CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

University of Nottingham
UK | CHINA | MALAYSIA

Discussion Paper No. 2023-13

Abigail Barr,
Anna Hochleitner and
Silvia Sonderegger

December 2023

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[§]

December 8, 2023

Abstract

In this paper we study the relationship between inequality and social instability. While the argument that inequality can be damaging for the cohesion of a society is old, the empirical evidence is mixed. We use a novel approach to isolate the causal relationship running from inequality to instability. Specifically, we conduct a laboratory experiment. In the experiment, two groups are interacting with each other repeatedly and have an incentive to cooperate even though cooperation comes at the cost of inter-group inequality. In the second half of the experiment, we vary the extent of the inequality implied by cooperation. Our results show that increasing such inequality has a destabilising effect; the disadvantaged group attacks the status quo. We show that this behaviour is consistent with a simple theoretical framework incorporating disadvantageous inequality aversion and myopic best response. Moreover, we find that a worsening of the absolute situation of the disadvantaged group or a sudden rather than gradual increase in inequality exacerbate the destabilising effect of inequality. Finally, we show that history matters, with people responding differently to the same level of inequality now depending on their past experiences.

Keywords: Collective decision making, Conflict and Revolutions, Inequality

JEL Codes: C92, D01, D63, D74

We åre grateful to the ESRC (Research Training Support Grant) and CeDEx for funding this research. We thank CeDEx members and participants at the CCC conference 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 (Inglehart & Norris, 2016), 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).

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. To the best of our knowledge, we are the first to investigate this causal relationship experimentally.

Our research design builds on previous experimental work that explores the emergence of inequitable conventions (see e.g. Dale et al., 2002; Hargreaves-Heap & Varoufakis, 2002; Oprea

[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.

et al., 2011; Benndorf et al., 2016; Berger et al., 2022). It involves a repeated battle-of-the-sexes 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 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). In our experiment, one of the two possible actions results in a higher payoff than the other, creating an 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 inequality between groups. More specifically, groups tend to coordinate, with each group specialising in one action and, thus, one group becoming relatively disadvantaged. Then, to test the "inequality causes instability" hypothesis, we diverge from previous studies by exogenously varying the payoff difference between the actions across the remaining 50 periods.

Our findings indicate that increasing inequality destabilises the existing status quo and that this destabilisation tends to be initiated by members of the disadvantaged group. This pattern of behaviour is in line with the predictions of a simple theoretical framework that incorporates disadvantageous inequality aversion and myopic best response. In addition, by exogenously varying the dynamics of inequality, we find that stability is even lower if 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 higher levels of inequality in the past.

Our findings, most importantly, 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 with some studies finding 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; Somanthan, 2020). Our findings indicate that, under certain conditions at least, increasing inequality is sufficient to cause social instability.

As mentioned above, 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 larger patterns across longer periods of time (see e.g. Scheidel, 2017; Hoyer et al., 2022). 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 (Hoyer et al., 2022). Our experiment has in its abstraction similarities to this long-run view; by stripping

away all confounding sociopolitical factors, we show that increasing inequality has a direct, negative effect on cooperation across variably advantaged groups.[2]

Sticking with the long-run perspective, our paper also has links to the application of evolutionary game theory to the emergence of unequal conventions and inter-group inequalities. Young (1993) describes conventions as a "*pattern of behaviour that is customary, expected and self-enforcing*" (p.57). Several theorists have modelled conventions as the outcomes of evolutionary coordination processes (Lewis, 1967; Baronchelli, 2018; Young, 1996) 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 studies 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.[3] In the absence of information about individual behaviour, people tend to focus on expectations about group behaviour (Dale et al., 2002). Group affiliation (e.g. 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 the theoretical and experimental work described in the previous paragraph is whether and how such unequal equilibria can be disturbed. Despite their often arbitrary origin, theoretical work shows that unequal conventions can perpetuate over long periods even if they are inefficient (Hwang et al., 2018; Belloc & Bowles, 2013). 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., 2018; Baronchelli, 2018). Here, 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; they focus on situations of common interest, while we intentionally introduce an element of conflict.

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

[2]We consciously decided not to give elites 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.

[3]The endogenous emergence of inequality has also been found in other contexts such as repeated public good games (see e.g. Gächter et al., 2017).

details of the data collection. Section 5 sets out our experimental results. Finally, Section 6 concludes.

2 Theoretical framework

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). After each period the individuals learn their own individual earnings and the average earnings for their fellow group members conditional on those members' 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 cooperate 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 of inter-group inequalities.[4] Below we show how individuals choose their actions in each period and how such a strategy profile can emerge

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: $h > l \geqslant 0$.

[4]Using group affiliation as a coordination device also requires the least amount of cognitive effort. While alternating between action A and B 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.

endogenously over time.

2.2 Comparative statics

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 $0 \leqslant \theta_i \leqslant 4$.[5]

Payoffs depend on both own and other's choices. Specifically, for an individual i in group g the payoff to 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 and 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

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

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)) and, 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 for B or more agents in the other group choosing A, action

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

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, $E(\lambda^{g'}_t)$, 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' usually choose action A, they will also expect them to do so in the next period, making B a likely best response. By choosing B the convention is then further strengthened.

More formally, a convention can be said to exist if one group specialises in A and the other group in B, i.e. $\lambda^g_t \to 1$ and $\lambda^{g'}_t \to 0$. 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) $\lambda^g_t \geqslant x > 0.5$, (ii) $\lambda^{g'}_t \leqslant y < 0.5$, (iii) $E(\lambda^g_t) \geqslant x$, and (iv) $E(\lambda^{g'}_t) \leqslant y$.*

Note that, holding $E(\lambda^{g'}_t)$, 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 Benndorf et al., 2016, for a similar approach):

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

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 λ_t^g, $E(\lambda_t^g) \geqslant x$, $\lambda_t^{g'}$, and $E(\lambda_t^{g'}) \leqslant y$. What will happen to the stability of this convention if the inequality between 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 the convention, as inequality increases, more and more members of the disadvantaged group will prefer A over B and the convention will attenuate. 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.[6] 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 high levels of inequality aversion θ_i. 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

[6] 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.

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

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 *i. Increases in inequality destabilise an existing convention and lower* SSI_t. *ii. Deviations are always initiated by the disadvantaged group. This, in turn, can induce reactions from the advantaged group once* $E(\lambda_t^{g'}) \geqslant \frac{h_t}{l_t+h_t} > 0.5$.

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 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.[7] 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.

[7] 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.

In the first 50 periods, we hold payoffs for 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.

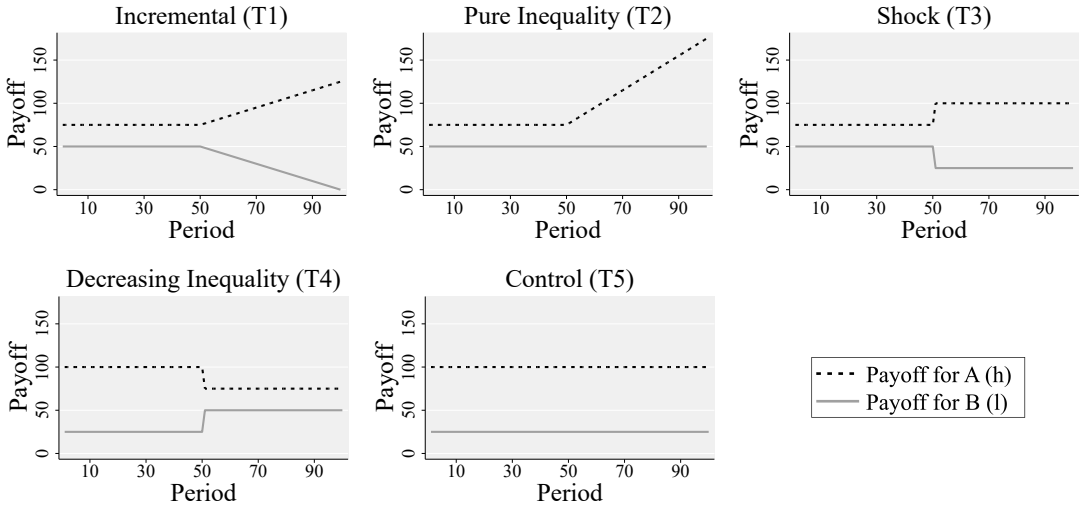
3.2 Treatments

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 for 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 ($\Delta_t = h_t - l_t = 25$),[8] 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.

