes to tables A14 and A15.

P.4 Treatment payoff neutrality: punishment and reward

Note: In treatment payoff neutrality, B's rational self-interested choice is to toss a fair coin, i.e. leave Prob(S_2) = 50%, and to take all payoff. A has no decision rights in either S_1 or S_2.

Dependent Dummy variable →	S_1 (1) vs. Fair Coin (0)		S_2 (1) vs. Fair Coin (0)		All Three Categories	
B participants → x_j ↓	111 (vs.)	111 (14 vs.)	117 (vs.)	129 (40 vs. 98)	134	149
Kohlberg class 1	−0.009 (0.050)	−0.021 (0.044)	−0.039 (0.053)	−0.050 (0.050)	−0.018 (0.052)	−0.028 (0.048)
Kohlberg class 2	0.018 (0.043)	0.031 (0.040)	0.063 (0.041)	0.069^c (0.040)	0.033 (0.042)	0.039 (0.040)
Kohlberg class 3	−0.029 (0.065)	−0.034 (0.058)	0.083^c (0.047)	0.077^c (0.046)	0.047 (0.050)	0.041 (0.048)
Kohlberg class 4	0.092^b (0.040)	0.085^b (0.043)	0.099^b (0.040)	0.101^a (0.038)	0.108^b (0.043)	0.109^a (0.041)
Kohlberg class 5	0.052 (0.048)	0.051 (0.045)	−0.057 (0.042)	−0.055 (0.040)	−0.030 (0.044)	-0.026^c (0.041)
Kohlberg class 6	−0.037 (0.049)	−0.021 (0.048)	-0.111^b (0.044)	-0.107^a (0.041)	−0.088 (0.045)	-0.084^b (0.043)
Dummy LIE	−0.095 (0.059)	−0.088 (0.055)	−0.096 (0.080)	−0.102 (0.081)	-0.106^c (0.100)	−0.103 (0.097)
Dummy SABOTAGE	0.002 (0.073)	0.011 (0.072)	−0.011 (0.078)	−0.014 (0.073)	0.008 (0.090)	0.008 (0.084)
expected punishment, ranging from 0 to 30 ECU	0.010^a (0.003)	0.010^a (0.003)	0.012^b (0.005)	0.013^a (0.005)	0.014^a (0.005)	0.015^a (0.005)
expected reward, ranging from 0 to 30 ECU	0.002 0.004	0.002 (0.004)	0.013 (0.009)	0.014^c (0.009)	0.012 (0.008)	0.013 (0.008)
Risk aversion	-0.084^a (0.018)	-0.078^a (0.016)	-0.052^b (0.027)	-0.058^a (0.022)	-0.080^a (0.026)	-0.083^a (0.022)
Envy	-0.110^c (0.065)	-0.107^c (0.063)	−0.110 (0.071)	−0.100 (0.069)	-0.161^b (0.074)	-0.148^b (0.071)
Age	−0.006 (0.021)	−0.005 (0.021)	0.003 (0.010)	0.003 (0.010)	0.005 (0.011)	0.005 (0.011)
Gender:female	0.108 (0.096)	0.093 (0.086)	−0.074 (0.075)	−0.091 (0.070)	−0.023 (0.080)	−0.040 (0.075)
Education	0.231	0.254	0.376^b	0.381^b	1.557^a	1.544^a
Social and Behavioral Sciences	−0.014	0.016	0.430^a	0.430^a	1.481^a	1.469^a
Business and Economics	0.206	0.237	0.440^a	0.456^a	1.567^a	1.562^a
Sciences	0.217^c	0.249^b	0.411^a	0.422^a	1.574^a	1.565^a
Engeneering			0.198	0.192	0.623^c	0.602^c
Law					0.624^c	0.621^c
Languages	0.575^a	0.588^a	0.666^a	0.635^a	1.858^a	1.794^a
Count R^2/Nagelkerke's R^2	0.88	0.90	0.83	0.83	0.58	0.57

Table A15: Ethical determinants of B participants' departures from rational self-interest, and the type of altruistic behavior they adopt (marginal effects).

Note: Significance levels of z-tests are indicated by $a : p < .01$, $b : p < .05$, $c : p < .10$.

