Group identities can undermine social tipping after intervention

Sönke Ehret[1,†,*], Sara M. Constantino[2,3,4,†,*], Elke U. Weber[2,5,6], Charles Efferson[1,‡,*], & Sonja Vogt[1,7,8,‡,*]

1 Faculty of Business and Economics, University of Lausanne, Switzerland
2 School of Public and International Affairs, Princeton University, USA
3 School of Public Policy and Urban Affairs, Northeastern University, USA
4 Department of Psychology, Northeastern University, USA
5 Andlinger Center for Energy and the Environment, Princeton University, USA
6 Department of Psychology, Princeton University, USA
7 Centre for Development and Environment, University of Bern, Switzerland
8 Nuffield College, University of Oxford, United Kingdom
[†] These authors contributed equally.
[‡] These authors jointly supervised the research.
[*] Address correspondence to sonkeklaus.ehret@unil.ch, sara.constantino@gmail.com, charles.efferson@unil.ch, and sonja.vogt@unil.ch.

Abstract

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 policy makers do not anticipate. To examine this, we implemented an experiment around the 2020 U.S. elections. Participants faced consistent incentives to coordinate their choices. Once 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 payoff-irrelevant relabelling generated extreme differences. Control groups developed norms slowly before intervention but transitioned to new norms rapidly after intervention. 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 U.S., longstanding hostility towards same-sex marriage unravelled in a few years[2]. Germany began subsidising 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 of time. A family doing so understood that other families were also abandoning the practice and thus would probably not insist on women with bound feed as future wives for their sons. This change in beliefs about others created a positive feedback that accelerated abandonment[6, 1].

Social tipping has generated enormous interest as a way to trigger behaviour change[7] in many domains related to public health[8, 9], social justice[10, 11], resource conservation[12, 13], and climate change[14, 15, 16, 17]. Given such widespread interest, researchers and practitioners have a responsibility to investigate the conditions that support or undermine tipping[18, 19, 20]. Accordingly, we examine group identities as a mechanism hypothesised to interfere with tipping[21, 22, 23].

Proof of concept exists for tipping. Observational data show that cultural evolutionary processes support multiple norms, and punctuated cultural change certainly occurs[24, 25, 26, 27, 1, 28]. Experimental studies have also demonstrated that interventions can spark rapid transitions from one norm to another[29, 5]. Nonetheless, studies of tipping around gender-based violence[30, 11, 31, 32], political revolutions[33], and lab experiments[5] suggest important limits on our ability to identify when tipping is possible and how to maximise the chance of tipping[20]. The associated risk is that policy makers 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[24, 1]. 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.

A policy maker seeking rapid social change aims to trigger this dynamic. When conformity or coordination support a status quo norm inconsistent with policy objectives, a policy maker can promote an alternative norm by incentivising her preferred behaviour in a subset of the population only[29, 22].. 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, 38, 39].

If enough people exposed to the intervention change behaviour, conformity or coordination can switch from supporting the status quo to supporting the policy maker's alternative. Individuals who do not change 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 norm quickly, even without additional input from the policy maker. Behaviour change is partly exogenous, because some people change their behaviour due to exposure to the intervention, and partly endogenous, because some people change 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 standardised measure of how popular the policy maker'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 policy maker'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 policy makers 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[41, 5, 20]. 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 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[45, 46, 22, 47, 5, 48, 40]. People differ in many dimensions critical to behaviour change. These differences interact with the policy maker's choices to shape the potential for tipping[22, 47, 48, 40] 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[49, 42] and cooperate within the group to compete against other groups[50, 44, 51].

Whatever the past function of group identification, models suggest that once in place it can have an outsize influence on cultural evolution[21, 22, 23]. Outgroup aversion[21] represents a special challenge for the policy maker. 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 policy maker promoting a single behaviour for the entire population[22]. To illustrate, imagine a population subdivided into two groups. Members of one group value low-emission transport. 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 policy maker 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 policy maker's efforts could even strengthen the tie between group identity and the behaviour in question[52]. In our example, this would mean the policy maker's efforts increase the value of pollution for some people and solidify their resistance to change. Whatever the details, the policy maker 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 incentivised online experiment around the 2020 election for U.S. President. A U.S. 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 choice options with images designed to activate partisan political identities (Fig. 1). Partisan loyalties provide an important component

of identity in contemporary U.S. 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 choice options (Supplementary Fig. 2), and in this sense our treatment manipulation was payoff-irrelevant.

