nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-022-01440-5

Group identities can undermine social tipping after intervention

Sönke Ehret[1,9] ✉, Sara M. Constantino[2,3,4,9] ✉, Elke U. Weber[2,5,6], Charles Efferson[1,10] ✉ and Sonja Vogt[1,7,8,10] ✉

Social tipping can accelerate behaviour change consistent with policy objectives in diverse domains from social justice to climate change. Hypothetically, however, group identities might undermine tipping in ways that policymakers do not anticipate. To examine this, we implemented an experiment around the 2020 US federal elections. The participants faced consistent incentives to coordinate their choices. Once the participants had established a coordination norm, an intervention created pressure to tip to a new norm. Our control treatment used neutral labels for choices. Our identity treatment used partisan political images. This simple pay-off-irrelevant relabelling generated extreme differences. The control groups developed norms slowly before intervention but transitioned to new norms rapidly after intervention. The identity groups developed norms rapidly before intervention but persisted in a state of costly disagreement after intervention. Tipping was powerful but unreliable. It supported striking cultural changes when choice and identity were unlinked, but even a trivial link destroyed tipping entirely.

Social change can stagnate for a long time and then unfold suddenly and unexpectedly. Foot binding persisted in China for centuries, only to disappear in a generation[1]. In the United States, long-standing hostility towards same-sex marriage unravelled in a few years[2]. Germany began subsidizing solar panels in the 1990s, but initial adoption was slow. Interactions among friends and neighbours accelerated the spread of the technology, and by 2016, Germany was generating more solar power per capita than any other country[3].

This kind of punctuated cultural change occurs when a population tips from one social norm to another[1,4]. Social tipping is a flamboyant form of cultural evolution in which many people suddenly change how they behave and how they think about the behaviour of others[5]. Foot binding provides a canonical example. Many families abandoned the practice in a short period. A family doing so understood that other families were also abandoning the practice and thus would probably not insist on women with bound feet as future wives for their sons. This change in beliefs about others created a positive feedback that accelerated abandonment[1,6].

Social tipping has generated enormous interest as a way to trigger widespread behaviour change[7] in many domains related to public health[8,9], social justice[10,11], resource conservation[12,13] and climate change[14–17]. Given such widespread interest, researchers and practitioners have a responsibility to investigate the conditions that support or undermine tipping[18–20]. Accordingly, we examine group identities as a mechanism hypothesized to interfere with tipping[21–23].

Proof of concept exists for tipping. Observational data show that cultural evolutionary processes support multiple norms, and punctuated cultural change certainly occurs[1,24–28]. Experimental studies have also demonstrated that interventions can spark rapid transitions from one norm to another[5,29]. Nonetheless, studies of tipping around gender-based violence[11,30–32], political revolutions[33] and lab experiments[5] suggest important limits on our ability to identify when tipping is possible and how to maximize the chance of tipping[20]. The associated risk is that policymakers misinvest in poorly designed or pointless efforts to activate tipping. However, when tipping is possible, it holds clear policy implications[7].

The same mechanisms contribute to the slow and fast phases of punctuated cultural evolution[1,24]. A tendency to conform and incentives to coordinate choices can both motivate people to behave like others. If a behaviour is rare, conformity, coordination or a mix of both keep it rare. This is the slow phase. If the behaviour becomes sufficiently common, for whatever reason, conformity or coordination switch from obstructing to accelerating change. This initiates the fast phase. Once sufficient change occurs, the population crosses a tipping point and quickly transitions to a new cultural regime without further interference.

Policymakers seeking rapid social change aim to trigger this dynamic. When conformity or coordination supports a status quo norm inconsistent with policy objectives, policymakers can promote an alternative norm by incentivizing their preferred behaviour in a subset of the population only[22,29]. Alternative norms might include the abandonment of female genital cutting[34], giving up smoking[35], driving electric cars[7], not aborting fetuses because they are female[36] and eating chicken instead of pangolin[13]. Interventions can take many forms, ranging from taxes and subsidies[5] to entertaining narratives with educational messaging[37–39].

If enough people exposed to the intervention change their behaviour, conformity or coordination can switch from supporting the status quo to supporting the policymaker's alternative. Individuals who do not change their behaviour as a direct consequence of the intervention see others changing behaviour and conclude that an alternative has become preferable to the status quo. When this happens, the population should complete the transition to a new

[1]Faculty of Business and Economics, University of Lausanne, Lausanne, Switzerland. [2]School of Public and International Affairs, Princeton University, Princeton, NJ, USA. [3]School of Public Policy and Urban Affairs, Northeastern University, Boston, MA, USA. [4]Department of Psychology, Northeastern University, Boston, MA, USA. [5]Andlinger Center for Energy and the Environment, Princeton University, Princeton, NJ, USA. [6]Department of Psychology, Princeton University, Princeton, NJ, USA. [7]Centre for Development and Environment, University of Bern, Bern, Switzerland. [8]Nuffield College, University of Oxford, Oxford, UK. [9]These authors contributed equally: Sönke Ehret, Sara M. Constantino. [10]These authors jointly supervised this work: Charles Efferson, Sonja Vogt. ✉e-mail: sonkeklaus.ehret@unil.ch; sara.constantino@gmail.com; charles.efferson@unil.ch; sonja.vogt@unil.ch

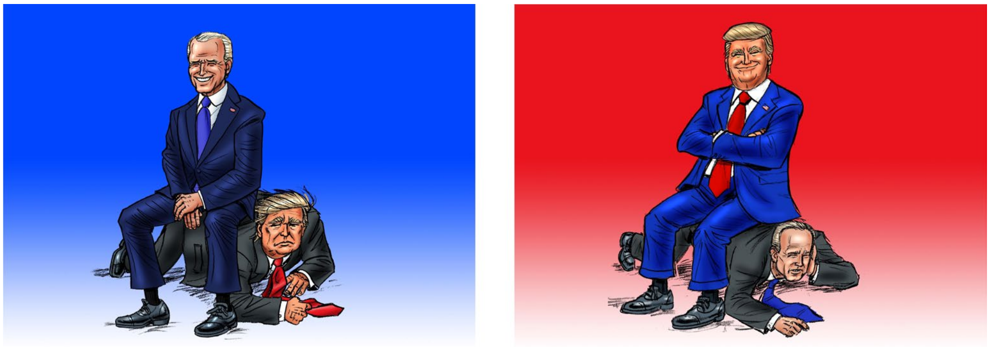

Fig. 1 | The two images used to label the buttons in the identity treatment. Instead of clicking on a button labelled with @ or #, as in the neutral treatment, participants in the identity treatment had to choose by clicking on one of two buttons with these images embedded in the buttons themselves (Supplementary Figs. 2 and 3).

norm quickly, even without additional input from the policymaker. Behaviour change is partly exogenous, because some people change their behaviour due to exposure to the intervention, and partly endogenous, because some people change their behaviour due to conformity or coordination after the population crosses the tipping point. Put differently, the direct effects of the intervention spill over and indirectly influence those never exposed to the intervention or those who were exposed but did not initially respond[22,40]. Spillovers are a standardized measure of how popular the policymaker's alternative eventually becomes (Methods).

Spillovers imply that endogenous social forces can produce substantial behaviour change, and tipping thus offers the hope of using the policymaker's limited resources efficiently. This possibility is important because many contemporary social problems are daunting in scale[7,13]. Moreover, promoting social change is an attempt to engineer culture, and even policymakers with the purest of intentions cannot escape the practical and ethical dilemmas this implies. To the extent that tipping produces change, change originates from within the population. The hope is that endogenous change moderates concerns about paternalistic intrusions in a society's culture and the associated risk of backlash[10,22].

The challenge is that conformity and coordination incentives rarely operate in isolation. Rather, they interact with other motives that could undermine tipping[5,20,41]. These motives often centre on group identities and the symbolic markers people use to display group affiliation[42]. People experience positive affect towards ingroup markers and the values these markers represent, together with negative affect towards outgroup markers and their associated values[43,44]. When these affective responses are linked to policy-relevant behaviours, group identities might obstruct tipping that would otherwise occur[22].

We thus hypothesized that group identities represent a form of heterogeneity that can undermine tipping in specific settings. Broadly speaking, heterogeneity may or may not hinder tipping. The details are all-important. The distribution of preferences in the population, heterogeneity in how people respond to information about others and heterogeneous social networks can all interfere with tipping, but they do not necessarily do so[5,22,40,45–48]. People differ in many dimensions critical to behaviour change. These differences interact with the policymaker's choices to shape the potential for tipping[22,40,47,48] as a way to effect change. Heterogeneity based on group identities holds particular interest because human psychology has a strong parochial streak, arguably based on an evolutionary history in which affiliation with a group helped people learn from others[42,49] and cooperate within the group to compete against other groups[44,50,51].

Whatever the past function of group identification, models suggest that once in place it can have an outsize influence on cultural evolution[21–23]. Outgroup aversion[21] represents a special challenge for the policymaker. Outgroup aversion means that groups define themselves in part by differentiating themselves from other groups. If they use policy-relevant behaviours to do so, outgroup aversion may disrupt the efforts of a policymaker promoting a single behaviour for the entire population[22]. To illustrate, imagine a population divided into two groups. The members of one group value low-emission transport. The members of the other group 'roll coal', which means they modify their vehicles to increase carbon emissions to differentiate themselves from the first group[21]. For people in the second group, a policymaker who wants the entire society to tip to low-emission transport may represent an existential threat to their shared group identity. In extreme cases, the policymaker's efforts could even strengthen the tie between group identity and the behaviour in question[52]. In our example, this would mean that the policymaker's efforts increase the value of pollution for some people and solidify their resistance to change. Whatever the details, the policymaker promotes a behaviour that is inconsistent with the group identities of at least some people, and these identities are subject to conspicuous and sufficiently strong outgroup aversion[21,53].