Notes to tables A14 and A15. Columns 2-5. Binary Logit regressions with robust errors. S_1 vs. Fair Coin: B participants who opt into S_1 against – column 1: B participants who toss a fair coin and take all payoff, and against – column 2: *all* B participants who toss a fair coin. S_2 vs. Fair Coin: B participants who opt into S_2 against – column 1: B participants who toss a fair coin and take all payoff, and against – column 2: *all* B participants who toss a fair coin. **Column 6-7.** All Three Categories: multinomial logit model of all categories where baseline category Fair Coin are – column 1: B participants who toss a fair coin and take all payoff, and – column 2: *all* B participants who toss a fair coin. **Moral judgement variables** – *Kohlberg class 1* ranges from -1.635 to 2.688 with median 0.130; *Kohlberg class 2* from -1.183 to 2.903 with median -0.162, *Kohlberg class 3* from -1.713 to 2.076 with median 0.317, *Kohlberg class 4* from -1.296 to 3.056 with median -0.052, *Kohlberg class 5* from -1.754 to 2.895 with median 0.239, and *Kohlberg class 6* from -2.235 to 2.952 with median 0.235. **Expected punishment** ranges from 0 to 30 ECU with mean 3.441. **Expected reward** ranges from 0 to 30 with mean 1.967. **Missing values**. There are three B participants whose *Kohlberg class 4* scores are not available because they did not fill in any item corresponding to this class, one tossing a fair coin, and two opting into S_2. These make up for the difference between the observations indicated and the number of observations in appendix M.3 and M.4. For binary Logits, we report the Count R^2 as before, for multinomial Logits, we continue to report Nagelkerke's R^2.