Each session had the following structure regardless of treatment. We formed an experimental group of either all Republicans or all Democrats. 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 payoffs simply favoured coordinating; they did not favour coordinating on a specific option (Table 1a). 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 payoffs to favour changing from the status quo choice, whatever this may have been, to the other choice option, which we call the "alternative" (Table 1b; Supplementary Information § 4.2). Non-targeted participants retained their original payoffs (Table 1c). Participants then continued playing repeatedly under this new incentive structure.

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

Our intervention created heterogeneity in material incentives. After intervention, targeted participants faced material incentives that favoured behaviour change in the precise sense that the alternative was dominant for self-regarding players (Table 1b). Non-targeted participants faced material incentives that simply favoured, for self-regarding players, coordinating on either option. 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 payoffs, and some would experience a strict increase.

Crucially, however, people do not simply care about their own material payoffs [59, 60, 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 payoffs simply by choosing the alternative. Non-targeted players could earn small or intermediate payoffs depending on whether they coordinated with their partners. Thus, if an experimental group were to tip fully to the alternative, targeted players would experience persistent inequality to their advantage and 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, however, controlled for this possibility by always using the same payoff 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 randomise 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 the labels activated partisan identities, the effects should have depended on how 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 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 only incentivised change via money. Finally, imagine that people traded money against identity concerns in heterogeneous ways. Heterogeneity implies that some players in identity sessions might have changed behaviour, while others might not. The result might have been no norm at all after intervention.

We first present results from a pre-registered analysis of spillovers (Methods) that reflects a central concern from the policy maker's perspective. Spillovers[22] are a normalised measure of how common the policy maker's alternative is in the long-run while accounting for the size of the policy maker's intervention (Methods). Spillovers do not explicitly account for choice dynamics; they quantify the final outcome net the policy maker'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 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 also address whether identity concerns among our participants were based on pre-existing identities already in place when participants began the experiment versus identities that emerged within the context of the experiment itself.

Results

Spillovers and individual choice.

In 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 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 normalised 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% CI $= [-0.38, 0.06]$, $p = 0.20$, 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, like an intervention, that sets behaviour change in motion. The resulting outcome goes 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 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 Information § 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 players faced an equilibrium-selection problem. Neutral labels did not help, and players simply had to develop an idiosyncratic local norm via repeated play with feedback. Political labels provided 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, 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, 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 non-targeted participants in the neutral treatment 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 payoffs. In our neutral treatment, the absence of a focal point[66] meant that players needed time to develop status quo norms. Neutral groups, however, were able to transition rapidly to an alternative norm when circumstances changed. Tipping and its payoff consequences are easily seen as a rapid increase in payoffs after intervention in neutral sessions (Fig. 5a). Identity sessions show the opposite pattern. Players established status quo norms quickly. However, with an alternative running counter to their pre-existing identities, 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 if and how choices changed after November 7, 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 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, 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 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 examine effects related to the inequality the intervention created. As explained, our design controlled for average effects by using the same payoff 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 summarised tolerance to hierarchy and inequality (Supplementary Information § 7.1 and § 8.2), and we also measured preferences for economic equality (Supplementary Information § 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 had an effect on the probability of choosing the alternative behaviour at the end of the post-intervention phase (Supplementary Table 17). This result may appear surprising because empirical research shows that people are averse to inequality, with considerable heterogeneity within and across cultures[62, 67, 59, 68, 69, 60, 61]. Why, then, do we find no evidence that inequality aversion affected the probability of adopting the alternative despite its introduction of inequalities?

Using the Fehr-Schmidt model[62], we show that inequality aversion has two coun-

tervailing effects in a setting like ours (Supplementary Information § 7.2), and the effect that supports tipping is likely to dominate. The model distinguishes between two types of inequality. Advantageous inequality means 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 only earn 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, 68, 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 Information § 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 only experience disadvantageous inequality 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 Information § 7.2).