To examine this kind of dynamic, we implemented an incentivized online experiment around the 2020 election for US president. A US sample participated in repeated play of coordination games. We designed our control treatment to be maximally favourable for tipping. The experimental treatment was identical with one exception. We relabelled the choice options with images designed to activate partisan political identities (Fig. 1). Partisan loyalties provide an important component of identity in contemporary US politics[54,55], and party affiliation has become increasingly sectarian[56,57]. Crucially, our partisan images had no explicit material consequences. They simply provided a labelling system to distinguish the choice options (Supplementary Figs. 2 and 3), and in this sense our treatment manipulation was pay-off-irrelevant.

Each session had the following structure regardless of treatment. We formed an experimental group of either all Republicans or all Democrats. The participants repeatedly played a coordination game with two choice options (Methods). In each period, each participant was randomly matched with another participant in the same experimental group. Monetary pay-offs simply favoured coordinating; they did not favour coordinating on a specific option (Table 1, pre-intervention (all)). The participants were anonymous, unable to communicate and had no prior information about the people with whom they were playing. To coordinate consistently, they had to establish a norm via repeated play with feedback (Methods).

Once an experimental group had established a 'status quo' norm, which meant that one of the two choice options had become sufficiently common, we implemented an intervention (Methods). We targeted a random sample of participants and changed their pay-offs

Table 1 | Participant pay-offs

	Pre-intervention (all)		Post-intervention (T)		Post-intervention (NT)	
	SQ	Alt	SQ	Alt	SQ	Alt
SQ	200	50	200	50	200	50
Alt	50	200	350	350	50	200

The matrices show row player pay-offs in points as a function of row and column choices. The status quo (SQ) choice was the choice associated with the norm that emerged during the pre-intervention phase. Given a status quo choice, the alternative (Alt) was simply the other choice option. Pay-offs were the same for everyone in the pre-intervention phase and did not favour any particular equilibrium. Because status quo norms evolved throughout the pre-intervention phase, we refer to choices for this phase as 'status quo' and 'alternative' from an ex post perspective (that is, after the experimental group had settled on a status quo norm). The intervention encouraged behaviour change by introducing new pay-offs that favoured the alternative option among targeted (T) players, regardless of the partner's choice. These pay-offs held for the entire post-intervention phase. Non-targeted (NT) players retained their original pay-offs post-intervention.

to favour changing from the status quo choice, whatever this may have been, to the other choice option, which we call the 'alternative' (Table 1, post-intervention (targeted); Supplementary Section 4.2). The non-targeted participants retained their original pay-offs (Table 1, post-intervention (non-targeted)). The participants then continued playing repeatedly under this new incentive structure.

The experiment consisted of two treatments randomly assigned to the experimental groups. In the neutral treatment, the choice options in the game were labelled with neutral symbols, @ and #. In the identity treatment, the choice options were labelled with two partisan images (Fig. 1). The labels were simply embedded in the buttons that the participants had to press to indicate a choice while playing the game (Supplementary Figs. 2 and 3). The labels had no other role.

Our intervention created heterogeneity in material incentives. After intervention, the targeted participants faced material incentives that favoured behaviour change in the precise sense that the alternative was dominant for self-regarding players (Table 1, post-intervention (targeted)). The non-targeted participants faced material incentives that simply favoured, for self-regarding players, coordinating on either option. The material incentives were heterogeneous, but in a way that supported behaviour change. Moreover, behaviour change after intervention was socially beneficial in the narrow Pareto sense[58]. For example, if everyone in an experimental group were to adopt the alternative, no one would experience a decline in monetary pay-offs, and some would experience a strict increase.

Crucially, however, people do not simply care about their own material pay-offs[59–61]. Some people find inequality aversive[62], and our intervention created two classes of player with the potential for systematic inequalities. Targeted players after intervention could earn large pay-offs simply by choosing the alternative. Non-targeted players could earn small or intermediate pay-offs depending on whether they coordinated with their partners. Thus, if an experimental group were to tip fully to the alternative, the targeted players would experience persistent inequality to their advantage and the non-targeted players persistent inequality to their disadvantage[62]. Anticipating aversion to such outcomes might affect behaviour change among one or both classes of player.

Our design controlled for this possibility by always using the same pay-off matrices regardless of treatment. This validates treatment comparisons even if inequality aversion was affecting behaviour. Moreover, our design also captured a characteristic of many interventions. Any intervention that does not reach everyone in a population creates potential inequalities that did not previously exist. To attenuate such inequalities, for example, many studies in economic development randomize the introduction of an intervention to different points in time while attempting to intervene everywhere eventually[63,64].

Aside from material concerns, which encompass both self-regard and aversion to inequality, our political labels added another currency of potential value, but they did so only in our identity treatment. Assuming that the labels activated partisan identities, the effects should have depended on how the participants traded money against identity concerns. First, imagine that money dominated identity concerns for everyone. Whatever the degree of behaviour change in our neutral treatment, behaviour change should have been exactly the same in our identity treatment because the material incentives were the same in both treatments. Second, imagine that identity concerns dominated money for everyone. No one should have changed behaviour in the identity treatment because the intervention incentivized change via money only. Finally, imagine that people traded money against identity concerns in heterogeneous ways. Heterogeneity implies that some players in the identity sessions might have changed behaviour, while others might not. The result might have been no norm at all after intervention.

We first present the results from a pre-registered analysis of spillovers (Methods) that reflects a central concern from the policymaker's perspective. Spillovers[22] are a normalized measure of how common the policymaker's alternative is in the long run while accounting for the size of the policymaker's intervention (Methods). Spillovers do not explicitly account for choice dynamics; they quantify the final outcome net the policymaker's effort. Negative spillovers are in $[-1, 0)$ and arise when the final proportion choosing the alternative in an experimental group is less than the proportional size of the intervention. Non-negative spillovers are in $[0, 1]$ and arise when the final proportion choosing the alternative behaviour is equal to or larger than the intervention.

We then present the results from pre-registered analyses of individual decision-making (Methods). Finally, we present additional results, based largely on exploratory analyses, that investigate the precise mechanisms at work in our experiment. These analyses compare behaviour before and after the election, and they examine the role of attitudes towards equality and inequality. Additional analyses address whether identity concerns among our participants were based on identities already in place when the participants began the experiment or on identities that emerged within the context of the experiment itself.

Results

Spillovers and individual choice. In the neutral sessions, some experimental groups converged on a status quo norm of choosing @, while others converged on #. We have no statistical evidence that the status quo norm was related to the shared political affiliations of participants in sessions together ($\chi^2(1, N = 35) = 2.08$, $P = 0.15$). In the identity sessions, although the same kind of flexibility was possible, all Republican sessions converged on triumphant Trump, and all Democrat sessions converged on triumphant Biden. With the status quo established in a session, the spillover is a normalized measure of how common the alternative choice was at the end of the session (Methods).

Spillovers were large and significantly positive in our neutral treatment. In contrast, our identity treatment produced a large and highly significant reduction in spillovers relative to this benchmark (Fig. 2 and Table 2). Average spillovers were negative but not significantly different from zero in our identity treatment (Table 2, linear combination 'Intercept + Identity = 0'; $F(1, 66) = 1.7$; 95% confidence interval, $(-0.38, 0.06)$; $P = 0.20$; Cohen's $f = 0.16$), which means we have no evidence that behaviour change exceeded the size of the intervention itself. A core principle associated with social tipping is that a tendency for people to behave like others can amplify the effects of some event (such as an intervention) that sets behaviour change in motion. The resulting outcome goes

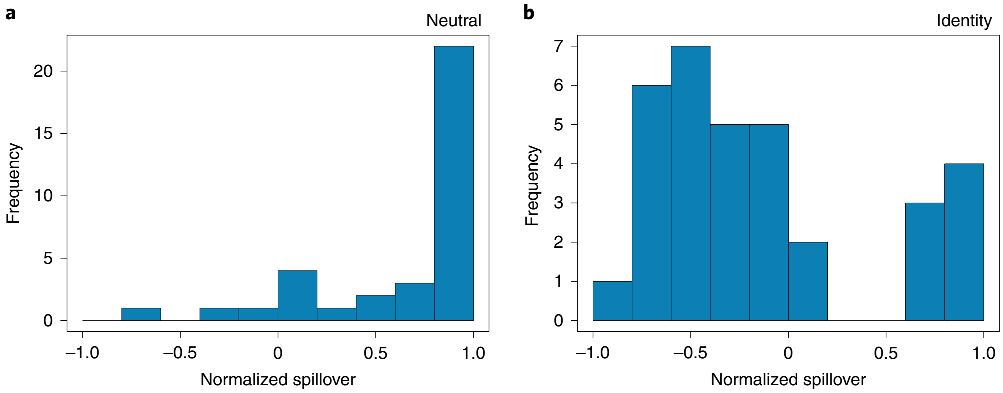

Fig. 2 | Distributions of normalized spillovers by treatment. a, The distribution of spillovers in the neutral treatment. **b**, The distribution of spillovers in the identity treatment. The spillover[22] is a normalized measure of how common the alternative becomes in an experimental group (Methods), and it can take any value in [−1, 1]. Negative values occur when the final proportion choosing the alternative behaviour is less than the proportional size of the intervention. Positive values occur when the final proportion choosing the alternative behaviour is greater than the proportional size of the intervention. The difference in spillovers by treatment is large and highly significant (Table 2).

Table 2 | Spillovers by treatment

	Spillovers
Intercept	0.69 (0.07)
	$P<0.001$, (0.55, 0.83)
Identity	−0.82 (0.12)
	$P<0.001$, (−1.06, −0.57)

Spillovers[22] take values in [−1, 1] and provide a normalized measure of long-run behaviour in a population while accounting for the size of the intervention (Methods). The results are from an ordinary least squares regression that models spillovers as a function of treatment (Fig. 2). Cohen's $f=0.83$. Identical conclusions follow from alternative estimation methods based on a beta regression model (Supplementary Table 15). Spillovers were large and positive in the neutral treatment (Intercept), and a simple relabelling of the choice options in the identity treatment resulted in a large reduction in spillovers (Identity). The *P* values are from two-sided *z* tests. Robust standard errors are shown in parentheses after the estimates, and 95% confidence intervals are shown after the *P* values.

well beyond the size of the event that initiated change. This happened in our neutral treatment, where spillovers reached 69% of the maximum conceivable value on average (Table 2, 'Intercept'). However, simply labelling the choice options in ways that misaligned behaviour change with pre-existing identities destroyed spillovers entirely (Table 2, 'Identity').

