Economic polarization and antisocial behavior: An experiment

Economic Polarization and Antisocial Behavior
an experiment*

Maria Bigoni† Stefania Bortolotti‡ Efşan Nas Özen§

Abstract

Economic inequality may fuel frustration, possibly leading to anger and antisocial behavior. We experimentally study a situation where only the rich can reduce inequality while the poor can express their discontent by destroying the wealth of a rich counterpart with whom they had no previous interaction. We test whether the emergence of such form of antisocial behavior depends only on the level of inequality, or also on the conditions under which inequality occurs. We compare an environment in which the rich can unilaterally reduce inequality with one where generosity makes them vulnerable to exploitation by the poor. We observe that the poor engage in forms of antisocial behavior more often when reducing inequality would be safe for the rich. These results cannot be rationalized by inequality aversion alone, while they are in line with recent models that focus on anger as the result of the frustration of expectations. Indeed, we find that the rich are expected to be more generous in the safe scenario than in the risky one, but in fact this hope is systematically violated.

JEL classification: C91, D63, D83, D84, D91

Keywords: expectations, frustration, inequality aversion, money-burning, punishment.

*We would like to thank Björn Bartling, Pierpaolo Battigalli, Alexander Cappelen, Gabriele Camera, Marco Casari, Elena Cettolin, Conchita D'Ambrosio, Martin Dufwenberg, Catherine Eckel, Diego Gambetta, Werner Güth, Johanna Mollerstrom, Nikos Nikiforakis, Hans-Theo Normann, Ernesto Reuben, Arthur Schram, Simeon Schudy, Ferdinand von Siemens, Matthias Sutter, Bertil Tungodden, Daniel J. Zizzo, participants at the EWEBE Conference Bertinoro, WESSI Florence, ESA European Meeting in Vienna, SEET Workshop in Lecce, i-See Workshop in Abu Dhabi, 2018 IMEBESS Conference Florence, EEG Inaugural Conference in Bonn, 2018 ESA World Meeting in Berlin, 2018 Lisbon Meeting on Economics and Political Science, 2019 Asia-Pacific ESA Meeting in Abu Dhabi, Bergamo Winter Symposium in Economics, 2019 EUI workshop on Experiments in Social Sciences, 2019 Stavanger Workshop on Trust and Cooperation in Markets and Organizations, and seminar participants at the University of Bologna, University of Torino, La Sapienza University, Max Planck Institute Bonn for Research on Collective Goods, Chapman University, Wageningen University, and the Luxembourg School of Finance for comments and suggestions on previous versions of this paper. We gratefully acknowledge financial support from the Italian Ministry of Education [SIR grant no. RBSI14I7C8]. Significant parts of this research were developed while the second author (Bortolotti) was at the University of Cologne and at the Max Planck Institute for Research on Collective Goods; she thanks those institutions for the financial and logistic support. The usual disclaimer applies.

†Bigoni: Department of Economics, University of Bologna, Piazza Scaravilli 2, 40126 Bologna, Italy & IZA; maria.bigoni@unibo.it.

‡Bortolotti: Department of Economics, University of Bologna, Piazza Scaravilli 2, 40126 Bologna, Italy & IZA; stefania.bortolotti@unibo.it.

§Nas Özen: World Bank; snasozen@worldbank.org. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations.

1 Introduction

Most people dislike disadvantageous inequality, to the point that they might be ready to burn money in order to reduce it (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Zizzo, 2003; Balafoutas et al., 2013). Yet, not all inequalities are born alike: whether a certain distribution of wealth is deemed fair and acceptable can greatly depend on the process that generated it (Konow, 2000; Alesina and Angeletos, 2005; Alesina and La Ferrara, 2005; Cappelen et al., 2007; Faravelli, 2007; Cappelen et al., 2013; Fehr, 2015; Almås et al., 2016; Cassar and Klein, 2016; Bortolotti et al., 2017; Cappelen et al., 2017).

Here we consider an environment that gives rise to a polarized income distribution (Esteban and Ray, 1994; Permanyer, 2018) with only two classes, the rich and the poor. Inequality in opportunities is imposed exogenously but inequality in outcomes is determined endogenously by the behavior of the rich, who could achieve perfect equality through a generous act; the poor, instead, have no means to enforce any form of redistribution.

Within this framework, we compare a scenario in which being generous entails no risk, with one where reducing inequality exposes the rich to the risk of exploitation by the poor; we focus on whether the poor react differently to these two different situations. To study this aspect, we allow the poor to express their discontent by destroying all the money at stake ("antisocial behavior"). Importantly, we are not interested in the poor's reaction as a form of direct punishment, but rather as an extreme act of protest which is undertaken before even knowing whether the rich who will be harmed would in fact be generous or not. In this sense, we consider a form of antisocial behavior that is directed towards a random person from a category of people – the "elite" – and not toward someone who is directly responsible for the suffered harm.

Our goal is to investigate three main questions: (i) whether the two environments – safe vs. risky – generate different levels of inequality in outcomes; (ii) whether being exposed to an extremely polarized economy induces subjects to undertake actions that are both individually and socially costly; (iii) whether this reaction depends only on the level of the experienced inequality, or whether the situation in which the polarization emerged plays a role as well.

It is hard to collect clean evidence on this phenomenon from observational data. Individual expectations are difficult to measure and keeping inequality constant across contexts can also prove challenging. In addition, it would be difficult to control for strategic and monetary motives triggering the antisocial behavior: episodes of vandalism and violent forms of protest may be rationalized as a way to affect political leaders' decisions (Lohmann, 1993). To exclude this sort of strategic motives, we analyze the insurgence of antisocial behavior in a tightly controlled laboratory environment.

To this aim, we devised a new zero-sum two-by-two game – the *Inequality Game* – where a Strong player must decide whether to be "generous" or "defensive", while her Weak counterpart must choose between "collaborate" and "exploit". The game has a unique Nash Equilibrium, in which Strong chooses defensive (which is the dominant action) and Weak collaborates, and they earn 90% and 10% of the pie, respectively. Inequality is therefore ingrained in the structure of the game and arises endogenously; however, there exists a perfectly equal outcome that can be reached if the Strong player is generous and the Weak one collaborates.

In a between-subjects design, we manipulate whether being generous is *Risky* or *Safe* for the Strong player, by switching from a simultaneous, to a sequential version of the game. In the simultaneous version (Risky treatment), well-intentioned Strong players who choose generous in an attempt to reach the equitable outcome face the risk of being left with only 10% of the pie, if the Weak player chooses to exploit them. Hence, the choice of the dominant action on behalf of the Strong player might be driven by the fear of exploitation, and not only by a greedy ambition. In the sequential version of the game (Safe treatment), instead, the Weak player moves first; if he/she collaborates, the Strong players can choose between a perfectly equal and a highly unequal distribution of resources, without facing any risk of exploitation.

To study the extent of antisocial behavior in these scenarios, we introduce the possibility – for both Strong and Weak players – to "exit" the game, destroying all the surplus potentially generated by the encounter. This captures a situation where Weak players do not have the chance to voice their complaints to try to improve their prospects, and burning money by quitting the game might be the only way to express their discontent. The timing of such

decision is crucial: the destruction of resources (exit) happens before knowing whether the Strong counterpart will be generous or not,[1] hence one may end up harming an *innocent* and well-intentioned opponent. The Inequality Game is one shot in order to rule out reciprocity and reputational mechanisms; however, to allow players to gain experience with the game, we let them interact repeatedly for ten periods, with fixed roles and perfect stranger matching.

We report three main results. First, the realized level of inequality is not significantly different between treatments. In other words, the Strong players are not more generous when it is safe to be so. Second, exit emerges only after some experience of the game and takes place more often in the *Safe* than in the *Risky* treatment. Third, a closer look at individual experiences indicates that the Weak players' decision to exit in one period is strongly correlated with their experience in previous periods: being repeatedly matched with Strong players who never act generously is positively associated with the use of exit. Interestingly, this effect is much more pronounced in the *Safe* treatment where the Strong players could unilaterally equalize the payoffs. This finding suggests that the choice to exit is not only driven by the realized and observed level of inequality, but also by the context in which it emerged.

To better understand why we observe more exit in *Safe*, we consider the role of expectations. In particular, a long tradition in the psychological literature suggests that disappointment of expectations may lead to frustration and hence anger (Potegal et al., 2010, Chapter 5). We conjecture that the Weak players in the *Safe* treatment expect a generous action from their Strong counterpart more often than the Weak players in the *Risky* treatment do, hence the same level of inequality can generate different degrees of frustration and anger.

According to this psychological paradigm, anger can in turn result in costly and economically inefficient actions that are not necessarily motivated by strategic concerns – i.e., a threat to improve the current situation – or targeted toward the person responsible for the disappointing outcome. Intuitively, that would mean that the lack of generosity by the Strong might ignite more exit when it is less expected and hence frustration is higher. This idea has recently been incorporated into formal models of psychological game theory by Battigalli et al. (2019b,a). We

[1] Strong players can choose to exit the game as well, but given the set-up of our game, we almost never observe such behavior.

show that, within our framework, the model by Battigalli et al. (2019b) would imply that the same level of inequality can ignite different levels of discontent and hence different reactions, depending on what the less well-off expected for their own future earnings in the first place.

Our hypothesis is that the Weak exhibit a stronger reaction to inequality in the *Safe* than in the *Risky* scenario because these two environments generate different expectations on the behavior of the Strong players. The idea is that frustration can grow over time: one enters the first period with some hope for redistribution, but then observes, over and over again, that redistribution never takes place. This frustration cannot be unleashed against a previous partner, whose actions are known, as exit can only affect the outcome of the period that is about to be played, where the player faces a completely unknown counterpart. If the Weak expect their Strong counterpart to be more generous in the *Safe* than in the *Risky* treatment, then – if faced with the same level of inequality – they would exit more often because they experience greater frustration and discontent. To directly assess whether the *Risky* and the *Safe* treatments in fact induce different expectations on players' behavior, hence on the realized outcomes, we elicit beliefs by means of an incentivized procedure. This "Belief Experiment" involved a new set of participants, who never took part in the *Inequality Game* but had to read the instructions and guess the actual choices made by the participants in our "Main Experiment."

The Belief experiment provides support for the idea that the mismatch between expectations and realized outcomes could be an important determinant of the decision to harm a Strong stranger. Results confirm that subjects expect the Strong players to be generous more often in the *Safe* than in the *Risky* environment. Since subjects in the two experiments are drawn from the same pool of participants, it is reasonable to assume that participants in the Main Experiment also had higher hopes for a more equal distribution of earnings in *Safe* than *Risky*.

Taken together, the two experiments suggest that a polarized distribution of resources may ignite extreme reactions in the form of antisocial behavior targeted at the elite in general when the specific person who is directly responsible for the unequal distribution of wealth cannot be targeted. In addition, initial expectations about the likelihood of an even outcome can play an

important role in driving the reaction to inequality: discontent, anger and aggressiveness will emerge when rosy expectations are frustrated by reality.

The paper is organized as follows. Section 2 discusses the main elements of novelty of our study, while Section 3 introduces the Inequality Game, illustrates the design of the Main Experiment and describes the experimental procedures. Section 4 presents the results of the Main Experiment. Section 5 details the design and results of the Belief Experiment. Section 6 discusses our findings in light of different theoretical models and Section 7 concludes.

2 Related literature

With respect to the existing literature, the most important element of novelty of our Main Experiment is the timing of the exit decision. By forcing subjects to exit before the interaction takes place, we remove any strategic motivation behind this – individually and socially costly – choice. This sets us apart from other forms of direct and indirect punishment which have been extensively studied in the literature.[2]

In contrast to the Ultimatum Game (Güth et al., 1982), where proposers may have an incentive to increase the offer to the responders if they fear that the latter will reject, exiting in our setting cannot provide any motive for the counterpart to share the resources more equally, not even off the equilibrium path. The timing of exit also sets us apart from the Power-to-Take Game (Bosman and van Winden, 2002; Bosman et al., 2005), where the take-authority might hesitate to take too much since the responder may react by destroying part of the endowment, so that punishment works as a deterrent against opportunistic behavior. Similarly, in experiments involving *money burning* (Zizzo and Oswald, 2001; Zizzo, 2003), the spiteful and inefficient punishment behavior arises only after receiving information about the earnings generated in the betting stage; even though the strategic component of punishment is only limited in their set-up, information about other's actions and actual earnings is available to players when they engage in antisocial behavior. A similar reasoning applies to the joy-

[2]See for instance Güth et al. (1982); Fehr and Gächter (2000); Bosman and van Winden (2002); Fehr and Fischbacher (2004); Bosman et al. (2005); Nikiforakis (2008); Ule et al. (2009); Balafoutas et al. (2014); Güth and Kocher (2014).

of-destruction game (Abbink and Sadrieh, 2009) which has being used to study inter-group conflict and vendettas (Abbink et al., 2018; Abbink and Herrmann, 2011; Abbink, 2012; Bolle et al., 2014; Prediger et al., 2014).[3]

As we already pointed out, the form of antisocial behavior we consider is directed towards a category of people and not toward someone who is directly responsible for the suffered harm. In this respect, the situation we analyze is also different from the one modeled by Bartling and Fischbacher (2012) and Bartling et al. (2015), where the direct attribution of responsibility is the main driver of punishment. Similarly, expectation-based models such as the one proposed by Rabin (1993) do not predict any exit. The timing of the exit decision marks a difference also from two recent experimental studies (Aina et al., 2018; Persson, 2018) testing the theoretical framework by Battigalli et al. (2019b). In fact, in both studies the antisocial behavior follows the realization of the outcome.

The only study with a form of punishment that is close to ours is Lacomba et al. (2014), which investigates post-conflict behavior where conflict is created through a Tullock contest. The study includes one treatment in which losers of the contest can decide to burn money before knowing how much the winner of the contest will appropriate. The primary difference between Lacomba et al. (2014) and our study lies in the research question we answer. While Lacomba et al. (2014) design is not suited to see how expectations affect antisocial behavior, we explicitly manipulate the expectations of Weak players on Strong players behavior. In addition, when they decide whether to burn money, the losers in Lacomba et al. (2014) know they will inevitably be poorer than the winners, whereas the players in our design make their decision in a position where an equitable outcome is possible.