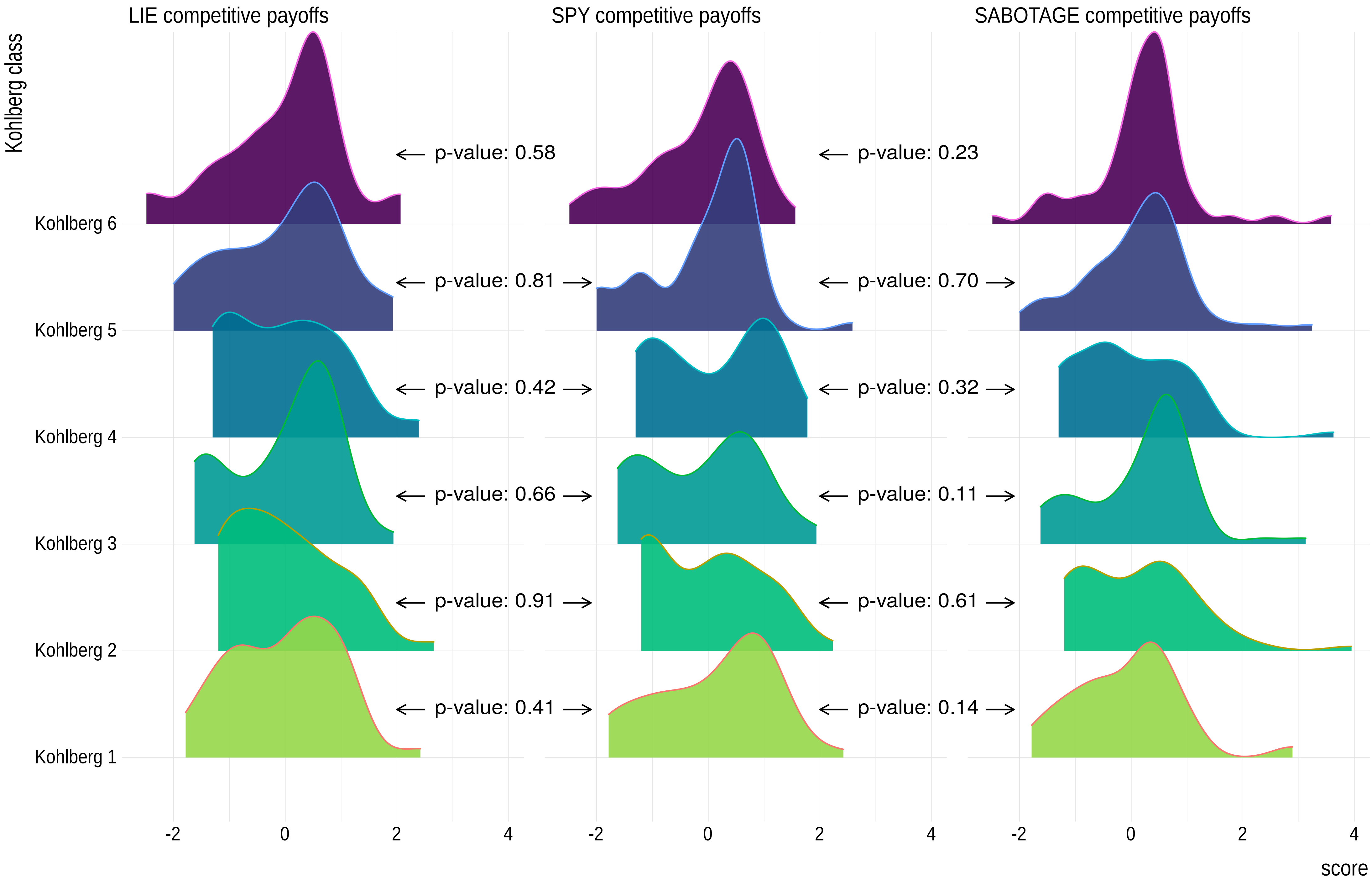


Figure A15: *B* participants' Kohlberg scores do not differ significantly across the main treatments LIE, SPY, and SABOTAGE with Competitive Payoffs (Wilcoxon Rank sum tests).