Interestingly, political labels facilitated coordination before intervention (Fig. 3 and Supplementary Section 7.4) by providing focal points[65,66]. The pre-intervention game involved material incentives that favoured neither of the two pure-strategy equilibria, and the players faced an equilibrium-selection problem. Neutral labels did not help, and the players simply had to develop an idiosyncratic local norm via repeated play with feedback. Political labels provided the participants with a shared non-monetary basis for ranking the equilibria, and this allowed players to converge quickly with minimal fuss.

Just as surely as political labels facilitated coordination before intervention, they hindered tipping after intervention. After intervention, the experimental groups in the neutral condition immediately started changing their behaviours, and the alternative behaviour was dominant by the end of the post-intervention phase (Fig. 3a). Under political labels, the experimental groups persisted in a state of chronic disagreement. Some players chose the status quo behaviour; some chose the alternative—miscoordination was common and persistent (Fig. 3b).

To investigate these effects in greater detail, we analysed individual choices (Methods) before and after intervention, by treatment, for both targeted and non-targeted players (Fig. 4). Under neutral labels before intervention, we have no evidence that targeted and non-targeted participants chose the alternative at different rates (Table 3, Model 1, (Neutral, T, Pre-int)). Similarly, we have no evidence that targeted and non-targeted participants in the identity treatment made different choices on average before intervention (Table 3, Model 1 linear combination in Fig. 4). Under political labels, both targeted (Table 3, Model 1, (Identity, T, Pre-int)) and non-targeted participants (Table 3, Model 1, (Identity, NT, Pre-int)) showed highly significant reductions in the probability of choosing the alternative behaviour relative to the omitted category—namely, non-targeted participants in the neutral treatment before intervention. This latter result confirms the idea that political labels facilitated coordination before intervention by providing players with focal points.

Post-intervention, both targeted (Table 3, Model 1, (Neutral, T, Post-int)) and non-targeted (Table 3, Model 1, (Neutral, NT, Post-int)) participants in the neutral treatment exhibited an increased probability of choosing the alternative relative to non-targeted participants in the neutral treatment before intervention. Targeted players showed a larger increase than non-targeted players (Table 3, Model 1 linear combination in Fig. 4), but the large and highly significant increase among non-targeted players demonstrates the power of endogenous social interactions to amplify the effects of a delimited intervention.

Targeted participants in the identity treatment also exhibited highly significant changes in behaviour (Table 3, Model 1, (Identity, T, Post-int)) in the wake of the intervention, but the effect was weaker than it was among targeted participants in the neutral treatment (Table 3, Model 1 linear combination in Fig. 4). These results suggest that targeted participants in the identity treatment varied in terms of how they traded money against identity concerns. For some, switching to the alternative choice in the identity treatment was sufficiently aversive to prevent behaviour change, but for others this was not the case.

Non-targeted participants in the identity treatment exhibited a significant but relatively small degree of behaviour change between pre- and post-intervention (Table 3, Model 1 linear combination in Fig. 4). In particular, after intervention these participants chose the alternative behaviour at a rate that was statistically indistinguishable from that of non-targeted participants in the neutral treatment

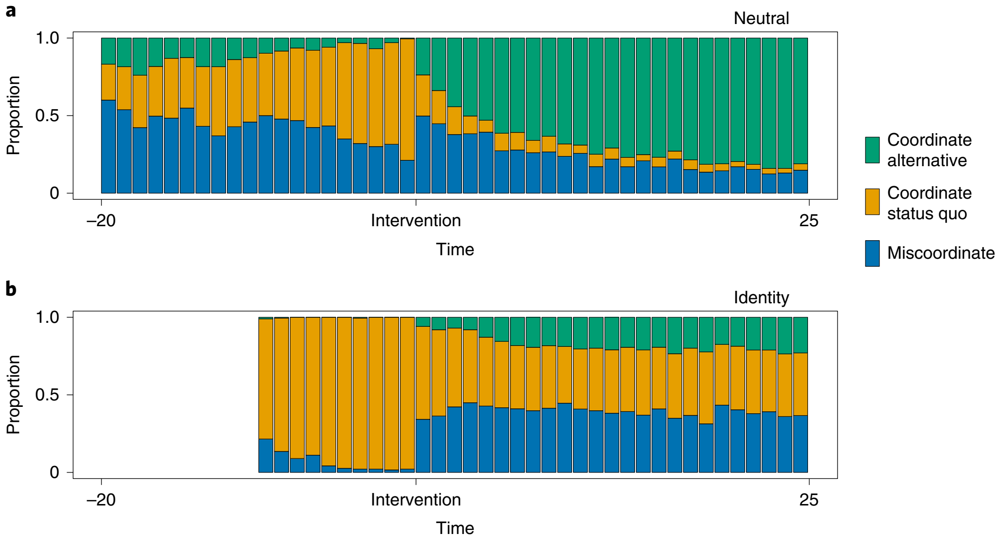

Fig. 3 | Choice dynamics by treatment. a,**b**, The status quo behaviour was the choice associated with the norm that emerged in the pre-intervention phase of a session. With a status quo established, the alternative behaviour was simply the other choice option, which was always favoured by the intervention (Table 1). Here we show the proportion of choices, over all relevant sessions, in which the participants coordinated on the status quo (orange), coordinated on the alternative (green) or miscoordinated (blue) for each period. 'Time' refers to the period of play, which is centred on the intervention period (0). In the neutral sessions, the participants were relatively slow to converge on the status quo before intervention and relatively fast to converge on the alternative after intervention (**a**). In the identity sessions, the participants converged quickly before intervention, requiring fewer overall trials to meet the intervention criteria, but persisted in a state of chronic disagreement after intervention (**b**). For reference, under random matching, the maximum possible expected rate of miscoordination is 0.5.

before intervention (Table 3, Model 1, (Identity, NT, Post-int)). Additionally, they were highly significantly less likely to choose the alternative behaviour post-intervention than their non-targeted counterparts in neutral sessions after intervention (Table 3, Model 1 linear combination in Fig. 4) and their targeted counterparts in identity sessions after intervention (Table 3, Model 1 linear combination in Fig. 4).

These results show that social tipping provided a powerful route to behaviour change in the neutral treatment, but it proved to be equivalently unreliable in the identity treatment. This difference also had stark consequences for participant pay-offs. In our neutral treatment, the absence of a focal point[66] meant that players needed time to develop status quo norms. The neutral groups, however, were able to transition rapidly to an alternative norm when circumstances changed. Tipping and its pay-off consequences are easily seen as a rapid increase in pay-offs after intervention in the neutral sessions (Fig. 5a). The identity sessions show the opposite pattern. The players established status quo norms quickly. However, with an alternative running counter to their pre-existing identities, the players were collectively unable to respond, and they accumulated substantial opportunity costs (Fig. 5b).

Effects of a federal election. Because we ran the study from late October through mid-December 2020, we were able to analyse whether and how choices changed after 7 November, the day major news networks called the election. We had no pre-registered hypotheses about associated effects, but multiple possibilities exist. For example, the actual outcome of the election could have provided all participants with a shared focal point[66] rooted in reality. If so, all sessions in the identity treatment, whether Democrat or Republican, would have converged before intervention on triumphant Biden, an especially compelling possibility given that we did not tell the participants they were together with other supporters of the same party. In addition, participants in the identity treatment after the election could have been more willing to change their behaviour after intervention. With the election settled, the participants could have been less likely to interpret choosing a specific partisan image as an endorsement of the associated election result. This, in turn, might have allowed the participants to disinvest emotionally and simply treat the images as a labelling system to facilitate coordination and make money. Alternatively, the conclusion of the election could have exacerbated outgroup aversion[21], with winners gloating and losers defensive. If so, behaviour change in the identity treatment should have declined after the election.

We have no evidence that any of this happened on average. As was true before the election, all sessions in the identity treatment converged before intervention on the image consistent with party loyalties. In particular, Republican sessions continued to converge on triumphant Trump. More broadly, when comparing before and after the election, we have no statistical evidence for variation in average tendencies to choose the alternative behaviour after conditioning on treatment, targeted status and pre- versus post-intervention (Table 3, Models 1 and 2).

Inequality aversion. We also examined effects related to the inequality that the intervention created. As explained, our design controlled for average effects by using the same pay-off matrices in both treatments. Individual participants might nonetheless have behaved differently from each other because of their attitudes towards inequality. At the recruitment stage (Methods), we measured the social dominance orientation of each participant, which summarized tolerance of hierarchy and inequality (Supplementary Sections 7.1 and 8.2), and we also measured preferences for economic equality (Supplementary Sections 7.1 and 8.2). We used these variables to test for heterogeneity in choices after intervention.