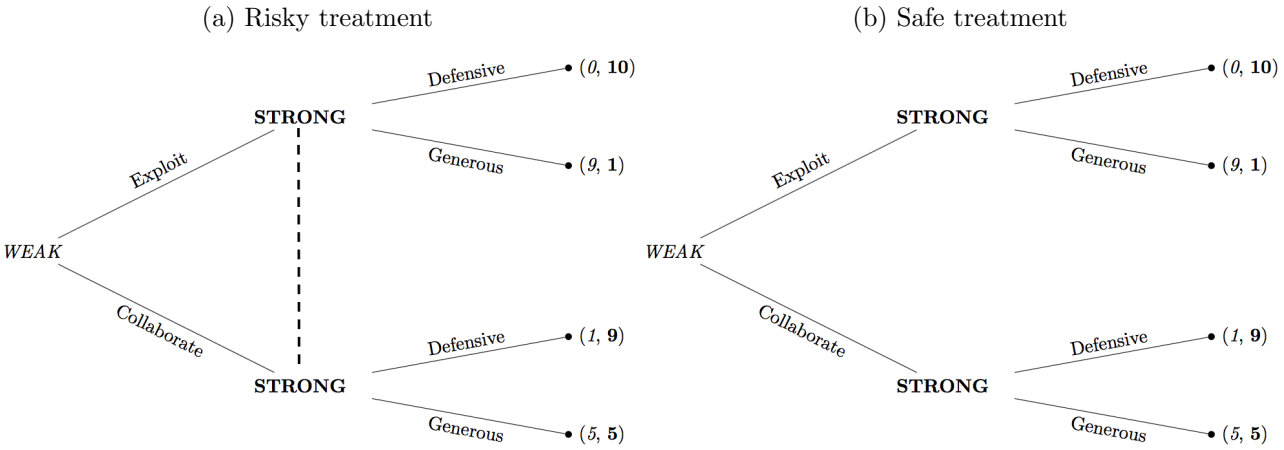
3 Inequality Game and Exit: Design

In the Main Experiment, we implemented a two-by-two between-subjects design. We exogenously manipulated two dimensions: the degree of strategic risk and the availability of an exit option. In the remainder of this section, we first describe the *Safe* and *Risky* variants of the

[3] Eckel et al. (2016) study a similar topic but relies on a different paradigm.

Inequality Game in their baseline version (*Control* treatments). We then introduce the exit option and illustrate the difference between the *Control* and the *Exit* treatment; finally we provide details on the experimental procedures.

Figure 1: The Inequality Game – Control treatments



The Inequality Game. To endogenously generate inequality, we developed a two-by-two asymmetric zero-sum game that involves a Strong and a Weak player.[4] The Strong player can choose between Defensive and Generous and the Weak one between Exploit and Collaborate: the payoffs (expressed in Euro) are reported in Figure 1. In the *Risky* treatment, both players decide at the same time, while in the *Safe* treatment, the Weak player decides first. In the latter treatment, we used the strategy method for Strong players, so that they made a decision for both nodes.

The simultaneous version of the Inequality Game (*Risky*) is dominance-solvable and has a unique Nash Equilibrium outcome (Collaborate, Defensive). The sequential version of the game (*Safe*) has a unique subgame-perfect Nash equilibrium (SPNE) which yields exactly the same outcome. The equilibrium payoffs are €9 for the Strong and €1 for the Weak player. Even though the payoffs in equilibrium are highly unequal, it is important to notice that a perfectly equitable outcome exists. The equal split, however, can be achieved only if the Strong player chooses a strictly dominated action. The two treatments – *Risky* and *Safe* – fundamentally

[4]The instructions were framed neutrally and the players were referred to as Red and Blue.

differ in the way the equitable outcome can be reached. A fair-minded Strong player can play Generous in the *Risky* game in the hope to reach the equal split (Collaborate, Generous). However, a self-interested Weak player could anticipate that and play Exploit, hence leaving the Strong player with only 10% of the total wealth (Exploit, Generous). Strong players can thus choose Defensive not only because they are self-interested but also out of strategic concerns. This is not the case in the *Safe* version of the game where the equitable outcome can be reached by the Strong player without any risk.

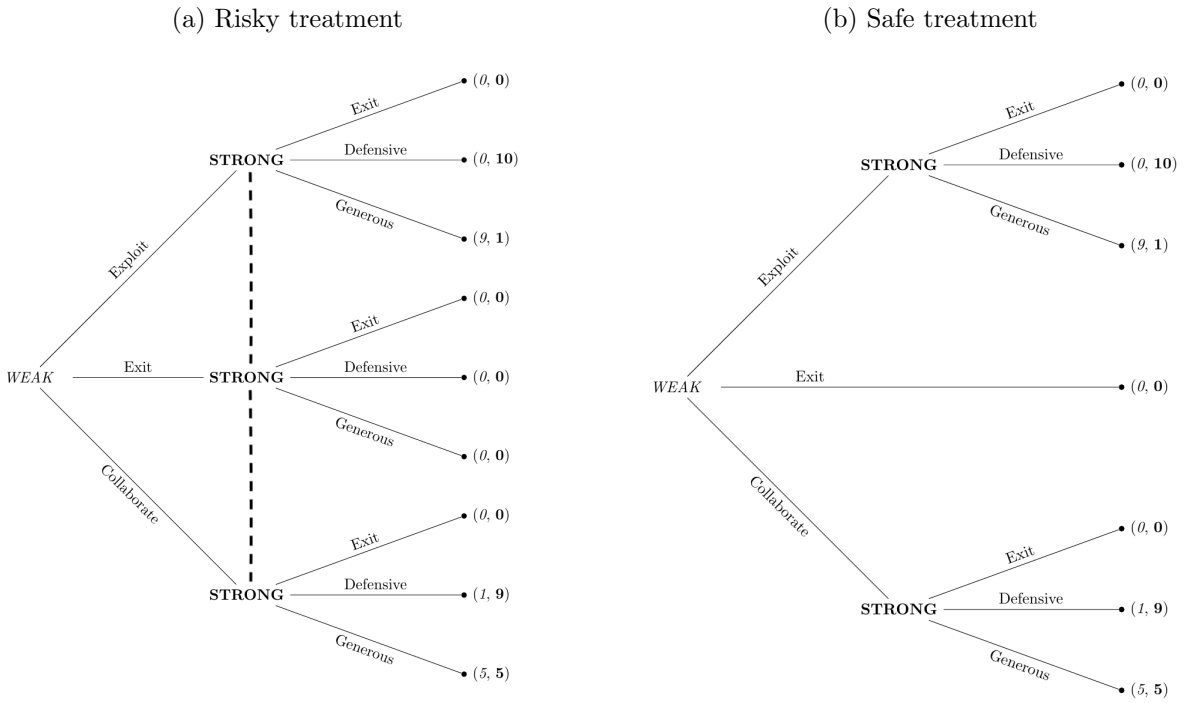
The exit option. In the *Exit* treatments, all participants – regardless of their role – were given the chance to exit the game before making any decision for the current period (Figure 2). If at least one of the two participants in the pair decided to exit, both players earned €0. Hence, the exit option is *harmful for both players* and *socially costly* as it generates a Pareto-dominated outcome. The choice to exit could only be taken before playing the game and, therefore, before having any information about the action taken by the other player. Subjects who did not exit were informed whether their counterpart chose to exit only at the end of the period.[5] This elicitation procedure ensures that in the *Risky* treatment the decisions of the two players are still simultaneous and players have one additional action in their choice set.

Both Strong and Weak do not have any incentive to use the exit option as it always implies some cost and can bring no material benefit. As a consequence, according to a standard game-theoretical approach, we should not observe any behavioral difference across the four treatments.

Repetitions, feedback, and matching. To allow subjects to gain experience, the *Inequality Game* was repeated for a total of 10 periods divided in two phases of equal length. At the end of each period, participants received feedback about the action adopted by their counterpart, and the payoffs in the pair. Roles were fixed over the entire duration of the experiment and there were exactly 10 Strong and 10 Weak players in each session. We used a perfect-strangers

[5]Knowing that the opponent did not exit might affect the player's second order beliefs on the opponent's expectations, in directions that are difficult to predict or interpret. To avoid this potential confounding, we chose to use the strategy method and not provide this information to the players before they take their move.

Figure 2: The Inequality Game – Exit treatments

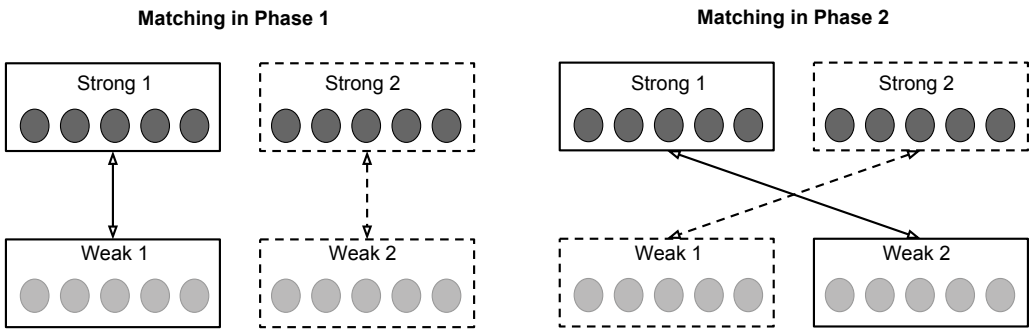


matching protocol and players were never matched together more than once.

Feedback and matching were designed so to strip away any possibility of forming an individual reputation, and hence, to rule out any form of direct or indirect reciprocity. Four sets of 5 players were formed at the beginning of the experiment: two sets of Strong players and two sets of Weak players. In Phase 1, each set of Strong players was matched with a set of Weak players, to form a 10-player "matching-group". In the five periods of Phase 1, each Strong player was paired once and only once with each Weak player in his/her matching-group. It is important to stress that when participants decided they had no information about their counterpart and his/her history of play. This is particularly relevant in the *Exit* treatments where deciding whether to exit or not, participants had no information on the history of play of their counterpart, and they could only rely on their own previous experiences with different counterparts.

At the end of Phase 1, participants were informed about the average earnings for the Strong and the Weak players in their matching-group and, in *Exit* treatments, they were also informed

Figure 3: Matching in Phase 1 and Phase 2



about the total number of exits by the Strong and by the Weak players in their matching-group. Regardless of the treatment, participants did not receive any feedback on the outcomes realized in the other matching-group. In Phase 2, each set of Strong players was matched with the set of Weak players they had not met in Phase 1 (Figure 3). This implies that, at the beginning of Phase 2, subjects had some aggregate information on the history of play of the other players in their own matching-group in Phase 1, but no information on the set of players they would be matched with in the next five periods.

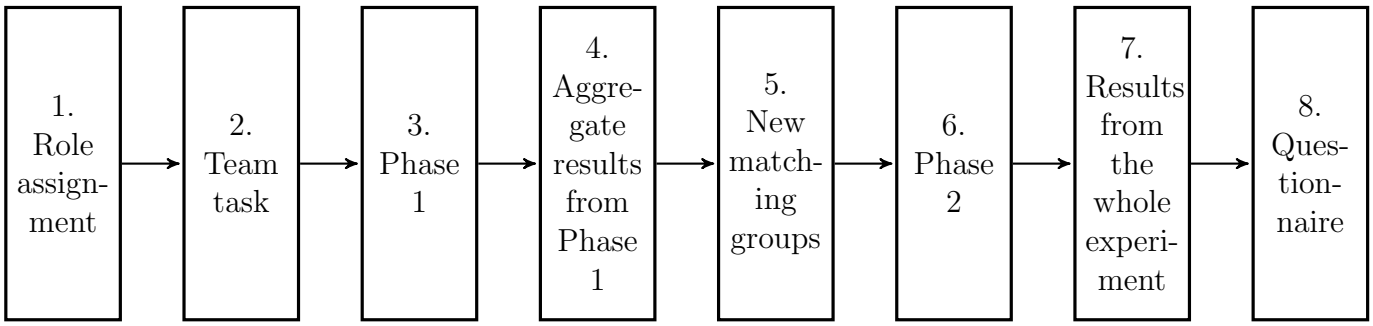
Procedures. 240 subjects equally divided into 12 sessions – 3 for each treatment – participated in the experiment that was conducted at the Cologne Laboratory for Economic Research (CLER) in May 2017. Participants were recruited via ORSEE (Greiner, 2015) and the experiment was programmed with zTree (Fischbacher, 2007). Upon arrival, participants were randomly assigned to visually isolated cubicles. Instructions were read aloud to ensure common knowledge and a paper copy of the instructions was distributed to participants.[6] An alphanumeric code was distributed together with the instructions and participants were asked to enter it on their computer at the beginning of the experiment. The code revealed the role – Strong or Weak player – assigned to the participant. After color assignment, subjects took part in a team task, which was the same in all treatments. Subjects were divided into two teams of ten, depending on their role. The task consisted of solving math problems to reveal a picture

[6]The experiment was run in English and that was announced in the recruitment message.

hidden on the subjects' screen.[7] This team task was meant to facilitate the understanding of the rules by focusing participants on the different roles in a subtle way. We are aware that such task might foster a sense of "group identity" which has been shown to be a potential trigger of antisocial behavior (Gangadharan et al., 2019). While we cannot exclude that the overall exit level has been affected by the team task, there are no reasons to believe its effect was different across treatments.

Figure 4 summarizes the eight steps of the experiment. In the first two steps, the role was assigned and the team task was performed. Step 3 included five periods of the stage game (Phase 1) and was followed by aggregate results at the matching-group level. In step 5, subjects were moved to a new matching-group and in step 6 five more periods of the *Inequality Game* were played (Phase 2). Aggregate results about Phase 2 were then provided.

Figure 4: Timeline of the experiment



At the end of the experiment, participants had to fill in a computerized questionnaire, which included some socio-demographic questions and a personality test (Ashton and Lee, 2009). To reduce any hedging problem, we paid only two periods. At the end of the experiments, one period from phase 1 and another period from phase 2 were selected at random for payment. Payments ranged from €6 to €26, with an average of €15.50, including a €4 show-up fee. A session lasted 50 minutes on average.

[7]Participants were asked to add up three two-digit numbers and every time a member of the team submitted a correct answer, one more piece of the picture behind the box was revealed. If the team task was successfully completed within 150 seconds, each team member earned €2; all teams succeeded.

4 Inequality Game and Exit: Results

This section is organized around two main parts. In the first, we present the aggregate results and treatment effects. In the second part, we dig deeper into individual-level behavior and focus on the use of the exit option conditional on the personal experience in the game.

4.1 Aggregate behavior and treatment effect

First, we report aggregate behavior for the two *Control* treatments to see if our novel game endogenously generates inequality between Strong and Weak players and if there is any difference between the *Safe* and *Risky* treatment. Second, we focus on *Exit* treatments by studying whether and how frequently the exit option is adopted and if there is a difference between the two versions of the game. We then dig deeper into individual-level behavior in *Exit* treatments. Finally, we assess the aggregate effect of the introduction of the exit option on behavior and outcomes, by comparing results from the *Exit* and *Control* treatments.

