Barr, Abigail; Hochleitner, Anna; Sonderegger, Silvia

Working Paper

Does increasing inequality threaten social stability? Evidence from the lab

CeDEx Discussion Paper Series, No. 2024-01

Provided in Cooperation with:
The University of Nottingham, Centre for Decision Research and Experimental Economics (CeDEx)

Suggested Citation: Barr, Abigail; Hochleitner, Anna; Sonderegger, Silvia (2024) : Does increasing inequality threaten social stability? Evidence from the lab, CeDEx Discussion Paper Series, No. 2024-01, The University of Nottingham, Centre for Decision Research and Experimental Economics (CeDEx), Nottingham

This Version is available at:
https://hdl.handle.net/10419/300402

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

University of Nottingham
UK | CHINA | MALAYSIA

Discussion Paper No. 2024-01

Abigail Barr,
Anna Hochleitner and
Silvia Sonderegger

March 2024

Does increasing inequality threaten social stability? Evidence from the lab

CeDEx Discussion Paper Series
ISSN 1749 - 3293

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The Centre for Decision Research and Experimental Economics was founded in 2000, and is based in the School of Economics at the University of Nottingham.

The focus for the Centre is research into individual and strategic decision-making using a combination of theoretical and experimental methods. On the theory side, members of the Centre investigate individual choice under uncertainty, cooperative and non-cooperative game theory, as well as theories of psychology, bounded rationality and evolutionary game theory. Members of the Centre have applied experimental methods in the fields of public economics, individual choice under risk and uncertainty, strategic interaction, and the performance of auctions, markets and other economic institutions. Much of the Centre's research involves collaborative projects with researchers from other departments in the UK and overseas.

Please visit http://www.nottingham.ac.uk/cedex for more information about the Centre or contact

Samantha Stapleford-Allen
Centre for Decision Research and Experimental Economics
School of Economics
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: +44 (0)115 74 86214
Samantha.Stapleford-Allen@nottingham.ac.uk

The full list of CeDEx Discussion Papers is available at

http://www.nottingham.ac.uk/cedex/publications/discussion-papers/index.aspx

Does increasing inequality threaten social stability? Evidence from the lab

Abigail Barr,[†] Anna Hochleitner,[‡] Silvia Sonderegger[§]

March 4, 2024

Abstract

We study the relationship between inequality and social instability. While the argument that inequality can be damaging for the cohesion of a society is well established, the empirical evidence is mixed. We use a novel approach to isolate the causal relationship running from inequality to social instability. We run a laboratory experiment in which two groups interact repeatedly and have an incentive to coordinate even though coordination comes at the cost of inter-group inequality. Then, we vary the extent of the inequality implied by coordination. Our results show that increasing inequality has a destabilising effect; the disadvantaged initiate the destabilisation; and a worsening of the absolute situation of the disadvantaged exacerbates the destabilising effect of increasing inequality. These findings are in line with a simple model incorporating inequality aversion and myopic best response. Finally, we show that history matters. People respond differently to the same level of current inequality depending on their past experiences. More specifically, a history of stability facilitates the re-emergence of coordination in more unequal environments, and a sudden increase in inequality is more destabilising than a gradual increase.

Keywords: Collective decision making, Conflict and Revolutions, Inequality

JEL Codes: C92, D01, D63, D74

We thankfully acknowledge funding by the ESRC (Research Training Support Grant) and CeDEx. We are grateful to Joël Berger, Sonja Vogt, and Charles Efferson for exchanges during the design stage of the study. We are particularly thankful to Malte Baader, Markus Eberhardt, Simon Gächter, and Joël van der Weele for their input and comments on earlier versions of this paper. Finally, we thank participants at the CCC conference, and in seminars at the University of Nottingham, NHH Bergen, and the University of Zurich for comments and feedback. The study received ethical approval from the Nottingham School of Economics Research Ethics Committee on 05/04/2019.

[†]University of Nottingham, U.K. Email: abigail.barr@nottingham.ac.uk.

[‡]NHH Bergen (SNF and FAIR), Norway. Email: anna.hochleitner@snf.no.

[§]University of Nottingham, U.K. Email: silvia.sonderegger@nottingham.ac.uk.

1 Introduction

Does increasing inequality threaten social stability? Examining this question has a long tradition in political philosophy. In his final composition, *The Laws*, Plato warned that inequality heightens the danger of civil disintegration and even war and, because of this, he advised that both "*extreme poverty and wealth must not be allowed to arise in any section of the citizen-body*" (Plato, cited in Cooper et al., 1997, 744d). Throughout history this argument has been taken up and explored from a diverse range of perspectives, for example, in the works of Machiavelli, Montesquieu, and Marx (for a discussion see Lichbach, 1989). At the core of these arguments is the notion of a society divided into the disadvantaged, who seek to challenge the existing status quo, and the advantaged, who seek to defend it. And, to this day, inequality continues to be put forward as an explanation for manifestations of sociopolitical instability such as the Arab Spring (Roubini, 2011), the widespread rise of populism (Norris & Inglehart, 2019), and large scale protests like the international "occupy" movement in 2011 and the "yellow vest" protests in 2018 in France, in which millions raged against a system that appeared to be leaving them behind (Satz, 2019).[1] This notion is also reflected in public opinion, with a representative survey of US citizens finding that 74% believe inequality increases crime and 67% believe that it decreases societal trust (Lobeck & Støstad, 2023).

While the link between inequality and instability seems unequivocal, there exist many historical counter-examples in which significant inequalities did not lead to unrest (Cramer, 2005). One possible explanation for this is that societies with high inequality also tend to have powerful elites, who use their powers to oppress resistance and, thereby, maintain the stability of the prevailing system (Lichbach, 1989). Another explanation is that the disadvantaged turning against the established order requires an enormous degree of coordination to overcome the inherent collective action problems (Olson, 1965; Collier, 1999; Blattman & Miguel, 2010; Edmond, 2013; Cantoni et al., 2019).

It is thus not surprising that, despite a large empirical literature dedicated to the topic, evidence for the "inequality causes instability" hypothesis is very mixed. *"The empirical problem is in fact extreme"* (Cramer, 2005, p.11) with the econometric analyses often being marred by data quality issues and problems relating to isolating the effect of inequality from those of other aspects of the sociopolitical environment.

In this paper we obviate these problems by taking a novel and distinct approach; we test the "inequality causes instability" hypothesis using a specially designed, incentivised lab experiment. The lab inevitably constitutes an artificial environment, but the control that it affords allows us to identify the causal relationship that we are interested in by excluding variation in all other aspects of the sociopolitical environment.[2] To the best of our knowledge, we are the first to investigate this causal relationship experimentally.[3]

Our research design builds on previous experimental work that explores the emergence of in-

[1] All examples are, of course, highly complex and had many contributing factors (Grossman, 2019). However, it is noteworthy that inequality keeps reappearing as an ex-post explanation for such events.

[2] See Smith (1982) for a discussion about experiments as microeconomic systems and the role of control and measurement.

[3] A recent experimental study by Radkani et al. (2023) revealed that inequality and deprivation increase stealing and reduce cooperation. However, their experimental design does not embody the notion of a society divided into the disadvantaged, who seek to challenge the existing status quo, and the advantaged, who seek to defend it.

equitable conventions (see e.g. Dale et al., 2002; Hargreaves-Heap & Varoufakis, 2002; Oprea et al., 2011; Benndorf et al., 2016; Berger et al., 2022). It involves a repeated battle-of-the-sexes (BoS) game that is played between members of two groups across 100 periods. In any given period, total gains are maximised if the groups successfully coordinate on choosing distinct actions, reflecting the idea that society benefits from its members specialising in different skills and then engaging in exchange. However, while such specialisation is beneficial to society as a whole, it can give one group an advantage over the other if the actions are not remunerated equally (see e.g. Henrich & Boyd, 2008). Then, over time, external shocks, such as the recent advancements in AI or other technological changes, may shift the relative returns to specialising in specific actions and, in turn, this may widen or close the gap between groups (see e.g. Acemoglu, 2002; Gmyrek et al., 2023).

In our experiment, successful coordination involves an unequal payoff, with one of the possible actions delivering a higher payoff than the other. This captures the inherent tension between society-level prosperity and cross-group equality. Over the first 50 periods we hold the payoffs from each action constant and, in line with the studies mentioned above, observe the endogenous emergence of unequal conventions between groups. Each group specialises on one action with the result that one group becomes relatively advantaged while the other becomes relatively disadvantaged. Then, across the remaining 50 periods, we exogenously vary the payoff difference between the actions. It is this novel departure from the to previous literature that allows us to test the "inequality causes instability" hypothesis.

Our hypotheses are informed by a simple model incorporating disadvantageous inequality aversion and myopic best response. In line with the theoretical predictions, we find that increasing inequality destabilises the existing status quo, and that the destabilising process is initiated by members of the disadvantaged group. By exogenously varying the dynamics of inequality, we find that the negative effect of inequality on stability is exacerbated when i) the increase in inequality is sudden rather than gradual, ii) the situation of the disadvantaged group deteriorates not only in relative but also in absolute terms, and iii) groups have experienced more instability in the past.

Our findings contribute to research on the inequality-instability nexus, where we provide new evidence in support of the "inequality causes instability" hypothesis using a novel methodology. As discussed above, previous evidence has been mixed. Some studies find a positive relationship between income inequality and measures of sociopolitical instability (see e.g. Hsieh & Pugh, 1993; Alesina & Perotti, 1996; Fajnzylber et al., 1998; Kennedy et al., 1998; Stewart, 2000; Østby, 2008), while others conclude that inequality is neither necessary nor sufficient for social conflict and stress the importance of other factors such as absolute deprivation (Lichbach, 1989; Fearon & Laitin, 2003; Dube & Vargas, 2013; Somanthan, 2020).[4] Our findings indicate that, under certain conditions at least, increasing inequality is sufficient to cause social instability.

As mentioned, in empirical work based on observational data relating to specific historical examples, it has often proven difficult to isolate the effect of inequality on instability from the effects of other aspects of the sociopolitical environment. One response to this problem has been to look for

[4]In recent work, Acemoglu et al. (2020) suggest that the relationship between inequality and stability may be non-monotonic. Moreover, they stress the importance of population pressures for instability.

patterns across longer periods of time (see e.g. Scheidel, 2017; Hoyer et al., 2022; Turchin, 2023). From this perspective, the long-term dynamics of social instability appear cyclical in nature. Extended periods of stability are interspersed with waves of sociopolitical instability. Attacks on the existing order have thereby been linked to periods of growing economic inequality and absolute poverty at the lower end of the distribution and power struggles at the upper end of the distribution (Hoyer et al., 2022; Turchin, 2023). In its abstraction, our experiment is similar to this long-run view. However, by stripping away all confounding sociopolitical factors, we show that increasing inequality has a direct, negative effect on coordination between variably advantaged groups.[5]

Sticking with the long-run perspective, our paper also has links with the evolutionary game theory literature that studies the emergence of unequal conventions and inter-group inequalities. Young (1993) describes a convention as a "*pattern of behaviour that is customary, expected and self-enforcing*" (p.57). Several theorists have modelled conventions as the outcomes of a coordination process (Lewis, 1967; Young, 1996; Ellison, 2000; Baronchelli, 2018; Newton, 2021) that can lead to persistent inequalities between groups (Axtell et al., 2001; Binmore et al., 2003; Henrich & Boyd, 2008; Bowles et al., 2014). Inequality can be (close to) inevitable - a natural consequence of specialising on different tasks even in the absence of underlying differences between individuals (Mookherjee & Ray, 2002). The experimental literature mentioned above (see Holm, 2000; Dale et al., 2002; Hargreaves-Heap & Varoufakis, 2002; Oprea et al., 2011; Benndorf et al., 2016; Berger et al., 2022) confirm that efficient, but unequal equilibria do emerge. In the absence of information about individual behaviour, people tend to focus on expectations about group behaviour (Dale et al., 2002). Group affiliation (such as gender in Holm (2000)) can then serve as a device to avoid miscoordination but at the same time give rise to discriminatory practices and inequalities (Hargreaves-Heap & Varoufakis, 2002).

An interesting question that arises from this theoretical and experimental work is whether and how such unequal equilibria can be disturbed. Despite their often arbitrary origin, from a theoretical perspective unequal conventions can perpetuate over long periods even if they are inefficient (Hwang et al., 2024; Belloc & Bowles, 2013). The empirical evidence corroborates this prediction. Even when a convention is undesirable for most people, behavioural change can be very difficult to initiate (Andreoni et al., 2021) and the inertia can carry over from one generation to the next (Schotter & Sopher, 2003). Change is possible, however, and can occur as a result of external shocks, errors, or conscious deviations (Belloc & Bowles, 2013; Acemoglu & Jackson, 2014; Hwang et al., 2024; Baronchelli, 2018). Here, the theory indicates that, if a change is to occur, it takes the form of a social tipping point, meaning that, once a crucial threshold is reached, change is sudden rather than incremental (Young, 2015). Previous experimental studies provide evidence in support of both the notion of social tipping points (Andreoni et al., 2021; Centola et al., 2018) and the idea that external changes can disturb established equilibria (Brandts & Cooper, 2006). We are also interested in whether and when conventions can be overturned. However, there is an important difference between the studies cited above and ours; the studies above focus on situations of common interest,

[5]Within the experiment, we consciously decided not to give the advantaged the option to redistribute income. Thus, our results focus on the consequences of increasing inequality when no action is (or can be) taken to prevent it and our experiment can be viewed as a baseline upon which to build.

while we intentionally introduce an element of conflict.[6]

The remainder of this paper is structured as follows. Section 2 outlines our theoretical framework. Section 3 describes the experimental design and hypotheses. Section 4 presents details of the data collection. Section 5 sets out our experimental results, while Section 6 provides further descriptive analysis and discussion. Finally, Section 7 concludes.

2 Theoretical motivation

2.1 The game

To explore the effect of increasing inequality on social stability, we develop a theoretical framework focusing on repeated interactions between two groups $g \in \{Y, G\}$ within a society. In each period t, two randomly chosen individuals from different groups interact with one another, each having to choose between two possible actions $a \in \{A, B\}$. If they coordinate on choosing different actions, A results in a higher payoff h, while B results in a lower payoff l. If they choose the same action, both individuals receive zero (see Figure 1). At the end of each period, the individuals learn their own individual earnings and the average earnings of their fellow group members conditional on their choices. The game is thus a variant of the battle-of-the-sexes game (Luce & Raiffa, 1957). This game constitutes an excellent environment in which to study the relationship between increasing inequality and social stability as it incorporates both strong incentives for individuals to coordinate and an inherent element of conflict.

Figure 1 presents the stage game, which has two pure asymmetric Nash equilibria (NE) (A, B) and (B, A), as well as one symmetric mixed Nash equilibrium. While in a repeated game there exist many possible strategy profiles, we are particularly interested in strategy profiles in which members of different groups specialise in distinct actions and play the same pure NE in each period. Using group affiliation as a salient marker allows individuals to avoid miscoordination but comes at the cost

Figure 1: The BoS stage game

		g=Y: A	g=Y: B
g=G	A	0, 0	h, l
	B	l, h	0, 0

Note: $h > l \geqslant 0$.

[6]Our paper is also related to the literature on inequality within public good games (see e.g. Cherry et al., 2005; Gächter et al., 2017). However, a crucial distinction is that this literature focuses on the effect of inequality on cooperation within groups while we are exploring the stability of coordination across groups.

of inter-group inequalities.[7] Below we show how individuals choose their actions in each period and how such a strategy profile can emerge endogenously over time.[8]

2.2 Comparative statics under inequality aversion and myopic best response

Here, we lay out a simple theoretical framework to guide our empirical analysis. Motivated by the large literature on the prevalence of inequality aversion and given our research questions, we consider a model where individuals are inequality averse. This does not imply that the predictions we derive (and which our experiment was designed to test) are unique to this framework. There are other models that may deliver similar predictions. However, for our purposes this is irrelevant given that our objective is not to test a specific theory but to identify specific causal relationships. (Readers that are less interested in theoretical underpinnings can skip directly to the experimental design in Section 3).

When making the choice between actions A and B in a given period, individuals choose the action that maximises their expected utility. Thus, they choose A if $E(u_{A,t}) \geqslant E(u_{B,t})$ and B otherwise. We assume that the utility from choosing an action depends on the expected payoff from that action ($\pi_t \in \{h_t, l_t, 0\}$), as well as an individual's level of disadvantageous inequality aversion θ_i, which is drawn for each individual from the distribution proposed by Fehr & Schmidt (1999), such that $0 \leqslant \theta_i \leqslant 4$.[9]

Payoffs depend on both own and other's choices. For an individual i in group g the expected payoff from A depends on the shares of players in the other group $g' \neq g$ that choose A ($\lambda_t^{g'}$) and B ($1-\lambda_t^{g'}$) in the current period. We assume that individuals are myopic in the sense that they expect the same share of players to choose A in the current period as did so in the last period $E(\lambda_t^{g'}) = \lambda_{t-1}^{g'}$. Since, in the very first period, individuals cannot turn to past experience, we assume that, in that period, each individual's belief about $\lambda_t^{g'}$ is a draw from a uniform distribution, $E(\lambda_1^{g'}) \sim U(0,1)$.

Definition: Myopic best response *At time t, individual i of group g selects the action that maximises their expected utility conditional on $E(\lambda_t^{g'}) = \lambda_{t-1}^{g'}$, i.e. on the expectation that the share of individuals in group $g' \neq g$ selecting action A in that period equals the share selecting action A in the previous period.*

Using this definition, we can formulate the individual decision rule comparing $E(u_{A,t})$ and $E(u_{B,t})$. If individuals are inequality averse, their utility from choosing B depends negatively on both the strength of their inequality aversion (θ_i) and the size of the inequality ($\Delta_t = h_t - l_t$). An individual of group g then chooses A iff

[7] Using group affiliation as a coordination device also requires the least amount of cognitive effort. While a strategy profile where groups alternate between action A and B is theoretically possible and avoids inequalities between groups, it is very difficult to establish such a rule, especially when, as in our experiment, individuals are randomly interacting with new draws from the other group in each period and communication is not possible.