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 for B is held constant across all 100 periods, while under

Figure 2: Experimental treatments



[8]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.

Incremental (T1) the payoff for B declines as inequality increases after $t = 50$. Thus, if a convention emerges and holds: under T2, the disadvantaged become relatively more disadvantaged but no 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 small changes in inequality lead to habituation, making it less likely that individuals deviate from a convention. In contrast, when change takes the form of a one-time shock, individual adjustment cannot occur 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.[9]

Finally, we designed the two treatments in the lower row (T4 and T5) to be further comparators to T1—T3 and to support an investigation into how different histories of inequality affect current decision-making. Participants under *Control* (T5) face the same level of inequality in the second half of the experiment as those under *Shock* (T3), but differ in terms of their previous experience; while those under *Shock* (T3) have a history of low inequality, those under *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 2i, 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*).*

[9] Under T3, in periods 51 to 100, the payoff difference is 75 and l=25. Under T1 and T2, across periods 51 to 100, the average payoff difference is 76, average l=25.5.

Hypothesis 3 relates to the underlying dynamics of this destabilisation. In line with Proposition 2ii, 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 the disadvantaged group facing an absolute deterioration in their situation? 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 more rapidly 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) then the theory predicts that social stability will decline since some members of the disadvantaged group will switch to A.

While the theory can accommodate history dependence, it is largely silent about what form it will take. Consider for example the disadvantaged group. Previous exposure to higher inequality may result in *lower* inequality aversion on average (an *habituation* effect), and thus in increased stability in the present, *ceteris paribus*. Or, to the contrary, it may result in *higher* inequality aversion (an *indignation* effect), thus decreasing present stability. A similar observation applies to the advantaged group, where exposure to higher inequality could generate an *entitlement* effect – and thus higher aversion to disadvantageous inequality – or a *guilt* effect – leading to lower inequality aversion.

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

A possible implication of history dependence is that we may observe reversals of initial conventions, with the advantaged and disadvantaged groups switching roles – successful revolutions within the the context of the experiment. This possibility already arises in the canonical model but may be facilitated by history dependence.

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.[10] 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 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.2), personality, using a short version of the Big Five (Gerlitz & Schupp, 2005) and social preferences, using a social value orientation (SVO) task (Murphy et al., 2011; Murphy & Ackermann, 2014).[11] Finally, we collected information on demographics.

The experimental sessions lasted between 60 and 75 minutes. The final payoffs consisted of four elements: the aggregate profits across all periods of the BOS; the payoff relating to one decision randomly selected from each of the risk elicitation and SVO tasks; and a show-up fee of £3. Average earnings were £9.56 per hour (SD 2.44).

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. Controlling for group size shows that the latter does not affect results.

[10] 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.

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

5 Results

5.1 H1 — Emergence of an unequal convention

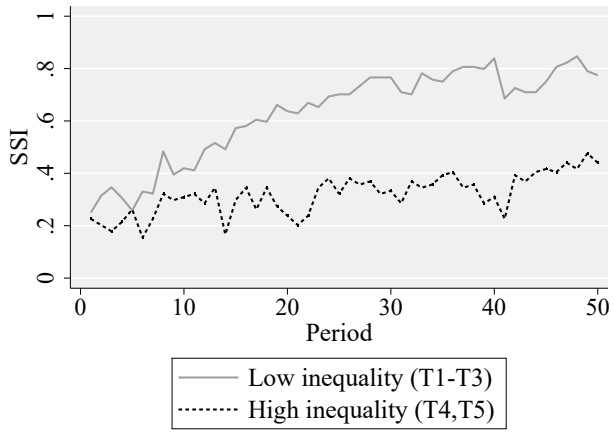
The necessary basis for all further analysis is that coordination on an unequal convention emerges during the first 50 periods of the experiment.

Figure 3 presents the average SSI across "societies" for the first 50 periods under the two 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 lower under high inequality ($SSI = 0.63(0.81)$ for $\Delta = 25$ versus $SSI = 0.32(0.45)$ for $\Delta = 75$), both across all periods (1–50), as well as in the last 5 periods as an approximation of equilibrium play (46–50).

Table 2 presents three regressions, two random effects and one fixed effects, each taking SSI as the dependent variable. The explanatory variables of specific interest are Period, a high inequality dummy, and the interaction between the two. The results indicate that the difference between high and low inequality treatments observed in Figure 3 is statistically significant.[12] Table 2 shows that SSI increases significantly across periods in the low inequality treatments, indicating the emergence of an unequal convention. Under high inequality, by contrast, *SSI* increases at a significantly lower rate. These results are robust to the inclusion of demographic controls in column (2) and fixed effects in column (3).[13] Thus, we find evidence in support of Hypothesis 1, a stable, inequitable convention is less likely to emerge when inequality is higher.

A high SSI could, in principle, reflect an equilibrium where groups coordinate within each

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



[12] 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.
[13] None of the demographic controls are significant (see Appendix A, Table A.1).

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

	(1)	(2)	(3)
High inequality (β_1)	-0.132***	-0.156***	
	(0.049)	(0.056)	
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 clusters	30	30	30
N periods	50	50	50
R^2	0.36	0.41	0.33

Note: Results of three panel regressions 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). Demographic controls include the difference in the share of female participants as well as in average measures of Big 5, risk aversion, prosociality 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.

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, 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 Appendix A, Figure A.1).[14] 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 inequality is higher coordination on an unequal convention is significantly less likely.*

Since the low equality 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 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, $p < 0.001$).

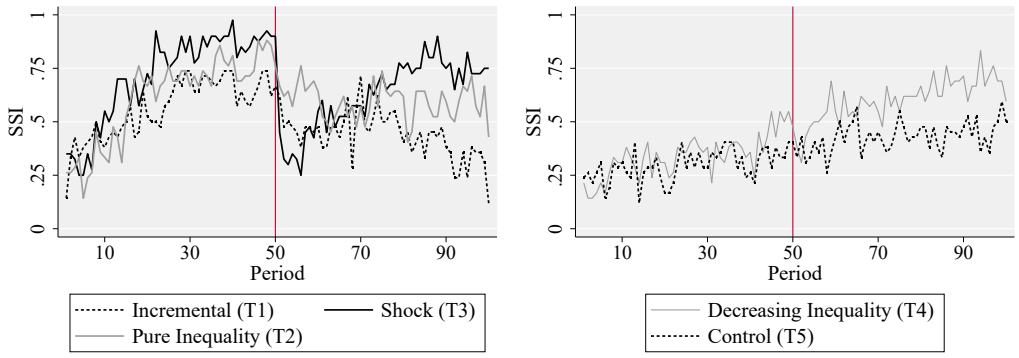
[14] This is also reflected in the decline in switches between actions for individuals across periods (see Appendix A, Figure A.3).

5.2 H2 — Inequality threatens stability

As we have seen, high inequality impedes the emergence of a stable convention. The next question we address is: 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 $t = 50$, 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 $t = 50$ (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 ($p < 0.001$).

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=50 in the data ($p = 0.84$ and $p = 0.86$).

Figure 4: Stability under different inequality dynamics



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 dummy 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. 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

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
Observations			2400		
N clusters			30		
N periods			80		
R^2			0.20		

Note: Results of 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.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the "society" level.

attractive as inequality increases (see Appendix A, Table A.4).[15]

Again, the behavioural pattern in the treatments without an increase in inequality (T4,T5) is quite different. Table 3 shows that in both T4 and T5 where, respectively, inequality declined and was stable, stability was significantly higher in the periods 51–100 compared to periods 1—50. As expected, the largest increase in stability occurred under T4, the only treatment involving a decline in inequality after $t = 50$.