Table 1: Summary statistics

Treatment	**Defensive %**	**Collaborate %**	**DC-outcome %**	**Exit %**	**Strong's share of tot. surplus**
Control – Risky	0.86	0.81	0.69	–	0.85
Control – Safe	0.79	0.92	0.72	–	0.82
Exit – Risky	0.86	0.85	0.74	0.07	0.85
Exit – Safe	0.81	0.82	0.66	0.12	0.83

Notes: frequency of exit is reported only for the Weak players as exit was observed only once among the Strong players. For comparability across treatments, the share of surplus to the Strong player is based on periods without exit only.

Behavior in *Control* treatments. To assess the possible differences in behavior and resulting inequality in our novel game, we concentrate on the *Control* treatments first. Table 1 shows that Weak players chose Collaborate in 81% of the cases in the *Risky* and in the 92% of the cases in the *Safe* treatment.[8] To test the significance of this difference, we rely on a

[8] In the first period, Weak players chose to collaborate in 80% of the cases in the *Risky*, and 90% of the cases in the *Safe* treatments. The pattern of Weak players' actions across periods is provided in Figure A1 in the Appendix.

random effect panel linear regression, in which the unit of observation is the average by session and period (N = 60) and the only regressor is a dummy for the *Safe* treatment. The estimated coefficient is significantly different from zero ($p = 0.009$).[9] The majority of Weak players consistently chose Collaborate in all ten repetitions of the stage game in the *Safe* treatment (70%), but not in the *Risky* one (47%).

Strong players chose Defensive 77% of the times in the first period of the *Risky* treatment, and this number increased to 86% if we consider all periods. In the *Safe* treatment, 77% of the Strong players opted for Defensive in the first period and 79% played Defensive over all periods ($p = 0.32$, all periods).[10] For the distribution of choices across periods and treatments see Figure A2 in Appendix.

Average earnings for Strong players are 85% of the total surplus in *Risky* and 82% in *Safe* ($p = 0.199$), remarkably close to the earnings for Strong players predicted by the Nash equilibrium (90%). Two observations are in order: (i) the *Inequality Game* successfully managed to generate high levels of inequality as a standard game-theoretical approach (NE) would suggest; (ii) Strong's behavior does not seem to depend on the presence or absence of strategic risk.

Adoption of the exit option. In line with standard game-theoretical predictions, we hardly observe any exit behavior by the Strong players. The exit option was implemented in only one out of 600 encounters. Instead, a non-negligible fraction of Weak players used the exit option in both treatments (Figure 5).[11] In the first phase, the share of exit is similar in the two treatments (6% and 7% in *Risky* and *Safe*, respectively), yet the gap widens in the second phase when Weak players exit more than twice as often in *Safe* than in *Risky* (17% vs 7%). Overall, the share of Weak players choosing the exit option is almost twice as high in *Safe* (12%) compared to *Risky* (7%). In both treatments, about 30% of the Weak players used the exit option at least once: Weak players that exit more than once are 17% in *Risky* and 30% in *Safe*.[12] To formally

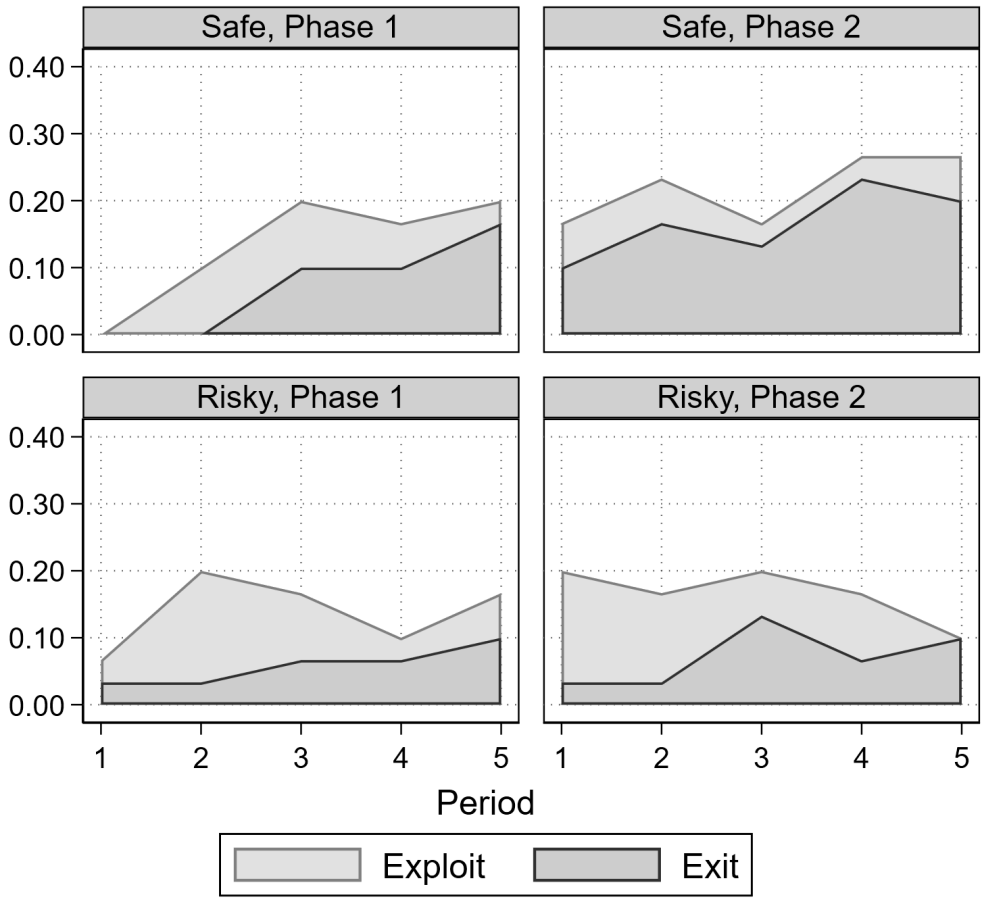
[9] The same specification is adopted below, to assess the significance of the treatment effect on the adoption of the Defensive action, and on surplus.

[10] For the sequential game (*Safe*) we consider only the Collaborate contingency (irrespectively of the actual choice of Weak – Exploit or Collaborate). In the branch where Weak chooses to Exploit, 95% of the Strong players chose Defensive.

[11] Table 1 only reports the frequency of exit by the Weak players.

[12] The maximum observed number of exists for a single player over the 10 periods is 4 in *Risky* and 6 in *Safe*.

Figure 5: Actions of Weak players in the *Exit* treatments



test if there is a gap in exit between the two treatments, we run a panel linear regression where the dependent variable is the average exit per period and session (Model 1 in Table 2). We find evidence that the share of exit increases over periods. Importantly, this increase is significantly more prominent in the *Safe* treatment (*Period x Safe* in the regression), hence leading to a positive treatment effect over time.

Result 1 *Weak players' adoption of the exit option increases with experience, and more so in the Safe compared to the Risky treatment.*

Weak players' behavior. A reason why the exit option is chosen more frequently in the *Safe* than in the *Risky* treatment may be that, in the latter, aversion to disadvantageous inequality may induce Weak players to choose Exploit rather than the exit option, if they expect Strong to play Generous sufficiently often (see Appendix B for predictions under the assumption of

Table 2: Behavior in the *Exit* treatments.

	Weak player			Strong player
	Exit	Exploit	Collaborate	Defensive
	Model 1	Model 2	Model 3	Model 4
Safe tr. (d)	-0.038	-0.042	0.080	-0.071
	(0.051)	(0.056)	(0.072)	(0.054)
Period	0.006	-0.003	-0.003	0.003
	(0.005)	(0.005)	(0.006)	(0.006)
Period × Safe	0.017**	0.002	-0.019**	0.003
	(0.007)	(0.007)	(0.008)	(0.008)
Constant	0.033	0.104***	0.862***	0.847***
	(0.036)	(0.039)	(0.051)	(0.038)
N.obs.	60	60	60	60
R^2 (overall)	0.332	0.035	0.176	0.101

Notes: Models 1 to 4 report results from panel linear regressions with session-level random effects. In Model 1, the dependent variable is the average share of exits by session and period. In Model 2, the dependent variable is the average share of Weak choosing Exploit by session and period. In Model 3, the dependent variable is the average share of Weak choosing Collaborate by session and period. In Model 4, the dependent variable is the share of Strong playing Defensive. For the sequential game (*Safe*) we consider only the Collaborate contingency (irrespectively of the actual choice of Weak – Exploit or Collaborate). Symbols ∗∗∗, ∗∗, and ∗ indicate significance at the 1%, 5% and 10% level, respectively.

inequality aversion). In the *Safe* treatment, instead, if the Weak player chooses to Exploit, the Strong counterpart can react by playing Defensive, which would increase, rather than decrease inequality. Hence exit is the only alternative for Weak players who want to avoid a highly unequal outcome. To test if this can explain the difference across treatments, we run a panel linear regression where the dependent variable is the average Exploit per period and session (Model 2 in Table 2). We fail to provide support to the idea that Weak players choose to play Exploit significantly more often in *Risky* than in *Safe*.

Another way to look at the same phenomena and rule out any "substitution effect" is to test if Collaborate is more frequent in *Safe* than in *Risky*. Model 3 in Table 2 tests if this is verified in the data. The dependent variable is the average of Collaborate choices per period and session. We do not find support for the idea that Weak players collaborate more often in *Safe* than *Risky*. If anything, the share of Weak players who play Collaborate decreases over time with the decline more marked for the *Safe* treatment, and the difference is statistically significant (see *Period x Safe*). In other words, the reason why exit is more prevalent in the

Safe treatment does not lie in a form of substitution between exit and Exploit. The exit option seems to be adopted by players who – in the *Risky* treatment – would have played Collaborate.

So far, we have established that Weak players use exit more often in the *Safe* treatment compared to the *Risky* one, and that this difference is not just driven by a substitution effect. The remainder of this section investigates the possible causes of this treatment difference in the use of exit.

Strong players' behavior. We now focus on the Strong players to see whether their behavior is different across treatments. It may in fact be possible to explain the treatment difference in the use of exit by the Strong players' behavior, if Strong players were more prone to behave altruistically (i.e., out-of-equilibrium) in the *Risky* rather than in the *Safe* treatment. If this is true, the difference in exit could simply be the result of different levels of inequality endogenously generated in the game. However, our data do not support this hypothesis: Strong players chose Defensive 86% of the times in the *Risky* and 81% in the *Safe* treatment.[13] Model 4 in Table 2 reports a panel regression where the dependent variable is the average number of Defensive plays per session and period. While we fail to find any treatment difference, it is interesting to notice that, if anything, Strong players are slightly more likely to play Generous in *Safe* than *Risky*. This is quite in line with the idea that Strong players should be more likely to play Generous when there is no risk of being exploited by their counterpart, which could in principle decrease the Weak players' propensity to exit. To conclude, the exit gap between treatments cannot be explained by differences in the behavior of Strong players.

Consequences of the introduction of the exit option. To understand the impact of the exit option on Strong and Weak players' behavior, we compare the *Exit* treatments with the *Control* treatments where the exit option is not available. As described in the previous section, the availability of the exit option should not affect the behavior of the Strong players.

[13] In the *Risky* treatment, Strong players can choose between Defensive and Generous and make only one decision in each period. In the *Safe* treatment instead, we use the contingent response method and the Strong players have to decide for each possible node of the game. Since the Collaborate node is selected in the vast majority of the instances (92%), we only report data for the Collaborate node, irrespectively of which node was actually reached.

Considering both the *Risky* and the *Safe* treatments, together, we observe that 82% of Strong players in *Control* compared to 84% in *Exit* play Defensive.[14] Models 3 and 6 in Table 3 provide further evidence that the introduction of an exit option does not change the behavior of the Strong players in either the *Risky* or the *Safe* treatments. This finding suggests that Strong players do understand that they should not react to the introduction of the exit option, as a kind action cannot dissuade the counterpart from exiting the game; indeed, the decision to exit is taken before even seeing the decision of the Strong player.

Result 2 *Strong players' behavior is not statistically different across treatments and it is not affected by the introduction of the exit option.*

Overall, the fraction of Weak players choosing Collaborate is 87% in the *Control* treatments, and 84% in the *Exit* treatments.[15] Table 3 reports results for OLS estimations for Weak player behavior, and the dependent variables are Exploit (Model 1 and 4) and Collaborate (Model 2 and 5), separately for *Risky* and *Safe*. Also notice that in the *Risky* treatment, the introduction of the Exit option induces the Weak players to Collaborate more often, and to Exploit less often, as compared to the case when exit is not feasible. Such an effect is absent from the *Safe* treatments, where – if anything – the introduction of the exit option induces the Weak players to Collaborate less as they acquire experience.

4.2 Individual history and exit

One possible explanation for the difference in exit behavior across treatments could be the individual history observed by each player. Even though there is no difference across treatments in the behavior of Strong players at the aggregate level, it is still important to check for the impact of individual experience. To control for individual-level history, at any period t, we focus on the subsample of players who had never been matched before with a Strong player who chose to be Generous (Figure 6). In the initial period of the game, we include all players as none of them has yet observed any action and none of them is exposed to a different history. In

[14] In the first period, 77% of the Strong players played Defensive under *Control* compared to 78% under *Exit*.
[15] In the first period, these fractions are 85% and 97% under *Control* and *Exit*, respectively.

Table 3: Comparison between the *Exit* and *Control* treatments.