[8] Our set-up also has parallels to the continental divide game developed by Van Huyck et al. (1997), who show the importance of initial conditions and best responses in dynamic games with multiple equilibria.

[9] To simplify the model we abstract from advantageous inequality aversion. If, in line with Fehr & Schmidt (1999), individuals are more averse to disadvantageous inequality (α_i) than they are to advantageous inequality (β_i), θ_i can be interpreted as the net difference between α_i and β_i.

$$\underbrace{(1 - E(\lambda_t^{g'}))h_t}_{E(u_{A,t})} \geqslant \underbrace{E(\lambda_t^{g'})(l_t - \theta_i \Delta_t)}_{E(u_{B,t})} \tag{1}$$

or

$$\theta_i \geqslant \underbrace{\frac{E(\lambda_t^{g'})(l_t + h_t) - h_t}{\Delta_t E(\lambda_t^{g'})}}_{\equiv \theta^*_{g,t}}. \tag{2}$$

Rearranging Equation (1), we derive a group-specific threshold $\theta^*_{g,t}$ that an individual's inequality aversion needs to exceed in order for A to be chosen (see Equation (2)). From this, we can derive a number of comparative statics. Holding everything else constant, $\theta^*_{g,t}$ increases with $E(\lambda_t^{g'})$ and l_t, implying a lower probability of i choosing A. This captures the notion that with a higher payoff of B or more agents in the other group choosing A, action A becomes less attractive. By contrast, $\theta^*_{g,t}$ decreases with the extent of the inequality Δ_t and h_t. So, action A becomes more attractive the higher its payoff is both in absolute terms (h_t) and compared to the payoff for action B (Δ_t). The comparative statics are summarised in Proposition 1. The proof follows directly from Equation 2.

Proposition 1: *Everything else being equal, the probability of an individual of group g choosing A* ($\theta_i \geqslant \theta^*_{g,t}$) *decreases with the expected share of A choices in the other group,* $E(\lambda_t^{g'})$*, as well as the payoff for B* (l_t)*, but increases with the extent of inequality* Δ_t *and the payoff for A* (h_t)*.*

2.3 Emergence of a convention

As mentioned above, we are particularly interested in equilibria where each group specialises in a different action. We refer to this type of equilibrium as a *convention*. Young (1993) defines conventions as "*customary, expected and self-enforcing*" (p.57), which translates perfectly to a repeated pure NE. If individuals in group g' know that individuals in group g tend to choose action A, they also expect them to do so in the next period, making B a likely best response. By choosing B, g' individuals then ensure that the convention is further strengthened, since this incentivises g people to choose A.

More formally, a convention can be said to exist if one group specialises in A and the other group in B, i.e. $\lambda_t^g \to 1$ and $\lambda_t^{g'} \to 0$. We refer to the group specialising in the high paying action, A, as the advantaged group and to the group specialising in the low paying action, B, as the disadvantaged group. For a convention to be said to exist, the share of people choosing A in the advantaged group must surpass some defined threshold x, while in the disadvantaged group the share choosing A must lie below some defined threshold y. The weakest form of inter-group specialisation would involve the majority of members in the advantaged group choosing A ($0.5 < x \leqslant 1$), while the majority of members in the disadvantaged group choose B ($0 \leqslant y < 0.5$). Finally, as conventions are self-enforcing, we require participants to expect that the convention will still be followed in the next period.

Definition: (x,y)-convention with g dominance *We say that a (x,y)-convention with g dominance holds at t when (i)* $\lambda_t^g \geqslant x > 0.5$*, (ii)* $\lambda_t^{g'} \leqslant y < 0.5$*, (iii)* $E(\lambda_t^g) \geqslant x$*, and (iv)* $E(\lambda_t^{g'}) \leqslant y$*.*

Note that, holding $E(\lambda_t^{g'})$, l_t, h_t, and Δ_t constant, it is more likely that an individual chooses B the lower their level of disadvantageous inequality aversion, θ_i. For this reason, the group with the higher average level of disadvantageous inequality aversion is more likely to become the advantaged one.

Taking an established convention as the status quo, we can then look at how different factors affect that convention's stability. As a measure of stability we construct a social stability index (SSI_t) that compares the share of individuals choosing action A across groups in a given period (see Bendorf et al., 2016, for a similar approach):

$$SSI_t = |\lambda_t^g - \lambda_t^{g'}|, \text{ with } 0 \leqslant SSI_{g,t} \leqslant 1 \tag{3}$$

If everyone chooses the same action, $SSI_t = 0$, indicating complete chaos. By contrast, $SSI_t = 1$ describes a situation of perfect compliance with the convention, with all the members of one group choosing A and all the members of the other group choosing B.

2.4 The effect of inequality on stability

Assume that a (x,y)-convention with g dominance exists such that both λ_t^g and $E(\lambda_t^g)$ are $\geqslant x > 0.5$ and both $\lambda_t^{g'}$ and $E(\lambda_t^{g'})$ are $\leqslant y < 0.5$. What happens to the stability of this convention as the payoff inequality from choosing different actions (Δ_t) increases? From the comparative statics, we see that an increase in inequality translates into a higher overall share of individuals choosing A. So, starting from an existing convention, as inequality increases, more and more members of the disadvantaged group start preferring A over B and the convention attenuates. To see that deviations must always be initialised by the disadvantaged group, note that $\theta^*_{g,t}$ is group-specific due to its dependence on expectations. With most members of g' choosing B, a member of the advantaged group g would only deviate from the convention and choose B iff

$$\theta_i \leqslant \theta^*_{g,t} = \underbrace{\frac{E(\lambda_t^{g'})(l_t + h_t) - h_t}{E(\lambda_t^{g'})\Delta_t}}_{< 0 \text{ as } E(\lambda_t^{g'}) \leqslant y < 0.5} . \tag{4}$$

As Equation (4) shows, $\theta^*_{g,t}$ is negative for members of the advantaged group independent of the degree of inequality Δ_t, due to $E(\lambda_t^{g'}) \leqslant y < 0.5$. So, as by definition $\theta_i \geqslant 0$, the condition for choosing B will never be met for members of the advantaged group.[10] Intuitively, for a member of the advantaged group, with more than 50% of individuals in g' choosing B ($E(\lambda_t^{g'}) \leqslant y < 0.5$), a deviation would mean giving up getting h with a high probability in favour of getting zero with an even higher probability.

Let us now turn to members of the disadvantaged group g'. From Equation (2), we see that as Δ_t increases, the threshold for choosing A ($\theta^*_{g,t}$), i.e., for deviating from the convention, declines. This means that, eventually, the condition for choosing A will be met for individuals with sufficiently

[10] If we allow for advantageous inequality aversion and depart from Fehr & Schmidt (1999) by assuming that the latter is larger than disadvantageous inequality aversion, deviations could be initiated by the advantaged group.

high levels of inequality aversion θ_i.[11] By deviating to A, members of the disadvantaged group are knowingly taking on a high risk of receiving zero, ($E(\lambda_t^g) > 0.5$), because inequality has passed the threshold that they are willing to accept. From Equation (3) we see that, as the share of individuals choosing A in g' increases, social stability SSI_t declines and, if inequality continues to increase, a growing number of individuals in g' will choose A, further destabilising the existing convention.

In addition, the increase in $\lambda_t^{g'}$ can, in turn, induce reactions from the advantaged group. The first individuals in g who will react are the ones with the lowest levels of disadvantageous inequality aversion, θ_i, as deviating implies forgoing h_t for l_t. As the lowest possible level of inequality aversion in our model is $\theta_i = 0$, the minimum proportion of members of g' required to induce a deviation by a member of the advantaged group is given by

$$0 \geqslant \frac{E(\lambda_t^{g'})(l_t + h_t) - h_t}{E(\lambda_t^{g'})\Delta_t} \longrightarrow E(\lambda_t^{g'}) \geqslant \frac{h_t}{l_t + h_t}. \tag{5}$$

As by definition $\Delta_t > 0$, it follows that $\frac{h_t}{l_t+h_t} > 0.5$. In other words, as long as the expected share of individuals in g' who choose A is not above 50%, there will be no reaction from the advantaged group. However, once the share of individuals in g' choosing A surpasses the threshold defined in Equation (5), individuals in the advantaged group will start deviating to B, further destabilising the convention and accelerating additional deviations by the disadvantaged group.

Proposition 2: Inequality causes instability *Increases in inequality destabilise an existing convention and lower* SSI_t.

Proposition 3: Initiators *Deviations are always initiated by the disadvantaged group. This, in turn, can induce reactions from the advantaged group once* $E(\lambda_t^{g'}) \geqslant \frac{h_t}{l_t+h_t} > 0.5$.

2.5 Abstracting from inequality aversion

As mentioned above, our results are not necessarily unique to an environment where individuals are inequality averse. For instance, a setup that focuses on quantal response equilibria (McKelvey & Palfrey, 1995) may deliver similar predictions to those derived above, even in the absence of inequality aversion. The underlying idea is that players make errors when choosing which pure strategy to play. Fixing the distribution of action choices by individuals of the other group, the probability that a given agent selects action A depends positively on the payoff from A and negatively on the payoff from B. This implies that an increase in inequality may induce disadvantaged players to choose A more often.

[11] Note that for individuals without inequality aversion ($\theta_i = 0$), it is never optimal to deviate from an established convention as long as there is a positive payoff from choosing B.

3 Experimental design

3.1 Basic set-up

At the beginning of the experiment, each individual is assigned to one of two groups (group size N=7) that interact repeatedly over 100 periods. We use the minimal group paradigm (Billig & Tajfel, 1973), so group affiliation has no deeper meaning and is determined arbitrarily, in our case, via the random draw of a coloured ball (green or yellow). Group affiliation (green/yellow) is fixed for the duration of the experiment.

In each period, a member of the green group interacts with a randomly selected member of the yellow group and plays a battle-of-the-sexes (BoS) game. In the game, each subject chooses either action A or B as described above (see Figure 1). At the end of a period, each subject receives feedback on their individual outcome, the average outcome for members of their own group who chose A and the average outcome for members of their own group who chose B. This feedback structure captures the idea that individuals can observe not only their own personal experience but also the experiences of socially proximate others.[12] Note that this set-up implies that individuals do not directly observe λ_t for the other group. However, they could infer it from the average payoffs from choosing actions A and B within their own group. Arguably, the feedback structure we apply, simplifies decision-making for participants, as it allows them to directly assess which of the two actions is the more profitable for members of their group at any given moment.

In the first 50 periods, we hold payoffs from actions A and B constant to allow a convention to emerge. After $t = 50$, we introduce exogenous payoff changes in order to explore the effects of inequality and different inequality dynamics on the stability of the convention. At the start of the experiment, participants are made aware that payoffs may change at any time during the experiment, but are not told how and when they will change.

3.2 Treatments

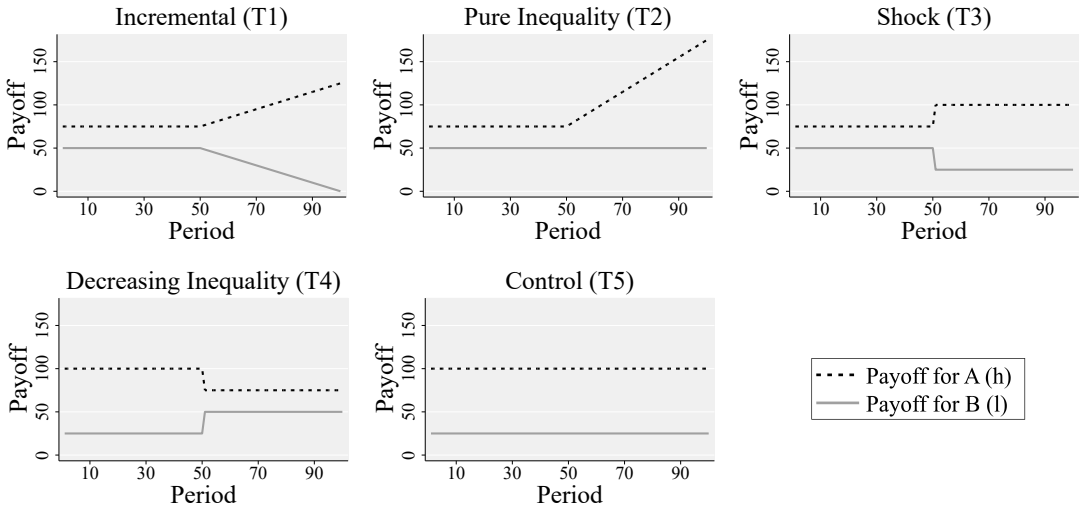
We ran five treatments that varied both in terms of initial payoff inequality for $t \leqslant 50$ and the dynamics of the inequality after $t = 50$. Figure 2 presents the payoffs from A and B under each of the treatments in each period.

The three treatments in the upper row (T1, T2 and T3) all start with the same relatively low level of inequality in the first 50 periods ($\Delta_t = h_t - l_t = 25$),[13] and then involve an increase in inequality in the second 50 periods. These treatments allow us to address our primary research question about whether increasing inequality affects social stability.

[12] While feedback structure is often modelled in evolutionary game theory as a random sample of decisions and outcomes relating to x other players (Young, 1996, 1993), it is reasonable to assume that the sampling process is non-random in the presence of separation into groups. Previous research shows that there is a higher probability of learning from in-group members and that segregated information networks are widespread (McPherson et al., 2001; Henrich & Boyd, 2008; DiPrete et al., 2011) and some have argued that this combination of individual learning and social environment is crucial for the perpetuation of inter-group inequalities (Bowles et al., 2014). Providing feedback on own group outcomes only can be viewed as an extreme case of such segregation.

[13] The payoffs were calibrated with reference to previous work by Berger et al. (2022) and the results of a pilot in which we tested whether participants would coordinate on an unequal convention under this calibration.

Figure 2: Experimental treatments



Treatments T1, T2, and T3 were also designed to support a deeper investigation into whether and how specific aspects of the dynamics of inequality affect social stability. The specifics of *Incremental* (T1) and *Pure Inequality* (T2) were inspired by the work of Fearon & Laitin (2003); Somanthan (2020); Hoyer et al. (2022), who show that absolute deprivation and the impoverishment of parts of the population are key drivers of social unrest. Under *Incremental* (T1) and *Pure Inequality* (T2), the extent of inequality, measured by the absolute difference in payoffs across the two actions, in each period is identical. However, under *Pure Inequality* (T2), the absolute payoff from B is held constant across all 100 periods, while under *Incremental* (T1) the payoff from B declines as inequality increases after $t = 50$. Thus, if a convention emerges and holds: under T2, the disadvantaged become relatively more disadvantaged but they do not become more disadvantaged in absolute terms; under T1 the disadvantaged become more disadvantaged both relatively and absolutely; and the level of relative disadvantage, i.e., the payoff difference, in each period is held constant across the two treatments.

Second, the specifics of *Shock* (T3), were inspired by the work on gradualism (see e.g. Andreoni & Samuelson, 2006; Weber, 2006), i.e., the notion that, over time, entrenchment and habituation effects deliver an amount of inertia, so that, in the presence of small environmental changes, coordination continues to be sustained despite the increasingly adverse environment. In contrast, when change takes the form of a one-time shock, habituation and inertia do not have time to kick in, and divergence from the convention is more likely. While both T1 and T2 incorporate gradual increases in inequality, *Shock* (T3) incorporates a sudden and pronounced change in inequality at $t = 51$, while holding the level of inequality across periods 51 to 100 at approximately the same average level as in T1 and T2 and the payoff for B across periods 51 to 100 at approximately the same average level as in T1.[14]

Finally, we designed the two treatments in the lower row (T4 and T5) to be further comparators

[14]Under T3, in periods 51 to 100, $\Delta = 75$ and $l = 25$. Under T1 and T2, across periods 51 to 100, the average Δ is 76, average $l = 25.5$.

to T1—T3 and to support an investigation into how different histories of inequality affect current decision-making. Participants in *Control* (T5) face the same level of inequality in the second half of the experiment as those in *Shock* (T3), but differ in terms of their previous experience; while those in *Shock* (T3) have a history of low inequality, those in *Control* (T5) have only ever experienced high inequality. Finally, under *Decreasing Inequality* (T4), inequality decreases to the level initially faced by participants in T1, T2 and T3 having, previously, been much higher.

3.3 Hypotheses

Here, we use our theoretical framework to derive predictions about how behaviour will vary across treatments. First, while we expect a convention to emerge before $t = 50$, when inequality is greater (T4 and T5), we know from Proposition 1 that $\theta^*_{g,t}$ is lower. Consequently, both the probability that a convention emerges and the strength of the convention will be lower.

Hypothesis 1 — Emergence of an unequal convention: *An unequal convention is less likely to emerge during the first 50 periods under treatments with higher initial payoff inequality (T4, T5) compared to treatments with lower initial payoff inequality (T1, T2, T3) and, if it does emerge, compliance is lower.*

Hypothesis 2 relates to our primary research question about a causal link running from inequality to stability. From Proposition 2, an increase in inequality will decrease stability.

Hypothesis 2 — Inequality threatens stability: *An increase in inequality (T1, T2, T3) causes a decline in stability (lower* SSI*).*

Hypothesis 3 relates to the underlying dynamics of this destabilisation. In line with Proposition 3, the disadvantaged will be the first to deviate from an established convention.

Hypothesis 3 — Instability is driven by the disadvantaged: *Deviations from an established convention are initiated by the disadvantaged group.*

What happens when an increase in inequality is associated with an absolute deterioration of the situation for the disadvantaged? Proposition 1 states that when l_t is lower, the probability of choosing A is higher, implying that, holding everything else constant, deviations from the convention will be faster under *Incremental* (T1) compared to *Pure inequality* (T2).