The upshot is that in our experiment the effects of inequality aversion and ordinary coordination incentives were likely redundant. Our intervention created pressure to tip with large unconditional payoffs 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 payoffs 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 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

payoff-irrelevant manipulation to activate affective responses based on pre-existing identities. That said, choice dynamics in identity sessions diverged from dynamics in neutral sessions almost immediately (Fig. 3). Treatment differences in behaviour change thus resulted in principle both from the link between political labels and pre-existing identities and from the divergent dynamics 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 to neutral sessions, 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 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 neutral sessions.

Broadly, participants in identity sessions may have experienced the normative force of the status quo more strongly than their counterparts in neutral sessions for two different reasons. On the one hand, status quo norms were consistent with pre-existing identities in identity sessions. As discussed, status quo norms and the party loyalties of participants were perfectly correlated in the identity treatment, but they were unrelated in the neutral treatment. On the other hand, when comparing identity sessions to neutral sessions, status quo norms in identity sessions developed quickly, produced homogeneous behaviour before intervention, and allowed 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 subjects before the experiment, we measured two forms of affective polarisation[54, 55, 70] for each participant. One measure quantified polarisation in terms of Democrats versus Republicans in general and the other in terms of Biden versus Trump specifically (Supplementary Information § 7.3.1). Increasing polarisation was statistically associated with a reduced tendency to choose the alternative after intervention in the identity treatment but not 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 Information § 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 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 Information § 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 Information § 7.4). Lastly, as explained above, the distinctive dynamics in identity sessions might have allowed 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 identity sessions should have responded to early feedback about others more strongly than in neutral sessions. This follows simply from the fact that this feedback was the only way for participants to draw the inferences in question. We did not find any statistical evidence for this possibility (Supplementary Information § 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 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 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 do not examine would occur when the policy maker 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 policy maker because the policy maker's behavioural objectives correlate neatly with the pre-existing subdivision of the population.

We focus instead on settings in which the policy maker'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 (e.g. 10%) than what we used in this study, small to moderate amounts of outgroup aversion may actually help destabilise 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 U.S. 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 marginalised groups pretending to adhere to the norms of a dominant group in step with the policy maker.

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, 22, 23], and our results are consistent with this idea. Identity policing can manifest itself in at least two ways relevant to policy. First, the policy maker's target population may not be strongly subdivided, but the policy maker herself represents an aversive outgroup. In the 1950's, 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 colonisers, and the ban on cutting actually seemed to increase commitment to the practice as a hallmark of cultural

identity[52].

Second, the target population is subdivided, people care greatly about protecting the group identities that result, and the policy maker's behavioural objectives do not fit well with this landscape of pre-existing identities. Climate change, for example, is one of several politically polarising issues in the contemporary U.S. 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.

For 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 policy maker 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 (link). The ad first attempted to decouple masks from the partisan baggage they had acquired in the U.S. 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 conservative Republicans attach to CNN[73] did not help. Regardless, the strategy is clear. The ad did not address the partisan divide in the U.S. 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 emphasise alternative rites of passage[74] designed to allow families to integrate their daughters in 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 policy makers 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 policy makers with an unusual challenge. Because tipping has impressive potential, strategies to provoke tipping will presumably remain a part of the policy maker'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 the policy maker's preferences are misaligned with those of the citizens under her 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 polarisation 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 U.S. between October 28 and December 16, 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 participants online via Prolific. At the recruitment stage, we screened potential participants based on their self-reported political affiliations and responses to two questions about political preferences. In particular, we asked about Biden and Trump using feelings thermometers[55, 54, 70, 79]. We used these responses to recruit participants to the main study who were either (i) warm about Biden and cold about Trump or (ii) cold about Biden and warm about Trump (Supplementary Information § 5.2). Altogether, we recruited 566 participants in category (i), all of whom reported being Democrats, and 235 participants in category (ii), all of whom

reported being Republican. 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 into the final sample.

For the main experiment, we formed experimental groups of either all Republicans or all Democrats. 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 minimise participant dropout (see below). The Supplementary Information (§ 5 and § 6.1) provides additional details and analyses related to recruitment, sample composition, and dropout.