	Risky treatments			Safe treatments		
	Exploit	Collaborate	Defensive	Exploit	Collaborate	Defensive
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Exit tr. (d)	-0.173***	0.140**	0.020	-0.024	0.029	0.042
	(0.064)	(0.064)	(0.048)	(0.048)	(0.066)	(0.081)
Period	-0.017***	0.017***	0.005	-0.002	0.002	0.010*
	(0.006)	(0.006)	(0.005)	(0.004)	(0.005)	(0.005)
Exit tr. x Period	0.013	-0.019**	-0.002	0.001	-0.023***	-0.003
	(0.009)	(0.008)	(0.007)	(0.006)	(0.007)	(0.007)
Constant	0.278***	0.722***	0.827***	0.087**	0.913***	0.733***
	(0.045)	(0.045)	(0.034)	(0.034)	(0.047)	(0.057)
N.obs.	60	60	60	60	60	60
R^2 (overall)	0.269	0.134	0.028	0.024	0.361	0.061

Notes: Models 1 to 6 report results from panel linear regressions with session-level random effects. In Models 1 and 4, the dependent variable is the average share of Weak players choosing Exploit by session and period. In Models 2 and 5, the dependent variable is the average share of Weak players playing Collaborate. In Models 3 and 6, the dependent variable is the average share of Strong players playing Defensive. For the sequential game (*Safe*) we consider only the Collaborate contingency (irrespectively of the actual choice of Weak – Exploit or Collaborate). Symbols $***$, $**$, and $*$ indicate significance at the 1%, 5% and 10% level, respectively.

any subsequent period t, we only include Weak players who have never seen a generous action of their matched partner from period 1 until period $t-1$.[16] Figure 6 presents the prevalence of the choice to exit over time for this subset of Weak players who share a common history. Conditional on having observed the same (extreme) history, exit is much more prominent in *Safe* than in *Risky*. If anything, after controlling for individual-level histories, the gap between the two treatments is even more pronounced and it manifests itself already in Phase 1.[17]

Table 4 shows the marginal effects from panel probit regressions on the exit choices of Weak players, with one observation per subject and period, random effects at the individual level, and standard errors clustered at the session level. We include the number of times the Weak player was matched with a Strong player who chose Generous in the earlier periods (*Observed Generous*). Recall that choosing Generous signals the Strong player's intention to share equally.

[16] That is the case if the Strong players in previous interactions always chose Defensive. However, it can also be the case that a Weak player chose to exit in one of the previous $t-1$ periods. In fact, in such a case the Weak player is not given any information about the behavior of the counterpart. This feature of our design prevents a Weak player from updating his beliefs about Strong players' behavior in case of exit.

[17] We corroborate these findings through regressions. Table A1 in Appendix A reports the marginal effects from probit regressions on exit choices of Weak players, with random effects at the subject level. Models 1, 2, and 3 clearly show an incremental treatment effect such that Weak players who always observed Defensive in all previous periods until $t-1$ are increasingly more likely to exit in period t. On the other hand, no such effect is visible for the remaining Weak players, as seen in Models 4, 5, and 6 (see also Figure A3 in Appendix A).

Models 1 and 2 in Table 4 show that Weak players who have observed Generous in the previous periods are in fact significantly less likely to exit, and this effect is more pronounced in the *Safe* treatment, especially in Phase 2.

At the end of Phase 1, players were informed about the average earnings for the Weak members of their own group and the average earnings for the 5 Strong players of the matched set. In Models 3 and 5, we include the ratio between these two averages (*Payoff ratio (ph1)*). A ratio of one implies equal earnings across the two groups. A ratio smaller than 1 indicates that Strong players were ahead and the smaller the ratio, the larger the inequality between the two groups. The idea behind this regressor is that Weak players who see a larger ratio (i.e., less inequality) in the first phase might be less likely to use the exit option in the second phase. Both Models 3 and 5 show that Weak players are less likely to exit in the *Safe* as the payoff ratio of Weak players in Phase 1 increases.

Before the beginning of Phase 2, players also receive information on the number of times the exit option was adopted by the members of their own and their matched set in Phase 1. In Models 4 and 5 we study whether observing a higher number of exits by fellow Weak players in Phase 1 induces Weak players to exit more often in Phase 2. Results suggest that this sort of bandwagon effect is not present in our data.

Result 3 *Similar individual-level experiences induce more exit in the Safe than in the Risky treatment.*

5 The Drivers of Exit: Expectations

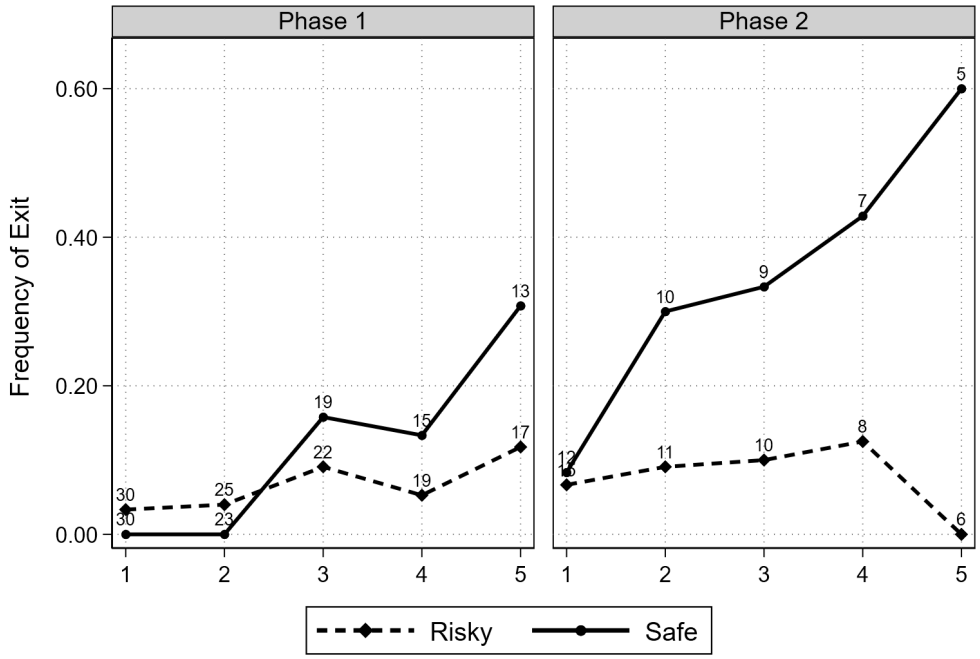
We have documented a treatment difference in exit behavior. Antisocial behavior in the form of exit grows over time and this is true only for the *Safe* treatment, where in Phase 2, exit is more than twice as frequent than in the *Risky* treatment. The gap in exit between treatments is particularly prominent when we focus only on Weak players whose past Strong counterparts were never generous. Given that the realized level of inequality for this subset of players is

Table 4: Individual-level history and the exit option (marginal effects)

	Exit option was chosen (Yes=1 and No=0)				
	Only Weak players				
	Phase 1 only	Phase 2 only			
	Model 1	Model 2	Model 3	Model 4	Model 5
Safe tr. (d)	-0.064	0.117*	0.363***	0.047	0.442***
	(0.083)	(0.069)	(0.119)	(0.121)	(0.135)
Period	0.026**	0.021**	0.018*	0.019*	0.021**
	(0.011)	(0.010)	(0.010)	(0.010)	(0.010)
Period × Safe	0.029	0.012	0.005	0.003	0.008
	(0.018)	(0.016)	(0.015)	(0.015)	(0.016)
Observed Generous	-0.051*	-0.027*			-0.040***
	(0.031)	(0.015)			(0.012)
Obs. Generous × Safe	-0.071	-0.086**			-0.045
	(0.078)	(0.034)			(0.038)
Payoff ratio (ph.1)			0.187		0.263
			(0.404)		(0.530)
Payoff ratio (ph.1) × Safe			-1.590***		-1.249**
			(0.495)		(0.610)
Exit by other Weak in Ph.1				-0.054	-0.042
				(0.035)	(0.030)
Exit by other Weak in Ph.1 × Safe				0.022	-0.029
				(0.045)	(0.033)
Individual characteristics	Yes	Yes	Yes	Yes	Yes
N.obs.	240	300	300	300	300
BIC	186.2803	245.8072	250.1436	250.8238	260.3437

Notes: Models 1 to 5 report the marginal effects from panel probit regressions on exit choices of Weak players, with random effects at the subject level. The dependent variable takes value 1 if Weak chooses exit and 0 otherwise. Model 1 includes only Phase 1, Models 2 to 5 include Phase 2 only. Controls for individual characteristics include age and the number of mistakes made in the control questions, and a set of dummies for: male, political orientation (indicating self-reported right-wing political views), non-German subjects, field of study (social sciences, hard sciences, and humanities). Standard errors robust for clustering at the session level (in parentheses). BIC stands for Bayesian Information Criterion. Symbols $***$, $**$, and $*$ indicate significance at the 1%, 5% and 10% level, respectively.

Figure 6: Frequency of exit for Weak players who never observed Generous



Notes: The horizontal axis reports the period within each phase, and the vertical axis reports the frequency of exit. Panel on the left provides the frequencies for Phase 1, and the panel on the right for Phase 2. The solid line is for the *Safe* treatment, whereas the dashed line is for the *Risky*. Labels on the lines provide the number of observations corresponding to that frequency. The number of observations decreases across periods since Weak players who observe Generous at time t are excluded from the analysis starting from time $t+1$.

exactly the same under the *Risky* and the *Safe* treatment, this finding suggests that the choice to exit is not simply driven by an aversion to inequality.

Our preferred interpretation builds on the model proposed by Battigalli et al. (2019b), to suggest that the exit divide can be explained by a mismatch between expectations and realized outcomes in the game where Strong players could easily opt for the equal outcome. To see the implications of Battigalli et al. (2019b)'s concept of "simple anger" in our set-up, here we consider a basic strategic context, and we focus on an out-of-equilibrium situation, where players can hold incorrect beliefs. Specifically, we need to look at our framework not just as a sequence of 10, distinct one-shot games, but as a single multi-stage game involving several one-shot interactions with different opponents. For simplicity, let us focus on the *Safe* environment, and consider a two-period version of our game. We now have two Weak players w_1 and w_2 who

meet two Strong players s_1 and s_2, and switch opponents at the end of the first interaction. Let us focus on the Weak players.

Following Battigalli et al. (2019b), we define player w_i's Frustration $F_{w_i}(h, \alpha_{w_i})$ at history h, given w_i's first order beliefs α_{w_i}, as "the gap, if positive, between w_i's initially expected payoff and the currently best expected payoff he believes he can obtain" (p.9). More formally:

$$F_{w_i}(h, \alpha_{w_i}) = \left[\mathbb{E}[\pi_{w_i}; \alpha_{w_i}] - \max_{a_{w_i} \in A_{w_i}(h)} \mathbb{E}[\pi_{w_i}|(h, a_{w_i}); \alpha_{w_i}]\right]^+$$

where $[x]^+ = max\{x, 0\}$ and $A_{w_i}(h)$ denotes the set of actions available to player w_i at h.

At any stage, players are assumed to choose the action that would maximize their expected utility, which in the case of simple anger (SA) does not depend on the attribution of blame, and player w_i's expected utility at history h is then represented by the following function:

$$u^{SA}_{w_i}(h, a_{w_i}, \alpha_{w_i}) = \mathbb{E}[\pi_{w_i}|(h, a_{w_i}); \alpha_{w_i}] - \theta_{w_i} F_{w_i}(h, \alpha_{w_i}) \mathbb{E}[\pi_{s_j}|(h, a_{w_i}); \alpha_{w_i}]$$

This implies that player w_i's propensity to hurt his/her opponent s_j is proportional to w_i's frustration.[18] By definition, there is no frustration at the root, hence this model cannot predict exit before the first interaction takes place.

Suppose that the Weak player w_1 planned to Collaborate in both interactions, and expected w_2 to do the same, and both Strong players to choose to be Generous at all nodes if their current Weak opponent chooses to Collaborate, and to be Defensive otherwise. Now, consider the case in which in the first interaction w_1 Collaborates but his opponent s_1 plays Defensive. At the beginning of the second interaction, w_1 does not know the outcome of the interaction between w_2 and s_2, but he knows he will interact with s_2, and he still expects s_2 to be Generous if he Collaborates. At this stage, player w_1 experiences a positive level of frustration, since the best expected payoff he believes he can obtain (that is 3) is lower than the initially expected one

[18] One could also think that unmet expectations could increase player w_i's propensity to reduce future frustration. Indeed, w_i might exit to avoid future frustration and not in the hope to hurt his/her opponent. While this is an interesting perspective, it cannot explain the difference in the exit rate across treatments in our study (see for instance Figure 6). This is why we do not investigate this mechanism further, in this study.

(which is 5, in our simple example).[19] This can justify the choice to Exit before the second interaction if his sensitivity to frustration, denoted by θ_{w_1}, is high enough. Formally, this is true if Exit maximizes his expected utility:

$$u_{w_1}^{SA}(h, a_{w_1}, \alpha_{w_1}) = \mathbb{E}[\pi_{w_1}|(h, a_{w_1}); \alpha_{w_1}] - \theta_{w_1} F_{w_1}(h, \alpha_{w_1}) \mathbb{E}[\pi_{s_2}|(h, a_{w_1}); \alpha_{w_1}]$$

At this stage, in our simple example, the expected utility from Exit is 0.5, while the expected utility from Collaborating is $3 - \theta_{w_1} \times 2 \times 5$; hence, the Weak player w_1 will Exit if $\theta_{w_1} > 0.25$.

This reasoning can be easily extended to our more complicated framework with ten repeated interactions, and to the *Risky* environment, and illustrates how frustration – hence anger and antisocial behavior – should be more prominent when the Weak have more optimistic expectations over the generosity of the Strong players, and after histories of play in which these expectations are repeatedly violated, increasing the gap between the payoff expected at the beginning and the best payoff a player can expect from that point on.

Notice that the proposed explanation hinges on the hypothesis that Weak players have incorrect beliefs over the strategy adopted by their opponents. More specifically, they are more *optimistic* about Strong players' propensity to be Generous in *Safe* than in *Risky*. To test this conjecture, we run a follow-up experiment with a new sample of participants who did not take part in the Main Experiment. We intentionally did not collect a measure of beliefs in the Main Experiment to avoid any possible interaction with the main decisions in the game.

Experimental design. We invited a new set of 122 subjects and we asked them to read the instructions of the Main Experiment. Each subject was exposed to either the *Safe* or the *Risky* version of the *Inequality Game*. They all read the instructions for the relevant treatment with an exit option. After reading the instructions, participants were asked to make two guesses: the number of Strong players who selected Defensive in the first period, and the number of Weak players who selected Collaborate in the first period out of 10 players who did not exit.

[19] The best expected payoff after this history is 3 if player w_1, who earned 1 in the first interaction, expects to earn 5 in the second one and knows he will be paid for either one of the two interactions.

Both estimates had to be integer numbers between 0 and 10. The belief elicitation task was incentivized according to a quadratic scoring rule, based on the comparison between the answers given and data from previous sessions of the Main Experiment (see Instructions in Appendix C).[20] In the model by Battigalli et al. (2019b), frustration emerges from the comparison between actual experiences and ex-ante expectations, which are well approximated by the beliefs elicited from a set of subjects who did not play the game, but only read the instructions. In the Belief experiment, we decided to elicit beliefs only on the first period, and not to the subsequent ones, to keep the design as simple and clean as possible.