Hypothesis 4 — The effect of the disadvantaged becoming even more disadvantaged: *The destabilising effect of increasing inequality is exacerbated when the situation of the disadvantaged group deteriorates in absolute terms. A convention will destabilise to a greater degree under Incremental (T1) than under Pure Inequality (T2).*

Finally, we discuss the possibility of history dependence. Our theoretical model considers the canonical case where an individual's inequality aversion θ_i is fixed and thus history-invariant. However, it is easy to see that our predictions continue to hold if we introduce history dependence. Denoting the inequality aversion of an individual i who has been exposed to history $\mathcal{H}$ as $\theta_i(\mathcal{H})$, an individual of group g chooses A iff $\theta_i(\mathcal{H}) \geqslant \theta^*_{g,t}$ where the threshold level of inequality aversion, $\theta^*_{g,t}$, is defined in Equation (2). Hence, fixing $\mathcal{H}$, if $\theta^*_{g,t}$ decreases enough (for instance, due to higher

inequality, Δ_t), the theory still predicts that social stability will decline since some members of the disadvantaged group will switch to A.

While the theory can be adjusted to accommodate history dependence, it is largely silent about what form it will take and, hence, what effect it will have. Consider, for example, the disadvantaged group. Over time, the experience of coordinating on an unequal convention may result in *lower* inequality aversion among the disadvantaged (an *habituation* effect) and, thus, in increased stability in the present, *ceteris paribus*. Or, to the contrary, it may result in *higher* inequality aversion among the disadvantaged (an *indignation* effect) and, thus, in reduced present stability. A similar observation applies to the advantaged group, where the experience of coordinating on an unequal convention may generate an *entitlement* effect and, thus, greater aversion to disadvantageous inequality, or a *guilt* effect and, thus, lower disadvantageous inequality aversion.

Hypothesis 5 — History dependence: *The effect of inequality on stability might depend on the history of the game.*

History dependence is of interest, in part, because it has implications for the likelihood of a successful revolution, i.e, a convention reversal involving the advantaged and disadvantaged groups switching places. This possibility already exists in the canonical model. However, if the advantaged experience guilt and/or the disadvantaged become indignant, the likelihood of a successful revolution increases. In Section 6.1, after investigating each of the hypotheses set out above, we take a brief look at the frequency of successful revolutions and what might be driving them.

4 Sample and data collection

We conducted the experiment between May and September 2019 in the CeDEx laboratory at the University of Nottingham. The experiment was programmed using z-Tree (Fischbacher, 2007). Students from the University of Nottingham were recruited via the Online Recruitment System for Economic Experiments (ORSEE) (Greiner, 2015). Interactions were anonymous and communication was not allowed during the experiment. Upon arrival, participants were randomly assigned to computer terminals and informed about the procedure. In order to establish common knowledge among all participants, instructions were read aloud (see Appendix B). After answering control questions, the participants were randomly assigned to either a yellow or a green group by each of them drawing a ball from a bag. Seven participants (in a few cases, six or eight) were assigned to each group.[15] In each period participants were then randomly (re)matched with a member of the other group (either green or yellow) and played the BoS game. We collected data relating to six "societies", each comprising of one yellow and one green group, per treatment, 30 "societies" in total (see Table 1).

In addition to the individual choice data, to investigate the possible mediating effect of relative grievance in the causal link between inequality and instability (Cramer, 2005; Blattman & Miguel, 2010), we elicited emotional affect, focusing on the dimension of valence, i.e., positive versus negative

[15]Two "societies" in T3 and one "society" in T4 consisted of groups of 6 participants and one "society" in T4 consisted of groups of 8 participants. The results presented below are robust to controlling for group size.

Table 1: Participants and treatment overview

Treatment	**Participants**	**"Societies"**	**h,l at** $t \leqslant 50$	**h,l at** $t > 50$
Incremental (T1)	84	6	75,50	75+x,50-x
Pure Inequality (T2)	84	6	75,50	75+2x,50
Shock (T3)	80	6	75,50	100,25
Decreasing Inequality (T4)	84	6	100,25	75,50
Control (T5)	84	6	100,25	100,25

Note: $x \in \{1, 50\}$ for periods 51 to 100. Note that two "societies" in T3 consisted of 12 instead of 14 participants. In T4, we had one "society" of 12 and one of 16 participants. Our results are robust to controlling for group size.

feelings (Russell, 2003), after every 10th period. To do this, we used a simple and fast variant of a pictorial assessment scale developed by Desmet et al. (2001).

After the final period, participants received feedback on total payoffs. We then elicited risk aversion, using a version of Holt & Laury (2002) adjusted to our context (see Appendix B.3), personality, using a short version of the Big Five (Gerlitz & Schupp, 2005) and social preferences — from which we derive a proxy for inequality aversion —, using a social value orientation (SVO) task developed by Murphy et al. (2011); Murphy & Ackermann (2014) (see Appendix B.2).[16] Finally, we collected information on demographics.

The experimental sessions lasted 60–75 minutes. Final payoffs consisted of four elements: aggregate profits across all periods of the BoS game; the payoff from one decision in each of the risk and SVO tasks, randomly selected; and a show-up fee of £3. Average earnings were £9.56 per hour.

5 Results

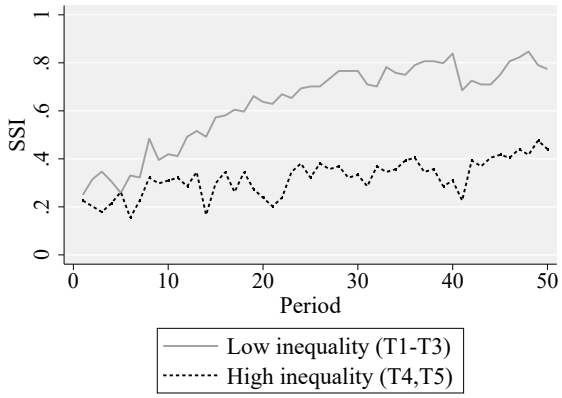
5.1 H1 — Emergence of an unequal convention

The necessary basis for all further analysis is that coordination on an unequal convention emerges during the first 50 periods of the experiment. Figure 3 presents the average SSI across "societies" for the first 50 periods under the two initial levels of inequality. While in treatments with low inequality (solid gray) we see a clear upward trend in stability, indicating increasing coordination on a convention, under high inequality (dashed black), stability only increases marginally across periods. Stability is substantially and significantly lower under high inequality, both across all periods (1–50) ($SSI = 0.63$ for $\Delta = 25$ versus $SSI = 0.32$ for $\Delta = 75$), and in the last five periods (46–50) ($SSI = 0.81$ for $\Delta = 25$ versus $SSI = 0.45$ for $\Delta = 75$).

Table 2 presents three regressions, two random effects and one fixed effects, each taking a "society", m, in a period, t, as an observation and SSI as the dependent variable. The explanatory variables of specific interest are period, a high inequality indicator variable, and the interaction between the

[16] The Big Five measure key dimensions of personality, namely openness, conscientiousness, extroversion, agreeableness and neuroticism (see Appendix B.4). The SVO task was implemented using the z-Tree code developed by Crosetto et al. (2012).

Figure 3: Stability (SSI) in $t \leqslant 50$



two.[17] The results indicate that the difference in trends between the high and low inequality treatments observed in Figure 3 is statistically significant. Table 2 reveals that SSI increases significantly across periods in the low inequality treatments, indicating the emergence of an unequal convention. Under high inequality, SSI also increases but at a significantly lower rate. These results are robust to the inclusion of demographic controls in column (2) and fixed effects in column (3).[18] Thus, we find evidence in support of Hypothesis 1, a stable, inequitable convention is less likely to emerge when the inequality implied by the convention is higher.

A high SSI could, in principle, reflect an equilibrium where groups coordinate within each period but switch the identity of the group who chooses A between periods. However, when we examine the proportion of A choices within groups across periods we see that, in most "societies" with low inequality (16/18), the yellow and green groups converge on playing the same pure NE in every period. This means that one group increasingly specialises in playing action A and the other in action B (see Figure A.1).[19] Finally, there is no significant difference in the frequency of the two pure NE across groups, i.e., neither colour is more likely to become advantaged.

Result 1: *With low inequality most "societies" coordinate on an unequal convention by the end of the first 50 periods. When the inequality implied by the convention is higher coordination on the convention is significantly less likely.*

Since the low inequality treatments (T1–T3) exhibit greater stability, they also feature higher average total earnings in periods 1–50. Subjects in low inequality treatments earned on average 30% more than subjects in high inequality treatments (t-test, $p < 0.001$). However, this came at a cost in terms of the inequality. At $t = 50$, in T1—T3, the average earnings of the disadvantaged group members were just 71% of those of advantaged group members (t-test, $p < 0.001$).

[17] A Levin-Lin-Chu unit root test indicates that the data are stationary ($p < 0.001$), supporting the use of static panel data models. This is true for $t \leqslant 50$, as well as for all 100 periods.
[18] See Table A.1 for the estimated coefficients on the demographic controls.
[19] This is also reflected in the decline in switches between actions for individuals across periods (see Figure A.3).

5.2 H2 — Inequality threatens stability

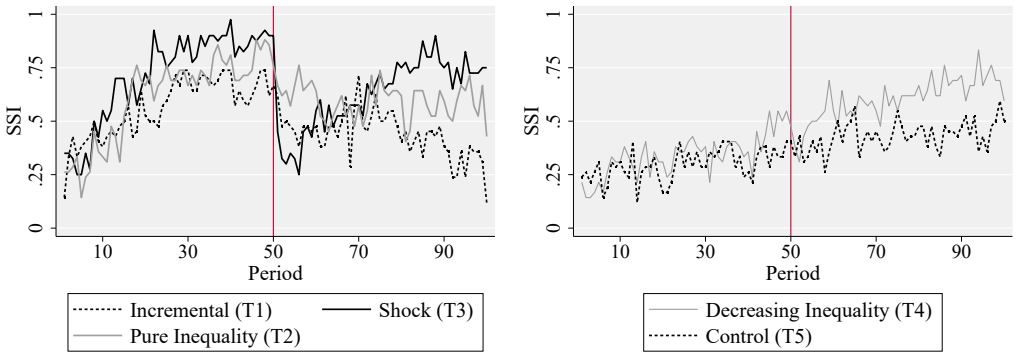
As we have seen, high inequality impedes the emergence of a stable convention. Next, we address our primary question: how does an increase in inequality affect the stability of a convention that is already in place? To answer this question, we need to examine behaviour across all 100 periods. The left graph in Figure 4 plots the average SSI across all periods for the treatments where inequality increases after period 50 (T1–T3). Average stability decreases substantially at $t = 51$, indicating a marked change in behaviour. To analyse this more formally, we test for a structural break in the SSI series for each treatment at $t = 51$ (Bai & Perron, 1998, 2003), using the Stata code developed by Ditzen et al. (2021). The analysis indicates that there is a significant structural break in all three treatments ($p < 0.001$).

Table 2: The effect of inequality level on convention emergence ($t \leqslant 50$)

	(1)	(2)	(3)
High inequality (β_1)	-0.132***	-0.123*	
	(0.049)	(0.073)	
Period (β_2)	0.011***	0.011***	0.011***
	(0.001)	(0.001)	(0.001)
High inequality x period (β_3)	-0.007***	-0.007***	-0.007***
	(0.002)	(0.002)	(0.002)
Constant (β_0)	0.355***	0.378***	0.302***
	(0.040)	(0.137)	(0.024)
Wald tests (p-values)			
$H_0 : \beta_2 + \beta_3 = 0$	0.005	0.009	0.005
Demographic Controls	No	Yes	FE
N observations	1500	1500	1500
N subjects	416	416	416
N societies	30	30	30
N periods	50	50	50
R^2	0.36	0.41	0.33

Note: Estimates from three panel regressions for $t \leqslant 50$. The dependent variable is the SSI for a given "society" in a given period. *High inequality* is a binary variable that takes the value 0 if $\Delta = 25$ (T1–T3) and 1 if $\Delta = 75$ (T4, T5). Demographic controls include the differences in and sums of the share of female participants, average Big 5 scores, risk aversion, inequality aversion, and age between the yellow and green group. Columns (1) and (2) are random effects estimations. Column (3) is a "society" fixed effects estimation. * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the "society" level.

Figure 4: Stability under different inequality dynamics



The right graph in Figure 4 plots the average SSI across all periods for the treatments in which inequality does not increase after period 50 (T4, T5). The right graph is markedly different to the left graph. In both T4 and T5, the positive trend in stability across the first 50 periods extends into the later 50 periods and we cannot reject the null that there are no structural breaks at $t = 51$ ($p = 0.79$ and $p = 0.74$ for T4 and T5 respectively).

Still focusing on Hypothesis 2, Table 3 presents results from estimating a single random effects model in which an observation is a "society", m, in a period, t, and the dependent variable is SSI. We exclude the first 20 periods from the analysis because, across these periods, stability was still low (average $SSI = 0.37$), indicating that the conventions had yet to become established. In the model, the explanatory variables are a full set of treatment indicators, a dummy variable that takes the value 1 if $t > 50$, and zero otherwise, and interactions between the treatment indicators and this binary variable. The standard errors are clustered at the "society" level. The table presents the difference in average SSI between $t \leqslant 50$ and $t > 50$ for each treatment that is implied by the single estimation. Table 3 shows that, under all three treatments where inequality increased (T1–T3), stability was significantly lower in periods 51–100 compared to 21–50.[20] Thus, we have evidence that an increase in inequality reduces stability. The individual choice data indicates that the decline in stability is driven by more individuals choosing action A, which, as explained in the theoretical section, becomes more attractive as inequality increases (see Table A.3).

Again, the behavioural pattern in the treatments with no increase in inequality (T4,T5) is different. Table 3 shows that in both T4 and T5 where, respectively, inequality declined or was unchanging, stability was significantly higher in periods 51–100 compared to periods 1—50. As expected, the largest increase in stability occurred under T4, the only treatment in which inequality declined.

Result 2: *An increase in inequality decreases stability.*

Table 3: Stability under increasing inequality

	(1) Incremental	(2) Pure Inequality	(3) Shock	(4) Decreasing inequality	(5) Control
After t=50	-0.22***	-0.15*	-0.23***	0.21***	0.10**
	(0.07)	(0.09)	(0.08)	(0.06)	(0.05)
Constant	0.65***	0.74***	0.86***	0.40***	0.33***
	(0.14)	(0.05)	(0.05)	(0.10)	(0.06)
Effect differences (p-values)					
Incremental	-	0.56	0.87	<0.001	<0.001
Pure inequality		-	0.48	0.001	0.01
Shock			-	<0.001	<0.001
Decreasing inequality				-	0.16
N observations			2400		
N subjects			416		
N societies			30		
N periods			80		
R^2			0.20		

Note: Estimates from a single random effects panel regression. The dependent variable is the SSI in a given "society" and period. *After* $t = 50$ is a dummy variable that takes the value 0 for the periods before $t = 50$ and 1 for periods 51–100. Each column presents the marginal effect for a different treatment. The p-values in the middle part of the table indicate the significance of cross-treatment differences in the marginal effects. The first 20 periods are excluded from the analysis. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the "society" level.

[20] These results are robust to including demographic controls and "society" fixed effects (see Table A.2).

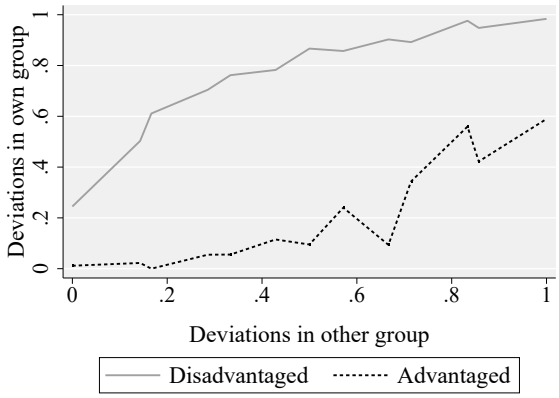
5.3 H3 — Instability is driven by the disadvantaged

Next, we investigate which group initiates the deviations from the established convention. An individual deviates from a convention if they are a member of a disadvantaged group and choose A, or they are a member of an advantaged group and choose B.[21]

On average, members of disadvantaged groups are 28% more likely to deviate compared to members of advantaged groups (t-test, $p < 0.001$). Figure 5 plots the average share of deviations after $t = 50$ in an individual's own group (y-axis) against the average share of deviations in the other group (x-axis) separately for disadvantaged (pale grey, solid) and advantaged (black, dashed) group members. So, for example, for advantaged group members, deviations in own group are defined as the share of individuals in the group choosing B, while deviations in the other, i.e. the disadvantaged group, are defined as the share of individuals choosing A.

Figure 5 confirms that members of disadvantaged groups are considerably more likely to deviate overall. Further, in line with Hypothesis 3, the figure shows that deviations are initiated by members of disadvantaged groups — even when the share of deviators in the advantaged group is zero, the share of deviators in the disadvantaged group is above 20%. In contrast, deviations by members of advantaged groups are rare up to the point where approximately 60% of disadvantaged group members are deviating. After that, in line with the concept of a tipping point, deviations by advantaged group members increase rapidly.[22] Finally, on average, the first period in which an individual deviates is significantly earlier for those in disadvantaged groups compared to advantaged group members (t-test, $p < 0.001$).

Figure 5: Deviations in treatments with increasing inequality (T1–T3) for $t \geqslant 50$



Note: Advantaged (disadvantaged) group refers to the group specialising in action A (B) between periods 21–50.

[21] To construct an indicator of which group is advantaged and which disadvantaged, we look at whether the majority of group members chose action A (advantaged group) or B (disadvantaged group) in periods 21–50. As before, the first 20 periods are excluded as the convention needs time to emerge. If the majority in both groups of a society chose A even in these periods, the group that ends up choosing A more often in periods 45–50 is defined as the advantaged one.