In both treatments, for both targeted and non-targeted players, we find no statistical evidence that these variables affected the probability of choosing the alternative behaviour at the end of the post-intervention phase (Supplementary Table 17). This result may seem surprising because empirical research shows that people are

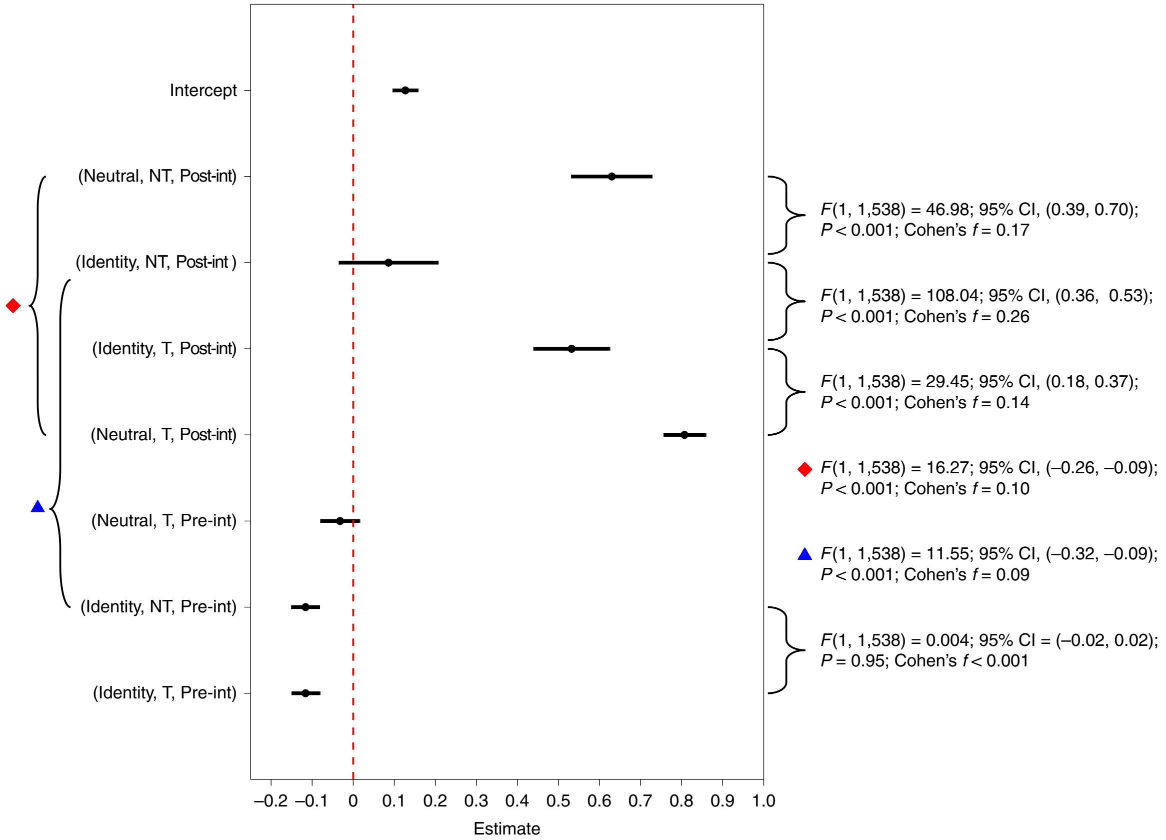

Fig. 4 | Choice of alternative behaviour by treatment. The effect sizes and 95% confidence intervals (CIs) are from Model 1 in Table 3, $N = 1{,}546$ observations. The P values are based on two-sided z-tests. No adjustments were made for multiple comparisons. The omitted category consists of the neutral treatment, non-targeted participants, pre-intervention (Neutral, NT, Pre-int), and all other effects are relative to this benchmark. The curly brackets show the results from various linear combinations discussed in the main text, all of which are based on the cluster-robust standard errors in Table 3. The red dashed vertical line represents no effect. To present these linear combinations graphically, we show the effects in a different order than in Table 3.

averse to inequality, with considerable heterogeneity within and across cultures[59–62,67–69]. Why, then, do we find no evidence that inequality aversion affected the probability of adopting the alternative despite the inequalities introduced by the intervention?

Using the Fehr–Schmidt model[62], we show that inequality aversion has two countervailing effects in a setting like ours (Supplementary Section 7.2), and the effect that supports tipping is likely to dominate. The model distinguishes between two types of inequality. Advantageous inequality means that the focal decision maker has more than others, while disadvantageous inequality means the opposite. In our setting, aversion to advantageous inequality could prevent targeted participants from switching to the alternative after intervention. The participants in question would dislike earning 350 when others earn only 200 or 50 (Table 1). To prevent all targeted participants from switching, however, advantageous inequality aversion would have to be much stronger than typically found in empirical research[67–69]. Moreover, with each targeted participant who switches to the alternative, advantageous inequality aversion has to be even stronger to prevent the remaining targeted participants from switching. In this way, advantageous inequality aversion is unlikely to prevent initial changes in behaviour among targeted participants, and its limited influence should decline quickly (Supplementary Section 7.2).

As participants switch to the alternative, aversion to disadvantageous inequality becomes increasingly relevant, and it always supports the alternative. A participant, whether targeted or not, can experience disadvantageous inequality only by choosing the status quo when matched with a targeted participant choosing the alternative. Hence, the only way to reduce expected disadvantageous inequality is to switch from the status quo to the alternative. Disadvantageous inequality aversion thus supports behaviour change, and its importance should increase through time as participants abandon the status quo (Supplementary Section 7.2).

The upshot is that in our experiment the effects of inequality aversion and ordinary coordination incentives were probably redundant. Our intervention created pressure to tip with large unconditional pay-offs for targeted players and coordination incentives for non-targeted players. Inequality aversion would have only supplemented this pressure, perhaps with little scope for accelerating behaviour change beyond the effects of unconditional pay-offs and coordination incentives. In any case, our neutral treatment reveals that our intervention led to rapid tipping when choice and identity were not linked. Our political labels linked choice and identity and activated a mix of mechanisms that hindered tipping. We now address this mix in detail.

Pre-existing and endogenous identities. We conducted a number of exploratory analyses to evaluate how treatment differences arose from pre-existing identities, of importance to the participants outside the experiment, versus endogenous identities that emerged within the context of the experiment. Specifically, our intention with the identity treatment was to use a pay-off-irrelevant manipulation to activate affective responses based on pre-existing identities. That said, the choice dynamics in the identity sessions diverged from the dynamics in the neutral sessions almost immediately (Fig. 3). Treatment differences in behaviour change thus resulted in principle

Table 3 | Participant chooses the alternative behaviour

	Choose alternative behaviour	
	Model 1	**Model 2**
Intercept	0.13 (0.02)	0.14 (0.02)
	P < 0.001, (0.10, 0.16)	*P* < 0.001, (0.10, 0.18)
Election		−0.02 (0.03)
		P = 0.47, (−0.08, 0.04)
(Neutral, T, Pre-int)	−0.03 (0.02)	−0.07 (0.04)
	P = 0.19, (−0.08, 0.02)	*P* = 0.07, (−0.15, 0.01)
(Neutral, NT, Post-int)	0.63 (0.05)	0.73 (0.06)
	P < 0.001, (0.53, 0.73)	*P* < 0.001, (0.61, 0.85)
(Neutral, T, Post-int)	0.81 (0.03)	0.80 (0.04)
	P < 0.001, (0.76, 0.86)	*P* < 0.001, (0.72, 0.88)
(Identity, NT, Pre-int)	−0.12 (0.02)	−0.12 (0.02)
	P < 0.001, (−0.15, −0.08)	*P* < 0.001, (−0.17, −0.07)
(Identity, T, Pre-int)	−0.12 (0.02)	−0.13 (0.02)
	P < 0.001, (−0.15, −0.08)	*P* < 0.001, (−0.17, −0.09)
(Identity, NT, Post-int)	0.09 (0.06)	0.03 (0.07)
	P = 0.17, (−0.04, 0.21)	*P* = 0.65, (−0.11, 0.17)
(Identity, T, Post-int)	0.53 (0.05)	0.54 (0.06)
	P < 0.001, (0.44, 0.63)	*P* < 0.001, (0.42, 0.66)
Election × (Neutral, T, Pre-int)		0.06 (0.05)
		P = 0.24, (−0.04, 0.16)
Election × (Neutral, NT, Post-int)		−0.15 (0.09)
		P = 0.11, (−0.33, 0.03)
Election × (Neutral, T, Post-int)		0.01 (0.05)
		P = 0.89, (−0.10, 0.11)
Election × (Identity, NT, Pre-int)		0.00 (0.03)
		P = 0.97, (−0.06, 0.06)
Election × (Identity, T, Pre-int)		0.02 (0.03)
		P = 0.51, (−0.04, 0.09)
Election × (Identity, NT, Post-int)		0.10 (0.12)
		P = 0.42, (−0.14, 0.34)
Election × (Identity, T, Post-int)		−0.02 (0.10)
		P = 0.87, (−0.20, 0.17)

The results of linear probability models (Methods) for individual choices in the final periods of the pre- and post-intervention phases are shown. 'Election' is a dummy indicating sessions after 7 November 2020, the day the election was called. Composite dummies are defined jointly over (1) treatment (Neutral versus Identity), (2) whether the participant was targeted (T) or not (NT), and (3) pre- versus post-intervention, with (Neutral, NT, Pre-int) as the omitted category. Model 1 was pre-registered. Model 2 is exploratory and additionally distinguishes between before (omitted category) and after the election. The results are robust to including day fixed effects (Supplementary Table 14), including more periods in the analysis (Supplementary Table 2) and alternative estimate methods with a random-effects logit model (Supplementary Table 16). The *P* values are from two-sided *z*-tests. Cluster-robust standard errors are shown in parentheses after the estimates, and 95% confidence intervals are shown after the *P* values.

both from the link between political labels and pre-existing identities and from the divergent dynamics that the two labelling systems induced. By focusing on overall treatment differences, our core analyses (Tables 2 and 3) effectively pool these two mechanisms. The task here is to disentangle them as much as possible.

Compared with the neutral sessions, the identity sessions converged on status quo norms quickly, and choices were relatively homogeneous at the end of the pre-intervention phase (Fig. 3). Both fast convergence and choice homogeneity could have revealed to participants within an identity session that everyone was relying on a common focal point. If so, participants in the identity sessions would have been able to infer that they shared political loyalties with others in the session in a way that was not possible in the neutral sessions.