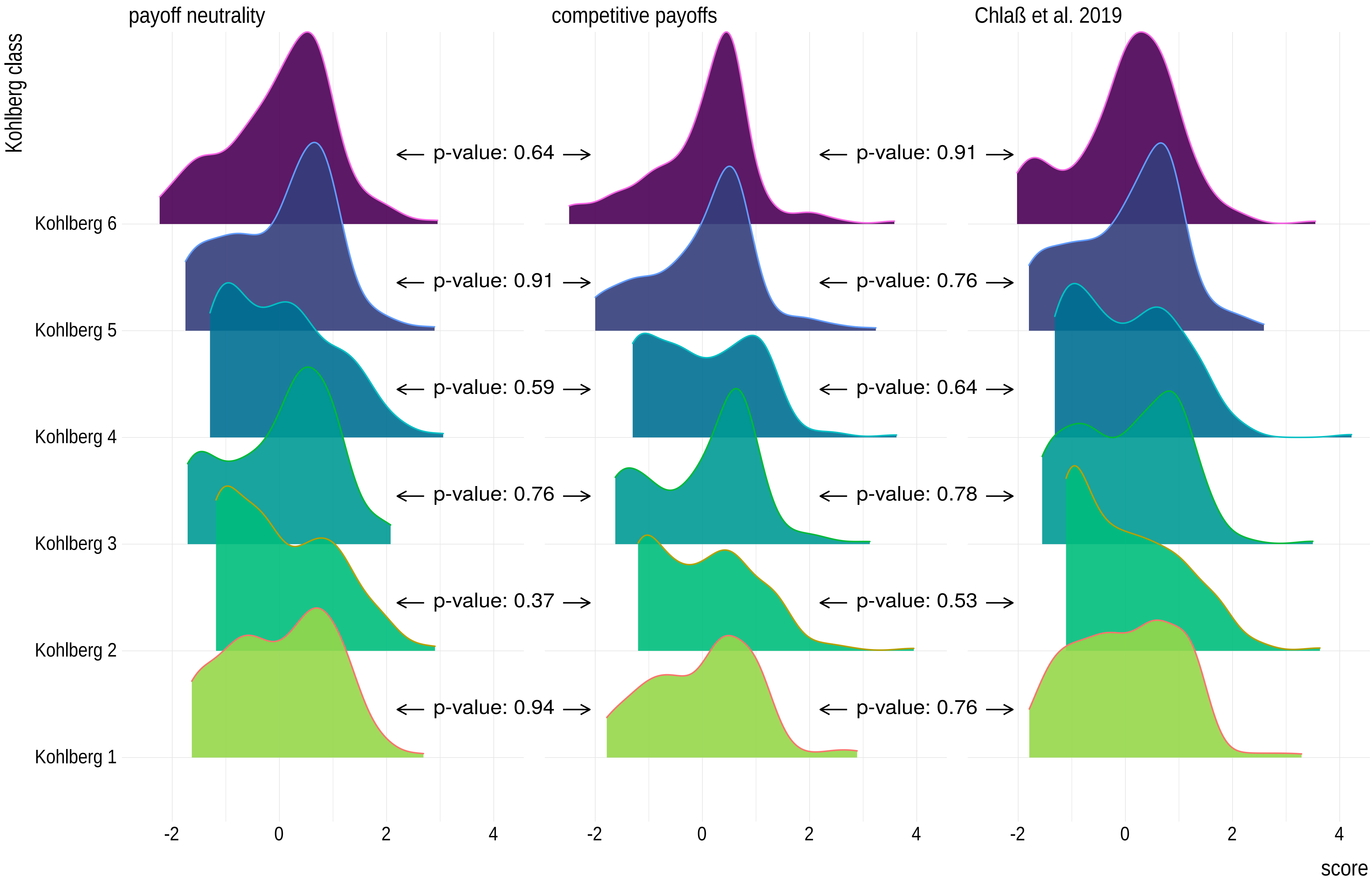


Figure A16: *B* participants' Kohlberg scores do not differ significantly across treatments payoff neutrality and competitive payoffs and do not differ from Chlass et al. (2019) (Wilcoxon Rank Sum tests).

Q B's procedural choice, $Prob(S_2)$, and choice of allocation by treatment

Figure A17: B's choice of Prob(S_2) and the allocation she chooses in S_2. LEFT: competitive payoffs – A has no decision rights in S_2; RIGHT: competitive payoffs with punishment/reward – A can punish B for Prob(S_2).