[22] The theoretical model predicts that members of the advantaged group should not deviate from a convention before $E(\lambda_t^{g'}) \leqslant \frac{h_t}{l_t+h_t} > 0.5$ is reached. Our data is in line with this prediction.

Result 3: *Deviations from a convention are initiated by disadvantaged group members. Members of the advantaged group tend not to deviate until a critical threshold is reached.*

5.4 H4 — The effect of the disadvantaged becoming even more disadvantaged

Hypothesis 4 states that the destabilisation of a convention is exacerbated if the increase in inequality is accompanied by the disadvantaged becoming even more disadvantaged in absolute terms. To investigate this, we compare stability in T1 and T2 for $t > 50$. Figure 4 suggests that, in line with Hypothesis 4, across these periods, SSI is lower under T1 compared to T2. Here, we investigate whether this apparent difference is statistically significant. Table 4 presents five regressions. In column (1), an observation is a "society" in a period (51—100) and the dependent variable is SSI. In columns (2) – (5), an observation is a choice by an individual in a period (51—100). The dependent variables are, in column (2), a binary variable that equals 1 if A is chosen and zero otherwise and, in columns (3) – (5), a binary variable that equals 1 if the individual chooses to deviate from the established convention in their "society" and zero otherwise. Models (2) – (5) are linear probability models. In columns (2) and (3), the choices made by all individuals are included in the sample. In columns (4) and (5), the samples are restricted to choices made by disadvantaged and advantaged group members respectively.[23] All regressions include SSI at $t = 50$, i.e. before any changes in

Table 4: The disadvantaged becoming even more disadvantaged ($t > 50$)

	SSI	A choices	Deviations		
	(1) All	(2) All	(3) All	(4) Disadvantaged	(5) Advantaged
Baseline = Incremental treatment (T1)					
Pure inequality (T2)	0.10	-0.05*	-0.19*	-0.23*	-0.14
	(0.08)	(0.02)	(0.11)	(0.12)	(0.09)
Shock (T3)	0.06	-0.04	0.01	-0.03	0.05
	(0.09)	(0.03)	(0.15)	(0.16)	(0.14)
Decreasing inequality (T4)	0.27***	-0.14***	-0.27***	-0.41***	-0.13
	(0.08)	(0.02)	(0.10)	(0.12)	(0.09)
Control (T5)	0.14**	-0.06***	-0.24***	-0.30***	-0.18**
	(0.07)	(0.02)	(0.09)	(0.10)	(0.08)
SSI at t=50	0.57***	-0.07***	-0.28***	-0.35***	-0.21**
	(0.07)	(0.02)	(0.09)	(0.10)	(0.09)
Constant	-0.12	0.71***	0.49***	0.70***	0.27**
	(0.13)	(0.04)	(0.13)	(0.17)	(0.11)
Observations	1500	20800	20800	10400	10400
N subjects	30	416	416	208	208
N societies	30	30	30	30	30
N periods	50	50	50	50	50
R^2	0.40	0.01	0.07	0.11	0.06

Note: Estimates from five random effects panel regressions for $t > 50$. The dependent variable in column (1) is the SSI in a given "society" and period and varies between 0 and 1. The dependent variable in column 2 is an individual's choice in a given period and takes the value 0 if an individual chooses B and 1 if they choose A. The dependent variable in columns (3), (4) and (5) is also binary, with a value of 0 if an individual does not deviate from a by $t = 50$ established convention and 1 if they do. The estimates in columns (2) to (5) are for linear probability models. Period is controlled for in all the regressions. * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the "society" level.

[23] This split-sample analysis needs to be viewed with a degree of caution owing to the split being based on an endogenous variable. We present it because, despite this concern, we believe that the results are informative. We do not present a split-sample analysis for A choices because, for the disadvantaged, a deviation and an A choice are the same thing and, for the advantaged, an A choice is precisely not a deviation and vice-versa.

inequality occur and before the periods under analysis, to control for differences in prior history.

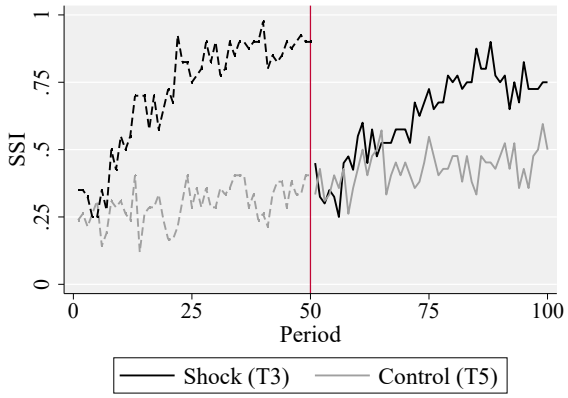
Here, we are specifically interested in the differences between *Incremental* (T1) and *Pure Inequality* (T2), which are captured by the coefficients on the *Pure Inequality* (T2) indicator variable. Column (1) indicates that while, in accordance with Hypothesis 4, SSI is lower under T1, where the disadvantaged become more so, the difference is statistically insignificant ($p = 0.22$). Columns (2) and (3) reveal that, under T1 there were more A choices and more deviations from the established convention, although, here, possibly owing to the larger sample size, the differences are weakly significant ($p = 0.07$ and $p = 0.08$ respectively). Finally, columns (4) and (5) indicate that it is the disadvantaged who drive the differences in A choices and deviations between T1 and T2.[24] Overall, we have weak but consistent evidence in support of Hypothesis 4.

Result 4: *If an increase in inequality is combined with a deterioration in the absolute outcomes of the disadvantaged, the destabilising effect of the increasing inequality is exacerbated.*

5.5 H5 — History dependence

To investigate the effect of past inequality and consequent past stability on current stability, first, we compare SSI in $t > 50$ under *Shock* (T3) and *Control* (T5). For $t > 50$, under *Shock* (T3) and *Control* (T5) participants face the same high level of inequality ($\Delta = 75$). However, this follows a history of low inequality and, as shown in Figure 6, consequently high stability under T3, while under T5 it follows a history of (already) high inequality and low stability. Focusing on $t > 50$, the figure indicates that, on average, stability is higher under T3 (t-test, $p < 0.001$). Under T3, after the shock, stability crashes but then rises again quickly and converges to a level not far below its pre-shock level. Under T3, for $21 \leqslant t \leqslant 50$, SSI is 0.86 and, for $71 \leqslant t \leqslant 100$, SSI is 0.74 (t-test, p<0.001). In contrast, in T5 stability remains low, while continuing to rise at a slow pace. Under T5,

Figure 6: Stability under Shock (T3) and Control (T5)



Note: Solid lines plot stability across periods when the level of inequality is ($\Delta = 75$) under each treatment. The dashed lines plot stability across prior periods when, under T3, the level of inequality is low ($\Delta = 25$), while under T5, it is already high ($\Delta = 75$).

[24] Table A.6 shows that these results are robust to the inclusion of demographic controls.

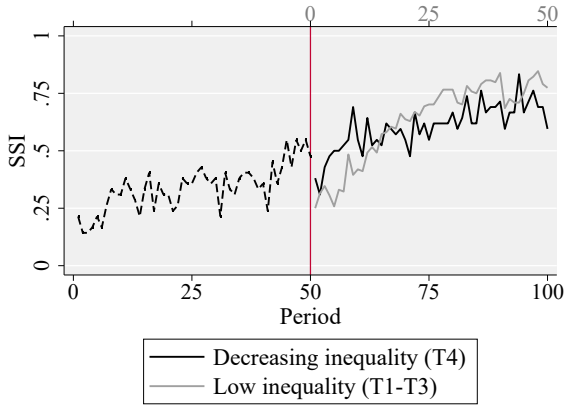
in $71 \leqslant t \leqslant 100$, SSI is 0.45. Regression analysis reveals that the upward trend in SSI for $t > 50$ is significantly steeper under T3 compared to T5 ($p = 0.03$, see Table A.7).

Result 5i: *A history of relatively low inequality and consequently high stability supports the re-establishment of stability under subsequently higher inequality.*

Next, we compare SSI in $t \leqslant 50$ under treatments with initially low inequality (T1–T3) and SSI in $t > 50$ under *Decreasing Inequality* (T4). Across all of these, inequality is low ($\Delta = 25$). However, under T1–T3 the participants enter the focal periods without any prior experience, whereas under T4 they enter having experienced high inequality and consequently low but slowly increasing stability. Note that this comparison does not allow us to distinguish between the effect of any experience and the effect of the specific experience offered by $t \leqslant 50$ under T4. This notwithstanding, we believe that the comparison is informative, at the very least, for future investigations. Figure 7 indicates that, even though average stability in the focal periods does not differ between T1–T3 in $t \leqslant 50$ and T4 in $t > 50$ (t-test, $p = 0.49$), the positive time trend is stronger under T1–T3 and regression analysis indicates that the difference in time trends is significant (Appendix A, Table A.8).

Result 5ii: *Compared to no history, a history of high inequality and consequent instability slows down the emergence of a convention even when current inequality is low.*

Figure 7: SSI under low inequality following different histories



Note: Upper horizontal axis indicates period under T4. Lower horizontal axis indicates period under T1–T3. Solid lines plot stability across focal periods when inequality is low ($\Delta = 25$). The dashed line plots stability in $t \leqslant 50$ under T4 when inequality is high ($\Delta = 75$).

Finally, we compare SSI in and shortly after $t = 75$ under *Incremental* (T1) and *Pure Inequality* (T2) to SSI in and shortly after $t = 51$ under *Shock* (T3). In $t = 75$ under T1–T2 and $t = 51$ under T3, inequality is equally high ($\Delta = 75$). However, in the first two, this level of inequality is reached via a gradual increase, while in the third it is reached via a sudden upwards shock. SSI is a considerable 0.16 greater in $t = 75$ under T1–T2 compared to $t = 51$ under T3. However, this difference is statistically insignificant (Mann-Whitney test, $p = 0.20$), probably owing to small sample size (N=18). If, instead, we compare $75 \leqslant t \leqslant 76$ under T1–T2 and $51 \leqslant t \leqslant 52$ under T3, we find that SSI is 0.20 greater under the former and the difference is statistically significant (Mann-Whitney test, $p = 0.05$, $N = 36$), despite inequality being higher on average. And shifting

from a two to five period focus, $75 \leqslant t \leqslant 79$ under T1–T2 and $51 \leqslant t \leqslant 55$ under T3, SSI is 0.24 greater under the former and the difference becomes highly significant (Mann-Whitney test, $p = 0.001$, $N = 90$), despite inequality becoming even higher on average. This also holds when we restrict the comparison to T1 and T3, which not only share the same level of inequality, Δ, but also the same h and l (Mann-Whitney test, $p = 0.07$, $N = 60$). Correspondingly, at the individual level, the probability of choosing action A in the five periods directly after the shock under T3 is greater compared to in $75 \leqslant t < 79$ under both T1 (t-test, $p = 0.04$) and T2 (t-test, $p < 0.001$).

Result 5iii: *A sudden upward shock to a specific level of inequality leads to greater instability than if the same level of inequality is reached gradually.*

6 Further analysis and discussion

Summarising Section 5, our key results are as follows. First, unequal conventions can emerge and persist. Second, an increase in the inequality implied by such a convention has a marked destabilising effect. Third, those who are disadvantaged by the convention initiate the destabilisation. Fourth, the destabilisation is exacerbated if the increase in the inequality implied by the convention is accompanied by a deterioration in the absolute position of the disadvantaged, and, fifth, history matters.

Now, before concluding and with a view to informing future research, we do three things. First, we present some descriptive analysis focusing on the substantial cross-society heterogeneity in social dynamics going on behind the aggregate SSI. Second, we review and add to the descriptive results derived from our data that speak to the validity of the theoretical framework set out in Section 2. Third, we delve a little deeper into the issue of history dependence, i.e., into how past experience of inequality and instability might affect current aversion to inequality and what our data reveals about this.

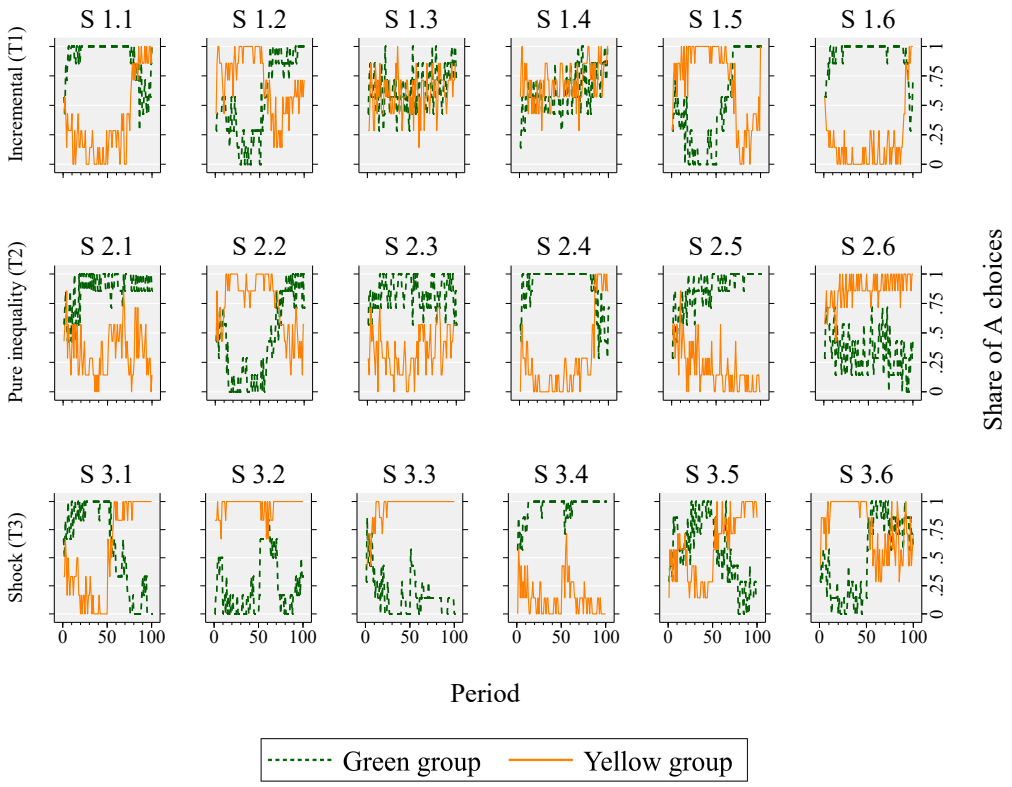
6.1 Cross-society heterogeneity in social dynamics

SSI is a useful aggregate measure of stability. However, behind SSI there is substantial cross-society heterogeneity. Figure 8 plots the share of A choices by yellow and green group members in each "society" under T1–T3.[25] Our analysis thus far indicates that deviations from an established convention increase as inequality increases. However, Figure 8 reveals that, across "societies" (S), such deviations can lead to markedly different outcomes.

Under *Shock* (T3) (bottom row), the increase in inequality at $t = 51$ is followed by multiple periods of chaos in one "society" (S3.6), a reversal of the previous convention, i.e. a successful revolution, in two (S3.1 and S3.5), and the destabilisation but subsequent re-emergence of the previous convention in three (S3.2, S3.3 and S3.4). Under *Incremental* (T1) and *Pure Inequality* (T2) we also see reversals (e.g. S1.2, S2.2). However, the strength of the new convention is much lower and, as

[25] See Figure A.6 for the behaviour in treatments with non-increasing inequality (T4 and T5).

Figure 8: Behaviour in treatments with increasing inequality (by "society")



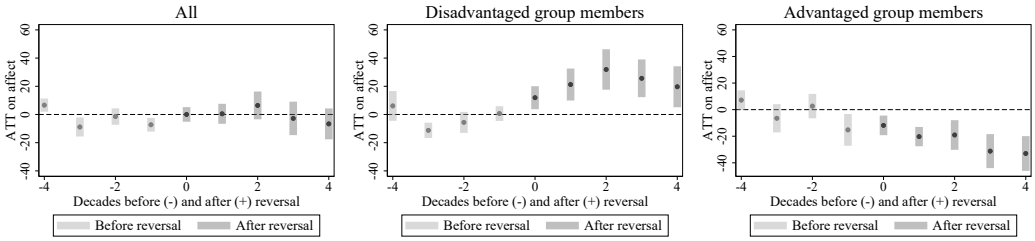
inequality continues to increase, most "societies" that reversed end up in chaos with both groups choosing A with $p > 0.5$. Further, there are fewer reversals under T2 compared to T1 (2 vs. 4).

Sample size is very restricted here — only 16 observations.[26] However, exploratory analysis reveals two correlates of reversals that might be worth investigating further. Reversals were more likely to occur in "societies" where there were larger cross-group differences in average risk aversion (Wilcoxon rank-sum test, $p = 0.05$) and average neuroticism (Wilcoxon rank-sum test, $p = 0.10$).

Finally, the dynamic relationship between affect and reversals is worthy of note. Here, we look at affect responses as a function of reversals through a staggered diff-in-diff analysis (see Callaway & Sant'Anna, 2021), where a successful reversal is used as the (non-exogenous) treatment (see Table A.9). Figure 9 presents the results of this descriptive analysis graphically. The vertical location of a bar in the plot indicates affect in the treated sample relative to affect in the untreated, i.e., the never treated and the not yet treated sample at decade intervals prior to and after the successful reversal. The left graph shows that average affect across all the participants in the sample was quite stable. However, the middle and right graphs reveal marked differences in the dynamics of affect between members of disadvantaged and advantaged groups. In the time leading up to a successful reversal, disadvantaged group members that were heading for a successful reversal indicated marginally more negative emotions compared to disadvantaged group members who were not. Then, in the time

[26] The two "societies" that did not converge during $t < 50$ have been excluded.