Participants were recruited via ORSEE (Greiner, 2015) from the same pool as the one of the Main Experiment. We ran 2 sessions for each between-subjects treatment at CLER in April 2018. After reading the instructions, participants answered the same set of 10 control questions used in the Main Experiment. To ensure that participants carefully read and understood the instructions, we paid them €0.20 for each control question correctly answered at the first try. Only one of the two guesses selected at random at the end of the experiment was relevant for payments. Earnings ranged from €5.50 to €19, with an average of €15.50, including a €4 show-up fee. A session lasted 45 minutes on average.

Results for the Belief Experiment. Figure 7 reports the distribution of expectations divided by player type and treatment. Results show that subjects who read the instructions for the *Risky* treatment have different prior beliefs for both Strong and Weak player actions in the first period than subjects who read the instructions for the *Safe* treatment.

Panel (a) of Figure 7 shows the distribution of guesses for the number of Strong players who chose Defensive for the two treatments, and Panel (b) shows the distribution of guesses for the number of Weak players who chose Collaborate for the two treatments.[21] Mean guess for the number of Strong players who select Defensive in the first period is 8.2 for the *Risky* treatment, whereas it decreases to 7.5 for the *Safe* treatment ($p = 0.045$, Wilcoxon rank-sum

[20] In particular, we had a random draw of 10 Strong and 10 Weak players that was performed at the individual level to avoid informational spillovers across sessions.

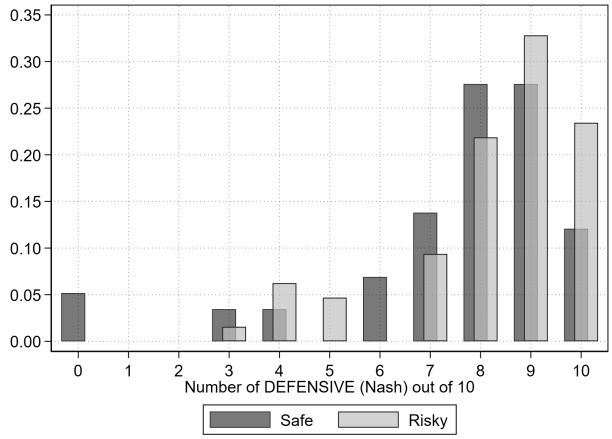
[21] For the *Safe* treatment, subjects make their guess on Strong player actions conditional on the Weak player selecting Collaborate.

test). In other words, ex ante, subjects who saw the *Safe* treatment expected Strong players to choose Generous more often than their counterparts who saw the *Risky* treatment.

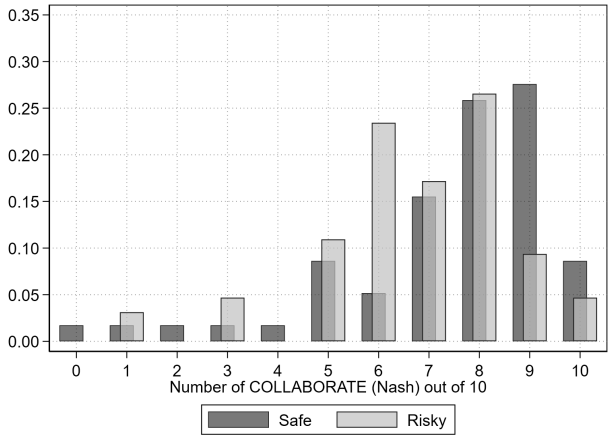
Mean guess for the number of Weak players who select Collaborate in the first period is 6.8 for the *Risky* treatment and it is 7.4 for the *Safe* treatment ($p = 0.011$, Wilcoxon rank-sum test). Participants in the Belief Experiment can clearly recognize the fact that in the *Risky* game Weak players can try to exploit Strong players with the hope of securing a higher payoff for themselves.

Figure 7: Expectations about Strong and Weak player actions

(a) Strong players expected to choose Defensive



(b) Weak players expected to choose Collaborate



Altogether, these results suggest that subjects who see the *Safe* treatment perceive an equitable outcome as more likely than the ones who see the *Risky* treatment, as the Strong are expected to be more generous when they can unilaterally choose the equal split without any fear of being exploited.

Result 4 *Subjects expect the Strong players to play more generously in the Safe than in the Risky treatment.*

6 Discussion

Results from the Belief Experiment (Section 5) indicate that subjects recognize the strategic uncertainty faced by the Strong players in the *Risky* treatment, and expect them to be more generous in the *Safe* environment. The model by Battigalli et al. (2019b) would then suggest that the same negative experiences with the Strong players would induce more frustration – hence more anger – among the Weak in the *Safe* than in the *Risky* treatment. Increased frustration and anger would then lead to more antisocial behavior, in the form of exit. While we do not establish any direct link between anger and antisocial behavior, previous experiments have demonstrated the crucial role of emotions in shaping economic decisions. Pillutla and Murnighan (1996) showed that self-reported anger is a powerful predictor of rejections in a Dictator Game; Fehr and Gächter (2002) find that negative emotions toward defectors ignite altruistic punishment in a Public Goods Game. Similarly, Bosman and van Winden (2002) find that negative emotions drive destructive behavior in the power-to-take-game.

Results from the Main Experiment confirm this prediction, and indicate that the difference between the *Safe* and the *Risky* environments emerges specifically among those players who never observed a Generous move by their past Strong counterparts. Here we show that our evidence cannot be rationalized simply by fairness (Rabin, 1993; Fehr and Schmidt, 1999) or responsibility (Bartling and Fischbacher, 2012) models.

While models based on frustration and anger can provide a rationale for higher levels of exit in *Safe* than in *Risky*, alternative explanations based on fairness or responsibility cannot fully account for our results (proof in Appendix B). Unlike a standard game-theoretical approach, inequality aversion à la Fehr and Schmidt (1999) can explain the choice of Weak to exit. Indeed, sufficiently inequality averse Weak players may choose to exit both in the *Safe* and in the *Risky* version of the game, if they have pessimistic expectations about the behavior of the Strong player – i.e., they expect their counterpart to play Defensive. However, Fehr and Schmidt (1999)'s model does not predict the behavioral differences in the use of exit we observe across treatments. Indeed, the only way to rationalize the treatment difference in the Main Experiment is that the Weak players should expect Strong players to be nicer – i.e., choosing Generous more

often – in *Risky* than *Safe*. If that were the case, higher levels of exit in the *Safe* treatment could be explained by inequality aversion. Instead, results from the Belief Experiment suggests just the opposite. By contrast, in (Rabin, 1993)'s model of intention-based reciprocity Exit cannot be part of a fairness-equilibrium, in either version of the game. If the Weak player believes that the Strong opponent believes that he is choosing to Exit (second order belief), then any action by the Strong player is payoff-irrelevant and cannot be perceived by the Weak player either as kind or as unkind; as a consequence the Weak player does not have any motive to incur a material cost in order to increase or to reduce the Strong player's payoff.

An alternative explanation could rely on the idea that Strong players have a different degree of responsibility in the two treatments. Strong players choosing Defensive in the *Safe* treatment can clearly be held accountable for the unequal outcome. Intentions of Strong players are instead not entirely clear in the *Risky* treatment. Bartling and Fischbacher (2012) suggest that responsibility attribution plays an important role in the decision to punish. While the general idea of responsibility attribution as a trigger of punishment could be applied to our set-up, formal models in this line of research always assume that blame is the result of some harm (no harm, no blame). However, in the *Exit* treatments, one cannot punish a Strong player that is for sure to be blamed as the exit decision is taken before knowing the action – and hence the responsibility level – of the counterpart. This implies that the model proposed by Bartling and Fischbacher (2012) does not apply to our game. One could interpret responsibility more broadly, and suppose that the Weak players might punish their counterparts just for being part of the Strong group, which is held collectively responsible for the realized inequality. While we cannot exclude this interpretation, we are not aware of any formal theory of responsibility attribution towards a collective entity. In addition, the existing experimental evidence suggests that responsibility and blame are attributed mainly to pivotal decision makers, while subjects are less prone to punish players whose choices cannot be directly linked to the negative outcome (Bartling et al., 2015; Duch et al., 2015).

Finally, one may argue that in our Main Experiment, participants do not know what their current opponent has done in the past, but may form beliefs based on their previous interactions

with other Strong players and hence expect the next counterpart to behave as other people from his/her lot. Weak players could use this information to update their beliefs and exit due to statistical discrimination. Indeed, a Weak player with a strong aversion to inequality might decide to exit given his/her revised expectations about Strong players as a group. While we cannot exclude this argument as a motive to exit in general, it cannot account for the observed treatment difference. In fact, we observe that exit is more frequent in *Safe* than *Risky* precisely among subjects who have never observed their previous Strong opponents choosing Generous, and hence should have formed very similar beliefs about the behavior of the Strong players as a group.

7 Conclusion

The steady increase in economic inequality is considered one of the main societal challenges of our times. Social tensions have gained a prominent role in the public arena and are at the center of heated political debates. The media and the rise of popular movements such as Occupy Wall Street have increasingly given voice to these tensions. While the extent of inequality is of course a decisive factor in fueling social unrest, not all inequalities are born alike. What is deemed fair and acceptable can greatly depend on the process that led to inequality.

We experimentally study the rise of antisocial behavior, under two different scenarios. In the first scenario (*Risky* treatment), a reduction in inequality is difficult to achieve and the poor have low expectations about a more equal society. In the second scenario (*Safe*), the rich can unilaterally reduce inequality and that induces more optimistic expectations among the poor. In both cases, the poor can signal their disappointment for the unfairness of the situation by destroying all the surplus (exit option in the experimental set-up). Not only this behavior is costly for the punisher; it is also highly inefficient. Besides, it can only target strangers, whose reputation is unobservable and with whom one has never interacted before.

Our findings suggest that the mismatch between expectations and realized outcomes is a major contributor for the decision to engage in antisocial behavior targeted at others, whose

past is unknown. In other words, if the difference between expectations and reality is wider, we should expect much stronger reactions to inequality than what the absolute inequality level itself might suggest. This result relates our work to the recent literature on the rise of anti-elite populist movements, which suggests explanations based on the frustration triggered among the relatively "weaker" part of the population by economic insecurity shocks (see Guiso et al. 2017, and references therein). Passarelli and Tabellini (2017) also highlight the role of emotions in triggering political unrest and shaping public policies; interestingly enough, they attribute a crucial role to the so called "resignation effect". Negative emotions and protests arise when citizens think they have been treated unfairly, and this notion depends on the options available to the government. When feasible options are limited, citizens revise their expectations accordingly and are more likely to accept less favorable policies without protesting – a mechanism that bears some resemblance with the idea proposed by Battigalli et al. (2019b) and is in line with our results.

Clearly, our study only represents a first step toward a better understanding of these intricate dynamics. Further experiments could prove useful in isolating specific aspects of populist movements, such as the role of propaganda on expectations and disappointment thereof. Another interesting aspect that could be studied by extending our experimental setting is the role of middle class in promoting a healthier functioning of the society, as long theorized in the political science literature. Controlled laboratory evidence on the role of income bipolarization on antisocial behavior could be important in light of the increase in absolute bipolarization observed in the last decades (Roope et al., 2018).

References

Abbink, K. (2012). Laboratory Experiments on Conflict. In M. Garfinkel and S. Skaperdas (Eds.), *The Oxford Handbook of the Economics of Peace and Conflict*. New York: Oxford University Press.

Abbink, K. and B. Herrmann (2011). The moral costs of nastiness. *Economic Inquiry 49*(2),

631–633.

Abbink, K., D. Masclet, and D. Mirza (2018, 3). Inequality and inter-group conflicts: Experimental evidence. *Social Choice and Welfare 50*(3), 387–423.

Abbink, K. and A. Sadrieh (2009). The pleasure of being nasty. *Economics Letters 105*(3), 306–308.

Aina, C., P. Battigalli, and A. Gamba (2018). Frustration and Anger in the Ultimatum Game: An Experiment. IGIER Working Paper No. 621.

Alesina, A. and G.-M. Angeletos (2005). Fairness and Redistribution. *American Economic Review 95*(4), 960–980.

Alesina, A. and E. La Ferrara (2005). Preferences for Redistribution in the Land of Opportunities. *Journal of Public Economics 89*(5–6), 897–931.

Almås, I., A. W. Cappelen, and B. Tungodden (2016). Cutthroat Capitalism Versus Cuddly Socialism: Are Americans More Meritocratic and Efficiency-Seeking than Scandinavians? Discussion Paper No. 18/2016, NHH Dept. of Economics.

Ashton, M. C. and K. Lee (2009). The HEXACO–60: A Short Measure of the Major Dimensions of Personality. *Journal of Personality Assessment 91*(4), 340–345.

Balafoutas, L., M. G. Kocher, L. Putterman, and M. Sutter (2013). Equality, Equity and Incentives: An Experiment . *European Economic Review 60*, 32–51.

Balafoutas, L., N. Nikiforakis, and B. Rockenbach (2014). Direct and Indirect Punishment among Strangers in the Field. *Proceedings of the National Academy of Sciences 111*(45), 15924–15927.

Bartling, B. and U. Fischbacher (2012). Shifting the Blame: On Delegation and Responsibility. *The Review of Economic Studies 79*(1), 67–87.

Bartling, B., U. Fischbacher, and S. Schudy (2015). Pivotality and responsibility attribution in sequential voting. *Journal of Public Economics 128*, 133–139.

Battigalli, P., R. Corrao, and M. Dufwenberg (2019). Incorporating belief-dependent motivation in games. *Journal of Economic Behavior & Organization 167*, 185–218.

Battigalli, P., M. Dufwenberg, and A. Smith (2019). Frustration, aggression, and anger in leader-follower games. *Games and Economic Behavior 117*, 15–39.

Bolle, F., J. H. W. Tan, and D. J. Zizzo (2014). Vendettas. *American Economic Journal: Microeconomics 6*(2), 93–130.

Bolton, G. E. and A. Ockenfels (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review 90*(1), 166–193.

Bortolotti, S., I. Soraperra, M. Sutter, and C. Zoller (2017). Too Lucky to be True Fairness Views under the Shadow of Cheating. Discussion Paper No. 6563/2017, CESIFO.

Bosman, R., M. Sutter, and F. van Winden (2005). The impact of real effort and emotions in the power-to-take game. *Journal of Economic Psychology 26*(3), 407–429.

Bosman, R. and F. van Winden (2002). Emotional hazard in a power-to-take experiment. *Economic Journal 112*(476), 147–169.

Cappelen, A. W., T. Halvorsen, E. O. Sørensen, and B. Tungodden (2017). Face-saving or fair-minded: What motivates moral behavior? *Journal of the European Economic Association 15*(3), 540–557.