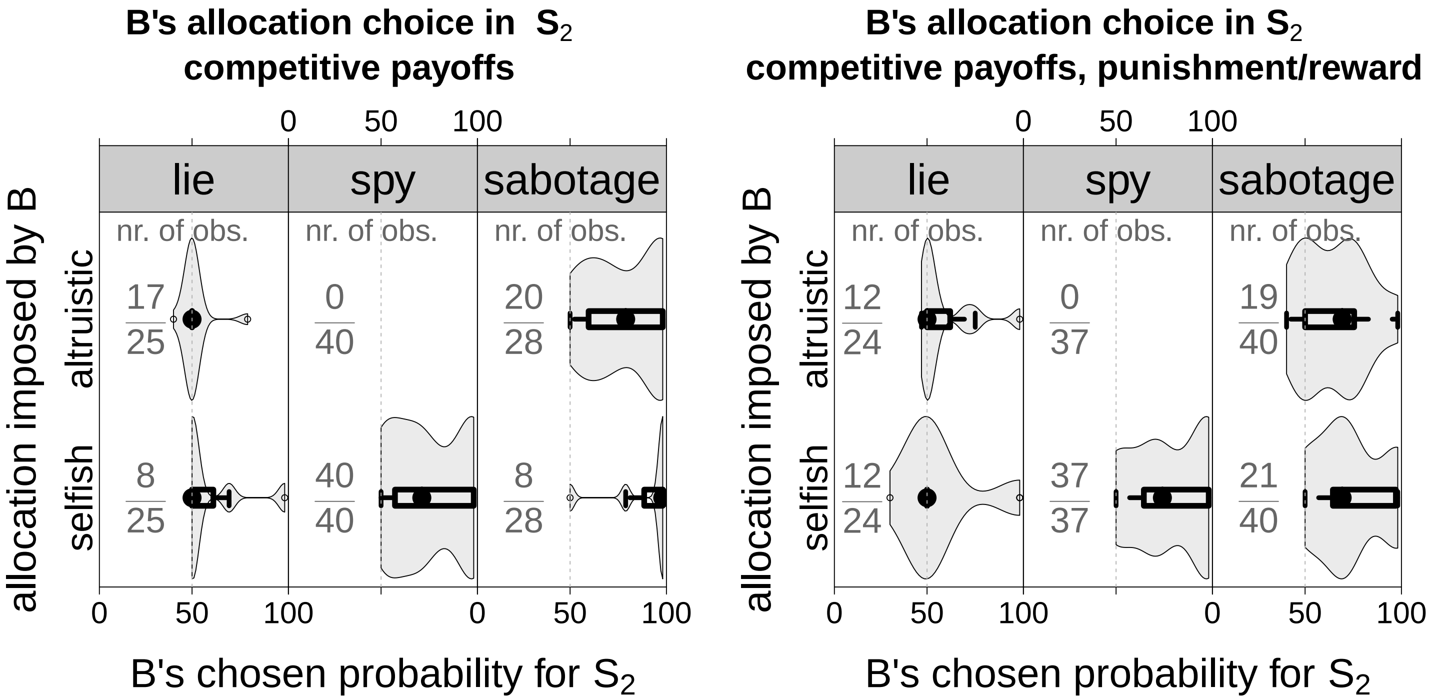


Figure A18: B's choice of Prob(S_2) and the allocation she chooses. LEFT: payoff neutrality [A has no decision rights in either S_1 or S_2]; RIGHT: payoff neutrality with punishment/reward [A has the same rights to punish and reward in S_1 and S_2].

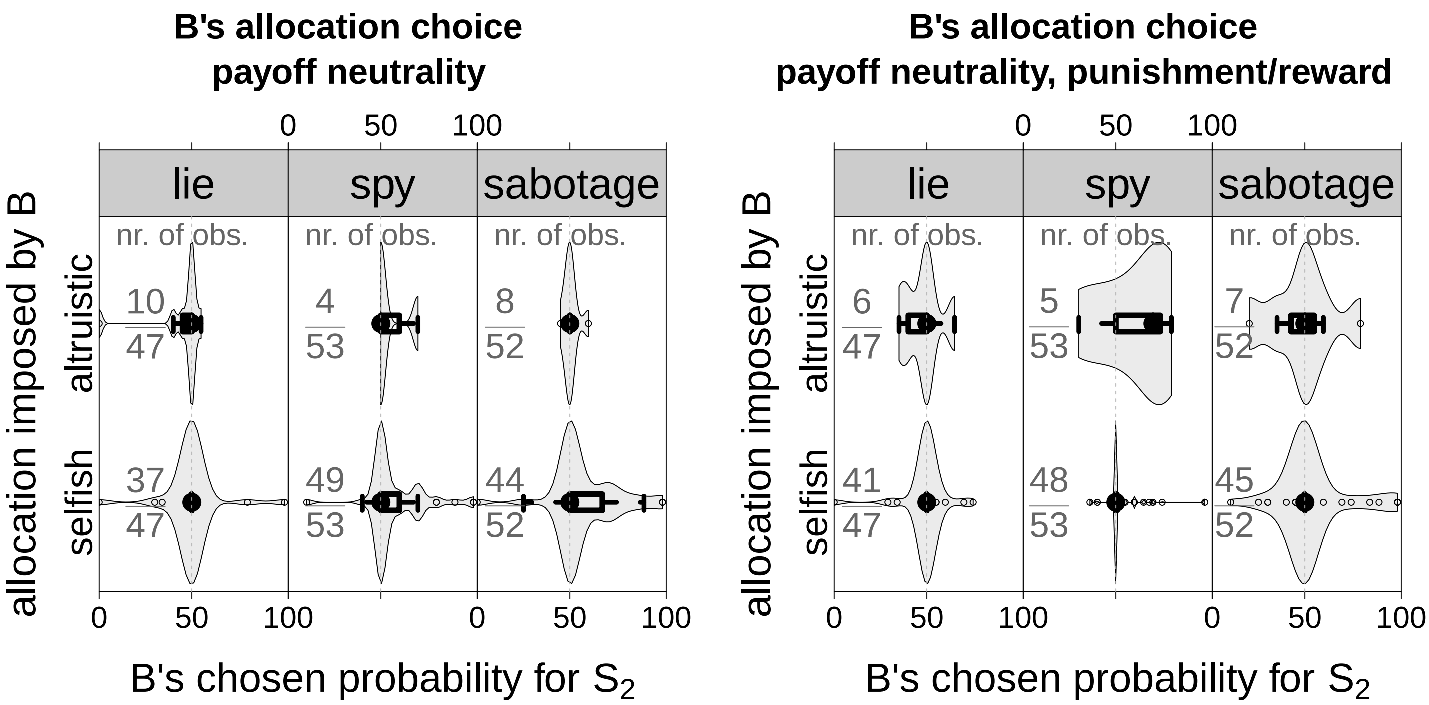


Figure A19: B's choice of Prob(S_2) and the allocation she chooses under payoff neutrality [A has no decision rights either in S_1 or S_2]; LEFT: S_1. RIGHT: S_2.