Figure 9: Affect dynamics as a function of reversals



Note: Event plot resulting from a staggered diff-in-diff regression, where the dependent variable is affect and the treatment variable is a successful reversal. The analysis is restricted to T1–T3. As a successful reversal is not an exogenous treatment but a retrospective sample split, this needs to be treated as a descriptive analysis. The time dimension is in decades, i.e., ten experimental period blocks, as affect is elicited once every ten periods. Advantaged (disadvantaged) describes groups where the majority chose A (B) in periods 21–50. Societies that did not converge on a convention at $t = 50$ are excluded. Whiskers represent 90% confidence intervals.

following a successful reversal, disadvantaged group members that had experienced a successful reversal indicated markedly more positive emotions compared to disadvantaged group members who had not. In marked contrast, in the time following a successful reversal, advantaged group members that had experienced a successful reversal indicated more negative emotions compared to advantaged group members who had not. While not conclusive, this analysis is consistent with the emotional responses of disadvantaged group members being the driver of successful reversals. This is also in line with the analysis of individual correlates of deviators. Table A.4 indicates that those experiencing more negative emotions are more likely to deviate from an established convention.

6.2 Failed reversals and myopic best response

The analysis of cross-society heterogeneity in social dynamics (above) reveals, not only several successful revolutions, but also some failed ones. In Figure 8 we see that in three "societies" under the *Shock* (T3) treatment (S3.2, S3.3 and S3.4), the disadvantaged group tried to reverse the existing convention, but ultimately failed.

Failed reversals are not consistent with myopic best response. If individuals in the disadvantaged group deviate to *A*, they do so despite knowing that they will receive zero with a very high probability. This being the case, if inequality remains the same (as it does under *Shock* (T3)), they should never revert back to B.

We acknowledge that expectations could be modelled in many ways and that the assumption of myopic best response constitutes a simplification, especially in the presence of large shocks. If expectations change following a significant shock to inequality, this might cause members of the disadvantaged group to try and overturn the convention. But, if the advantaged group keeps choosing A, members of the disadvantaged group may eventually correct their expectations and revert back to choosing B. Interestingly, failed reversals only appear in T3, lending support to the idea that deviations from myopic best response are linked to large shocks to inequality.

While myopic best response may not hold in the presence of large shocks, it allows us to focus on the equilibria before and after such shocks. Moreover, restricting expectations about the

other group's behaviour to current behaviour makes them traceable and is in line with the feedback structure of the experiment. Abstracting from forward-looking behaviour also yields the most conservative predictions for the effect of inequality on instability; if individuals endeavour to overturn a convention through leading-by-example this would speed up the destabilisation process.

6.3 The salience of inequality aversion

All of our key results are consistent with but do not depend on people being averse to disadvantageous inequality. However, there are several regularities in our data that are consistent with people having such an aversion and this aversion playing a behaviour-determining role in the experiment. First, when investigating inequitable convention emergence (section 5.1) and stability (section 5.2), we find that both are compromised in "societies" that are more inequality averse (see Tables A.1 and A.2). Second, when looking at who deviates from an established convention, we find that those who report more negative affect responses are more likely to deviate in $t > 50$ (see Table A.4). Third, analysis of the determinants of affect reveals that greater inequality is associated with more negative affect (see Table A.5).[27] Finally, in Figure 9 we saw how a successful revolution was preceded by a notable dip in affect among the disadvantaged and that, after the revolution, the same people reported much higher levels of affect compared to those who had not brought about such a revolution.

While all of these findings must be treated with caution owing to their descriptive nature, the individual-level correlations between experienced inequality, affect and choices to deviate, combined with the society-level correlations between inequality aversion and slow convention emergence and instability, are consistent with inequality aversion playing an important role in the inter-dynamics of inequality and social instability (Cramer, 2005; Blattman & Miguel, 2010).

6.4 History dependence revisited

Finally, we explore the implications of preferences being endogenous to past experiences (see e.g. Bowles, 1998; Henrich et al., 2010; Fehr & Hoff, 2011). Within our experiment past experience of an unequal convention could change individuals' levels of inequality aversion. Of relevance to the disadvantaged, previous studies have investigated both indignation, where past experience of being low in an income distribution increases current levels of inequality aversion (Hong & Bohnet, 2007; Buttrick & Oishi, 2017) and habituation, where similar experiences have the opposite effect (Lerner, 1978; Jost et al., 2003; Malahy et al., 2009; Trump, 2018). Relatedly, over time, the advantaged group may become accustomed to their relative position and, thus, more averse to the disadvantage associated with choosing B (Kraus & Keltner, 2013; Nishi et al., 2015; Côté et al., 2015) or may experience feelings of guilt and become more willing to choose B as a form of atonement (Beranek et al., 2015).

From Section 5.5, we see that current high inequality is less destabilizing in contexts that have experienced low inequality and, consequently, high stability in the past. This suggests that there may be a process of habituation and/or guilt at work.

Another related result is that a gradual increase in inequality is less destabilizing than a sudden

[27] See Figures A.4 and A.5 for how affect responses evolve over time.

Table 5: The effect of current and past inequality on affect

	(1) Overall	(2) Disadvantaged	(3) Advantaged
Δ (h − l)	-0.21***	-0.15*	-0.27***
	(0.05)	(0.08)	(0.08)
Δ lagged	0.09*	0.13*	0.05
	(0.05)	(0.07)	(0.09)
Period	0.04	0.03	0.05
	(0.03)	(0.05)	(0.05)
Treatments (Baseline = Control (T5))			
Incremental (T1)	3.46	9.72	-2.81
	(2.91)	(8.19)	(7.08)
Pure inequality (T2)	2.78	5.43	0.14
	(3.60)	(6.62)	(4.71)
Shock (T3)	2.09	4.05	0.13
	(3.90)	(8.23)	(6.42)
Decreasing inequality (T4)	2.85	6.91	-1.22
	(2.58)	(5.82)	(4.15)
Constant	8.34**	-5.68	22.36***
	(3.88)	(6.75)	(6.54)
Observations	3328	1664	1664
N subjects	416	208	208
N societies	30	30	30
N affect measurements	8	8	8
R^2	0.02	0.01	0.04

Note: Estimates of three random effects panel regressions. The dependent variable is affect and takes values between -50 and 50, with higher values indicating more positive affect responses. Affect is measured every tenth period. Δ indicates the average difference in h − l over the last 5 periods prior to the affect measurement. Lagged Δ indicates the first lag of h − l. Advantaged (disadvantaged) describes groups where the majority chose A (B) in periods 21–50. The first 20 periods and the associated affect measurements are excluded, resulting in 8 affect measurements that are used for the analysis. * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the "society" level.

upward shock to the same level of inequality (Result 5iii). Again, this is consistent with the idea of habituation among the disadvantaged. However, it might also be explained by a sharp increase in inequality serving as a signal — a wake-up call — that helps the disadvantaged overcome the collective action problem associated with overturning an established inequitable convention. Working against this signal-related alternative and providing further support for habituation among the disadvantaged, we see that, after controlling for current inequality, experiences of past inequality are associated with significantly more positive current affect responses (see Table 5). Advantaged group members appear not to be affected by past inequality, but do appear to be negatively affected by current inequality, which is consistent with feelings of guilt.[28]

7 Conclusion

In this paper we present the findings from a novel and unique experiment designed to investigate the causal relationship between inequality and social instability. The experiment involves a repeated game in which groups have an incentive to coordinate by specialising in different actions and coordination leads to stable but inequitable conventions. In the first half of the experiment inequitable conventions are allowed to emerge and, in line with previous studies, they do. In the second half, under three treatments the level of inequality associated with the conventions is exogenously increased

[28]The results hold when including demographic controls (see Table A.10) and also when controlling for the Payoff for B or actual earnings. The latter are highly significant and positively linked to affect.

and this leads to significant social destabilisation. Further, the deviations from the conventions underlying the destabilisation are initiated by members of the disadvantaged group.

We also identify a number of factors that intensify the destabilising effect of inequality. This is more pronounced when not only the relative but also the absolute position of the disadvantaged worsens, and when inequality increases suddenly rather than gradually. This may be due to a habituation effect, which makes current inequality more acceptable when it is reached gradually rather than suddenly. Consistent with this, we also find that, in the presence of high current inequality, the convergence to a stable convention is facilitated by a history of past stability. The importance of past experiences for current assessments of inequality is confirmed by an analysis of emotional responses during the experiment.

Taken together our findings provide unambiguous evidence of a casual relationship running from inequality to social instability, demonstrating the value of experiments as a tool for investigating factors that moderate this relationship, and providing both motivation and a strong foundation for further investigation.

References

Acemoglu, D. (2002). Technical change, inequality, and the labor market. *Journal of Economic Literature*, 40(1), 7–72.

Acemoglu, D., Fergusson, L., & Johnson, S. (2020). Population and conflict. *The Review of Economic Studies*, 87(4), 1565–1604.

Acemoglu, D. & Jackson, M. O. (2014). History, expectations, and leadership in the evolution of social norms. *The Review of Economic Studies*, 82(2), 423–456.

Alesina, A. & Perotti, R. (1996). Income distribution, political instability, and investment. *European Economic Review*, 40(6), 1203–1228.

Andreoni, J., Nikiforakis, N., & Siegenthaler, S. (2021). Predicting social tipping and norm change in controlled experiments. *Proceedings of the National Academy of Sciences*, 118(16), e2014893118.

Andreoni, J. & Samuelson, L. (2006). Building rational cooperation. *Journal of Economic Theory*, 127(1), 117–154.

Axtell, R. L., Epstein, J. M., & Young, H. P. (2001). The emergence of classes in a multiagent bargaining model. *Social Dynamics*, 27, 191–211.

Bai, J. & Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1), 47–78.

Bai, J. & Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1), 1–22.

Baronchelli, A. (2018). The emergence of consensus: a primer. *Royal Society Open Science*, 5(2), 172189.

Belloc, M. & Bowles, S. (2013). The persistence of inferior cultural-institutional conventions. *American Economic Review*, 103(3), 93–98.

Benndorf, V., Martinez-Martinez, I., & Normann, H.-T. (2016). Equilibrium selection with coupled populations in hawk–dove games: Theory and experiment in continuous time. *Journal of Economic Theory*, 165(2016), 472–486.

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

Berger, J., Vogt, S., & Efferson, C. (2022). Pre-existing fairness concerns restrict the cultural evolution and generalization of inequitable norms in children. *Evolution and Human Behavior*, 43(1), 1–15.

Billig, M. & Tajfel, H. (1973). Social categorization and similarity in intergroup behaviour. *European Journal of Social Psychology*, 3(1), 27–52.

Binmore, K., Samuelson, L., & Young, P. (2003). Equilibrium selection in bargaining models. *Games and Economic Behavior*, 45(2), 296–328.

Blattman, C. & Miguel, E. (2010). Civil war. *Journal of Economic Literature*, 48(1), 3–57.

Bowles, S. (1998). Endogenous preferences: The cultural consequences of markets and other economic institutions. *Journal of Economic Literature*, 36(1), 75–111.

Bowles, S., Loury, G. C., & Sethi, R. (2014). Group inequality. *Journal of the European Economic Association*, 12(1), 129–152.

Brandts, J. & Cooper, D. J. (2006). A change would do you good.... an experimental study on how to overcome coordination failure in organizations. *American Economic Review*, 96(3), 669–693.

Buttrick, N. R. & Oishi, S. (2017). The psychological consequences of income inequality. *Social and Personality Psychology Compass*, 11(3), e12304.

Callaway, B. & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2), 200–230.

Cantoni, D., Yang, D. Y., Yuchtman, N., & Zhang, Y. J. (2019). Protests as strategic games: experimental evidence from hong kong's antiauthoritarian movement. *The Quarterly Journal of Economics*, 134(2), 1021–1077.

Centola, D., Becker, J., Brackbill, D., & Baronchelli, A. (2018). Experimental evidence for tipping points in social convention. *Science*, 360(6393), 1116–1119.

Cherry, T. L., Kroll, S., & Shogren, J. F. (2005). The impact of endowment heterogeneity and origin on public good contributions: evidence from the lab. *Journal of Economic Behavior & Organization*, 57(3), 357–365.

Collier, P. (1999). Doing well out of war. *World Bank Report*, 28137.

Cooper, J. M., Hutchinson, D. S., et al. (1997). *Plato: complete works*. Hackett Publishing.

Côté, S., House, J., & Willer, R. (2015). High economic inequality leads higher-income individuals to be less generous. *Proceedings of the National Academy of Sciences*, 112(52), 15838–15843.

Cramer, C. (2005). *Inequality and conflict: A review of an age-old concern*. United Nations Research Institute for Social Development Geneva.

Crosetto, P., Weisel, O., & Winter, F. (2012). A flexible z-tree implementation of the social value orientation slider measure (Murphy et al. 2011)–manual. *Jena Economic Research Paper*, 62.

Dale, D. J., Morgan, J., Rosenthal, R. W., et al. (2002). Coordination through reputations: A laboratory experiment. *Games and Economic Behavior*, 38(1), 52–88.

Desmet, P., Overbeeke, K., & Tax, S. (2001). Designing products with added emotional value: Development and application of an approach for research through design. *The Design Journal*, 4(1), 32–47.

DiPrete, T. A., Gelman, A., McCormick, T., Teitler, J., & Zheng, T. (2011). Segregation in social networks based on acquaintanceship and trust. *American Journal of Sociology*, 116(4), 1234–83.

Ditzen, J., Karavias, Y., & Westerlund, J. (2021). Testing and estimating structural breaks in time series and panel data in stata. *arXiv preprint arXiv:2110.14550*.

Dube, O. & Vargas, J. F. (2013). Commodity price shocks and civil conflict: Evidence from colombia. *Review of Economic Studies*, 80(4), 1384–1421.

Edmond, C. (2013). Information manipulation, coordination, and regime change. *Review of Economic Studies*, 80(4), 1422–1458.

Ellison, G. (2000). Basins of attraction, long-run stochastic stability, and the speed of step-by-step evolution. *The Review of Economic Studies*, 67(1), 17–45.

Fajnzylber, P., Lederman, D., & Loayza, N. (1998). Determinants of crime rates in latin america and the world: an empirical assessment. *The World Bank Report, ISBN: 978-0-8213-4240-4*.

Fearon, J. D. & Laitin, D. D. (2003). Ethnicity, insurgency, and civil war. *American Political Science Review*, 97(1), 75–90.

Fehr, E. & Charness, G. (2023). Social preferences: fundamental characteristics and economic consequences. *CESifo Working Paper*.

Fehr, E. & Hoff, K. (2011). Introduction: Tastes, castes and culture: The influence of society on preferences. *The Economic Journal*, 121(556), 396–412.

Fehr, E. & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The 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.

Gächter, S., Mengel, F., Tsakas, E., & Vostroknutov, A. (2017). Growth and inequality in public good provision. *Journal of Public Economics*, 150(2017), 1–13.

Gerlitz, J.-Y. & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. *DIW Research Notes*, 4, 2005.

Gmyrek, P., Berg, J., & Bescond, D. (2023). Generative ai and jobs: A global analysis of potential effects on job quantity and quality. *ILO Working Paper*, 96.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association*, 1(1), 114–125.

Grossman, E. (2019). France's yellow vests - symptom of a chronic disease. *Political Insight*, 10(1), 30–34.

Hargreaves-Heap, S. & Varoufakis, Y. (2002). Some experimental evidence on the evolution of discrimination, co-operation and perceptions of fairness. *The Economic Journal*, 112(481), 679–703.

Henrich, J. & Boyd, R. (2008). Division of labor, economic specialization, and the evolution of social stratification. *Current Anthropology*, 49(4), 715–724.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., & Henrich, N. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science*, 327(5972), 1480–1484.

Holm, H. J. (2000). Gender-based focal points. *Games and Economic Behavior*, 32(2), 292–314.

Holt, C. A. & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5), 1644–1655.

Hong, K. & Bohnet, I. (2007). Status and distrust: The relevance of inequality and betrayal aversion. *Journal of Economic Psychology*, 28(2), 197–213.

Hoyer, D., Bennett, J. S., Whitehouse, H., François, P., Feeney, K., Levine, J., Reddish, J., Davis, D., & Turchin, P. (2022). Flattening the curve: Learning the lessons of world history to mitigate societal crises. *Working Paper*.

Hsieh, C.-C. & Pugh, M. D. (1993). Poverty, income inequality, and violent crime: a meta-analysis of recent aggregate data studies. *Criminal Justice Review*, 18(2), 182–202.

Hwang, S.-H., Naidu, S., & Bowles, S. (2024). Social conflict and the evolution of unequal conventions. *Journal of the European Economic Association*, (pp. jvae004).

Jost, J. T., Pelham, B. W., Sheldon, O., & Ni Sullivan, B. (2003). Social inequality and the reduction of ideological dissonance on behalf of the system: Evidence of enhanced system justification among the disadvantaged. *European Journal of Social Psychology*, 33(1), 13–36.

Kennedy, B. P., Kawachi, I., Prothrow-Stith, D., Lochner, K., & Gupta, V. (1998). Social capital, income inequality, and firearm violent crime. *Social Science & Medicine*, 47(1), 7–17.

Kraus, M. W. & Keltner, D. (2013). Social class rank, essentialism, and punitive judgment. *Journal of Personality and Social Psychology*, 105(2), 247–261.

Lerner, M. (1978). *The belief in a just world. A fundamental delusion.* New York: Plenum Press.

Lewis, D. (1967). *Convention: A philosophical study.* Cambridge: Harvard University Press.

Lichbach, M. I. (1989). An evaluation of "does economic inequality breed political conflict?" studies. *World Politics*, 41(4), 431–470.

Lobeck, M. & Støstad, M. N. (2023). The consequences of inequality: Beliefs and redistributive preferences. *CESifo Working Paper*.

Luce, R. D. & Raiffa, H. (1957). Games and decisions: Introduction and critical survey.(1957). *reprinted by Dover Publications. M. De Vos and D. Vermeir," Choice Logic Programs and Nash Equilibria in Strategic Games" Proceedings of the 13th CSL*, 99, 266–276.