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

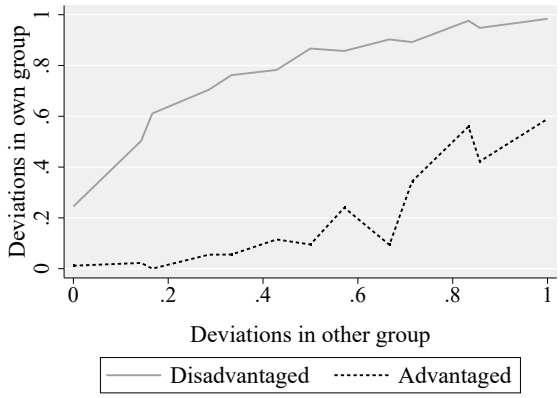
5.3 H3 — Instability is driven by the disadvantaged

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 switch to choosing A or they are a member of an advantaged group and switch to choosing B.[16] When comparing the probability of deviating across advantaged and disadvantaged groups, we find that the latter is on average 28% more likely to deviate (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 (dark grey, 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

[15] These results are robust to including demographic controls and to excluding only the first 10 periods (see Appendix A Tables A.2 and A.3).

[16] 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.

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.

share of individuals choosing A. Figure 5 shows that, in line with Hypothesis 3, deviations are initiated by members of disadvantaged group, who are also considerably more likely to deviate overall. Even when the share of deviators in the advantaged group is zero, the share of deviators in the disadvantaged group is above 20%.[17] In contrast, deviations by members of advantaged groups are very 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 increases rapidly.[18]

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

When examining deviations, we can also test whether — apart from being a member of the disadvantaged group — some individual characteristics make a participant more likely to deviate. We find that some characteristics do indeed predict deviations. Individuals who are more prosocial or more inequality averse, who have lower levels of risk aversion, and who have more negative affect responses are more likely to be in the vanguard of deviators (see Appendix A, Table A.5).[19] The link between negative affect and deviations provides support for the argument that relative grievances can lead to discontent and that such emotional factors, in turn, drive instability (Cramer, 2005; Blattman & Miguel, 2010).

[17] A t-test indicates that the first period in which an individual deviates is significantly earlier for those in the disadvantaged group compared to advantaged group members ($p < 0.001$).

[18] 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.

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

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

Hypothesis 4 states that deviations from a convention are more likely if the increase in inequality is accompanied by the disadvantaged becoming even more disadvantaged. To investigate this, we compare stability in T1 and T2 for $t > 50$. Table 4 presents three regressions. In column (1), an observation is a "society" in a period (51—100) and the dependent variable is SSI. In columns (2) and (3), an observation is a choice by an individual in a period (51—100), the dependent variables are, respectively, a dummy variable that equals 1 if A is chosen and zero otherwise and a dummy variable that equals 1 if the individual chooses to deviate from the established convention in their "society" and zero otherwise, and the models are linear probability models. In each case, the regressors are a full set of treatment indicators (T1 is the basis for comparison) and SSI at $t = 50$, i.e. before any changes in inequality occur and before the periods under analysis. SSI at $t = 50$ is included to control for *ex ante* cross-society variation in stability.

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. Table 4 reveals that the absolute decline in outcomes for disadvantaged that occurs under T1 but not T2 leads to weakly significantly more A choices ($p = 0.07$), weakly significantly more deviations from the established convention ($p = 0.08$) and insignificantly less social stability ($p = 0.22$). Thus, we have weak but consistent evidence in support of H4.

Result 4: *If an increase in inequality is combined with a deterioration in the absolute outcomes*

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

	(1) SSI	(2) A choices	(3) Deviations
Baseline = Incremental treatment (T1)			
Pure inequality (T2)	0.10	-0.05*	-0.19*
	(0.08)	(0.02)	(0.11)
Shock (T3)	0.06	-0.04	0.01
	(0.09)	(0.03)	(0.15)
Decreasing inequality (T5)	0.27***	-0.14***	-0.27***
	(0.08)	(0.02)	(0.10)
Control (T5)	0.14**	-0.06***	-0.24***
	(0.07)	(0.02)	(0.09)
SSI at t=50	0.57***	-0.07***	-0.28***
	(0.07)	(0.02)	(0.09)
Constant	0.06	0.72***	0.66***
	(0.07)	(0.02)	(0.09)
Observations	1500	20800	20800
N panels	30	416	416
N clusters	30	30	30
N periods	50	50	50
R^2	0.39	0.01	0.07

Note: Results of three random effects panel regressions for $t > 50$. The dependent variable in model (1) is the *SSI* in a given "society" and period and varies between 0 and 1. The dependent variable in model 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 model (3) is also binary, with a value of 0 if an individual does not deviate from an established convention in a given period and 1 if they do. Models (2) and (3) are linear probability models.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the "society" level.

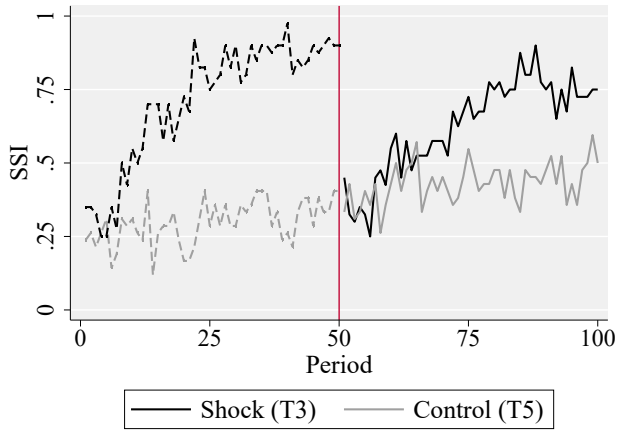
of the disadvantaged, the destabilising effect of the increasing inequality is exacerbated.

5.5 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, ultimately high stability under T3, while under T5 it follows a history of (already) high inequality and low stability. Turning to the subsequent $t > 50$, the figure indicates that stability is higher under T3 (t-test of null that average SSI is the same across treatments, $p < 0.001$). Under T3, after the shock, stability rises again quickly and converges to a level not far below pre-shock stability.[20] In contrast, in T5 stability remains low, while continuing to rise at a slow pace.[21] Regression analysis reveals that the upward trend in SSI for $t > 50$ is significantly steeper under T3 compared to T5 ($p = 0.03$, see Appendix A, Table A.7).

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

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 ($\Delta = 75$).

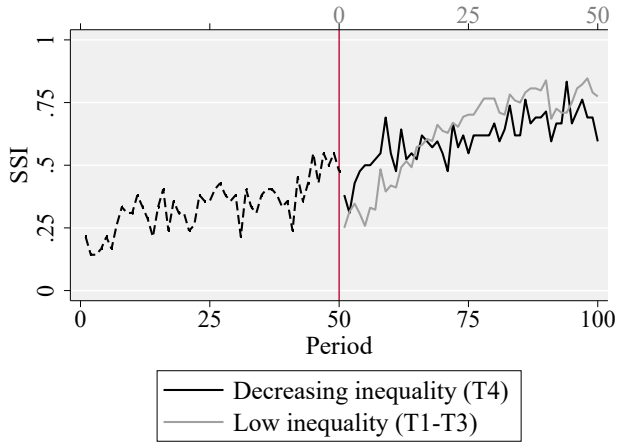
Next, we compare SSI in $t \leqslant 50$ under treatments with initially low inequality (T1–T3) and SSI in $t > 50$ under *Decreasing Inequality* (T4). Across all of these, inequality is low ($\Delta = 25$). However, under T1–T3 the participants enter the focal periods without any prior experience, whereas 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

[20] 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, $p<0.001$).
[21] Under T5, in $71 \leqslant t \leqslant 100$, SSI is 0.45.