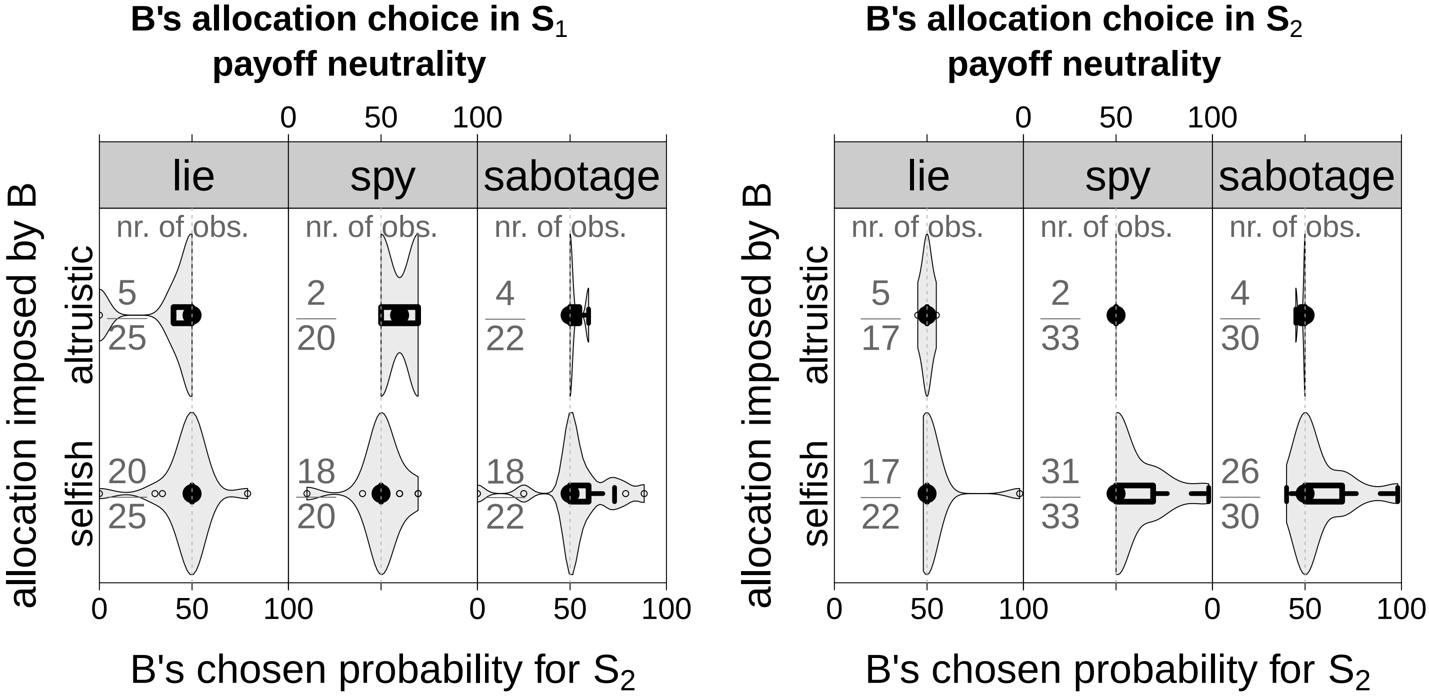
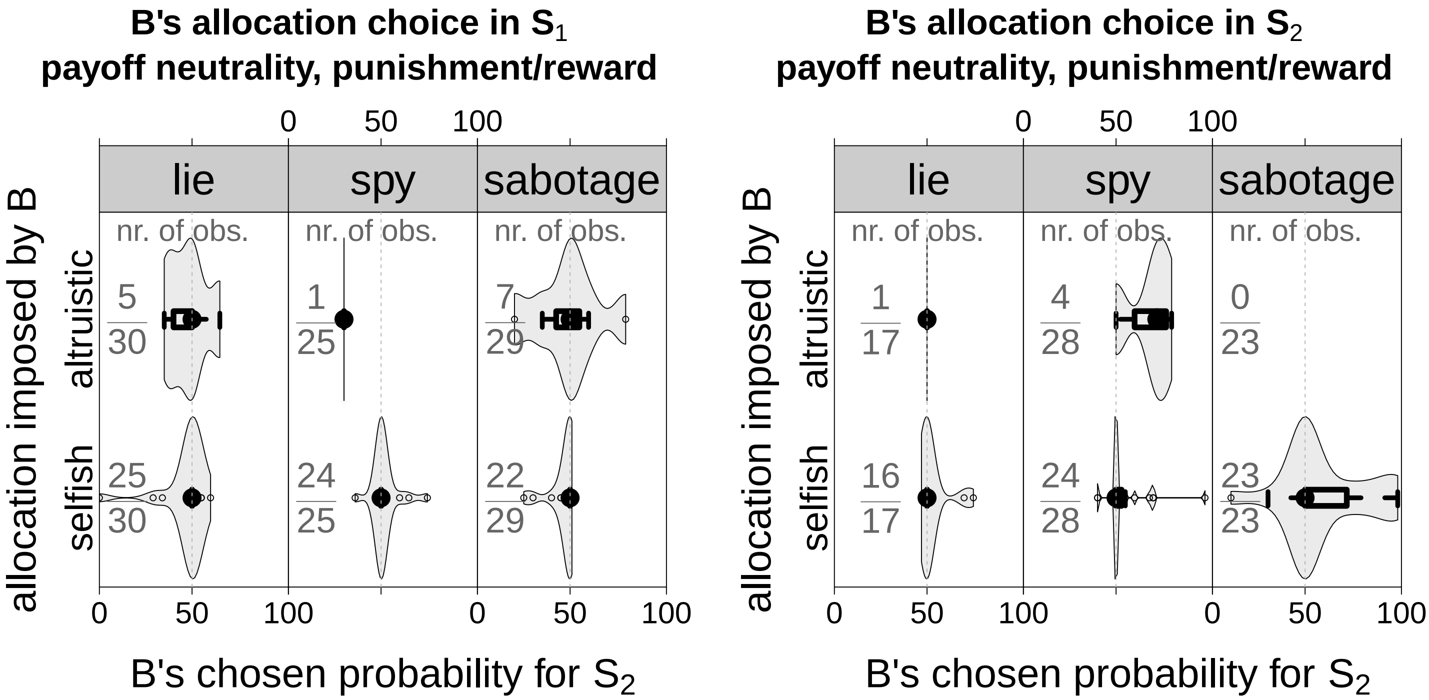