Malahy, L. W., Rubinlicht, M. A., & Kaiser, C. R. (2009). Justifying inequality: A cross-temporal investigation of us income disparities and just-world beliefs from 1973 to 2006. *Social Justice Research*, 22(4), 369–383.

McKelvey, R. D. & Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1), 6–38.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1), 415–444.

Mookherjee, D. & Ray, D. (2002). Is equality stable? *American Economic Review*, 92(2), 253–259.

Murphy, R. O. & Ackermann, K. A. (2014). Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review*, 18(1), 13–41.

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. (2011). Measuring social value orientation. *Judgment and Decision Making*, 6(8), 771–781.

Newton, J. (2021). Conventions under heterogeneous behavioural rules. *The Review of Economic Studies*, 88(4), 2094–2118.

Nishi, A., Shirado, H., Rand, D. G., & Christakis, N. A. (2015). Inequality and visibility of wealth in experimental social networks. *Nature*, 526(7573), 426–429.

Norris, P. & Inglehart, R. (2019). *Cultural backlash: Trump, Brexit, and authoritarian populism*. Cambridge University Press.

Olson, M. (1965). *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge: Harvard University Press.

Oprea, R., Henwood, K., & Friedman, D. (2011). Separating the hawks from the doves: Evidence from continuous time laboratory games. *Journal of Economic Theory*, 146(6), 2206–2225.

Østby, G. (2008). Polarization, horizontal inequalities and violent civil conflict. *Journal of Peace Research*, 45(2), 143–162.

Radkani, S., Holton, E., De Courson, B., Saxe, R., & Nettle, D. (2023). Desperation and inequality increase stealing: evidence from experimental microsocieties. *Royal Society Open Science*, 10(7), 221385.

Roubini, N. (2011). The instability of inequality. *RGE Global EconoMonitor*, October 17, 2011.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145–172.

Satz, D. (2019). What's wrong with inequality? *The IFS Deaton Review*, available at: https://www.ifs.org.uk/inequality/expert-comment/whats-wrong-with-inequality/.

Scheidel, W. (2017). *The great leveler: Violence and the history of inequality from the stone age to the twenty-first century.* Princeton University Press.

Schotter, A. & Sopher, B. (2003). Social learning and coordination conventions in intergenerational games: An experimental study. *Journal of Political Economy*, 111(3), 498–529.

Smith, V. L. (1982). Microeconomic systems as an experimental science. *The American Economic Review*, 72(5), 923–955.

Somanthan, R. (2020). Group inequality in democracies: Lessons from cross-national experiences. In J.-M. Baland, F. Bourguignon, J.-P. Platteau, & T. Verdier (Eds.), *The Handbook of Economic Development and Institutions* (pp. 137–152). Princeton: Princeton University Press.

Stewart, F. (2000). Crisis prevention: Tackling horizontal inequalities. *Oxford Development Studies*, 28(3), 245–262.

Trump, K.-S. (2018). Income inequality influences perceptions of legitimate income differences. *British Journal of Political Science*, 48(4), 929–952.

Turchin, P. (2023). *End times: elites, counter-elites, and the path of political disintegration.* Penguin.

Van Huyck, J. B., Cook, J. P., & Battalio, R. C. (1997). Adaptive behavior and coordination failure. *Journal of Economic Behavior & Organization*, 32(4), 483–503.

Weber, R. A. (2006). Managing growth to achieve efficient coordination in large groups. *American Economic Review*, 96(1), 114–126.

Young, H. P. (1993). The evolution of conventions. *Econometrica: Journal of the Econometric Society*, 61(1), 57–84.

Young, H. P. (1996). The economics of convention. *Journal of Economic Perspectives*, 10(2), 105–122.

Young, H. P. (2015). The evolution of social norms. *Annual Review of Economics*, 7(1), 359–387.

Appendices

A Additional analysis

A.1 Emergence of an unequal convention

Emergence by "society"

Over the first 50 periods, the majority of "societies" in treatments with low inequality (T1–T3) coordinate on an unequal convention. This is reflected in a significant and positive increase of SSI over time. Figure A.1 confirms that this stability is achieved not by groups taking turns between actions but through the emergence of the same repeated pure NE. Figure A.1 shows the share of A choices over time for the yellow and the green group in each society. In most "societies", yellow and green groups specialise on different actions. In only two "societies" play concurs with a mixed NE, which predicts for $\Delta = 25$ that participants choose A in 60% of periods.

Under high inequality (T4, T5), by contrast, the emergence of a pure NE during the first 50 periods is less common. Figure A.2 shows that in T4 and T5 there is much more chaos with groups never coordinating on different actions in several "societies". Even where there is a tendency for groups to specialise on different actions, it is much less pronounced than under low inequality.

Figure A.1: Emergence under low inequality by "society"

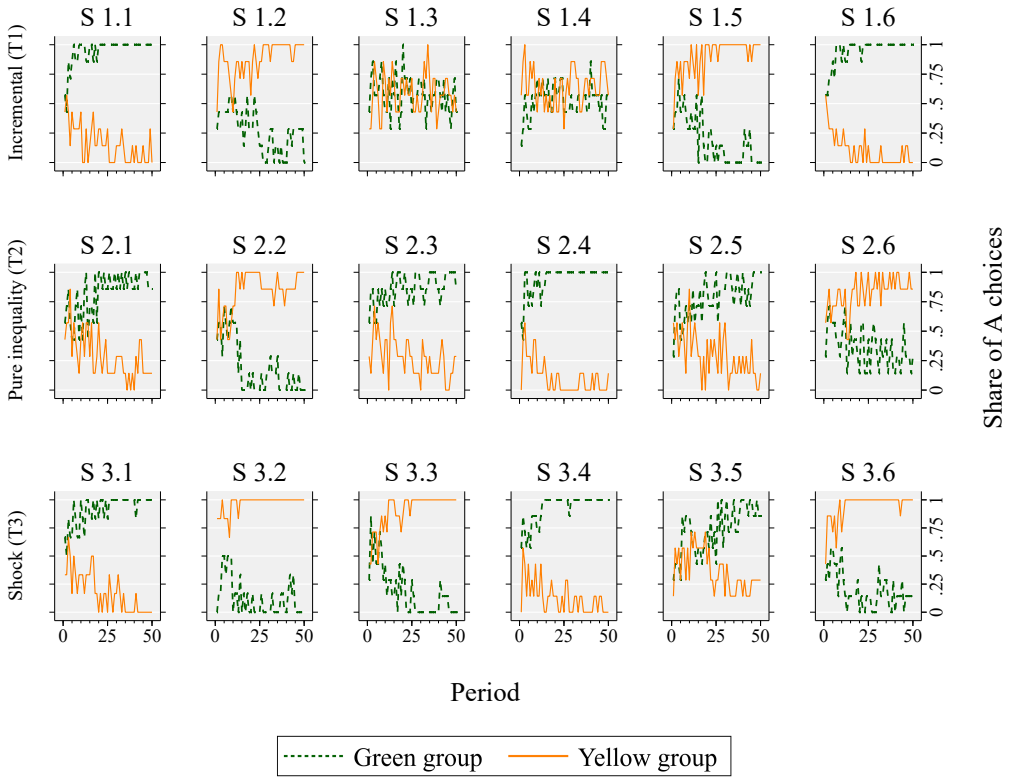
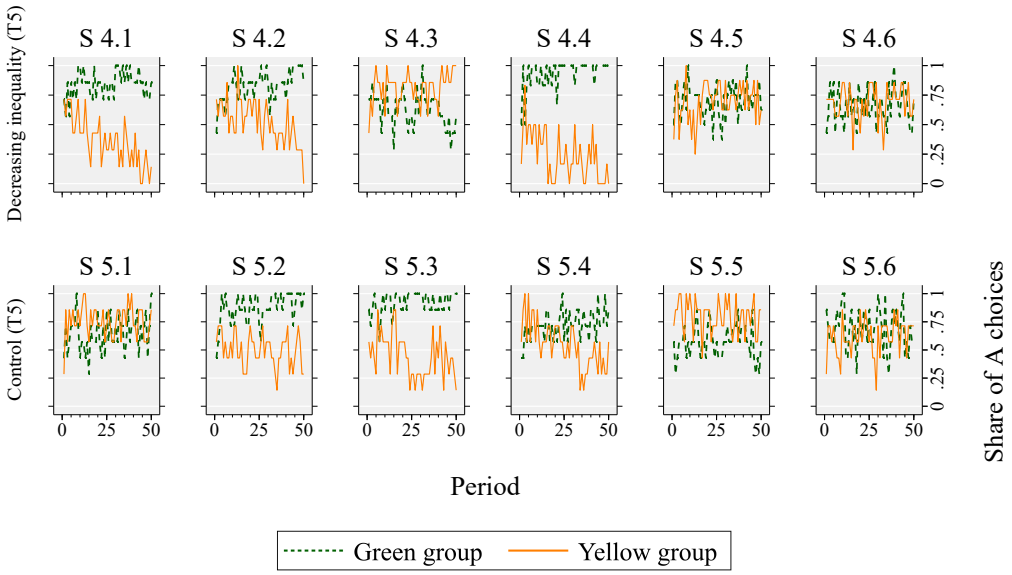


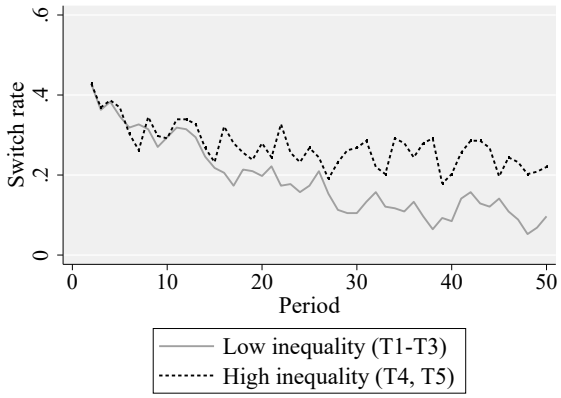
Figure A.2: Emergence under high inequality by "society"



Individual choices — reduction in switches over time

The emergence of an unequal convention is associated with a reduction in switches between actions A and B. Figure A.3 shows that the number of individuals who switch decreases during the first 50 periods of the experiment. Again, this pattern is more pronounced under low (T1–T3) than under high inequality (T4, T5).

Figure A.3: Decrease in switches between A and B over the first 50 periods



Emergence of an unequal convention - adding controls

Table A.1 reports the coefficients for all of the explanatory variables included in the estimation presented in column (2) of Table 2 in the main text. Focusing on the demographic controls, stability is higher if the total share of women or extroverted people in a "society" is higher and lower if the overall level of inequality aversion is higher. Moreover, stability is lower if the difference in consci-

entiousness between groups within a "society" is larger.

Table A.1: The effect of inequality level on convention emergence ($t \leqslant 50$) - including controls and fixed effects

	b	SE
High inequality	-0.123*	(0.073)
Period	0.011***	(0.001)
High inequality x period	-0.007***	(0.002)
Sum of controls		
Female	0.322*	(0.194)
Age	0.032	(0.020)
Inequality aversion	-0.670***	(0.243)
Risk aversion	-0.048	(0.053)
Openness	-0.094	(0.076)
Conscientiousness	-0.084	(0.052)
Extroversion	0.099*	(0.052)
Agreeableness	0.007	(0.127)
Neuroticism	0.036	(0.071)
Difference in controls		
Female	-0.131	(0.266)
Age	0.039	(0.052)
Inequality aversion	-0.673	(0.454)
Risk aversion	0.072	(0.093)
Openness	0.181	(0.144)
Conscientiousness	-0.289**	(0.122)
Extroversion	-0.021	(0.124)
Agreeableness	0.113	(0.148)
Neuroticism	-0.045	(0.127)
Constant	-0.430	(1.795)
N observations	1500	
N subjects	416	
N societies	30	
N periods	50	
R^2	0.53	

Note: Estimates of a random effects panel regression for $t \leqslant 50$. The dependent variable is the SSI for a given "society" in a given period and varies between 0 and 1. *High inequality* is a binary variable that takes the value 0 if $\Delta = 25$ (T1–T3) and 1 if $\Delta = 75$ (T4, T5). * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the "society" level.

A.2 Inequality threatens stability

Stability before and after t=50 — adding controls

Table A.2 shows that our findings that SSI declines in the second half of the experiment under treatments with an increase in inequality (T1–T3) and SSI increases in the second half of the experiment under treatments where inequality declines or is constant (T4 and T5), hold after controlling for demographic characteristics (Panel A) and after including fixed effects (Panel B).

Table A.2: Stability under increasing inequality (after $t = 50$) - including controls

	(1) Incremental	(2) Pure Inequality	(3) Shock	(4) Decreasing	(5) Control
Panel A: Controls					
After t=50	-0.22*** (0.07)	-0.15* (0.09)	-0.23*** (0.08)	0.21*** (0.06)	0.10** (0.05)
Constant	-1.64 (1.48)	-1.59 (1.45)	-1.30 (1.47)	-1.93 (1.41)	-1.85 (1.45)
Effect differences (p-values)					
Incremental	-	0.56	0.87	<0.001	<0.001
Pure inequality		-	0.48	0.001	0.001
Shock			-	<0.001	<0.001
Decreasing inequality				-	0.17
Sum of Controls					
Female			0.33** (0.13)		
Age			0.04** (0.02)		
Inequality aversion			-1.04*** (0.26)		
Risk aversion			-0.05 (0.05)		
Openness			-0.06 (0.06)		
Conscientiousness			0.02 (0.05)		
Extroversion			0.16*** (0.04)		
Agreeableness			-0.11 (0.11)		
Neuroticism			0.06 (0.05)		
Difference in Controls					
Female			-0.00 (0.20)		
Age			0.11** (0.05)		
Inequality aversion			-0.06 (0.36)		
Risk aversion			0.02 (0.08)		
Openness			0.36*** (0.11)		
Conscientiousness			-0.15 (0.10)		
Extroversion			-0.09 (0.10)		
Agreeableness			-0.09 (0.12)		
Neuroticism			-0.17 (0.13)		
N observations			2400		
N subjects			416		
N societies			30		
N periods			80		
R^2			0.50		
Panel B: Fixed effects					
After t=50	-0.22*** (0.07)	-0.15* (0.09)	-0.23*** (0.08)	0.21*** (0.06)	0.10** (0.05)
Constant	-1.64 (1.48)	-1.59 (1.45)	-1.30 (1.47)	-1.93 (1.41)	-1.85 (1.45)
Effect differences (p-values)					
Incremental	-	0.56	0.87	<0.001	0.001
Pure inequality		-	0.48	0.002	0.017
Shock			-	<0.001	0.001
Decreasing inequality				-	0.17
N observations			2400		
N subjects			416		
N societies			30		
N periods			80		
R^2			0.17		

Note: Estimates of a random effects panel regression (Panel A) and a fixed effects panel regression (Panel B). The dependent variable is the SSI in a given society and period and varies between 0 and 1. *After* $t = 50$ is a binary variable that takes the value 0 for the periods before $t = 50$ and 1 for periods 51–100. The first 20 periods are excluded from the analysis. At the top of each panel, each column presents the marginal effect for a different treatment. The p-values below that indicate the significance of cross-treatment differences in the marginal effects. * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the "society" level.

Probability of choosing A before and after t=50

As the SSI is based on the difference in A choices between the yellow and green group, an increase in the probability of choosing A for each individual is equivalent to a decrease in stability. Table A.3 shows the results of a single random effects panel model, taking a binary variable that takes the value

1 if an individual chooses A and 0 otherwise as the dependent variable and the *After* $t = 50$ dummy as the explanatory variable of interest. As in the analysis of the SSI, the regression includes treatment indicators and their interaction with *After* $t = 50$ as controls. Each column in Table A.3 presents the marginal effect of shifting from the first to the latter 50 periods for a specific treatment. In all three treatments where inequality increases (T1–T3), we see a significant increase in the probability of choosing A after $t = 50$. In contrast, under treatments where inequality is not increasing (T4, T5), the probability of choosing A decreases after $t = 50$, indicating an increase in stability.

Table A.3: Changes in the probability of choosing A after $t = 50$

	(T1) Incremental	(T2) Pure inequality	(T3) Shock	(T4) Decreasing inequality	(T5) Control
After t=50	0.12***	0.08***	0.08***	-0.10***	-0.03*
	(0.01)	(0.03)	(0.02)	(0.01)	(0.02)
Constant	0.38***	0.37***	0.38***	0.48***	0.50***
	(0.10)	(0.10)	(0.09)	(0.11)	(0.10)
Effect differences (p-values)					
Incremental	-	0.62	0.62	<0.001	<0.001
Pure inequality		-	0.90	<0.001	<0.001
Shock			-	<0.001	<0.001
Decreasing inequality				-	0.004
N observations			41600		
N subjects			416		
N societies			30		
N periods			80		
R^2			0.01		

Note: Estimates of a single random effects panel regression. The dependent variable is a binary variable that takes the value of 0 if an individual chooses action B and 1 if they choose action A. *After* $t = 50$ is a dummy variable that takes the value 0 for periods before $t = 50$ and 1 for periods 51–100. The regression controls for treatment indicators and their interaction with *After* $t = 50$. Each column presents the marginal effect for a different treatment. The p-values below that indicate the significance of cross-treatment differences in the marginal effects. The regression controls for group size. The first 20 periods are excluded from the analysis, as it takes time for a convention to emerge. * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the "society" level.

Deviations from an established convention

Table A.4 shows the results of random effects panel regressions with a binary variable that takes the value 1 if an individual deviates early, i.e., chooses an action that goes against the established convention, while the majority is still following it. The analysis is restricted to periods $T > 50$ under treatments T1–T3. Column (1) presented estimated based on a pooled analysis. Columns (2) – (4) present results by treatment, derived from a single model including interaction terms.