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, $t \leqslant 50$, on the one hand, and T4, $t > 50$, on the other (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 and 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 and 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 and T2 and $51 \leqslant t \leqslant 52$ under T1, we find that SSI is 0.20 greater under the former and the difference is weakly 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 and T2 and $51 \leqslant t \leqslant 55$ under T1, SSI is 0.24 greater under the former and the difference becomes highly significant (Mann-Whitney test, p=0.001, N=36), despite inequality becoming even higher on average. 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 T1 and T2 (t-test, $p = 0.04$).

Result 5iii: *A sudden upward shock to a given level of inequality leads to greater instability than*

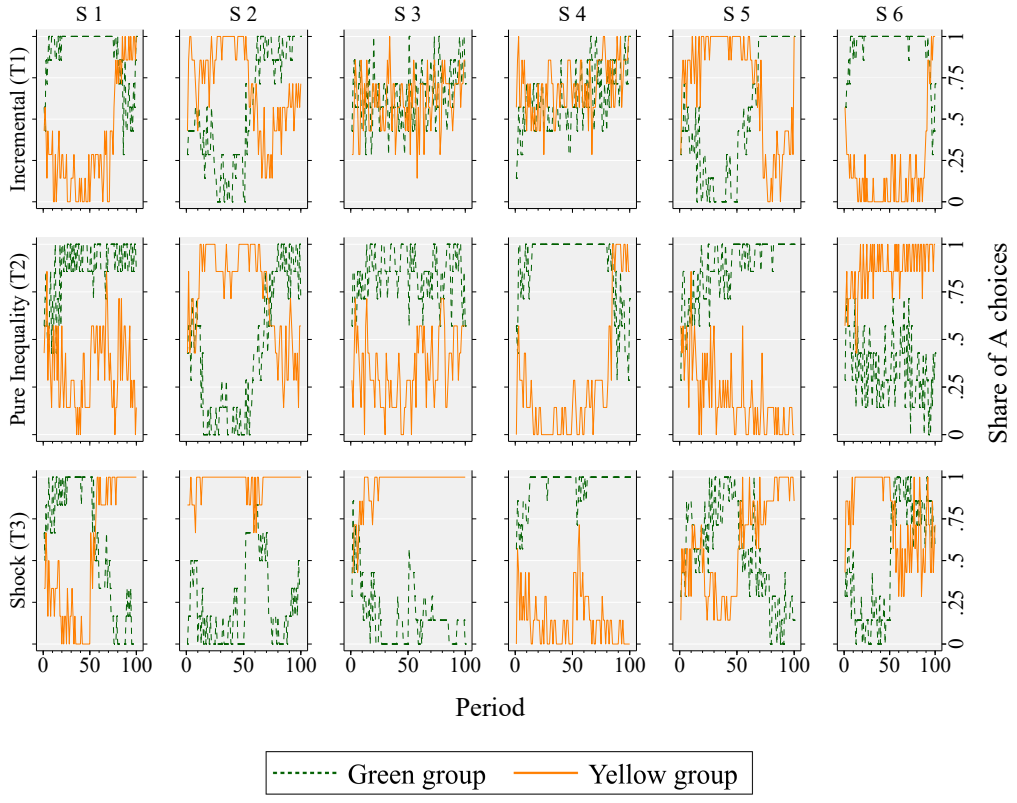
if the same level of inequality is reached gradually.

5.6 Convention reversals

SSI is a useful aggregate measure for stability. However, behind SSI there is substantial heterogeneity. Figure 8 plots the share of A choices by yellow and green group members in each "society" under T1–T3.[22] Our analysis thus far indicates that deviations from an established convention increase as inequality increases. However, Figure 8 reveals that, across "societies" (S), the deviations can lead to quite different outcomes.

Under *Shock* (T3) (bottom row), the increase in inequality at $t = 51$ is followed by multiple periods of chaos in one "society" (S6), a reversal of the previous convention in two (S1 and S5), and the re-emergence of the previous convention after a failed attempt to change it in three (S2, S3 and S4). Under *Incremental* (T1) and *Pure Inequality* (T2) we also see reversals (e.g. S2 in both treatments), however, the strength of the new convention is much lower and, as 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).

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



[22] See Appendix A, FigureA.6 for the behaviour in treatments with non-increasing inequality (T4 and T5).

6 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 experimental treatments the level of inequality associated with the conventions is exogenously increased and, in line with a simple theoretical model, this leads to significant social destabilisation. Further, and also in line with the theoretical model, the deviations from the conventions underlying the destabilisation are initiated by the disadvantaged group.

We also identify certain factors that intensify the destabilising effect of inequality. The destabilisation is more pronounced when not only the relative but also the absolute position of the disadvantaged group worsens, and when inequality increases suddenly rather than gradually. Finally, we find that past experiences of inequality and consequent stability or instability affect responses to current inequality levels and, hence, current social stability.

These findings provide unequivocal evidence of a casual relationship running from inequality to social instability, demonstrate the value of experiments as a tool for investigating factors that moderate this relationship, and provide a strong foundation for further investigation.

References

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.

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., 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.

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

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.

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*.

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. & 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.

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.

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.

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.

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. (2018). Social conflict and the evolution of unequal conventions. *Santa Fe Institute Working Paper*, 2018.

Inglehart, R. F. & Norris, P. (2016). Trump, brexit, and the rise of populism: Economic have-nots and cultural backlash. *HKS Working Paper*, (RWP16-026).

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.

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.

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.

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.

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.

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.

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.

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

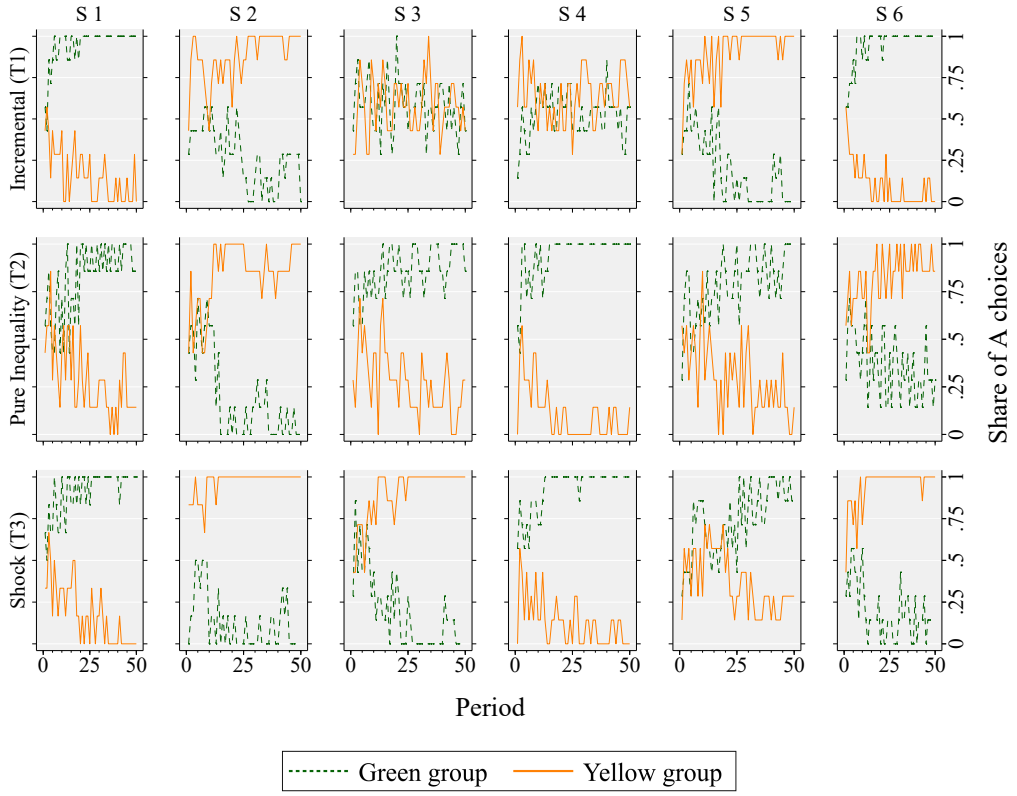
A.1 Emergence of an unequal convention

Emergence by "societies"

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, as shown in the main text. By looking at the individual "societies", we can confirm that stability is not achieved by taking turns between actions but that we observe 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, confirming that in most "societies", yellow and green groups specialise on different actions. Only in two "societies" play concurs with a mixed NE, which predicts for $\Delta = 25$ that participants choose A in 60% of periods.