Cappelen, A. W., A. D. Hole, E. O. Sørensen, and B. Tungodden (2007). The Pluralism of Fairness Ideals: An Experimental Approach. *American Economic Review 97*(3), 818–827.

Cappelen, A. W., J. Konow, E. Ø. Sørensen, and B. Tungodden (2013). Just Luck: An Experimental Study of Risk-Taking and Fairness. *American Economic Review 103*(4), 1398–1413.

Cassar, L. and A. H. Klein (2016). A Matter of Perspective: How Experience Shapes Preferences for Redistribution. Discussion paper, Mimeo.

Charness, G. and M. Rabin (2002). Understanding Social Preferences with Simple Tests. *Quarterly Journal of Economics 117*(3), 817–869.

Duch, R., W. Przepiorka, and R. Stevenson (2015). Responsibility attribution for collective decision makers. *American Journal of Political Science 59*(2), 372–389.

Eckel, C. C., E. Fatas, and M. J. Kass (2016). Sacrifice: An Experiment on the Political Economy of Extreme Intergroup Punishment. Working paper.

Esteban, J.-M. and D. Ray (1994). On the measurement of polarization. *Econometrica 62*(4), 819–851.

Faravelli, M. (2007). How Context Matters: A Survey Based Experiment on Distributive Justice. *Journal of Public Economics 91*(7–8), 1399–1422.

Fehr, D. (2015). Is Increasing Inequality Harmful? Experimental Evidence. WZB Discussion Paper.

Fehr, E. and U. Fischbacher (2004). Third-party punishment and social norms. *Evolution and Human Behavior 25*(2), 63–87.

Fehr, E. and S. Gächter (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Review 90*(4), 980–994.

Fehr, E. and S. Gächter (2002). Altruistic Punishment in Humans. *Nature 415*(6868), 137–140.

Fehr, E. and K. M. Schmidt (1999). A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics 114*(3), 817–868.

Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-made Economic Experiments. *Experimental Economics 10*(2), 171–178.

Gangadharan, L., P. J. Grossman, M. K. Molle, and J. Vecci (2019). Impact of social identity and inequality on antisocial behaviour. *European Economic Review 119*(C), 199–215.

Greiner, B. (2015). Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE. *Journal of the Economic Science Association 1*(1), 114–125.

Guiso, L., H. Herrera, M. Morelli, and T. Sonno (2017). Populism: Demand and Supply. SSRN Scholarly Paper ID 2924731, Social Science Research Network, Rochester, NY.

Güth, W. and M. G. Kocher (2014). More than Thirty Years of Ultimatum Bargaining Experiments: Motives, Variations, and a Survey of the Recent Literature. *Journal of Economic Behavior & Organization 108*, 396–409.

Güth, W., R. Schmittberger, and B. Schwarze (1982). An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior & Organization 3*(4), 367–388.

Konow, J. (2000). Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions. *American Economic Review 90*(4), 1072–1091.

Lacomba, J. A., F. Lagos, E. Reuben, and F. van Winden (2014). On the Escalation and De-escalation of Conflict. *Games and Economic Behavior 86*, 40–57.

Lohmann, S. (1993). A Signaling Model of Informative and Manipulative Political Action. *American Political Science Review 87*(2), 319–333.

Nikiforakis, N. (2008). Punishment and Counter-Punishment in Public Good Games: Can we Really Govern Ourselves? *Journal of Public Economics 92*(1-2), 91–112.

Passarelli, F. and G. Tabellini (2017). Emotions and political unrest. *Journal of Political Economy 125*(3), 903–946.

Permanyer, I. (2018). Income and social polarization: theoretical approaches. In C. D'Ambrosio (Ed.), *Handbook of Research on Economic and Social Well-Being*, pp. 434–459. Cheltenham, UK: Edward Elgar Publishing.

Persson, E. (2018). Testing the impact of frustration and anger when responsibility is low. *Journal of Economic Behavior & Organization 145*, 435–448.

Pillutla, M. M. and J. Murnighan (1996). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational Behavior and Human Decision Processes 68*(3), 208 – 224.

Potegal, M., G. Stemmler, and C. Spielberger (2010). *International Handbook of Anger: Constituent and Concomitant Biological, Psychological, and Social Processes*. New York, NY, US: Springer.

Prediger, S., B. Vollan, and B. Herrmann (2014). Resource scarcity and antisocial behavior. *Journal of Public Economics 119*(C), 1–9.

Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *The American Economic Review 83*(5), 1281–1302.

Roope, L., M. Nino-Zarazúa, and F. Tarp (2018). How polarized is the global income distribution? *Economics Letters 167*, 86 – 89.

Ule, A., A. Schram, A. Riedl, and T. N. Cason (2009). Indirect Punishment and Generosity Toward Strangers. *Science 326*(5960), 1701–1704.

Zizzo, D. J. (2003). Money burning and rank egalitarianism with random dictators. *Economics Letters 81*(2), 263–266.

Zizzo, D. J. and A. J. Oswald (2001). Are People Willing to Pay to Reduce Others' Incomes? *Annals of Economics and Statistics* (63-64), 39–65.

A Tables and Figures

Table A1: Individual histories and exit behavior

	Exit option (Only Weak players, Yes=1 and No=0)					
	Never observed Generous			Observed Generous at least once		
	Phase 1 only	Phase 2 only	All phases	Phase 1 only	Phase 2 only	All phases
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Period	0.022***	0.005	0.016**	0.044	0.045***	0.041**
	(0.008)	(0.012)	(0.006)	(0.042)	(0.016)	(0.016)
Safe	-0.099	-0.093	-0.086	-0.121	0.187	0.098
	(0.077)	(0.100)	(0.070)	(0.240)	(0.141)	(0.081)
Period × Safe	0.038*	0.060***	0.036**	0.015	-0.035	-0.026
	(0.021)	(0.018)	(0.017)	(0.044)	(0.026)	(0.026)
Phase 2 (d)=1			0.017			0.060***
			(0.026)			(0.018)
Phase 2 × Safe=1			0.104**			0.053
			(0.043)			(0.089)
Individual characteristics	Yes	Yes	Yes	Yes	Yes	Yes
N.obs.	213	93	306	64	177	248
BIC	155.6855	127.9139	233.9083	65.78362	149.1246	188.2652

Notes: Models 1 to 6 report the marginal effects from probit regressions on exit choices of Weak players, with random effects at the subject level. The dependent variable takes value 1 if Weak chooses Exit and 0 otherwise. Models 1 and 4 include Phase 1 only, Models 2 and 5 include Phase 2 only, Models 3 and 6 include both phases. In all models except Model 4, controls for individual characteristics include age and the number of mistakes made in the control questions, and a set of dummies for: male, political orientation (indicating self-reported right-wing political views), non-German subjects, field of study (social sciences, hard sciences, and humanities). In Model 4, controls for individual characteristics include age and the number of mistakes made in the control questions, and a set of dummies for: political orientation (indicating self-reported right-wing political views), non-German subjects, field of study (social sciences, hard sciences, and humanities). The difference in Model 4 is because only male subjects exited in Phase 1. Standard errors robust for clustering at the session level (in parentheses). Symbols $***$, $**$, and $*$ indicate significance at the 1%, 5% and 10% level, respectively.

Figure A1: Actions of Weak players in *Control* treatments

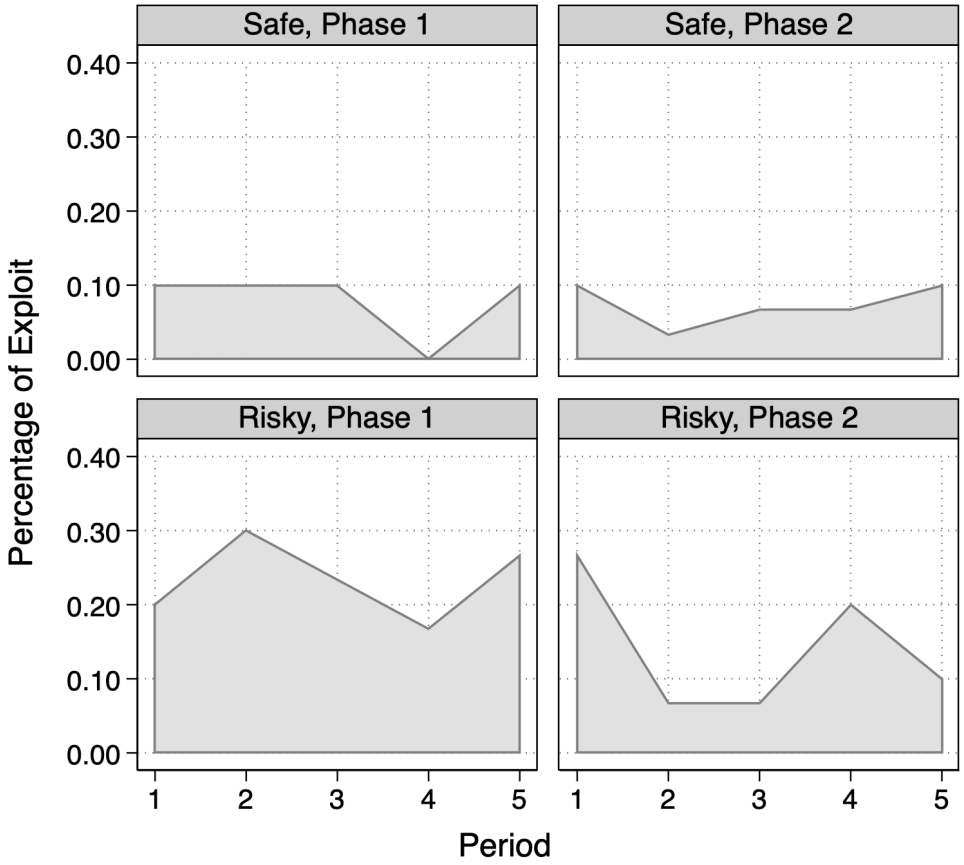


Figure A2: Actions of Strong players by treatment

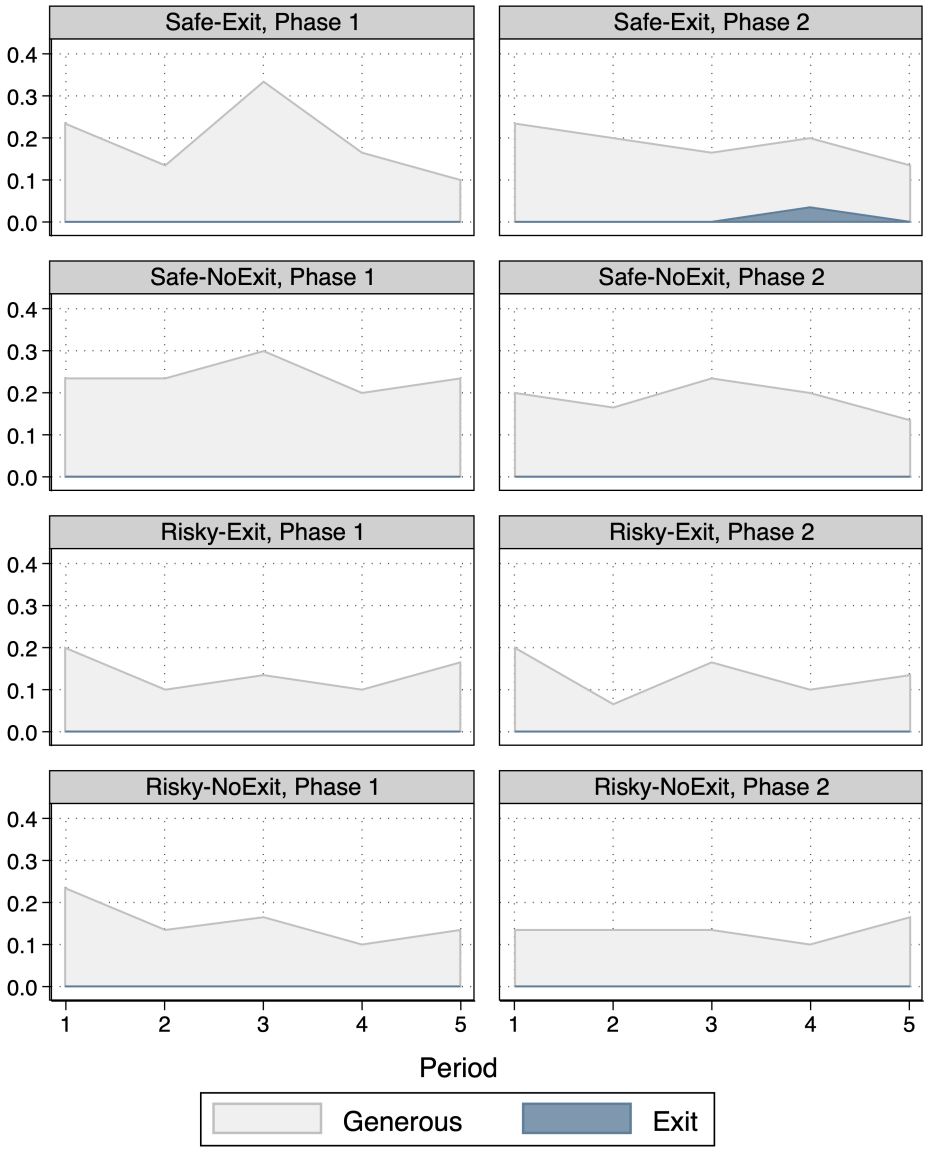
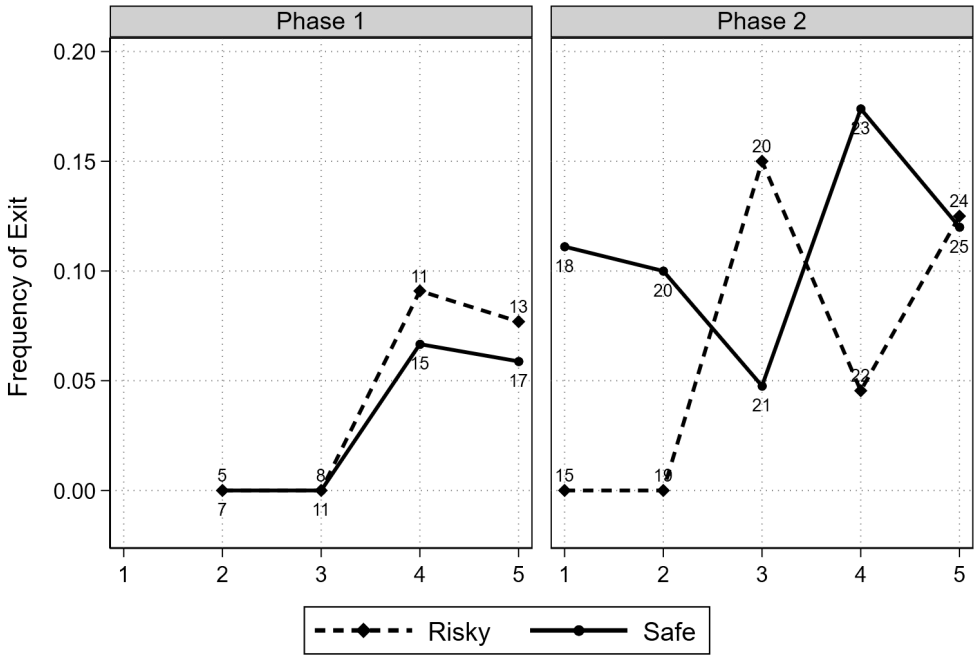


Figure A3: Frequency of exit for Weak players who saw at least one Generous choice



Notes: The horizontal axis reports the period within each phase, and the vertical axis reports the frequency of exit. Panel on the left reports the frequencies for Phase 1, and the panel on the right for Phase 2. The solid line is for the *Safe* treatment, whereas the dashed line is for the *Risky*. Labels on the lines provide the number of observations corresponding to that frequency. The number of observations increases across periods since the number of Weak players with a constant history of having observed Defensive from period 1 throughout period $t-1$ decreases whenever they are matched with a Strong player who plays Generous.