The results show that the probability of deviating is lower if the convention was more stable to begin with (higher SSI at $t = 50$). This is expected, as the risk of receiving zero following a deviation is higher the more stable the convention. After controlling for this, individuals who are more risk averse and report a higher negative affect are more likely to deviate. Finally, participants who are inequality averse are significantly more likely to deviate in T1.

Affect responses

Figure A.4 shows how affect develops over time across the first 50 periods, separately for members of groups that ultimately end up being advantaged and for members of group that ultimately end

Table A.4: Correlates of deviating from a convention (t > 50)

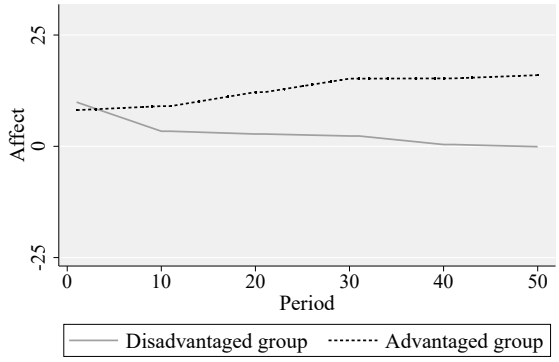
	High versus low inequality (1)	By treatment (2)	(3)	(4)
	Low inequality	Incremental	Pure inequality	Shock
SSI at t=50	-0.169***	-0.149***	-0.377***	0.018
	(0.030)	(0.037)	(0.120)	(0.036)
Female	0.003	0.018	0.021	-0.037
	(0.029)	(0.042)	(0.071)	(0.023)
Age	-0.002	-0.001	-0.004	-0.002
	(0.003)	(0.003)	(0.008)	(0.004)
Inequality averse	0.020	0.133*	0.019	-0.000
	(0.029)	(0.074)	(0.024)	(0.029)
Risk aversion	-0.014***	-0.009	-0.023***	-0.019***
	(0.004)	(0.007)	(0.006)	(0.006)
Big 5				
Openness	-0.000	0.013	0.001	-0.014
	(0.012)	(0.015)	(0.021)	(0.015)
Conscientiousness	0.009	-0.021	0.016	0.009
	(0.011)	(0.024)	(0.019)	(0.010)
Extroversion	0.014	0.028	0.020	0.003
	(0.011)	(0.021)	(0.019)	(0.012)
Agreeableness	0.004	-0.022***	-0.003	0.035***
	(0.010)	(0.007)	(0.025)	(0.006)
Neuroticism	-0.003	0.006	-0.016*	-0.002
	(0.007)	(0.013)	(0.008)	(0.011)
Negative affect	0.002***	0.002***	0.002**	0.000
	(0.000)	(0.001)	(0.001)	(0.000)
Period	-0.003***	-0.002	-0.002***	-0.004***
	(0.001)	(0.002)	(0.001)	(0.001)
Constant	0.480	0.192	0.372	0.665**
	(0.320)	(0.313)	(0.356)	(0.324)
N observations	20800		20800	
N subjects	416		416	
N societies	30		30	
N periods	50		50	
R^2	0.07		0.10	

Note: Estimates of two random effects panel regressions for t > 50. The dependent variable is a binary variable that takes the value of 1 if an individual deviates early, i.e., chooses an action that goes against the established convention, while the majority is still following it, and 0 otherwise. Both regressions include treatment indicators and their interactions with personal characteristics. Column (1) presents marginal effects for low inequality treatments (T1 – T3) pooled. Columns (2) – (4) present the marginal effects from a single regression separately for T1 – T3. Regressions control for group size. * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the "society" level.

up being disadvantaged. With growing stability and corresponding inter-group inequality, a gap in affect emerges. Advantaged groups show more positive emotional affect over time, while disadvantaged groups show more negative affect. Across all 50 periods, there is a significant difference in average affect (t-test, $p < 0.001$).

Figure A.5 shows how affect develops for advantaged and disadvantaged groups across all 100 periods under each treatment. Under 4 out of the 5 treatments, the affect gap between advantaged and disadvantaged groups persists throughout the experiment. The increase in inequality (T1, T2, T3) initially leads to more negative affect, while a decrease in inequality (T4) results in more positive affect responses. Later, we observe more positive responses in T2 and T3 and a closing of the gap between disadvantaged and advantaged groups in T1. These distinct patterns reflect the specific

Figure A.4: Affect responses for $t \leqslant 50$



Note: Affect is measured on a scale from -50 to 50, with higher values indicating more positive emotions. Affect is measured every tenth round.

dynamics of each "society" and the existence of reversals (see Section 6.1).

The relationship between affect and inequality is also evident in a regression analysis. Table A.5 indicates that individual affect is lower when inequality is higher. This remains the case after controlling for earnings in the previous period, i.e. the actual payoff that was realised as a consequence of the individual's and their matched partners choice in the BoS game ($\pi \in \{h, l, 0\}$).

Figure A.5: Affect responses across treatments

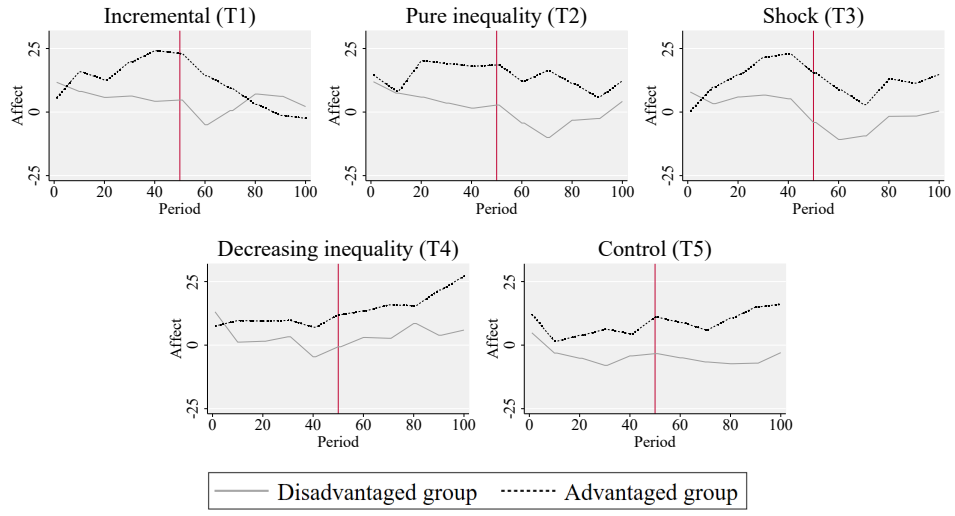


Table A.5: The effect of inequality on affect responses

	(1)	(2)	(3)
Δ (h − l)	-0.14***	-0.11***	-0.11***
	(0.03)	(0.03)	(0.03)
Period	0.05*	0.03	0.03
	(0.03)	(0.03)	(0.03)
Treatments (Baseline = Control (T5))			
Incremental (T1)	3.60	3.39	3.39
	(2.73)	(3.07)	(3.07)
Pure inequality (T2)	2.92	0.16	0.16
	(3.51)	(3.92)	(3.92)
Shock (T3)	2.15	0.50	0.50
	(3.67)	(3.80)	(3.80)
Decreasing inequality (T4)	1.78	1.47	1.47
	(2.65)	(3.17)	(3.17)
Earnings (π)		0.18***	0.18***
		(0.02)	(0.02)
Observations	3328	3328	3328
Demographic controls	No	No	Yes
N subjects	416	416	416
N societies	30	30	30
N affect measurements	8	8	8
R^2	0.02	0.11	0.11

Note: Estimates of random effects panel regressions. The dependent variable is affect and takes values between -50 and 50, with higher values indicating more positive affect responses. Affect is measured every tenth period. Δ indicates the difference between h − l in a given period. The first 20 periods and the associated affect measurements are excluded, resulting in 8 affect measurements that are used for the analysis. * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the "society" level.

A.3 The effect of the disadvantaged becoming even more disadvantaged

Table A.6 shows that, after including demographic controls, A choices and deviations by members of the disadvantaged group remain significantly higher under the incremental treatment (T1) compared to the pure inequality (T2).

Table A.6: The disadvantaged becoming even more disadvantaged ($t > 50$) - adding controls

	SSI	A choices	Deviations		
	(1) All	(2) All	(3) All	(4) Disadvantaged	(5) Advantaged
Baseline = Incremental treatment (T1)					
Pure inequality (T2)	-0.01	-0.05**	-0.19*	-0.24**	-0.12
	(0.04)	(0.02)	(0.10)	(0.12)	(0.09)
Shock (T3)	0.13***	-0.05	-0.00	-0.04	0.04
	(0.05)	(0.03)	(0.15)	(0.17)	(0.13)
Decreasing inequality (T4)	0.05	-0.14***	-0.28***	-0.42***	-0.14
	(0.05)	(0.02)	(0.10)	(0.12)	(0.09)
Control (T5)	0.15***	-0.06**	-0.24***	-0.31***	-0.20**
	(0.04)	(0.03)	(0.09)	(0.10)	(0.08)
SSI at t=50	0.49***	-0.06***	-0.28***	-0.34***	-0.22**
	(0.04)	(0.02)	(0.09)	(0.10)	(0.09)
Constant	-1.26**	0.91***	0.58***	0.94***	0.17
	(0.56)	(0.20)	(0.19)	(0.28)	(0.22)
Demographic controls	Yes	Yes	Yes	Yes	Yes
Observations	1500	20800	20800	10400	10400
N subjects	30	416	416	208	208
N societies	30	30	30	30	30
N periods	50	50	50	50	50
R^2	0.58	0.02	0.08	0.12	0.06

Note: Estimates from five random effects panel regressions for $t > 50$. The dependent variable in column (1) is the SSI in a given "society" and period and varies between 0 and 1. The dependent variable in column 2 is an individual's choice in a given period and takes the value 0 if an individual chooses B and 1 if they choose A. The dependent variable in columns (3), (4) and (5) is also binary, with a value of 0 if an individual does not deviate from a by $t = 50$ established convention and 1 if they do. The estimates in columns (2) – (5) are for linear probability models. Period is controlled for in all the regressions. Demographic controls in column (1) are the difference and sum of each characteristic between the yellow and green group within a "society". Demographic controls in columns (2) – (5) are gender, age, inequality aversion, risk aversion, and big 5 at the individual level. * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the "society" level.

A.4 History dependence

Coordination under high inequality given different histories

Under both the *Shock* (T3) and the *Control* (T5) treatment, participants face high inequality ($\Delta = 75$) for $t > 50$. The difference is that under T3, participants experienced high stability in the first 50 periods while participants under the *Control* (T5) treatment did not. In Table A.7, we regress the SSI on a time trend, treatment indicators, and their interaction. We thereby focus on the periods 51-100, where inequality is identical in both T3 and T5. While the *Control* (T5) treatment does not show a statistically different SSI at $t = 51$ from the *Shock* (T3) treatment, we can see that the increase in stability is significantly lower. This shows that past experiences of instability can undermine coordination even in environments with low inequality. The regression results are robust to the

inclusion of controls in column (2) and a fixed effects specification in column (3).

Table A.7: The effect of different histories on stability under high inequality for $t > 50$

	(1)	(2)	(3)
Baseline = Shock (T3)			
Control (T5)	0.385	0.347	
	(0.236)	(0.257)	
Period	0.010***	0.010***	0.010***
	(0.003)	(0.003)	(0.003)
Control (T5) x period	-0.008**	-0.008**	-0.008**
	(0.004)	(0.004)	(0.004)
Constant	-0.102	-2.864**	0.364***
	(0.213)	(1.321)	(0.088)
Observations	1500	1500	1500
N subjects	416	416	416
N societies	30	30	30
N periods	50	50	50
Controls	No	Yes	FE
R^2	0.13	0.55	0.05

Note: Estimates of three panel regressions for $t > 50$. The dependent variable is the SSI in a given "society" and period and varies between 0 and 1. Demographic controls include the difference in and sum of the share of female participants as well as in average measures of Big 5, risk aversion, inequality aversion and age between yellow and green groups within a "society". All regressions include treatment indicators for T1, T2, and T4. As the relevant comparison is between T3 and T5, they are omitted from the table. Columns (1) and (2) are random effects models, while column (3) reports results of a fixed effects estimation. * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the "society" level.

Coordination under low inequality given different histories

Table A.8 regresses the SSI in a given "society" and period on treatment indicators, a time trend, and their interaction. Our comparison of interest is between the *Decreasing Inequality* (T4) treatment for $t > 50$ and the treatments with low initial inequality (T1–T3) for $t \leqslant 50$. For these intervals, the treatments share low levels of inequality ($\Delta = 25$), but T4 has a history of instability that is absent in the other treatments. Column (1) in Table A.8 compares the SSI by treatment, taking T4 as the baseline. While for all treatments stability increases over time, we see that the time trend is more positive in T1–T3 (even though not statistically significant for T1). In columns (2) – (4), we pool all three treatments with initially low inequality. Again the latter show a significantly more positive time trend than T4. This result is robust to the inclusion of controls in column (3) and a fixed effects in column (4).

Table A.8: The effect of different histories on coordination under low inequality ($\Delta = 25$)

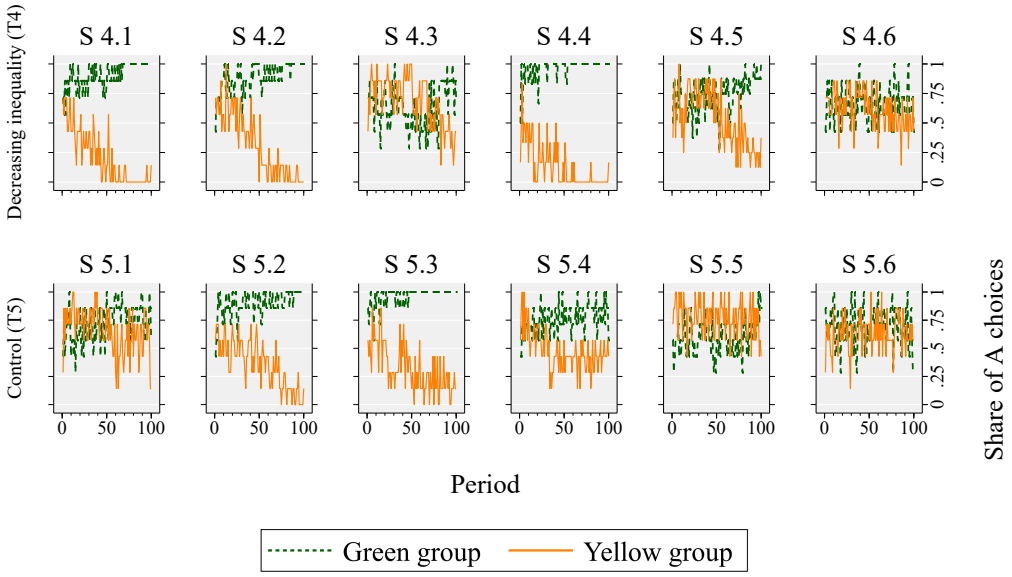
	By treatment	Treatments with low Δ pooled		
	(1)	(2)	(3)	(4)
Baseline = Decreasing inequality (T4)				
Incremental (T1)	-0.124			
	(0.147)			
Pure inequality (T2)	-0.182			
	(0.146)			
Shock (T3)	-0.091			
	(0.158)			
Shock (T3) x period	0.008***			
	(0.002)			
Incremental (T1) x period	0.003			
	(0.003)			
Pure inequality (T2) x period	0.007***			
	(0.002)			
Period	0.005***	0.005***	0.005***	0.005***
	(0.002)	(0.002)	(0.002)	(0.002)
Low inequality (T1-T3)		-0.132	-0.114	0.000
		(0.140)	(0.124)	(.)
Low inequality (T1-T3) x period		0.006***	0.006***	0.006***
		(0.002)	(0.002)	(0.002)
Constant	0.488***	0.488***	-0.415	0.367***
	(0.135)	(0.135)	(2.065)	(0.033)
Observations	1500	1500	1500	1500
N subjects	416	416	416	416
N societies	30	30	30	30
N periods	50	50	50	50
Controls	No	No	Yes	FE
R^2	0.23	0.20	0.49	0.16

Note: Estimates of four panel regressions. The dependent variable is the SSI in a given "society" and period and varies between 0 and 1. All regressions include a treatment indicator for T5. As the relevant comparison is between T4 and T1–T3, it is omitted from the table. Demographic controls include the difference in and the sum of the share of female participants as well as in average measures of Big 5, risk aversion, inequality aversion, and age between the groups interacting with each other in a "society". Column (1), (2) and (3) are random effects models, while column (4) reports results of a fixed effects estimation. * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the "society" level.

Behaviour in treatments with non-increasing inequality

Figure A.6 shows the share of A choices for yellow and green groups by "society" for treatments with non-increasing inequality (T4, T5). In particular, in the *Decreasing Inequality* (T4) treatment — but also to a lesser extent in the *Control* (T5) treatment — average stability is larger for $t > 50$ than in the first 50 periods. We do not see any attempts at reversals in these treatments.

Figure A.6: Behaviour in treatments with non-increasing inequality (by "society")



A.5 Convention reversals and affect

Table A.9 shows the average treatment effect on the treated (ATT) of a successful reversal on affect by period before and after treatment. The treatment is whether a "society" achieved a successful reversal in the second half of the experiment, focusing on T1–T3. The analysis compares affect in groups that experienced a successful reversal at a given time to affect in groups that never or had not yet experienced a successful reversal. As reversals are endogenous, the analysis is descriptive rather than causal. Column (1) presents results for all groups, columns (2) and (3) separately for disadvantaged and advantaged groups (defined with reference to Period 50).