Broadly, participants in the identity sessions may have experienced the normative force of the status quo more strongly than their counterparts in the neutral sessions for two different reasons. On the one hand, status quo norms were consistent with pre-existing identities in the identity sessions. As discussed, status quo norms and the party loyalties of the participants were perfectly correlated in the identity treatment, but they were unrelated in the neutral treatment. On the other hand, when comparing identity sessions with neutral sessions, status quo norms in the identity sessions developed quickly, produced relatively homogeneous behaviour before intervention and allowed the players to draw strong inferences about others. These differences all emerged within the context of the experiment, specifically in the pre-intervention phase, and they could have all intensified commitment to the status quo in ways that were independent of pre-existing identities. What evidence do we have for each type of mechanism?

For pre-existing identities, when recruiting participants before the experiment, we measured two forms of affective polarization[54,55,70] for each participant. One measure quantified polarization in terms of Democrats versus Republicans in general and the other in terms of Biden versus Trump specifically (Supplementary Section 7.3.1). Increasing polarization was statistically associated with a reduced tendency to choose the alternative after intervention in the identity treatment but not in the neutral treatment (Supplementary Table 19). This intuitive result provides our first clue that treatment differences stemmed at least in part from the link between our political labels and pre-existing identities.

When recruiting, we also implemented a priming experiment to manipulate the extent to which participants viewed party affiliations as central to their identities (Supplementary Section 7.5.1). We used this experimental prime as an instrumental variable[71] to analyse choices in the experiment proper. As the importance of party affiliation increased, the probability of choosing the alternative behaviour after intervention declined in the identity treatment, but we have no evidence for such a decline in the neutral treatment (Supplementary Table 22). This finding further indicates that resistance to behaviour change in the identity treatment resulted from our use of political labels to foreground identities with meaning to the participants outside the experiment. Finally, although we found no evidence of differences in average behaviour before versus after the election (Table 3, Model 2), treatment differences after the election were most pronounced in the immediate aftermath of the election, presumably when emotions were peaking and most susceptible to manipulation (Supplementary Fig. 13). Altogether, these results clearly indicate that pre-existing identities shaped observed treatment differences.

For differences that emerged within the context of the experiment, we also found limited but intriguing evidence for a secondary effect based on endogenous dynamics. Controlling for treatment, we estimated that fast convergence on the status quo before intervention was associated with increased resistance to the alternative choice after intervention (Supplementary Section 7.4). This suggests that, if neutral sessions had managed to converge more quickly on average than they actually did, they would have experienced less behaviour change than they actually did. We found no statistical evidence that homogeneity of choices before intervention had an independent effect on choices after intervention (Supplementary Section 7.4). Lastly, as explained above, the distinctive dynamics in the identity sessions might have allowed the participants to quickly draw strong inferences about shared political loyalties within sessions. If this heightened potential to draw inferences had an effect that was independent of pre-existing identities, participants in the identity sessions should have responded to early feedback about others more strongly than in the neutral sessions. This follows

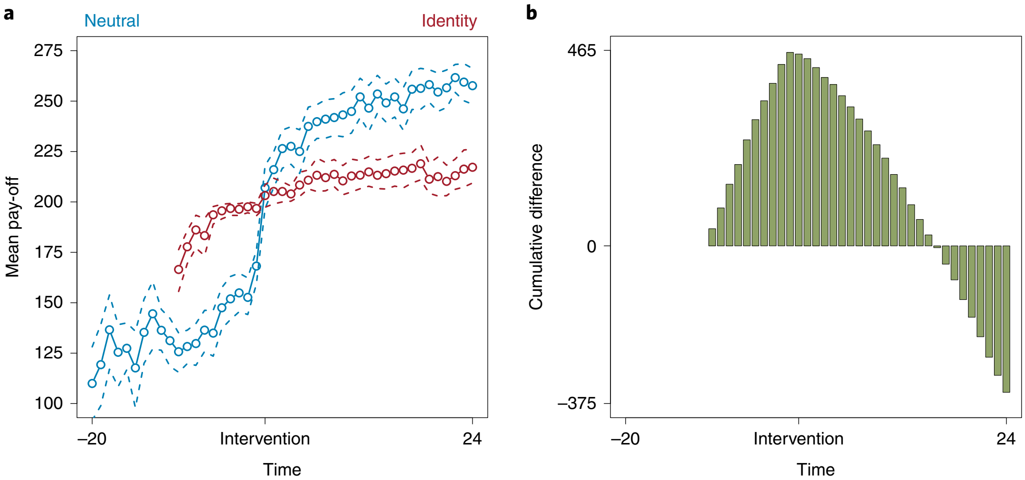

Fig. 5 | Pay-off dynamics. a, Mean pay-offs by treatment and period. The pay-offs are measured in experimental currency points (100 points = US$1). 'Time' refers to the period of play, which is centred on the intervention period (0). The dashed lines are 95% confidence intervals from a bootstrapping algorithm clustered at the level of the experimental group. Compared with the neutral treatment, political labels in the identity treatment provided a ready focal point[66] that allowed the participants to converge on a norm quickly before intervention. After intervention, however, chronic disagreement (Fig. 3) prevented the participants in the identity treatment from transitioning to new norms in the same way the participants did in the neutral treatment. **b**, The accumulated difference in mean pay-offs, identity minus neutral, shows the monetary opportunity costs that the participants in the identity sessions ultimately paid.

simply from the fact that this feedback was the only way for the participants to draw the inferences in question. We did not find any statistical evidence for this possibility (Supplementary Section 7.4).

These results show that our political labels might have hindered tipping in two distinct ways. First, because of their relation to identities already in place when the experiment began, political labels added value to the status quo and detracted value from the alternative. The evidence overall indicates that this mechanism was primarily responsible for treatment differences in behaviour change. Second, political labels had an immediate influence on cultural evolutionary dynamics, and the identity sessions and neutral sessions had already taken divergent paths by the time the intervention occurred. Exploratory analyses suggest that fast convergence to the status quo, which occurred mainly in the identity sessions, may have reduced tipping in a way that was partially distinct from pre-existing identities.

Discussion

Our results show that even a seemingly superficial link between identity and choice can restructure cultural evolution and undermine tipping that would otherwise occur. Although our results demonstrate just how easily this can happen, group identities may not always impede tipping and behaviour change. A possibility we did not examine would occur when the policymaker promotes different behaviours in different pre-existing groups already seeking to differentiate themselves. Imagine a policy initiative that promotes different behaviours for women and men. If many people adopt gendered patterns of behaviour, this tendency could easily help the policymaker because the policymaker's behavioural objectives correlate neatly with the pre-existing subdivision of the population.

We focused instead on settings in which the policymaker's objectives do not align well with pre-existing identities. Even then, however, an interest in protecting group identities does not necessarily impede behaviour change. If ingroup conformity is strong and the intervention considerably smaller (for example, 10%) than what we used in this study, small to moderate amounts of outgroup aversion may actually help destabilize the status quo norm and support behaviour change as a result[22]. In addition, incentives can sometimes favour signalling one's identity in covert ways that have little or no meaning to outgroup members, especially if the covert signallers belong to a disadvantaged minority[53]. Given the partisan nature of contemporary US politics[55,57], group identities in this study were probably not subject to covert signalling. When present, however, covert signalling might weaken the tendency for group identities to undermine tipping, though perhaps with members of marginalized groups pretending to adhere to the norms of a dominant group in step with the policymaker.

When outgroup aversion is strong and out in the open, however, theory suggests that the pressure to police the boundaries of group identities will tend to dominate cultural evolution[21–23], and our results are consistent with this idea. Identity policing can manifest itself in at least two ways relevant to policy. First, the policymakers' target population may not be strongly subdivided, but the policymakers themselves may represent an aversive outgroup. In the 1950s, for example, a council of local male leaders banned female genital cutting in the Meru District of Kenya. Citizens apparently saw these leaders as the puppets of colonizers, and the ban on cutting actually seemed to increase commitment to the practice as a hallmark of cultural identity[52].

A second possibility is that the target population is subdivided, people care greatly about protecting the group identities that result and the policymaker's behavioural objectives do not fit well with this landscape of pre-existing identities. Climate change, for example, is one of several politically polarizing issues in the contemporary United States in the sense that adopting a specific stance on the issue is part of what party loyalty requires[72]. Some people drive hybrids, and other people roll coal to show their contempt for people who drive hybrids[21]. Reducing emissions among the former may entrench resistance among the latter.

In both of these scenarios, identities are linked to particular choices in the policy domain in question, and the link adds value to the status quo choice for some or all individuals. Our results show that this implicit value can constrain behaviour change in general and endogenous change due to social tipping specifically. In situations of this sort, the policymaker might consider an intervention before the intervention[22]. The first intervention should weaken the link between identity and choice in the policy domain at hand to lay the groundwork for the intervention proper. With identity concerns less relevant because of this initial intervention, the intervention proper could then promote the alternative norm of primary interest. CNN adopted this approach with an ad about face masks during the COVID-19 pandemic (https://twitter.com/cnn/status/1289176533140029441?lang=en). The ad first attempted to decouple masks from the partisan baggage they had acquired in the United States in the early days of the pandemic. It began with a photo of a mask and said, “This is a mask. It prevents the spread of coronavirus. This is not a political statement. It’s a mask.” The ad then moved on to its primary behavioural objective and concluded with, “Please wear a mask.” We know of no evidence about the effectiveness of this ad, but presumably the limited credibility that conservative Republicans attach to CNN[73] did not help. Regardless, the strategy is clear. The ad did not address the partisan divide in the United States. It simply tried to decouple this divide from choices about wearing masks.

An extension of this approach centres on strategies that attempt to transfer identity concerns from the choice domain of interest to some other domain. For example, a number of initiatives promoting the abandonment of female genital cutting emphasize alternative rites of passage[74] designed to allow families to integrate their daughters into society without the harm of genital cutting. The hope is that families become increasingly willing to abandon cutting if they have suitable substitute behaviours and traditions. Substitutes effectively change the underlying coordination game by expanding the set of actions[75]. We do not know of much evidence on the value of such approaches, but a recent field experiment in Malawi showed that providing substitute behaviours can reduce early marriage and teenage pregnancy[76]. Broadly, future research should examine the effects of decoupling identity and choice when policymakers are attempting to influence the cultural evolution of social norms.