B Theoretical predictions

For the risky treatment without the exit option, standard game-theoretical predictions trivially suggest a unique Nash equilibrium in which the Weak player chooses Collaborate and the Strong player chooses Defensive. The sequential version of the game (*Safe* treatment) has a unique SPNE, which yields the same outcome. In both treatments, when the exit option is available, it should never be used in equilibrium.

Under the assumption of inequality aversion, we consider a utility function of the Fehr and Schmidt (1999) type, where utility for player i is given by

$$U_i(x) = \begin{cases} x_i - \beta(x_i - x_j) & if \quad x_i \geq x_j \\ x_i - \alpha(x_j - x_i) & if \quad x_i < x_j \end{cases}$$

where $x = x_i, x_j$ denotes a vector of monetary payoffs for players i and j and α and β represents the sensitivity toward disadvantageous and advantageous inequality. We assume that $\alpha \geq \beta$ and $0 \leq \beta < 1$.

We denote with $p_{collaborate}$ be the expected probability attached to the event that Weak plays Collaborate and $p_{defensive}$ the expected probability that Strong plays Defensive. We derive equilibrium predictions based on α, β, $p_{collaborate}$, $p_{defensive}$.

One threshold, γ, is relevant for deriving the theoretical predictions for the Strong players:

$$\gamma_1 = \frac{9 + 8\alpha - 10\beta}{5 + 8\alpha - 2\beta} \tag{1}$$

Three thresholds, θ, are relevant for deriving the theoretical predictions for the Weak players:

$$\theta_1 = \frac{4 - 8\beta}{5 + 2\alpha - 8\beta} \tag{2}$$

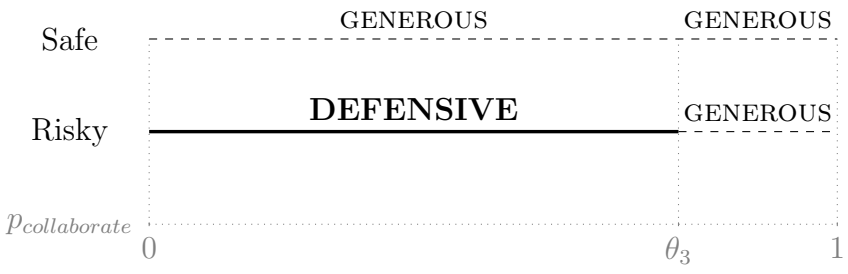
$$\theta_2 = \frac{9 - 8\beta}{9 + 10\alpha - 8\beta} \tag{3}$$

$$\theta_3 = \frac{5}{4 + 8\alpha} \tag{4}$$

Predictions for the Strong players under inequality aversion

Let us first consider the treatments without the exit option (*Control* treatments). It is immediate to see that Strong players with $\beta < 1/2$ always play Defensive in both treatments. Figure B1 summarizes the predictions for inequality-averse Strong players ($\alpha \geq \beta > 1/2$) for both versions of the game. In the *Safe* treatment, an inequality-averse Strong player ($\alpha \geq \beta > 1/2$) always plays Generous. In this case, the choice of the Strong players only depends on their inequality aversion and not on the beliefs about the Weak players. In the *Risky* treatment instead, the share of Strong players choosing Generous depends on both inequality aversion and beliefs about the Weak player behavior. In particular, a Strong player chooses Generous if $\alpha \geq \beta > 1/2$ & $p_{collaborate} > \gamma_1$. One can see from Figure B1 that inequality averse players that would play Generous in *Safe* may play Defensive in *Risky* because they expect a large enough fraction of the Weak players to play Exploit.

Figure B1: Predictions for inequality-averse Strong players ($\alpha \geq \beta > 1/2$)

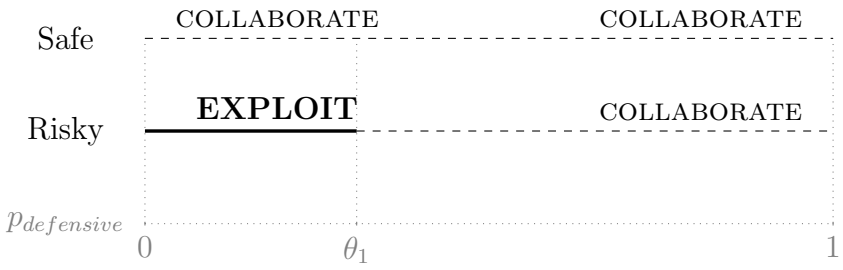


Considering the treatments with the exit option, the predictions for the Strong players are the same as for the *Control* treatments without the exit option. In *Safe*, Strong players will never choose to exit, since – for any value of β, with $0 \leq \beta \leq 1$ – the utility of Exit is 0, while they can get a utility strictly higher than 0 by choosing Defensive.[22] The same reasoning applies for the *Risky* treatment.

Predictions for the Weak players under inequality aversion

Figure B2 summarizes the predictions for inequality-averse Weak players in the *Control* treatments. In the *Safe* treatment, there is no value of α and β such that Weak plays Exploit. In the *Risky* treatment instead, a Weak player will play Exploit if $\beta < 1/2$ & $p_{defensive} < \theta_1$.

Figure B2: Predictions for Weak players who are not strongly averse to favorable inequality ($\beta < 1/2$) in *Control* treatments



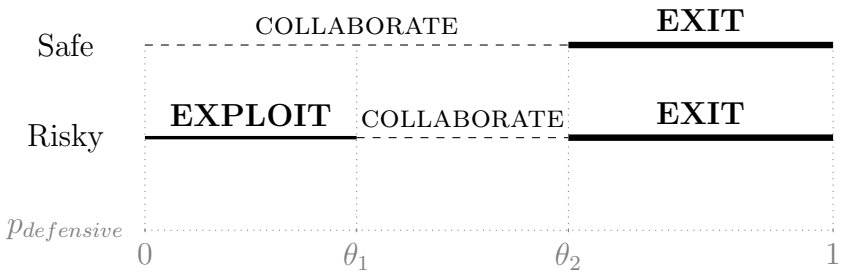
Moving to the treatments where the exit option was available, Weak players' behavior depends on their sensitivity to inequality and their expectations about $p_{defensive}$. In particular, we distinguish two cases based on the parameters of the utility function.

Case 1. For $\alpha < \frac{9}{22} \vee \beta > \frac{22\alpha - 9}{64\alpha - 8}$, the predictions are shown in Figure B3. In the *Safe* treatments, Weak players play Collaborate unless they expect Strong players to play Defensive with $p_{defensive} > \theta_3$, in which case they prefer to Exit. In the *Risky* treatments, Weak players Exit if they expect Strong players to play Defensive with $p_{defensive} > \theta_3$, as in *Safe*. However, players with $p_{defensive} \leq \theta_3$ might play either Collaborate or Exploit. If a Weak player expects

[22] Conditional on Weak player choosing Collaborate. If the Weak player chooses Exploit, and $\beta = 1$, the utility of Defensive would be exactly 0.

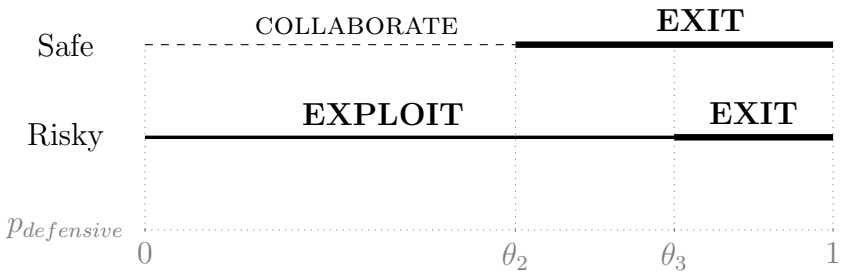
Defensive with a low enough probability, she would play Exploit. The intuition is as follows: the Weak player has a fairly good chance to be matched with a Strong player that will choose Generous and can hence exploit him by playing Exploit, since it would yield 9 for the Weak player.

Figure B3: Predictions for the Weak players in *Exit* treatments (case 1)



Case 2. For $\alpha > \frac{9}{22}$ and $\beta < \frac{22\alpha - 9}{64\alpha - 8}$, the predictions are shown in Figure B4. The predictions for the *Safe* treatment are the same as in Case 1: the Weak players will play Collaborate if $p_{defensive} \leq \theta_3$ and Exit otherwise. For the *Risky* treatment, Weak players choose Exploit if $p_{defensive} \leq \theta_2$, and Exit otherwise. One might notice that for large enough α and small enough β, some players that were willing to Exit in *Safe* are now willing to play Exploit. They will never play Collaborate as they are very sensitive to disadvantageous inequality and hence prefer to either Exit or try to exploit the Strong players.

Figure B4: Predictions for the Weak players in *Exit* treatments (case 2)



To sum up:

(i) The exit option does not affect the behavior of the Strong player;

(ii) The fraction of Strong players playing Generous in the *Risky* treatment is smaller than or equal to that in the *Safe* treatment;

(iii) Holding expectations and preferences constant across treatments, the fraction of Weak players playing Collaborate in the *Risky* treatment is smaller than in the *Safe* treatment. The fraction of Weak players playing Exploit or Exit should be larger in in the *Risky* treatment compared to the *Safe* treatment;

(iv) Prediction (iii) is reinforced if Weak players expect Strong players to play Defensive more frequently in the *Risky* treatment than in the *Safe* treatment.

C Instructions

Instructions[23]

Welcome to this study on economic decision-making. These instructions are a detailed description of the procedures we will follow. You earned €4.00 to show up on time. You can earn additional money during the study depending on the choices you and the other participants will make.

During the study you are not allowed to communicate with the other participants. We also ask you to switch off your mobile phone now. If you have a question at any time, please your hand and remain seated: someone will come to your desk to answer it.

As we proceed with the instructions, you will be asked to answer ten questions designed to verify your understanding of the instructions.

The study is divided into **two parts.** Your final earnings depend on the results of Part 1, and the results of Part 2. You will be paid privately and in cash at the end of the study.

Your color and your team

Together with these instructions, you received a **code.** Codes have been randomly distributed, and determine your color, which will be either **red**, or **blue.**

Your color defines which **team** you belong to: the RED or the BLUE team. Each team contains ten participants.

Your color and your team will remain the same throughout the whole study.

[23]Instructions for *Safe Exit* treatment. The instructions for the other treatments are available upon request from the authors.

- In **Part 1**, you will interact exclusively with participants of **your own team:** if you are red, you will only interact with other red participants, if you are blue you will only interact with other blue participants.

- In **Part 2**, you will interact exclusively with participants of **the other team:** if you are red, you will interact only with other blue participants, if you are blue you will only interact with other red participants.

We will now read instructions for Part1. Instructions for Part 2 will be distributed at the end of Part 1.

INSTRUCTIONS FOR PART 1

At the beginning of this Part, you will be asked to enter your code and you will learn your color and your team.

Once teams are formed, you will perform a team task. The task is to solve some math problems to reveal what is behind the big box you will see on your screen. You will be asked to add up three two-digit numbers. Every time a member of your team submits a correct answer, one more piece of what is behind the box will be revealed.

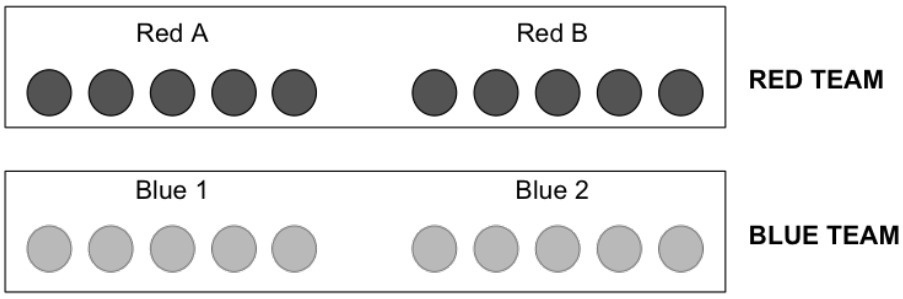
If you and your team members can uncover what is behind the box in **less than 150 seconds,** you will win **2 Euros each.** If you fail as a team, none of your team members will earn anything.

Before we start, we would like you to answer a few questions, to verify the full understanding of instructions.

INSTRUCTIONS FOR PART 2

Your set

In this Part, you will always interact only with participants of the other team. Each team is divided into two **sets** of 5 participants each, as illustrated in the following figure.



All participants in one set have the same color:

- if you are blue, all members in your set are blue;
- if you are red, all members in your set are red.

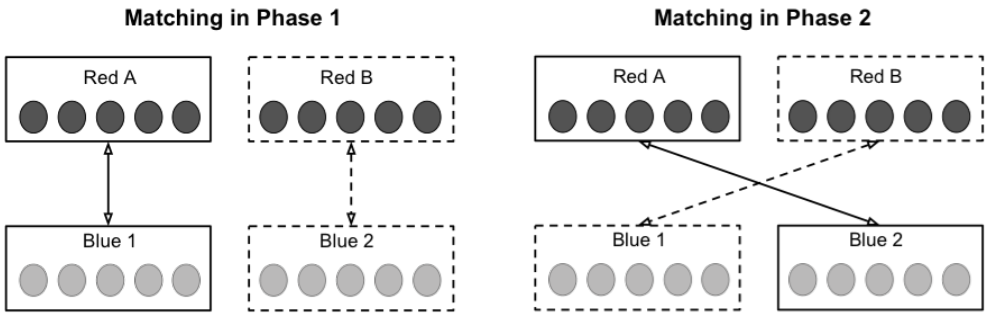
Your color and your set will remain the same, until the end of the study.