Table A.9: ATT of affect (treatment = successful reversal)

	Overall (1)	Disadvantaged (2)	Advantaged (3)
Before reversal			
Average pre	-2.75*	-2.53	-2.97
	(1.42)	(2.90)	(2.00)
T(-4)	6.63**	6.08	7.20
	(2.83)	(6.42)	(4.43)
T(-3)	-8.86 **	-11.24***	-6.48
	(4.10)	(3.27)	(6.45)
T(-2)	-1.48	-5.62	2.66
	(3.53)	(4.53)	(5.59)
T(-1)	-7.30	0.64	-15.24**
	(2.95)	(3.17)	(7.26)
After reversal			
Average post	-0.50	22.07***	-23.08***
	(4.50)	(5.96)	(5.28)
T(0)	0.03	11.91**	-11.84***
	(3.16)	(4.95)	(4.48)
T(1)	0.47	21.24***	-20.29***
	(4.31)	(6.90)	(4.42)
T(2)	6.42	31.91***	-19.07***
	(5.99)	(8.72)	(6.78)
T(3)	-2.77	25.65***	-31.20***
	(7.21)	(8.12)	(7.78)
T(4)	-6.66	19.66**	-32.98***
	(6.67)	(8.83)	(7.95)
N observations	1320	660	660

Note: Estimates of three staggered difference in difference estimations using a successful reversal as a treatment and never or not yet treated "societies" as the control group. The analysis is restricted to T1–T3, where reversals, not necessarily successful, are observed. The time dimension is decades as affect is only measured every tenth period. Advantaged (disadvantaged) describes groups where the majority chose A (B) in periods 21–50. "Societies" without a convention at $t = 50$ are excluded from the analysis. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the "society" level.

A.6 History dependence revisited

Table A.10 shows that the effect of inequality and past inequality on emotional affect persists after including demographic controls.

Table A.10: The effect of current and past inequality on affect - including controls

	(1) Overall	(2) Disadvantaged	(3) Advantaged
Δ (h − l)	-0.21***	-0.15*	-0.27***
	(0.05)	(0.08)	(0.08)
Δ lagged	0.09*	0.13*	0.05
	(0.05)	(0.07)	(0.09)
Period	0.04	0.03	0.05
	(0.03)	(0.05)	(0.05)
Treatments (Baseline = Control (T5))			
Incremental (T1)	4.03	8.50	-4.80
	(3.38)	(7.75)	(7.02)
Pure inequality (T2)	4.07	5.35	-2.68
	(3.60)	(6.22)	(4.92)
Shock (T3)	2.54	3.38	-0.96
	(3.81)	(8.13)	(6.54)
Decreasing inequality (T4)	3.84	7.57	-2.82
	(3.01)	(6.35)	(4.15)
Constant	-14.33	-43.69**	19.26
	(11.67)	(20.51)	(20.02)
Demographic controls	Yes	Yes	Yes
Observations	3328	1664	1664
N subjects	416	208	208
N societies	30	30	30
N affect measurements	8	8	8
R^2	0.04	0.10	0.06

Note: Estimates of three random effects panel regressions. The dependent variable is affect and takes values between -50 and 50, with higher values indicating more positive affect responses. Affect is measured every tenth period. Δ indicates the average difference in h − l over the last 5 periods prior to the affect measurement. Lagged Δ indicates the first lag of h − l. Advantaged (disadvantaged) describes groups where the majority chose A (B) in periods 21–50. The first 20 periods and the associated affect measurements are excluded from the analysis, resulting in 8 affect measurements that are used for the analysis. Demographic controls include gender, age, inequality aversion, risk aversion, and big 5. * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the "society" level.

B Instructions

B.1 Main experiment

All participants received a hard-copy of the instructions for the main experiment at the beginning of the session. The following text shows the instructions for *Incremental* (T1), *Pure Inequality* (T2) and *Shock* (T3). Instructions for *Control* (T4) and *Decreasing Inequality* (T5) involve different payoffs and are presented in brackets.

General instructions for participants

We warmly welcome you to this experimental study.

Please read the following instructions carefully. During this experiment, depending on your decisions and those of other participants, you can earn some money over and above your show-up fee of £3. It is very important that you read all the instructions carefully, so that you understand the potential consequences of your decisions. If you have any questions, please raise your hand and an experimenter will come to you. During the experiment, please do not try to communicate with any of the other participants and please do not use mobile phones. If you do not follow these rules, you will be excluded from the study and will not be paid. Below, we describe the experiment you are going to participate in during today's session. The anonymity of all the decisions you make during the experiment is guaranteed.

Detailed information about the study

Earnings

During the experiment, you will earn points. Then, at the end of the experiment, your total points will be converted into pounds using the following conversion rate: **100 points = £0.15** The resulting amount plus your show-up fee of £3 will be given to you in cash at the end of today's session.

Allocation into groups

At the beginning of the experiment, **you and each of the other participants in the session will be randomly assigned to a group** by the draw of a coloured ball (with probability 0.5). Half of the participants will be assigned to a **green** group, the other half to a **yellow** group. **Each participant will stay in the same group for the whole experiment**.

The decision situation

In this experiment you will play a game 100 times. We will refer to each time you play as a round. The game is played in pairs. **In each round, you will play the game with someone randomly**

selected from a differently coloured group. **New random selections will be made for each round**. So, if you are in a **green** group, you will be randomly and newly paired with a player from a **yellow** group in each round. And, if you are in a **yellow** group, you will be randomly and newly paired with a player from a **green** group in each round. **The group you are paired with stays the same** for all 100 rounds. You will never know the identity of the people you play each game with and they will never know yours.

In a round, you and your playing partner for that round each have to choose between **Action A** and **Action B**. The consequence for you of the action that you choose depends also on the action chosen by your partner. To understand exactly what this means, take a look at the **Decision table** below.

Decision table

		Choice of yellow player	
		Action A	Action B
Choice of green player	Action A	0 points 0 points	*b* points *a* points
	Action B	*a* points *b* points	0 points 0 points

The Decision table can be read as follows:

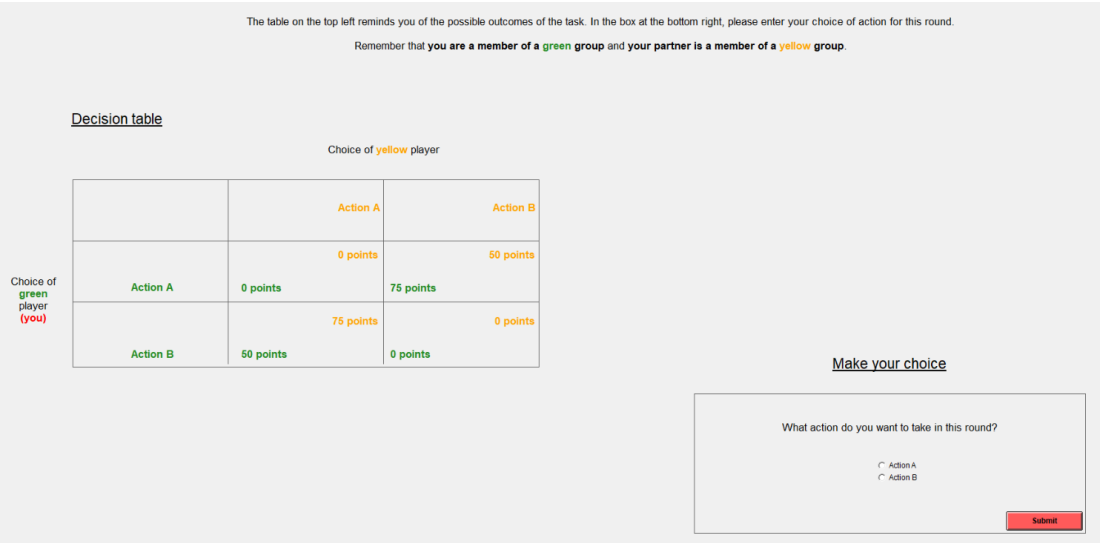
If the **green** player chooses A and the **yellow** player chooses A, then both players receive 0 points. If the **green** player chooses B and the **yellow** player chooses B, then both players receive 0 points.

If the **green** player chooses A and the **yellow** player chooses B, then the **green** player receives **a** points and the **yellow** player receives **b** points. If the **green** player chooses B and the **yellow** player chooses A, then the **green** player receives **b** points and the **yellow** player receives **a** points. During the experiment the amounts for **a** and **b** will vary, so keep an eye on the decision table when making your choices. At regular intervals we are going to highlight the decision table to help you remember to check it. However, the amounts of **a** and **b** could change at any time. So, you need to quickly check the decision table, before making your choice in each round.

As you can see, the lowest paying situation for the players is if both make the **same choice**, that is, if both players choose A or both players choose B.

Both players make their decision whether to choose A or B simultaneously. **So, in each round, you will not know your partner's decision before you make your decision.**

Below, is a screenshot of what you will see on your computer in the **first round** of the experiment.[29] The screen is set up assuming that you are a member of a **green** group. At the top of the screen is a reminder of your group and the group to which your playing partner for that round belongs. On the left-hand side is the Decision table (same as you saw above). Note that for example in this round **a=75 and b=50** [**a=100 and b=25**]. You indicate your choice for the round in the box in the bottom right hand corner of the screen. In each round you must choose either **Action A** or **Action B**.

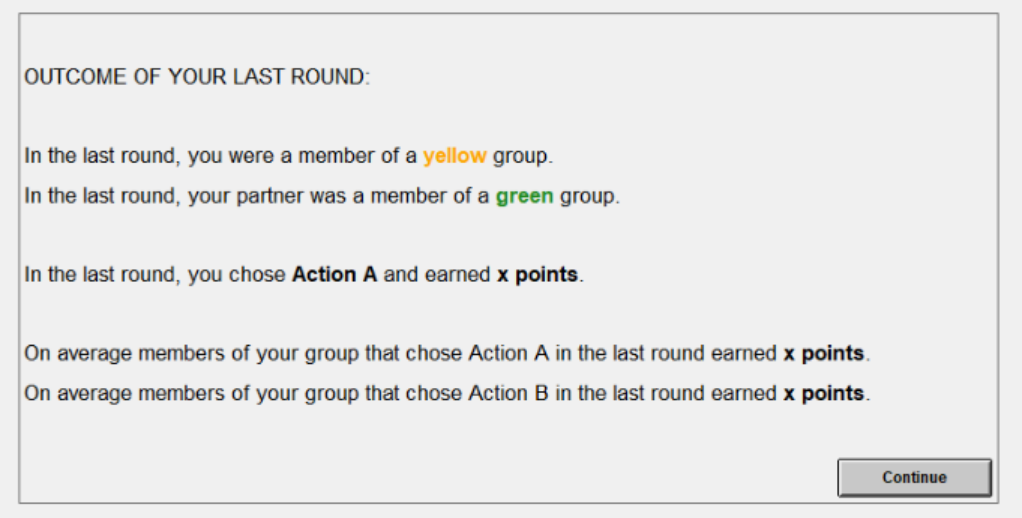


After each round, before you start playing the next round, you will be given the following information, summarising the last round:

- a reminder of your group affiliation and the group affiliation of the person you played with in the previous round
- the choice you made in the previous round and how many points you earned in the previous round
- how many points players from YOUR group who chose A in the previous round earned on average in that round
- how many points players from YOUR group who chose B in the previous round earned on average in that round.

Below is an **example** for the screen summarizing the last round, assuming that you are a member of a **yellow** group:

[29] The screenshot for the *Decreasing Inequality* (T4) and the *Control* (T5) treatment shows a=75 and b=50.



Note that where you see **"x points"** in the screenshot above, positive numbers will appear, as appropriate, during the experiment. After round 100 you will receive a summary of the whole experiment. Then, we will ask you to complete a **questionnaire**, which consists of three parts. You will receive more detailed information about the questionnaire after you have completed the experimental task. Finally, you will be paid.

Summary

Participants are randomly assigned to a **yellow** or a **green** group. Participants' group affiliation remain the same for all 100 rounds.

- In each round you will be newly, randomly matched with a participant from a differently coloured group. You will always be paired with someone from the **same other** group.
- You and your partner each have to choose either A or B and you do this simultaneously.
- If you and your partner both choose A or both choose B, each of you will earn zero. If one of you chooses A and the other B, the one choosing A earns **a** points and the one choosing B earns **b** points.
- The amounts for **a** and **b** can change during the experiment. It is thus important that you quickly check the decision table, before making your choice in each round.

If you have completely understood the instructions, please answer the control questions on screen. If you have any questions please raise your hand and an experimenter will come to you.

B.2 Social value orientation (SVO) and proxy for inequality aversion

We elicit social value orientation following Murphy et al. (2011). Participants see the following text and have to decide how to share tokens between themselves and another participant in six scenarios. The order of scenarios is thereby randomised.

In line with Murphy et al. (2011), we define individuals with an SVO angle of 37.4 as inequality averse. We did not administer the full SVO test including the secondary items that support a differentiation between different prosocial motives due to time constraints. We acknowledge that using the exact SVO angle can only serve as a proxy for inequality aversion (see also Fehr & Charness (2023)). When using SVO types instead of inequality aversion in our analysis, we find that

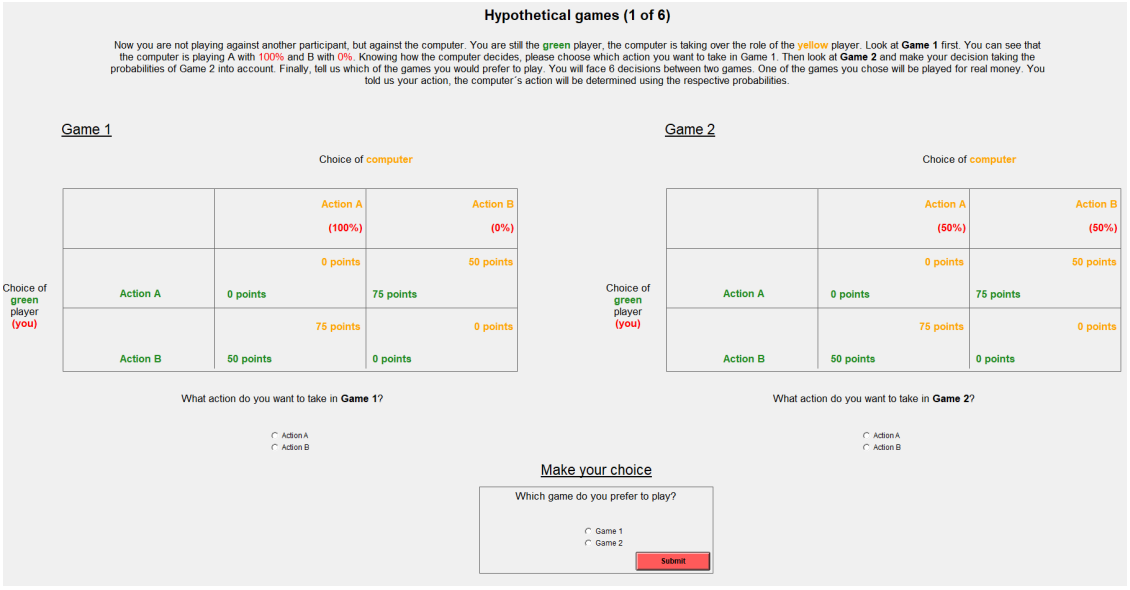
individuals classified as prosocial behave similarly to those classified as inequality averse.

B.3 Risk elicitation task

The instructions for this part of the experiment were provided on screen (see screenshot below). This is an example of the instructions for a green player. The instructions for a yellow player were adjusted accordingly. The text on top of the screen reads as follows:

Hypothetical games (1 of 6)

Now you are not playing against another participant, but against the computer. Further, as well as making its own choice between A and B, the computer makes your choice as well. The only choice you have is which game to play. You are going to be asked to make six choices. Each time, you will be choosing one out of two games. You are still the **green** player, the computer has taken over the role of the **yellow** player. Look at **Game 1** first. In **Game 1** the likelihood of the computer playing A is **100%** and the likelihood of it playing B is **0%** and the computer has chosen B for you. Then look at **Game 2**. In **Game 2** the likelihood of the computer playing A is **50%** and the likelihood of it playing B is **50%** and the computer has chosen A for you. Please tell us which of the games you would prefer to be played. Once you have chosen your six games, one will be played out for real money.



Note: Participants take decisions between Game 1 and Game 2 in six rounds. Game 1 always has a 100% probability of the other choosing A. Game 2 successively decreases the probability of the other choosing A from 50% in round 1, to 40% in round 2, until it reaches 0% in round 6.

B.4 Questionnaire

1) Below you see a number of statements. For each statement, please indicate how much you agree with this. I SEE MYSELF AS SOMEONE WHO...
[*Answer options were on a 7 point Likert scale from strongly disagree to strongly agree*]

- ...does a thorough job
- ...is communicative, talkative
- ...is sometimes somewhat rude to others
- ...is original, comes up with new ideas
- ...worries a lot
- ...has a forgiving nature
- ...tends to be lazy
- ...is outgoing, sociable
- ...values artistic experiences
- ...gets nervous easily
- ...does things effectively and efficiently
- ...is reserved
- ...is considerate and kind to others
- ...has an active imagination
- ...is relaxed, handles stress well

2) What is your gender? (male/ female/ other)

3) What is your age?

4) Which year of university are you in? (Undergraduate first year/second year/ third year/ fourth year or further/ Master/ PhD/ other)

5) Which faculty do you belong to? (Arts/ Engineering/ Medicine and Health Sciences/ Sciences/ Economics or Business School/ Social Sciences/ None of the above)

6) How many participants of this experiment have you known beforehand?

7) At which round did the payoffs change?

8) How often did you take the decision table into account before making your decision?

9) Did you hear any details about this experiment from other students before participating? (Yes/ No)

10) Do you have any other comments on this experiment? If yes, you can give us feedback here. If not, just type no into the box.

Thank you very much for participating in this experiment!

Your total payoff from this experiment is £x.

You will receive your payment in a moment. All payments will be rounded up to the next decimal. Please wait until you are called up.