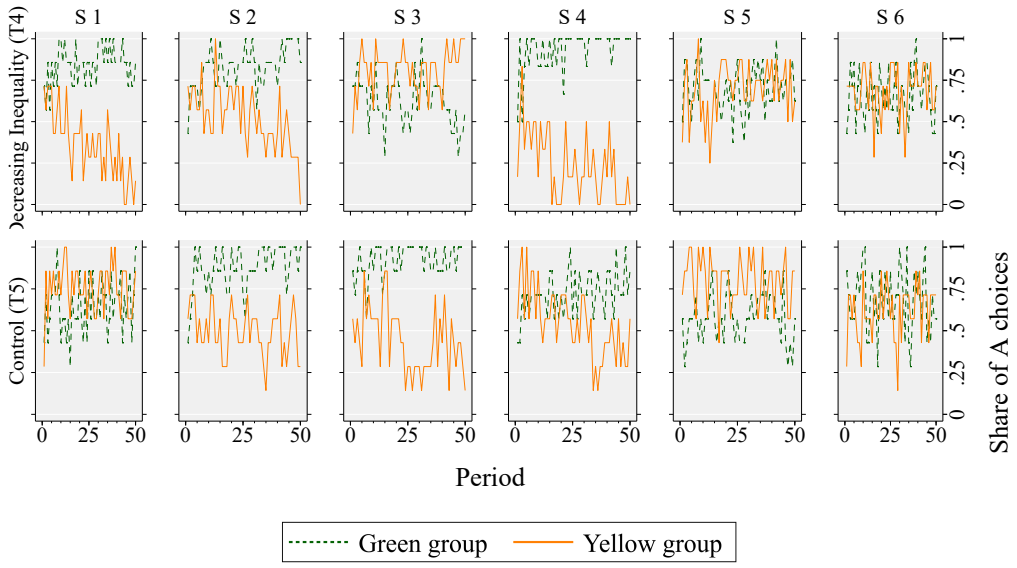
Under high inequality (T4, T5), by contrast, the emergence of a pure NE is less common. Figure A.2 shows that in T4 and T5 there is much more chaos with groups not achieving to coordinate on different actions in several "societies". Moreover, even if there is a tendency for

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



groups to specialise on different actions, it is much less pronounced than under low inequality.

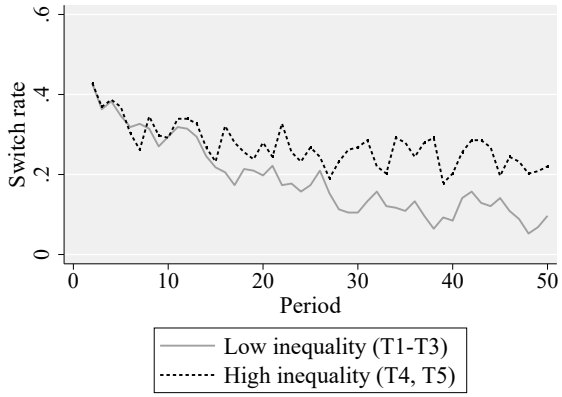
Figure A.2: Emergence under high inequality by "society"



Individual choices — reduction in switches over time

The emergence of an unequal convention is also reflected in a reduction of switches between actions A and B over the first 50 periods. 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 - controls

Table A.1 reports the coefficients for all control variables when estimating the effect of inequality on the stability of a convention in the first half of the experiment (see Table 2 in the

main text). As can be seen below, none of the demographic characteristics have a significant effect on the *SSI*.

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

	(1)
High inequality	-0.156***
	(0.056)
Period	0.011***
	(0.001)
High inequality x period	-0.007***
	(0.002)
Demographic controls	
Female	0.175
	(0.206)
Age	0.055
	(0.053)
Openness	0.020
	(0.097)
Conscientiousness	-0.087
	(0.115)
Extroversion	0.014
	(0.097)
Agreeableness	-0.009
	(0.149)
Neuroticism	-0.153
	(0.116)
Prosociality	-0.057
	(0.181)
Risk aversion	0.041
	(0.050)
Constant	0.378***
	(0.137)
Observations	1500
N clusters	30
N periods	50
R^2	0.41

Note: Results of three panel regressions 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). Demographic controls are the difference in each characteristic between the yellow and green group within a "society".

* 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 — robustness checks

Table A.2 shows that our findings that treatments with an increase in inequality (T1–T3) are characterised by a lower *SSI* in the second half of the experiment, while treatments without an increase are characterised by a higher *SSI* holds after controlling for demographic characteristics.

In the main text, we exclude the first 20 periods from the analysis of how stability changes after $t = 50$. The reason for this is that we want to test how an established convention is affected by increases in inequality. However, in the first 20 periods of the game, stability is still relatively low (average $SSI = 0.37$), indicating that a convention has not yet emerged. Using all periods before $t = 50$ can thus confound results, as it includes initial periods of miscoordination. Table A.3 shows that if we only exclude the first 10 periods as an initial

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

	(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	0.40	0.47**	0.70***	0.13	0.07
	(0.26)	(0.20)	(0.13)	(0.18)	(0.20)
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
Demographic controls					
Female			0.12		
			(0.22)		
Age			0.14***		
			(0.05)		
Openness			0.23***		
			(0.09)		
Conscientiousness			0.08		
			(0.10)		
Extroversion			0.00		
			(0.09)		
Agreeableness			-0.12		
			(0.12)		
Neuroticism			-0.10		
			(0.13)		
Prosociality			0.03		
			(0.19)		
Risk aversion			0.04		
			(0.05)		
Observations			2400		
N clusters			30		
N periods			80		
R^2			0.32		

Note: Results of a single random effects panel regression. The dependent variable is the *SSI* in a given society and period and varies between 0 and 1. *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 shows the results for a different treatment as the base category. The treatment indicators and their interaction with the *After* t = 50 dummy are omitted for reasons of conciseness. The p-values below the main table report differences in stability after t=50 across treatments. Demographic controls are the difference in each characteristic between the yellow and green group within a "society". The first 20 periods are excluded from the analysis, as it takes time for the convention to emerge and the first periods still exhibit a large degree of switching behaviour (average SSI = 0.37).

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the "society" level.

phase of emergence, where the average SSI = 0.29, all coefficients keep the same sign and all but the coefficient for the *Pure Inequality* (T2) treatment remain significant.

Probability of choosing A before and after t=50

The change in behaviour between the first and the second part of the experiment also becomes clear when looking at the individual choice data. 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.4 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.4 shows the results for taking a different treatment as the base category to present the effect for each treatment. As can be seen from Table A.4, the individual probability of

Table A.3: Changes in stability after $t = 50$ — excluding first 10 periods

	(T1) Incremental	(T2) Pure inequality	(T3) Shock	(T4) Decreasing inequality	(T5) Control
After t=50	-0.18***	-0.10	-0.18**	0.23***	0.12**
	(0.06)	(0.09)	(0.09)	(0.07)	(0.05)
Constant	0.61***	0.69***	0.81***	0.38***	0.31***
	(0.13)	(0.06)	(0.07)	(0.09)	(0.05)
N observations			2700		
N clusters			30		
N periods			90		
R^2			0.19		

Note: Results of a single random effects panel regression. The dependent variable is the *SSI* in a given "society" and period and varies between 0 and 1. *After* $t = 50$ is a dummy variable that takes the value 0 for the 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 shows the results for a different treatment as the base category. Treatment differences are omitted from the table for conciseness. The first 10 periods are excluded, as it takes time for the convention to emerge and the first periods still exhibit a large degree of switching behaviour (average SSI = 0.29).