Figure A20: B's choice of Prob(S_2) and the allocation she chooses under payoff neutrality with punishment/reward [A has the same decision rights to punish and reward in S_1 and S_2]; LEFT: S_1. RIGHT: S_2.



R Predictions: payoff neutral treatment

			BEHAVIOURAL PREDICTIONS						
			LIE		SPY		SABOTAGE		
			make selfish proposal in S_1	make selfish proposal in S_2	make selfish proposal in S_1	make selfish proposal in S_2	make selfish proposal in S_1	make selfish proposal in S_2	**same outcomes across LIE, SPY, SABOTAGE**
SOCIAL PREFERENCE MODELS	Outcome based	Self Interest	+	+	+	+	+	+	+
		Inequity Aversion	+	+	+	+	+	+	+
		Altruism	depends on degree of altruism	depends on degree of altruism	depends on degree of altruism	depends on degree of altruism	depends on degree of altruism	depends on degree of altruism	+
	Reciprocity - based	Falk & Fischbacher (2006)	+	+	+	+	+	+	+
		Dufwenberg& Kirchsteiger (2004)	+	+	+	+	+	+	+
	Guilt based	Battigalli & Dufwenberg (2007)	depends on sensitivity to guilt	depends on sensitivity to guilt	depends on sensitivity to guilt	depends on sensitivity to guilt	depends on sensitivity to guilt	depends on sensitivity to guilt	+
OUTCOME-BASED PROCEDURAL FAIRNESS MODELS	Inequity based	e.g. Bolton et al. (2005)	+	+	+	+	+	+	+
	Reciprocity - based	Sebald (2010)	+	+	+	+	+	+	+
PURELY PROCEDURAL FAIRNESS MODELS	equal decision rights	Chlaß et al. (2019)	+	+	+	+	+	+	+
	equal information rights	Chlaß et al. (2019)	+	+	depends on sensitivity to unequal information	depends on sensitivity to unequal information	+	+	−