We have shown that social tipping can offer a powerful but unreliable route to social change. This combination presents policymakers with an unusual challenge. Because tipping has impressive potential, strategies to provoke tipping will presumably remain a part of the policymaker’s repertoire. Because tipping is unreliable, interventions designed to trigger tipping may easily fail to do so. Researchers and practitioners thus require an empirically grounded understanding of when tipping is possible and how to spark tipping[5]. In particular, tipping offers the possibility of using limited resources efficiently, but it could also be extremely costly if policymakers’ preferences are misaligned with those of the citizens under their influence[22,77]. In such cases, an intervention could be worse than ineffective; it could bring a net social cost even if promoting a behaviour that appears to be a Pareto improvement.

Group identities are ubiquitous phenomena[43] that can encourage polarization along political, religious and ethnic lines[57]. Group identities can have positive consequences[42,78], but they can also inhibit efforts to change cultural norms. Understanding when and how group identities influence social tipping would allow for the design of interventions that appropriately consider the effects of identity concerns as we all confront the formidable challenges facing contemporary human societies.

Methods

Participants. We conducted the study with adult participants living in the United States between 28 October and 16 December 2020. The study was approved by the Institutional Review Boards at the University of Lausanne, the University of Bern and Princeton University. All participants provided informed consent.

We recruited the participants online via Prolific. At the recruitment stage, we screened potential participants on the basis of their self-reported political affiliations and responses to two questions about political preferences. In particular, we asked about Biden and Trump using feelings thermometers[54,55,70,79]. We used these responses to recruit participants to the main study who were either (1) warm about Biden and cold about Trump or (2) cold about Biden and warm about Trump (Supplementary Section 5.2). Altogether, we recruited 565 participants in category 1, all of whom reported being Democrats, and 235 participants in category 2, all of whom reported being Republicans. We did not intentionally recruit participants who were cold or warm about both candidates, but a small error allowed four Democrats who were cold about both candidates and one Democrat who was warm about both candidates into the final sample.

For the main experiment, we formed experimental groups of either all Republicans or all Democrats. The participants were anonymous, they could not communicate with others in the session and they had no information about the composition of the experimental group. All sessions began with 12 participants, and we relied on several protocols to minimize participant dropout (see below). Supplementary Sections 5 and 6.1 provide additional details and analyses related to recruitment, sample composition and dropouts.

Repeated game play and treatments. The participants repeatedly played coordination games for up to 45 periods. In each period, we randomly paired players within the experimental group to play. In the pre-intervention phase, everyone played the same coordination game (Table 1, pre-intervention (all)). The pre-intervention phase lasted a minimum of 10 periods. After crossing this threshold, the pre-intervention phase ended when at least 90% of players chose the same option in a period or when 20 periods had passed. Each session had a well-defined majority behaviour (that is, the status quo) at the end of the pre-intervention phase.

To begin the post-intervention phase, we applied a new pay-off matrix to a subset of players (Table 1, post-intervention (targeted)). The remaining players retained their original incentives (Table 1, post-intervention (non-targeted)). The intervention was randomly assigned to 50% of players in the experimental group at the start of the session (Supplementary Section 3.3). Because assignment to the targeted subset occurred at the beginning of sessions, sporadic dropouts before intervention meant that the targeted subset occasionally consisted of 40% or 60% of the group (see Supplementary Section 6.3 for the associated robustness checks).

Each period, each participant made a choice by clicking an on-screen button that was integrated with the display of the player’s pay-off matrix. The labels for the choices, whether neutral or political, were simply embedded in the buttons themselves (Supplementary Figs. 2 and 3). The treatments were identical apart from the difference in labels. Political labels were pre-tested to ensure that one label was appealing and the other aversive (Supplementary Section 3.4).

For feedback, the participants received three pieces of information at the beginning of each period after the first. Namely, each participant saw (1) the complete distribution of choices from the previous period among 10 randomly selected players in the experimental group, (2) the choice of the focal player’s partner in the previous period and (3) the points that the focal participant had earned in the previous period. Communicating the choices of 10 randomly selected players allowed us to continue a session when someone dropped out without disturbing our feedback protocol. Specifically, because the participants played in pairs, we required an even number of participants. Thus, if a player dropped out, we removed the player’s partner in that period, but only after the partner had made a choice. If more than two players exited the group, for whatever reason, we ended the session. In sum, each experimental group started with 12 participants, and we randomly selected 10 participants each period for feedback. Some experimental groups dropped to 10 participants during the session, and at that point we provided feedback by reporting the distribution of choices among all 10 remaining players. Dropouts were not related to treatment (Supplementary Section 6.1).

The points from the games were converted to dollars at a fixed rate. The total pay-off for each participant was calculated by summing the pay-offs from five randomly selected periods. The participants were informed about payment and other procedures before the start of the game (Supplementary Section 8).

Analyses. The initial data consisted of 28,303 observations from 908 participants in 77 groups. We removed nine groups that, due to dropouts, did not have at least one period post-intervention. This left 27,624 observations from 805 participants in 68 groups. The analyses were pre-registered (https://osf.io/84jpq) unless otherwise indicated.

Table 2 presents an analysis of spillovers[22]. Spillovers provide a normalized measure of how common the alternative behaviour ultimately becomes. Let ϕ_j be the proportion of participants in experimental group j targeted by the intervention. Let $\hat{q}_j$ be the proportion of participants in j choosing the alternative behaviour in the final period post-intervention. Spillovers in j, denoted Θ_j, are defined as

$$\Theta_j = \begin{cases} \frac{\hat{q}_j - \phi_j}{1-\phi_j} & \text{if } \hat{q}_j > \phi_j \\ \frac{\hat{q}_j - \phi_j}{\phi_j} & \text{otherwise.} \end{cases} \qquad (1)$$

Spillovers take values in $[-1, 1]$. If the spillover is positive, the final effect of the intervention is larger than the proportional size of the intervention. A negative spillover signifies the opposite (Supplementary Section 2.3).

To examine spillovers, we ran an ordinary least squares regression with spillovers as a function of treatments. The model is

$$\Theta_j = \beta_0 + \beta_1 u_j + \epsilon_j, \qquad (2)$$

where $j \in \{1, 2, \ldots, J\}$ indexes experimental group, $u_j \in \{0, 1\}$ indicates whether group j was in the neutral treatment ($u_j = 0$) or in the identity treatment ($u_j = 1$), and ϵ_j is a group error term. We used robust standard errors because ϵ_j may not be homoscedastic normal[80,81]. Alternative estimation methods lead to the same conclusions (Supplementary Table 15).

We used linear probability models (ordinary least squares) to examine individual choices (Table 3). Choice is a function of treatment, whether the participant in question was targeted or not, and whether the choice occurred in the final period of the pre-intervention or post-intervention phase. Restricting attention to the final periods of the two phases minimizes the role of transient dynamics and thus focuses on transitions between equilibria. The results hold with more periods (Supplementary Table 2).

Our pre-registered core model (Table 3, Model 1) is

$$\begin{aligned} c_i = {} & \beta_0 + \beta_1[u_i = 0 \wedge z_i = 1 \wedge \tau_i = 0] + \beta_2[u_i = 0 \wedge z_i = 0 \wedge \tau_i = 1] + \\ & \beta_3[u_i = 0 \wedge z_i = 1 \wedge \tau_i = 1] + \beta_4[u_i = 1 \wedge z_i = 0 \wedge \tau_i = 0] + \\ & \beta_5[u_i = 1 \wedge z_i = 1 \wedge \tau_i = 0] + \beta_6[u_i = 1 \wedge z_i = 0 \wedge \tau_i = 1] + \\ & \beta_7[u_i = 1 \wedge z_i = 1 \wedge \tau_i = 1] + \epsilon_i. \end{aligned} \quad (3)$$

The index $i \in \{1, 2, \ldots, I\}$ specifies observation at the level of an individual making a single choice, and $c_i \in \{0, 1\}$ indicates whether the associated choice was the status quo ($c_i = 0$) or the alternative ($c_i = 1$). The variable $u_i \in \{0, 1\}$ indicates whether observation i was associated with a participant in the neutral condition ($u_i = 0$) or the identity condition ($u_i = 1$). The variable $z_i \in \{0, 1\}$ indicates whether observation i was associated with a participant targeted by the intervention ($z_i = 1$) or not ($z_i = 0$), $\tau_i \in \{0, 1\}$ indicates whether the observation was from the pre-intervention ($\tau_i = 0$) or post-intervention phase ($\tau_i = 1$), and ϵ_i is an individual choice error term. The $[\cdot]$ are Iverson brackets, and $\wedge$ denotes logical 'and'. Iverson brackets return 1 if the condition within is met and 0 otherwise. To illustrate, $[u_i = 1 \wedge z_i = 0 \wedge \tau_i = 1]$ returns 1 if i was associated with a participant in the identity treatment ($u_i = 1$) who was not targeted ($z_i = 0$) and was making a post-intervention choice ($\tau_i = 1$). We refer to this variable as (Identity, NT, Post-int) in Table 3, and β_6 is the associated coefficient. The omitted category for the regression is $[u_i = 0 \wedge z_i = 0 \wedge \tau_i = 0]$—that is, (Neutral, NT, Pre-int).

We extended the core model with an exploratory analysis that added a dummy variable, with interactions, to indicate sessions after 7 November 2020 (Table 3, Model 2). For the individual choice models, we used cluster-robust standard errors[80,81], clustered at the level of the experimental group, to account for errors that are not homoscedastic normal and may be correlated within clusters. Alternative estimation methods lead to the same conclusions (Supplementary Section 6.6).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data are publicly available at the Open Science Framework, at https://doi.org/10.17605/OSF.IO/KN3A2.