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the "society" level.

choosing A differs significantly between the first and the second 50 periods for all treatments. In all three treatments where inequality increases (T1–T3), we see a significant increase in the probability of choosing A after $t = 50$. This shows again that stability is negatively affected by increasing inequality. In treatments where inequality is not increasing (T4, T5), by contrast, the probability of choosing A decreases after $t = 50$, indicating an increase in stability.

Table A.4: 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***	0.07***	0.07***	-0.11***	-0.04**
	(0.01)	(0.02)	(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)
N observations			41600		
N subjects			416		
N clusters			30		
N periods			100		
R^2			0.01		

Note: Results 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 the first 50 periods of the experiment and 1 for periods 51–100. The regression controls for treatment indicators and their interaction with *After* $t = 50$. Each column shows the results for a different treatment as the base category. Treatment differences are omitted from the table for conciseness. The regression controls for group size.

* 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.5 explores whether certain individual characteristics predict the probability to initialise deviations from an established convention. Initialising a deviation is thereby defined as choosing an action that goes against the established convention, while the majority of the own group is still following it. An individual is thus classified as an initiator if the majority

Table A.5: Probability of deviating from a convention ($t > 50$)

	(1) Pooled	(2) Incremental	(3) Pure inequality	(4) Shock
SSI at t=50	-0.169***	-0.147***	0.029	-0.375***
	(0.034)	(0.039)	(0.028)	(0.112)
Female	-0.001	0.005	-0.029	0.019
	(0.029)	(0.039)	(0.027)	(0.071)
Age	-0.003	-0.001	-0.003	-0.005
	(0.003)	(0.003)	(0.005)	(0.008)
Prosocial	0.042*	0.032*	0.047	0.030
	(0.023)	(0.017)	(0.031)	(0.044)
Risk aversion	-0.014***	-0.009	-0.018***	-0.022***
	(0.004)	(0.008)	(0.006)	(0.006)
Big 5				
Openness	0.000	0.011	-0.013	-0.000
	(0.012)	(0.015)	(0.014)	(0.021)
Conscientiousness	0.009	-0.016	0.011	0.015
	(0.011)	(0.022)	(0.010)	(0.018)
Extroversion	0.013	0.032	0.002	0.017
	(0.011)	(0.020)	(0.013)	(0.018)
Agreeableness	0.004	-0.022***	0.032***	-0.002
	(0.010)	(0.007)	(0.003)	(0.025)
Neuroticism	-0.003	0.008	-0.002	-0.018**
	(0.007)	(0.013)	(0.011)	(0.008)
Negative affect	0.002***	0.002**	0.001	0.002*
	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.263	-0.072	0.211	0.206
	(0.315)	(0.274)	(0.301)	(0.419)
Observations	20800	20800	20800	20800
N subjects	416	416	416	416
N clusters	30	30	30	30
N periods	50	50	50	50
R^2	0.07	0.09	0.09	0.09

Note: Results of random effects panel regressions for $t > 50$. The dependent variable is a binary variable that takes the value of 1 if an individual chooses action A despite the majority of their group choosing B or vice versa and 0 otherwise. Analysis is restricted to periods larger than 50, as the first 50 periods are treated as the emergence stage of a convention. All regressions control for treatment indicators and their interaction with personal characteristics. For reasons of conciseness, we only report the coefficients on personal characteristics taking T1–T3 as the baseline. Regressions control for group size.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the "society" level.

of their group still chooses B (A), but they decide to choose A (B). As the first 50 periods are treated as the emergence stage of a convention after which one group specialises in action A and the other in action B, the analysis is restricted to periods larger than 50.

Table A.5 shows the results of random effects panel regressions with a binary variable that takes the value 1 if an individual deviates while the majority in their group still follows the convention and 0 otherwise as the dependent variable and different individual characteristics as the explanatory variables. Column (1) reports results for a pooled regression across T1–T3, while columns (2) – (4) show results by treatment. 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 intuitive, as in these cases the risk of receiving zero is particularly high. Moreover, individuals who are more risk averse and report a higher negative affect are more likely to

Table A.6: Probability of deviating from a convention (t > 50)

	(1) Pooled	(2) Incremental	(3) Pure inequality	(4) Shock
SSI at t=50	-0.166***	-0.143***	0.007	-0.380***
	(0.031)	(0.038)	(0.031)	(0.122)
Female	0.002	0.003	-0.032	0.026
	(0.028)	(0.037)	(0.027)	(0.066)
Age	-0.003	-0.001	-0.002	-0.004
	(0.002)	(0.003)	(0.004)	(0.008)
Inequality averse	0.083***	0.078**	-0.043	0.099
	(0.028)	(0.033)	(0.031)	(0.073)
Risk aversion	-0.012***	-0.007	-0.018***	-0.021***
	(0.003)	(0.007)	(0.004)	(0.005)
Big 5				
Openness	-0.000	0.009	-0.016	-0.001
	(0.012)	(0.013)	(0.014)	(0.022)
Conscientiousness	0.008	-0.012	0.008	0.010
	(0.011)	(0.022)	(0.009)	(0.017)
Extroversion	0.014	0.032*	0.001	0.015
	(0.011)	(0.019)	(0.013)	(0.019)
Agreeableness	0.007	-0.022***	0.037***	0.006
	(0.010)	(0.007)	(0.004)	(0.030)
Neuroticism	-0.002	0.008	-0.001	-0.015*
	(0.007)	(0.013)	(0.012)	(0.008)
Negative affect	0.002***	0.002**	0.001	0.002*
	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.284	-0.038	0.328	0.219
	(0.292)	(0.260)	(0.289)	(0.385)
Observations	20800	20800	20800	20800
N subjects	416	416	416	416
N clusters	30	30	30	30
N periods	50	50	50	50
R^2	0.07	0.09	0.09	0.09

Note: Results of random effects panel regressions for t > 50. The dependent variable is a binary variable that takes the value of 1 if an individual chooses action A despite the majority of their group choosing B or vice versa and 0 otherwise. Analysis is restricted to periods larger than 50, as the first 50 periods are treated as the emergence stage of a convention. All regressions control for treatment indicators and their interaction with personal characteristics. For reasons of conciseness, we only report the coefficients on personal characteristics taking T1–T3 as the baseline. Regressions control for group size.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the "society" level.

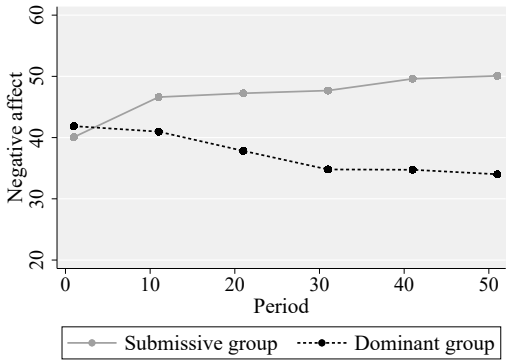
deviate. There is also some evidence that prosocial individuals are more likely to deviate (positive in all specifications, significant only in the pooled regression and in the one for T1). Using the SVO angle methodology suggested by Murphy et al. (2011), we find that 57% of participants can be classified as individualistic (i.e. maximising their own payoffs), while 43% can be classified as prosocial. By deviating from the convention, individuals are definitely not maximising their individual earnings which is in line with prosocial types being more likely to deviate. However, the interpretation is not completely straightforward as prosocial types can be either motivated by minimising inequality or maximising joint gains. When examining the primary items of the SVO scale, Murphy et al. (2011) show that a decision maker who is perfectly inequality averse would have an SVO angle of 37.4. Using this metric, we see in Table A.6 that participants who are perfectly inequality averse are significantly more likely to

deviate in the pooled regression and the regression for T1.[23]

Affect responses

Figure A.4 shows how negative affect develops over time across advantaged and disadvantaged groups. With growing stability but also growing inter-group inequality, we observe that a gap in affect emerges. advantaged groups show more positive emotional affect over time, while disadvantaged groups show more negative affect, leading to a significant difference over the first 50 periods (t-test, $p < 0.001$).