Notes. *1) Inequity aversion.* Denote B's earnings by x, and $A's$ earnings by y. An inequity averse B has utility $x - a \cdot \max\{(y-x), 0\} - b \cdot \max\{(x-y), 0\}$ where $a, a \leq 4$ and $b, b \leq 1$ are non-negative individual parameters. Allocation ($x = 100$, $y = 0$) yields B utility $100 - b \cdot 100$, allocation ($x = 0$, $y = 100$) utility $-a \cdot 100$, respectively. An inequity averse B with $b \leq 1$ therefore always prefers ($x = 100$, $y = 0$) over ($x = 0$, $y = 100$).

2a) Reciprocity, Falk and Fischbacher (2006). B chooses between an intentionally weakly kind, i.e. ($x = 0, y = 100$), and an intentionally selfish (unkind) allocation ($x = 100, y = 0$). A has no decision rights at all; she cannot reject (must accept) all allocations, can hence not be unkind to B, and B need hence not be kind to induce kindness. B therefore chooses the selfish allocation ($x = 100, y = 0$).

2b) Dufwenberg and Kirchsteiger (2004). There are only efficient strategies in the game (no strategy destroys the pie). Since A cannot reject (must accept) all allocations, she cannot be unkind to B, and B therefore chooses the selfish allocation ($x = 100, y = 0$).

3) Guilt aversion (Battigalli and Dufwenberg (2007)). Guilt matters only if B harms A and lets her down (disappoints A's expectations). A cannot harm B and her guilt payoff is therefore always Zero. A very guilt averse B who very much expects A to expect the generous allocation, might indeed offer ($x = 0, y = 100$). As long, however, as B's beliefs about A's payoff expectations are identical in S_1 and S_2, B makes the same choice in both situations. Looking at B's empirical punishment expectations, B

players expect A players to expect identical payoffs in S_1 and S_2.

4) Preferences for equal expected payoffs (Bolton and Ockenfels (2005). Ex-ante, B chooses between S_1 where she opts for allocation $(x = 100, y = 0)$ for sure, and S_2 where she also opts for allocation $(x = 100, y = 0)$ for sure. She therefore has no choice between more or less equal expected payoffs, the expected payoffs are degenerate in each situation, and she is indifferent between S_1 and S_2.

5) Preferences for kind procedures (Sebald 2010). Since A cannot reciprocate in either S_1 or S_2, B therefore chooses the selfish allocation $(x = 100, y = 0)$ always and S_1 and S_2 are therefore equally unkind, B is indifferent between S_1 and S_2.

6a) Purely procedural preferences for equal decision rights (Chlaß et al. 2019). In S_1 and in S_2, for any contingency of the game – that is, whether B chooses either L, or R, A's choices always leave her equally well off: if B chooses L, A's choices L and R both yield her Zero payoff and hence, she cannot prefer L over R or vice versa; if B chooses R, A's choices both yield her 100 ECU and hence, she cannot prefer L over R or vice versa either. She therefore has no decision rights and cannot look after her own self-interest. Therefore, it is not within B's power to either grant A, or impair the latter's decision rights. B is therefore indifferent between S_1 and S_2 and chooses $(x = 100, y = 0)$.

6b) Purely procedural preferences for equal information rights (Chlaß et al. 2019). In S_1, neither A nor B knows which action the opponent has chosen. Only B knows that the interaction structure is S_1. B can therefore distinguish two out of the four terminal nodes of the game. The same holds for S_2 in treatments LIE and SABOTAGE. A's cardinality over the terminal nodes of S_1 and S_2 is always One, since she does not know the interaction structure. In LIE and SABOTAGE therefore, B has no power to either grant A, or impair the latter's information rights. This does not hold for treatment SPY where in S_2, B knows A's choice, but A does not know how B has chosen and B can therefore distinguish all four terminal nodes of the game. If B dislikes having greater information rights than A in SPY, she prefers S_1 over S_2, or chooses S_2 and compensates A by opting for allocation $(x = 0, y = 100)$. If to the contrary, B prefers greater information rights, she prefers S_2 over S_1, and opts for allocation $(x = 100, y = 0)$. Note that B never takes away information rights from A; she always improves her own relative position in information rights. Note, too, that it is not within her power to grant A exactly *equal* information rights.