Code availability

The code for the analyses is publicly available at the Open Science Framework, at https://doi.org/10.17605/OSF.IO/KN3A2. To collect the interactive group data, we used the open-source otree software, version 3.3, accessible at otree.org. We used Qualtrics (October 2020 version) to collect the questionnaire data.

Received: 9 July 2021; Accepted: 29 July 2022;
Published online: 22 September 2022

References

1. Young, H. P. The evolution of social norms. *Annu. Rev. Econ.* **7**, 359–387 (2015).
2. Rosenfeld, M. J. Moving a mountain: the extraordinary trajectory of same-sex marriage approval in the United States. *Socius* **3**, 2378023117727658 (2017).
3. Rode, J. & Weber, A. Does localized imitation drive technology adoption? A case study on rooftop photovoltaic systems in Germany. *J. Environ. Econ. Manage.* **78**, 38–48 (2016).
4. Winkelmann, R. et al. Social tipping processes towards climate action: a conceptual framework. *Ecol. Econ.* **192**, 107242 (2022).
5. Andreoni, J., Nikiforakis, N. & Siegenthaler, S. Predicting social tipping and norm change in controlled experiments. *Proc. Natl Acad. Sci. USA* **118**(16) (2021).
6. Mackie, G. Ending footbinding and infibulation: a convention account. *Am. Sociol. Rev.* **61**, 999–1017 (1996).
7. Nyborg, K. et al. Social norms as solutions. *Science* **354**, 42–43 (2016).
8. Christakis, N. A. & Fowler, J. H. The spread of obesity in a large social network over 32 years. *N. Engl. J. Med.* **357**, 370–379 (2007).
9. Arnot, M. et al. How evolutionary behavioural sciences can help us understand behaviour in a pandemic. *Evol. Med. Public Health* **2020**, 264–278 (2020).
10. Cloward, K. *When Norms Collide: Local Responses to Activism against Female Genital Mutilation and Early Marriage* (Oxford Univ. Press, 2016); https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780190274917.001.0001/acprof-9780190274917
11. Platteau, J.-P., Camilotti, G. & Auriol, E. in *Towards Gender Equity in Development* (eds. Anderson, S. et al.) Ch. 15 (Oxford Univ. Press, 2018).
12. Castilla-Rho, J. C., Rojas, R., Andersen, M. S., Holley, C. & Mariethoz, G. Social tipping points in global groundwater management. *Nat. Hum. Behav.* **1**, 640–649 (2017).
13. Travers, H., Walsh, J., Vogt, S., Clements, T. & Milner-Gulland, E. Delivering behavioural change at scale: what conservation can learn from other fields. *Biol. Conserv.* **257**, 109092 (2021).
14. Barrett, S. & Dannenberg, A. Sensitivity of collective action to uncertainty about climate tipping points. *Nat. Clim. Change* **4**, 36–39 (2014).
15. Kopp, R. E., Shwom, R. L., Wagner, G. & Yuan, J. Tipping elements and climate–economic shocks: pathways toward integrated assessment. *Earth's Future* **4**, 346–372 (2016).
16. Farmer, J. D. et al. Sensitive intervention points in the post-carbon transition. *Science* **364**, 132–134 (2019).
17. Otto, I. M. et al. Social tipping dynamics for stabilizing Earth's climate by 2050. *Proc. Natl Acad. Sci. USA* **117**, 2354–2365 (2020).
18. Bicchieri, C. & Dimant, E. Nudging with care: the risks and benefits of social information. *Public Choice* **191**, 443–464 (2019).
19. Smith, S. R., Christie, I. & Willis, R. Social tipping intervention strategies for rapid decarbonization need to consider how change happens. *Proc. Natl Acad. Sci. USA* **117**, 10629–10630 (2020).
20. Efferson, C. Policy to activate cultural change to amplify policy. *Proc. Natl Acad. Sci. USA* **118**(23) (2021).
21. Smaldino, P. E., Janssen, M. A., Hillis, V. & Bednar, J. Adoption as a social marker: innovation diffusion with outgroup aversion. *J. Math. Sociol.* **41**, 26–45 (2017).
22. Efferson, C., Vogt, S. & Fehr, E. The promise and the peril of using social influence to reverse harmful traditions. *Nat. Hum. Behav.* **4**, 55–68 (2020).
23. Smaldino, P. E. & Jones, J. H. Coupled dynamics of behaviour and disease contagion among antagonistic groups. *Evol. Hum. Sci.* **3**, e28 (2021).
24. Henrich, J. Cultural transmission and the diffusion of innovations: adoption dynamics indicate that biased cultural transmission is the predominate force in behavioral change. *Am. Anthropol.* **103**, 992–1013 (2001).
25. Young, H. P. & Burke, M. A. Competition and custom in economic contracts: a case study of Illinois agriculture. *Am. Econ. Rev.* **91**, 559–573 (2001).
26. Rogers, E. M. *Diffusion of Innovations* (Simon and Schuster, 2010).
27. Eugster, B., Lalive, R., Steinhauer, A. & Zweimüller, J. The demand for social insurance: does culture matter? *Econ. J.* **121**, F413–F448 (2011).
28. Eugster, B., Lalive, R., Steinhauer, A. & Zweimüller, J. Culture, work attitudes, and job search: evidence from the Swiss language border. *J. Eur. Econ. Assoc.* **15**, 1056–1100 (2017).
29. Centola, D., Becker, J., Brackbill, D. & Baronchelli, A. Experimental evidence for tipping points in social convention. *Science* **360**, 1116–1119 (2018).
30. Bellemare, M. F., Novak, L. & Steinmetz, T. L. All in the family: explaining the persistence of female genital cutting in West Africa. *J. Dev. Econ.* **116**, 252–265 (2015).
31. Muthukrishna, M. Cultural evolutionary public policy. *Nat. Hum. Behav.* **4**, 12–13 (2020).
32. Novak, L. Persistent norms and tipping points: the case of female genital cutting. *J. Econ. Behav. Organ.* **177**, 433–474 (2020).
33. Kuran, T. Now out of never: the element of surprise in the East European revolution of 1989. *World Polit.* **44**, 7–48 (1991).
34. Shell-Duncan, B. & Hernlund, Y. *Female 'Circumcision' in Africa: Culture, Controversy, and Change* (Lynne Rienner, 2000).
35. Christakis, N. A. & Fowler, J. H. The collective dynamics of smoking in a large social network. *N. Engl. J. Med.* **358**, 2249–2258 (2008).
36. Schief, M., Vogt, S. & Efferson, C. Investigating the structure of son bias in Armenia with novel measures of individual preferences. *Demography* **58**, 1737–1764 (2021).
37. DellaVigna, S. & La Ferrara, E. in *Handbook of Media Economics* Vol. 1 (eds Anderson, S.P. et al.) 723–768 (Elsevier, 2015).
38. La Ferrara, E. Mass media and social change: can we use television to fight poverty? *J. Eur. Econ. Assoc.* **14**, 791–827 (2016).
39. Vogt, S., Zaid, N. A. M., Ahmed, H. E. F., Fehr, E. & Efferson, C. Changing cultural attitudes towards female genital cutting. *Nature* **538**, 506–509 (2016).
40. Schimmelpfennig, R., Vogt, S., Ehret, S. & Efferson, C. Promotion of behavioural change for health in a heterogeneous population. *Bull. World Health Organ.* **99**, 819–827 (2021).
41. Granovetter, M. Threshold models of collective behavior. *Am. J. Sociol.* **83**, 1420–1443 (1978).
42. Efferson, C., Lalive, R. & Fehr, E. The coevolution of cultural groups and ingroup favoritism. *Science* **321**, 1844–1849 (2008).
43. Tajfel, H. *Human Groups and Social Categories: Studies in Social Psychology* (Cup Archive, 1981).