The Part is divided into two **Phases.** At the beginning of each Phase, your set will be matched with another set of the opposite color. If you are in a **blue** set, you will be matched with a **red** set, and vice versa:

- set Red A will play with set Blue 1 in Phase 1, and with set Blue 2 in Phase 2;
- set Red B will play with set Blue 2 in Phase 1, then with set Blue 1 in Phase 2.

In other words, **in each phase, your set will be matched with a different set.**

Each Phase includes 5 rounds. Hence, Part 2 lasts **10 rounds** in total.



Matching

In each round, you will be paired with a participant of the opposite color. We will call this person your **"counterpart".**

- If you are **blue**, you will be paired with a **red** participant of your matched set.
- If you are **red**, you will be paired with a **blue** participant of your matched set.

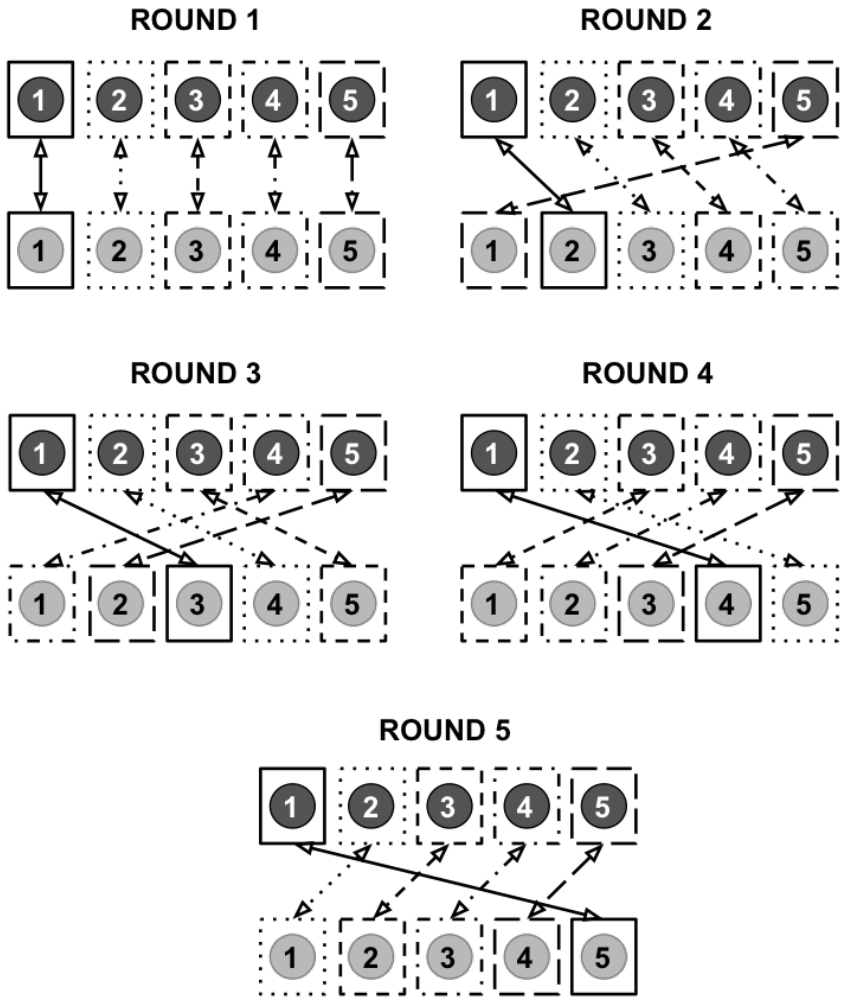
You will be paired with each and every participant in your matched set once and only once. **You can never be paired with the same participant twice, throughout the whole study.** The figures below illustrate an example of the pairing structure for the five rounds of each Phase.

In other words, in Part 2 you will be paired with each and every participant of the other team once and only once.

To see how your payoffs are determined in each round, please follow the next instructions.

The "Main Game"

In each round, you and your counterpart will play the "Main Game." Your payoff in each round depends on your choices and the choices of your counterpart.



If you are **red**, you must choose between **UP** and **DOWN**. If you are **blue**, you must choose between **LEFT** and **RIGHT**.

These choices determine your **payoff** and the payoff of your counterpart, as displayed in the following table:

In the table, the numbers in the bottom-left corner of each cell represent the payoff of the **red** person, and the numbers in the top-right corner represent the payoff of the **blue** person. All payoffs are expressed in €.

This payoff table is the same for all participants.

		Blue Player	
		Left	*Right*
Red Player	*Up*	(10,0)	(9,1)
	Down	(1,9)	(5,5)

To read the payoff corresponding to a specific pair of choices, you should

- find the row in the table that corresponds to the choice of the **red** person;
- move to the right to find the cell where this row crosses the column corresponding to the choice of the **blue** person.

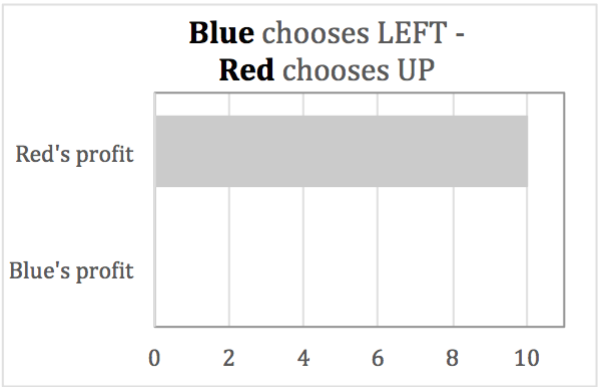
Blue moves first, and cannot condition his choice on the choice made by the **red** counterpart. **Red** moves after **blue**, and can condition his choice on the choice made by his **blue** counterpart.

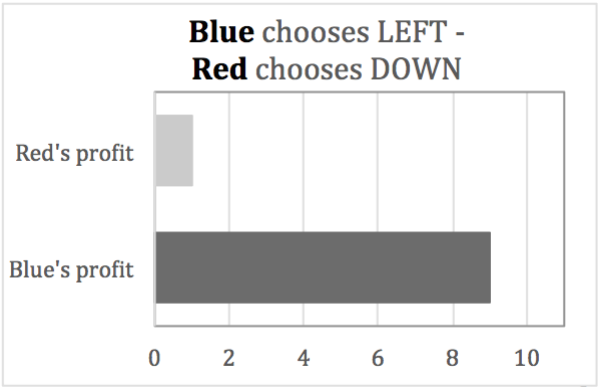
Consider the case in which blue chooses LEFT.

Red can choose between UP and DOWN.

- **red** chooses UP
  - **red** earns €10;
  - **blue** earns €0.
- If **red** chooses DOWN
  - **red** earns €1;
  - **blue** earns €9.

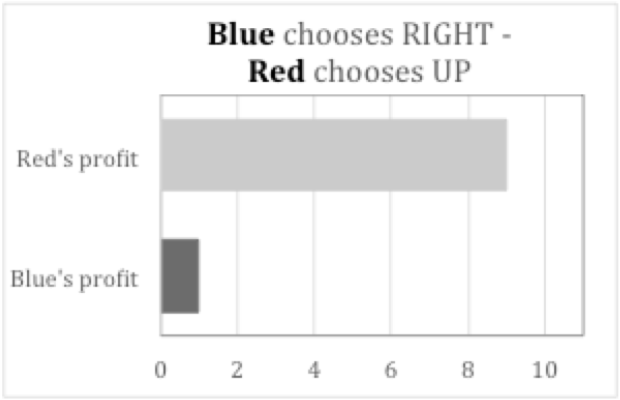
Consider now the case in which blue chooses RIGHT.

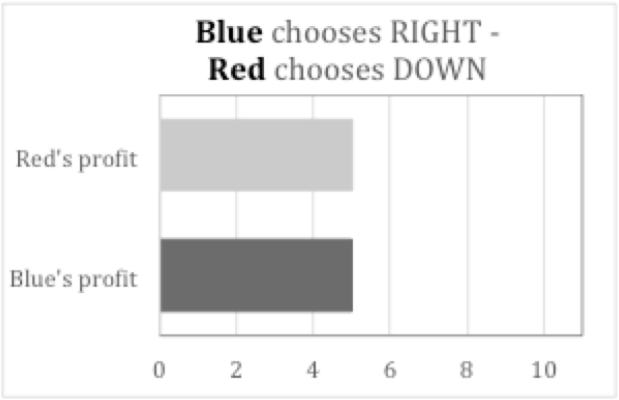




Red can choose between UP and DOWN.

- If **red** chooses UP
  - **red** earns €9;
  - **blue** earns €1.
- **red** chooses DOWN
  - **red** earns €5;
  - **blue** earns €5.





In practice, **blue** will have to answer **one question:**

- Which option do you choose: LEFT or RIGHT?

Red, instead, will have to answer **two questions:**

1. Which option do you choose if your blue counterpart selects RIGHT: UP or DOWN?

2. Which option do you choose if your blue counterpart selects LEFT: UP or DOWN?

Only one of the two choices made by red will be implemented. If **blue** selects RIGHT, the payoffs will be determined by **red**'s answer to the first question. If **blue** selects LEFT, the payoffs will be determined by **red**'s answer to the second question. **Red** will be informed about the relevant decision only after making both choices. It is therefore important for **red** to pay attention to both choices, as he does not know in advance which one will be relevant.

The "Exit" option

In each round, you will need to take another decision, before making your choice in the "Main Game." You will decide whether you want to EXIT this game, or STAY.

If you select EXIT, the Main Game will not be played. Regardless of the choices made by your counterpart, both of you will earn 0 in this round: If you choose EXIT, you do not have to make any choice in the Main Game

If you select STAY, the payoffs in this round will depend on the decision made by your counterpart.

- If your counterpart selects EXIT, the game will not be played. Regardless of the choices you made, both of you will earn 0 in this round.
- If your counterpart selects STAY, the payoffs will be determined by the choices you and your counterpart made in the Main Game.

You will be informed about the choice – to EXIT or STAY – of your counterpart only after taking your decision in the Main Game. If your counterpart chooses EXIT, your decision will not be relevant. **Remember that you will make this choice for each round separately.** In each round, both participants in the pair will have the chance to decide whether they would like to EXIT or STAY, before playing the Main Game, and hence **before knowing the choice made by their counterpart.**

Feedback information

After each round, you will receive information on whether your counterpart selected EXIT or STAY. In case both you and your counterpart chose STAY, you will be informed on the choice made by your counterpart in the Main Game. If you or your counterpart (or both) chose EXIT, you will not receive any information about the chosen option. You will also see

your payoff and the payoff of your counterpart.

After each Phase, that is after round 5 and after round 10, you will also receive information on

- the average payoff of the members of your set over all rounds of the Phase;
- the average payoff of the members of your matched set over all rounds of the Phase;
- how frequently the participants in your set selected EXIT in all rounds of the Phase;
- how frequently the participants in your matched set selected EXIT in all rounds of the Phase.

Remember that in Phase 2 you can never be paired with any member of the set you were matched with in Phase 1.

Your earnings in Part 2

At the end of Part 2, one round from each Phase will be selected, and your payoff in those two rounds will be paid to you.

Hence, your earnings in Part 2 depend on your choices and the choices of your counterpart in one randomly selected round of Phase 1 (rounds 1-5), and in one randomly selected round of Phase 2 (rounds 6-10).

———————— *new set of instructions* ————————

Instructions for belief elicitation sessions[24]

Welcome to this study on economic decision-making. These instructions are a detailed description of the procedures we will follow. You earned €4.00 to show up on time. You can earn additional money during the study depending on the choices you make.

During the study you are not allowed to communicate with the other participants. We also ask you to switch off your mobile phone now. If you have a question at any time, please raise your hand and remain seated: I will come to your desk to answer it.

As we proceed with the instructions, you will be asked to answer ten questions designed to verify your understanding of the instructions. You will receive 20 cents for each question you answer correctly at the first trial.

You will be paid privately and in cash at the end of the study.

In this experiment, you are asked to provide an estimate about decisions made by other people who took part in a previous study. This study was conducted in Cologne, at this laboratory.

Below we report the instructions we used in this previous study. We ask you to read them on your own.

It is important that you carefully follow these instructions and fully understand the original instructions. To verify your full understanding, we ask you to answer the same quiz we administered to the participants who took part in the previous study. You will receive 20 cents for

[24]Instructions for belief elicitation for the *Safe Exit* treatment. Instructions for the *Simultaneous Exit* treatment are available upon request from the authors.

each question you answer correctly at the first trial.

When everyone has completed this quiz, we will proceed and explain your task in today's study, and how your earnings are computed.

———————— *instructions for the original experiment here*[25] ————————

Your task.

You will be asked to guess the choices made by the participants in the first round of the previous study.

At the beginning of today's study, the computer will randomly draw the choices made in the first round by 20 of the subjects who took part in the previous study. Of these 20 participants, 10 were assigned the role of blue players, while the other 10 were assigned the role of red players. None of them chose to exit.

You need to answer two questions:

1. How many of the 10 blue players chose RIGHT in the first round?
2. How many of the 10 red players chose UP in the first round if their counterpart selected RIGHT?

For both questions, your answer should be an integer number between 0 and 10.

Your earnings.

[25] We reported the complete set of the original instructions, including both the team task (part 1) and the main game (part 2).

Your earnings can vary between 0 and 13 euro per question. The closer you get to the correct answer, the higher your earnings. Please see Table 1. You earn 13 euros if your guess coincides with the right answer, or if it departs from it by at most one unit (from above or below). If instead your guess departs from the correct answer by 2 units, you earn 11; if it departs from the correct answer by 3 units, you earn 8.5, and so forth and so on. If your guess departs from the correct answer by 6 or more units you earn nothing.

Table 1: Earnings table

Distance from the correct answer	**Earnings**
0 or 1	13
2	11
3	8.5
4	5
5	0.5
6 or more	0

You will be paid for one of the two guesses selected at random by the computer. You will know which guess will be relevant for your payment only at the end of the experiment. It is hence in your interest to pay attention to both decisions.

Please raise your hand if you have any questions and I will come to your desk to answer them.