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



Note: Negative affect is measured on a scale from 0 to 100, with higher values indicating more negative emotions. Negative affect is measured every tenth round.

Figure A.5 shows how negative affect develops for advantaged and disadvantaged groups over the whole experiment for the different treatments. We can see that overall the gap between advantaged and disadvantaged groups persists throughout the experiment. Moreover, affect depends on the level of inequality. An increase in inequality (T1, T2, T3) initially leads to an increase in negative affect, while a decrease in inequality (T4) results in more positive affect responses.[24]

A.3 History dependence

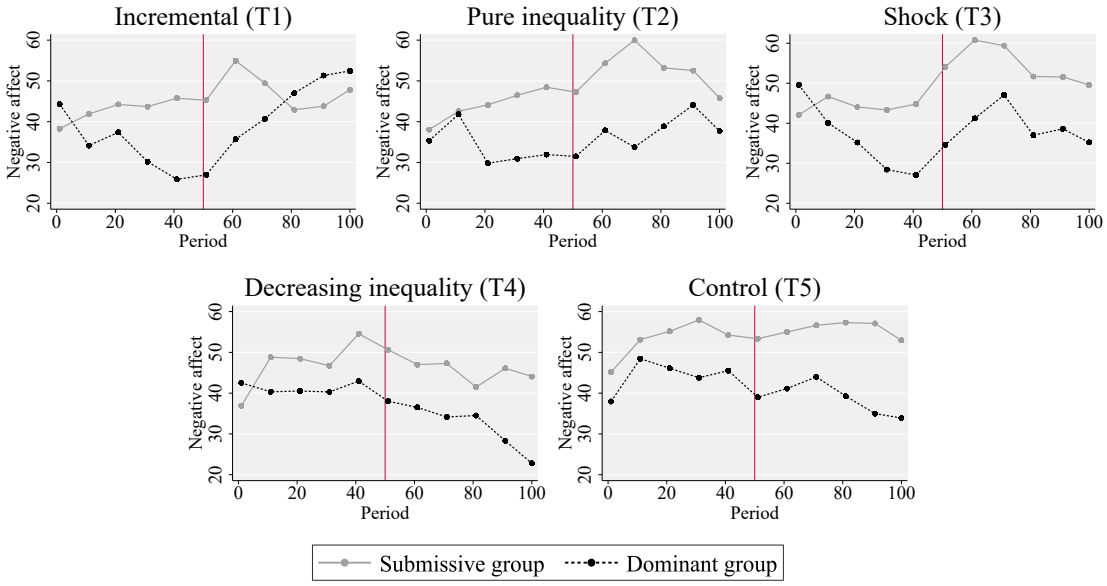
Coordination under high inequality given different histories

In both the *Shock* (T3) and the *Control* (T5) treatment, participants face high inequality ($\Delta = 75$) for $t > 50$. The difference is however that in T3, participants experienced high stability in the first 50 periods while participants in the *Control* (T5) treatment did not. In Table A.7, we regress the SSI as the dependent variable on a time trend, treatment indicators, and their

[23] We acknowledge that using the exact SVO angle is only a proxy. We did not administer the full SVO test including the secondary items that allow to differentiate between different prosocial motives due to time constraints.

[24] At later periods, we observe more positive responses in T2 and T3 and even a closure of the gap between disadvantaged and advantaged groups in T1. This has to do with the specific dynamics of each "society" and the existence of reversals (see Section 5.6).

Figure A.5: Affect responses across treatments



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 coordination under high inequality ($\Delta = 75$)

	(1)	(2)	(3)
Baseline = Shock (T2)			
Control (T5)	0.385	0.248	
	(0.236)	(0.296)	
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	-0.287	0.364***
	(0.213)	(0.270)	(0.088)
Demographic controls	No	Yes	FE
N observations	1500	1500	1500
N clusters	30	30	30
N periods	50	50	50
R^2	0.13	0.33	0.05

Note: Results 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 the share of female participants as well as in average measures of Big 5, risk aversion, prosociality 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	T1–T3 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)			
Period	0.005***			
	(0.002)			
Incremental (T1) x period	0.003			
	(0.003)			
Pure inequality (T2) x period	0.007***			
	(0.002)			
Shock (T3) x period	0.008***			
	(0.002)			
Low inequality (T1-T3)		-0.132	-0.111	
		(0.140)	(0.142)	
Time trend		0.005***	0.005***	0.005***
		(0.002)	(0.002)	(0.002)
Low inequality (T1-T3) x period		0.006***	0.006***	0.006***
		(0.002)	(0.002)	(0.002)
Constant	0.488***	0.488***	0.388*	0.367***
	(0.135)	(0.135)	(0.216)	(0.033)
Demographic controls	No	No	Yes	FE
N observations	1500	1500	1500	1500
N clusters	30	30	30	30
N periods	50	50	50	50
R^2	0.23	0.20	0.24	0.16

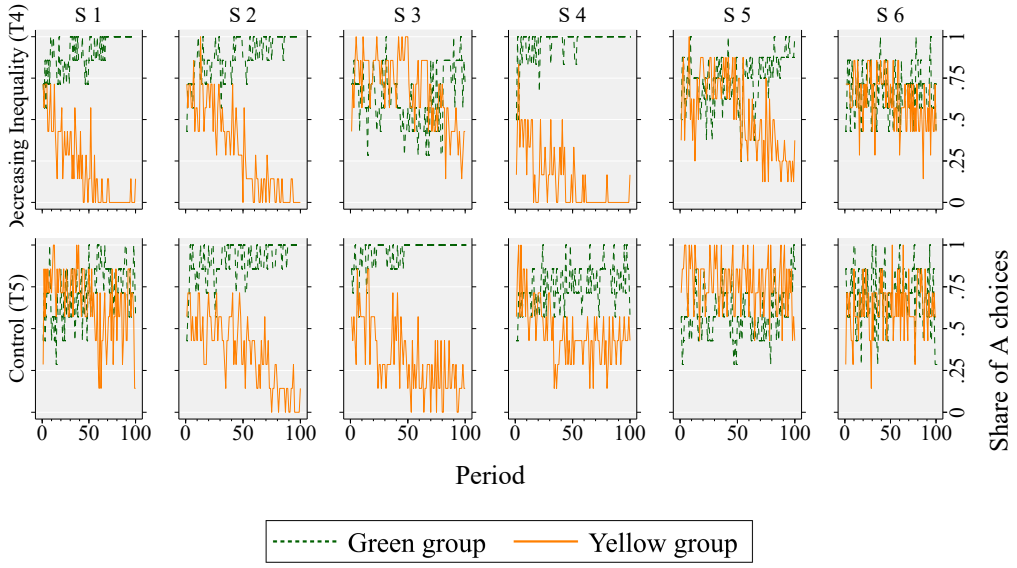
Note: Results of four panel regressions. The dependent variable is the *SSI* in a given "society" and period and varies between 0 and 1. All regressions include a treatment indicator for T5. As the relevant comparison is between T4 and T1–T3, it is omitted from the table. Demographic controls include the difference in the share of female participants as well as in average measures of Big 5, risk aversion, prosociality 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. As inequality is either decreasing or stable, it is not surprising that we do not see any attempts at reversals in these treatments.