44. De Dreu, C. K., Gross, J., Fariña, A. & Ma, Y. Group cooperation, carrying-capacity stress, and intergroup conflict. *Trends Cogn. Sci.* **24**, 760–776 (2020).
45. Young, H. P. Innovation diffusion in heterogeneous populations: contagion, social influence, and social learning. *Am. Econ. Rev.* **99**, 1899–1924 (2009).
46. Jackson, M. O. & López-Pintado, D. Diffusion and contagion in networks with heterogeneous agents and homophily. *Netw. Sci.* **1**, 49–67 (2013).
47. Gavrilets, S. The dynamics of injunctive social norms. *Evol. Hum. Sci.* **2**, e60 (2020).
48. Berger, J., Efferson, C. & Vogt, S. Tipping pro-environmental norm diffusion at scale: opportunities and limitations. *Behav. Public Policy* 1–26 https://doi.org/10.1017/bpp.2021.36 (2021).
49. Boyd, R. & Richerson, P. J. The evolution of ethnic markers. *Cult. Anthropol.* **2**, 65–79 (1987).
50. Choi, J.-K. & Bowles, S. The coevolution of parochial altruism and war. *Science* **318**, 636–640 (2007).
51. Handley, C. & Mathew, S. Human large-scale cooperation as a product of competition between cultural groups. *Nat. Commun.* **11**, 702 (2020).
52. Thomas, L. M. in *Female 'Circumcision' in Africa: Culture, Controversy, and Change* (eds Shell-Duncan, B. and Hernlund, Y.) Ch. 7 (Lynne Rienner, 2000).
53. Smaldino, P. E. & Turner, M. A. Covert signaling is an adaptive communication strategy in diverse populations. *Psychol. Rev.* **129**, 812–829 (2022).
54. Iyengar, S., Sood, G. & Lelkes, Y. Affect, not ideology: a social identity perspective on polarization. *Public Opin. Q.* **76**, 405–431 (2012).
55. Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N. & Westwood, S. J. The origins and consequences of affective polarization in the United States. *Annu. Rev. Polit. Sci.* **22**, 129–146 (2019).
56. McConnell, C., Margalit, Y., Malhotra, N. & Levendusky, M. The economic consequences of partisanship in a polarized era. *Am. J. Polit. Sci.* **62**, 5–18 (2018).
57. Finkel, E. J. et al. Political sectarianism in America. *Science* **370**, 533–536 (2020).
58. Bowles, S. *Microeconomics: Behavior, Institutions, and Evolution* (Princeton Univ. Press, 2004).
59. Henrich, J. et al. 'Economic man' in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behav. Brain Sci.* **28**, 795–815 (2005).
60. Cooper, D. J. & Kagel, J. H. Other-regarding preferences. *Handb. Exp. Econ.* **2**, 217–289 (2016).
61. Falk, A. et al. Global evidence on economic preferences. *Q. J. Econ.* **133**, 1645–1692 (2018).
62. Fehr, E. & Schmidt, K. M. A theory of fairness, competition, and cooperation. *Q. J. Econ.* **114**, 817–868 (1999).
63. Fairlie, R. W. & Robinson, J. Experimental evidence on the effects of home computers on academic achievement among schoolchildren. *Am. Econ. J. Appl. Econ.* **5**, 211–240 (2013).
64. Hanna, R., Duflo, E. & Greenstone, M. Up in smoke: the influence of household behavior on the long-run impact of improved cooking stoves. *Am. Econ. J. Econ. Policy* **8**, 80–114 (2016).
65. Schelling, T. C. *The Strategy of Conflict* (Harvard Univ. Press, 1960).
66. Crawford, V. P., Gneezy, U. & Rottenstreich, Y. The power of focal points is limited: even minute payoff asymmetry may yield large coordination failures. *Am. Econ. Rev.* **98**, 1443–1458 (2008).
67. Goeree, J. K. & Holt, C. A. Asymmetric inequality aversion and noisy behavior in alternating-offer bargaining games. *Eur. Econ. Rev.* **44**, 1079–1089 (2000).
68. Blanco, M., Engelmann, D. & Normann, H. T. A within-subject analysis of other-regarding preferences. *Games Econ. Behav.* **72**, 321–338 (2011).
69. Beranek, B., Cubitt, R. & Gächter, S. Stated and revealed inequality aversion in three subject pools. *J. Econ. Sci. Assoc.* **1**, 43–58 (2015).
70. Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M. & Ryan, J. B. Affective polarization, local contexts and public opinion in America. *Nat. Hum. Behav.* **5**, 28–38 (2021).
71. Angrist, J. D. & Pischke, J.-S. *Mostly Harmless Econometrics* (Princeton Univ. Press, 2009).
72. Fiorina, M. P. *Unstable Majorities: Polarization, Party Sorting, and Political Stalemate* (Hoover, 2017).
73. Stroud, N. J. & Lee, J. K. Perceptions of cable news credibility. *Mass Commun. Soc.* **16**, 67–88 (2013).
74. Hughes, L. Alternative rites of passage: faith, rights, and performance in FGM/C abandonment campaigns in Kenya. *Afr. Stud.* **77**, 274–292 (2018).
75. Gulesci, S. et al. A stepping stone approach to understanding harmful norms. CEPR Discussion PaperNo. DP15776. *SSRN* https://ssrn.com/abstract=3784002 (2021).
76. Hänni, S. & Lichand, G. *Harming to Signal: Child Marriage vs. Public Donations in Malawi* Working Paper (University of Zurich, Department of Economics, 2021).
77. Efferson, C., Vogt, S. & von Flüe, L. in *Oxford Handbook of Cultural Evolution* (eds Kendal, J. et al.) Ch. TBD (Oxford Univ. Press, 2023).
78. Chen, Y. & Li, S. X. Group identity and social preferences. *Am. Econ. Rev.* **99**, 431–457 (2009).
79. Mason, L. Ideologues without issues: the polarizing consequences of ideological identities. *Public Opin. Q.* **82**, 866–887 (2018).
80. Wooldridge, J. M. *Econometric Analysis of Cross Section and Panel Data* (MIT Press, 2010).
81. Arai, M. Cluster-robust standard errors using R. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.573.9323&rep=rep1&type=pdf (2011).

Acknowledgements

For helpful comments, we thank J. Barceló, S. Lonati, H. Nax, N. Nikiforakis, C. Zehnder and seminar participants at the Collegio Carlo Alberto, the University of Lausanne, NYUAD, Oxford University, Princeton University and the University of Zurich. We thank Clara Sfeir for research assistance. The study was funded by the Swiss National Science Foundation (grant no. 100018 185417/1 to C.E. and S.V.). E.U.W. and S.M.C. thank the High Meadows Environmental Institute at Princeton for funding. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

All authors designed the study. S.E. programmed the main experiment. S.E. and S.V. worked with a freelance artist to develop the images of Biden and Trump. S.E. and S.M.C. pre-tested the images, ran the initial surveys to identify partisan commitments and ran the experimental sessions. S.E., S.M.C. and C.E. analysed the data. All authors interpreted the results. S.E., S.M.C. and C.E. wrote the paper with input from E.U.W. and S.V., S.E., S.M.C. and S.V. wrote the Supplementary Information with input from E.U.W. and C.E.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-022-01440-5.

Correspondence and requests for materials should be addressed to Sönke Ehret, Sara M. Constantino, Charles Efferson or Sonja Vogt.

Peer review information *Nature Human Behaviour* thanks Paul Smaldino, Simon Siegenthaler and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature portfolio

Corresponding author(s): Charles Efferson

Last updated by author(s): Jun 19, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
□	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
□	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	To collect the interactive group data we used the open source otree software, version 3.3, accessible at otree.org. We used the Qualtrics software (October 2020 version) to collect questionnaire data.
Data analysis	The code for implementing the data analyses and graphing the results is available the Open Science Framework at http://dx.doi.org/10.17605/OSF.IO/KN3A2, together with associated documentation.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data is publicly available at the Open Science Framework, at http://dx.doi.org/10.17605/OSF.IO/KN3A2, together with associated documentation.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study is quantitative experimental, based on the choices made by human subjects.
Research sample	The study sample consists of general population participants recruited from Prolific Academic, a general purpose online experiments platform. As per demographics, we have 45 % self-reported female, a mean age of 41 years, 71 % are Democrats, and 29% Republicans. The study sample is not representative of the general population. Our sample was recruited to measure the effect of identities on tipping dynamics. It contained therefore groups with a salient identity background, in our case, political identities during the November 2020 US presential election. In order to make it possible to run online interactive group games at appropriate scale, the study sample has been recruited from a larger pre-screened participant panel (N=2,788), 55% Democrat, 45 % Republican.
Sampling strategy	We used convenience sampling, where participants could opt into a participant panel. The panel initially screened based on several criteria, such as self-reported partisanship, political preferences for presidential candidates, residence in the United States and completeness of responses. We then invited participants from the panel to participate in pre-scheduled group sessions. We conducted statistical power analyses to determine the minimum number of groups needed in order to conduct general linear model test on the group analysis level at 80% power and 0.05 significance. With one degree of freedom and an effect size (f2) of 0.2, these analyses yielded roughly 50 groups required in minimum. For reference, our actual empirical R^2 in the basic linear model is 40, and the corresponding empirical Cohen's f2 effect size is 0.66. We used the R software, pwr package, to conduct power analyses based on the pwr.f2.test function.
Data collection	To collect the group level data, we used a virtual lab setup consisting of two steps. We first built a participant panel. We then recruited individuals from this panel into pre-scheduled sessions to participate in interactive virtual lab sessions. The participant panel was recruited via Prolific Academic, a company for providing online convenience samples. We targeted respondents located in the United States, from the general population as available on prolific academic, without principal restrictions on demographic background. Sampling for the participant panel involved filling six recruitment cells, based on interlocking and equally-sized demographic strata: self-identified partisanship (we included only self-reported Republicans and Democrats), age (two groups: 18 to 38 years, and 39 years and older) and sex (male and female). To qualify, participants had to be U.S. nationals residing in the U.S. at the time of the study. We only included participants with a Prolific Academic study approval rate of 95% or higher. Nobody was present besides the participants(s) and the researcher and their assistants. The researcher was not blinded to the experimental condition nor the the study hypothesis.
Timing	Data collection commenced on Oct 28, 2020, and stopped on December 16, 2020.
Data exclusions	The raw study data-set contained 77 groups and 908 players. We removed groups which did not have at least one post-intervention phase period. This was necessary because our comparison of interest is between the pre-intervention and post-intervention phase behaviours in the individual level analyses, and spillovers in the post-intervention phase for the group-level analyses. We also removed the first two groups collected at the beginning of the study due to a technical error. This resulted in the total removal of 9 groups, leaving 68 unique groups and 805 players.
Non-participation	Out of 68 groups, 28 were affected by players dropping out, of which 10 groups ended prematurly. A dropout is defined by a player being unresponsive (making no choice / decision) for more than 150 seconds, in which case the player is automatically excluded from the study.
Randomization	Participants were randomly allocated into neutral / identity label groups. To control for the partisan composition of the groups, at the beginning of each week of the study we pre-determined the composition of groups (as Democrat or Republican) for a given day, in countervailing order on the same day.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
□	☒	Human research participants
☒	□	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

See above

Recruitment

To collect the group level data, we used a virtual lab setup consisting of two steps. We first built a participant panel which announced the study as study in two parts, containing the recruitment questionnaire and the group study. We then recruited individuals from this panel into pre-scheduled sessions to participate in interactive virtual lab sessions. Prolific academic is a platform concerned with recruiting participants for paid research studies, and by using this platform we may have predominantly selected subjects which were self-selected due heterogenous motives of partaking in research, as well as earning money. This potential self-selection may have biased our results in favor of finding evidence for tipping, in both experimental conditions, compared to participants who have less interest in partaking in research and / or are less reactive to monetary incentives. We did not highlight or make any explicit mention on the partisanship nature of the study.

Ethics oversight

The study was approved by the ethics boards of the University of Lausanne (# “HEDGE” and # “HEDGE 2”), the University of Bern (#162020), and Princeton University (IRB# 12733). Informed consent was obtained for the recruitment survey and the main experiment.

Note that full information on the approval of the study protocol must also be provided in the manuscript.