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



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

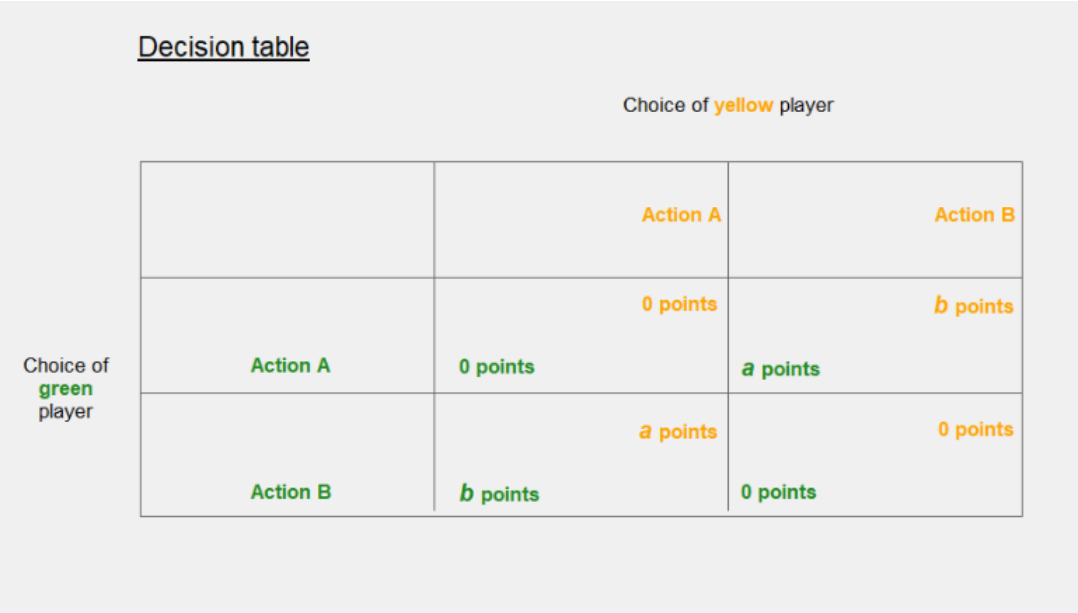
At the beginning of the experiment, **you and each of the other participants in the session will be randomly assigned to a group** by the draw of a coloured ball (with probability 0.5). Half of the participants will be assigned to a **green** group, the other half to a **yellow** group. **Each participant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



The Decision table can be read as follows:

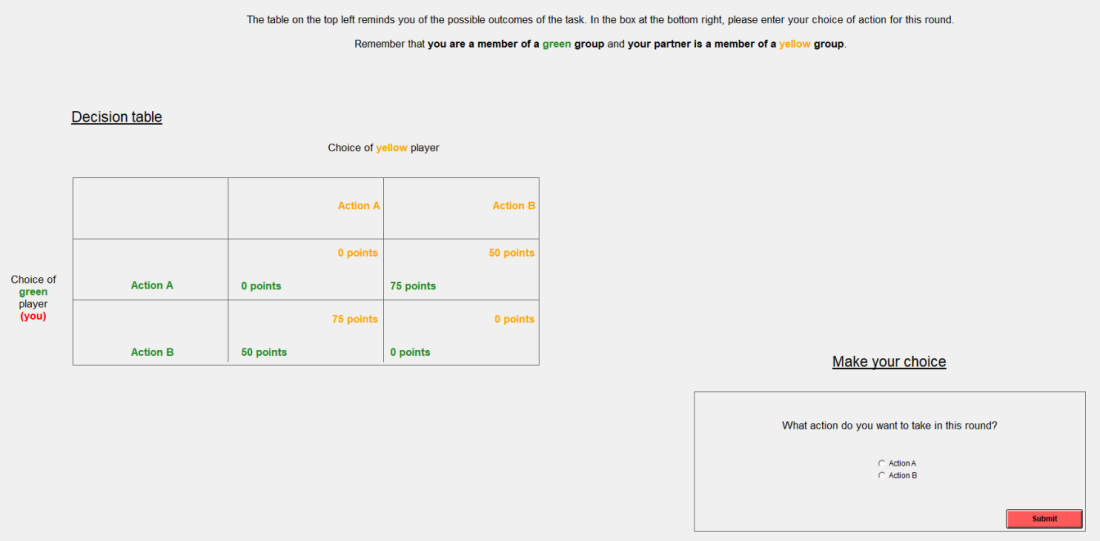
If the **green** player chooses A and the **yellow** player chooses A, then both players receive 0 points. If the **green** player chooses B and the **yellow** player chooses B, then both players receive 0 points.

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.[25] 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**.

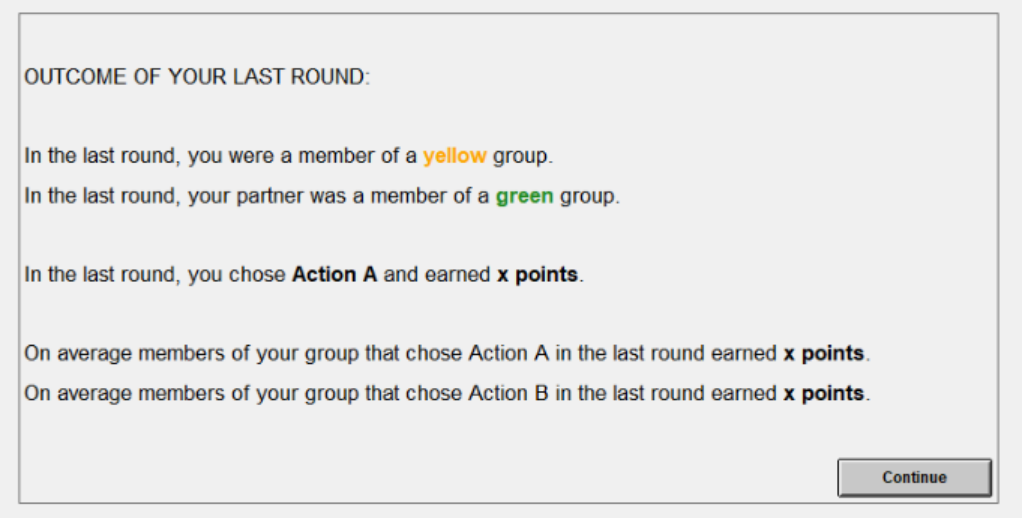


After each round, before you start playing the next round, you will be given the following information, summarising the last round:

- 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.

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

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



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.

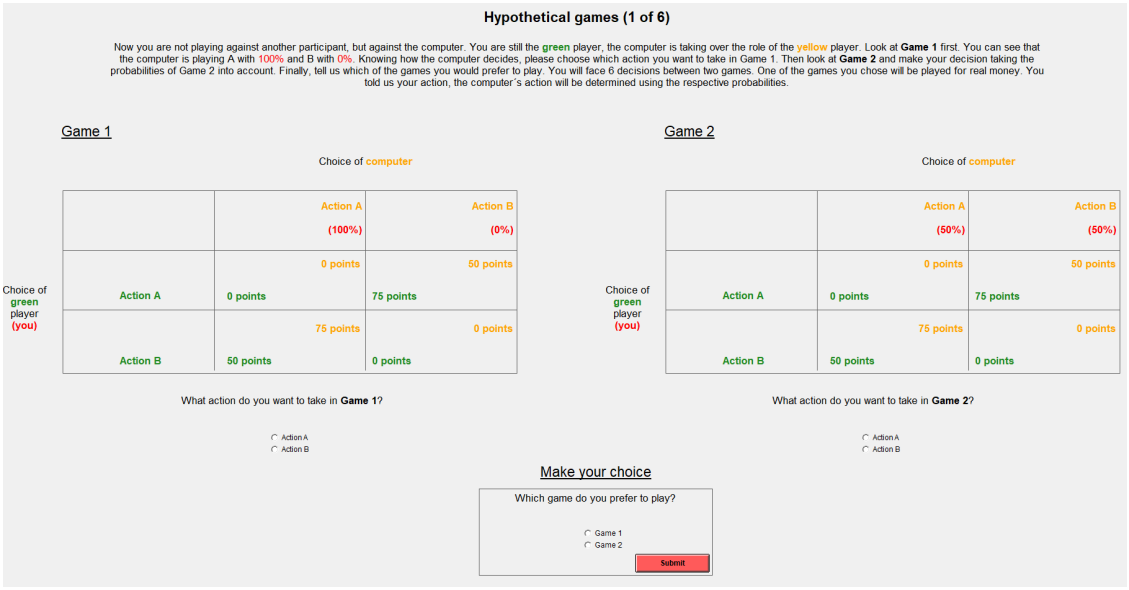
If you have completely understood the instructions, please answer the control questions on screen. If you have any questions please raise your hand and an experimenter will come to you.

B.2 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.



B.3 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.