Repeated game play and treatments.

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 1a). 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, i.e. the status quo, at the end of the pre-intervention phase.

To begin the post-intervention phase, we applied a new payoff matrix to a subset of players (Table 1b). The remaining players retained their original incentives (Table 1c). The intervention was randomly assigned to 50% of players in the experimental group at the start of the session (Supplementary Information § 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 Information § 6.3 for 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 payoff matrix. Labels for choices, whether neutral or political, where 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 Information § 3.4).

For feedback, participants received three pieces of information at the beginning of each period after the first. Namely, each participant saw (i) the complete distribution of choices, from the previous period, among 10 randomly selected players in the experimental group, (ii) the choice of the focal player's partner in the previous period,

and (iii) the points the focal participant earned in the previous period. Communicating the choices among 10 randomly selected players allowed us to continue a session when someone dropped out without disturbing our feedback protocol. Specifically, because 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 (see Supplementary Information § 6.1).

Points from the games were converted to dollars at a fixed rate. The total payoff for each participant was calculated by summing the payoffs from five randomly selected periods. Participants were informed about payment and other procedures before the start of the game (Supplementary Information § 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. Analyses were pre-registered (https://osf.io/84jpq) unless otherwise indicated.

Table 2 presents an analysis of spillovers[22]. Spillovers provide a normalised 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

$$\Theta_j = \begin{cases} \dfrac{\hat{q}_j - \phi_j}{1 - \phi_j} & \text{if} \quad \hat{q}_j > \phi_j \\ \dfrac{\hat{q}_j - \phi_j}{\phi_j} & \text{otherwise.} \end{cases}$$

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

To examine spillovers, we ran an Ordinary Least Squares (OLS) regression with

spillovers as a function of treatments. The model is

$$\Theta_j = \beta_0 + \beta_1 u_j + \epsilon_j,$$

where $j \in \{1, 2, ..., J\}$ indexes experimental group, $u_j \in \{0, 1\}$ indicates whether group j was in the neutral treatment ($u_j = 0$) or identity treatment ($u_j = 1$), and ϵ_j is a group error term. We used robust standard errors because the 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 (OLS) 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 minimises the role of transient dynamics and thus focuses on transitions between equilibria. Results hold with more periods (Supplementary Table 2).

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

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

The index $i \in \{1, 2, ..., I\}$ specifies observation at the level of an individual making a single choice, and $c_i \in \{0, 1\}$ indicates if the associated choice was the status quo ($c_i = 0$) or alternative ($c_i = 1$). The variable $u_i \in \{0, 1\}$ indicates if observation i was associated with a participant in the neutral condition ($u_i = 0$) or identity condition ($u_i = 1$). The variable $z_i \in \{0, 1\}$ indicates if 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 if 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 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 $[u_i = 0 \wedge z_i = 0 \wedge \tau_i = 0]$, i.e. (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 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 Information § 6.6).

Data Availability Statement

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

Code Availability Statement

The code for analyses is publicly available at the Open Science Framework, at `http://dx.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 the Qualtrics software (October 2020 version) to collect questionnaire data.

Acknowledgements

For helpful comments, we thank Joan Barceló, Sirio Lonati, Heinrich Nax, Nikos Nikiforakis, Simon Siegenthaler, Paul Smaldino, Christian Zehnder, two anonymous reviewers, and seminar participants at the Collegio Carlo Alberto, University of Lausanne, NYUAD, Oxford, Princeton University, and the University of Zurich. The study was funded by the Swiss National Science Foundation (Nr. 100018 185417/1) to CE and SV). 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. SE programmed the main experiment. SE and SV worked with a free-lance artist to develop the images of Biden and Trump. SE and SC pre-tested the images, ran initial surveys to identify partisan commitments, and ran the experimental sessions. SE, SC, and CE analysed the data. All authors interpreted the results. SE, SC, and CE wrote the paper with input from the other authors. SE, SC, and SV wrote the Supplementary Information with input from the other authors.

Competing Interests

The authors declare no competing interests.

Table 1 | Participant payoffs. Matrices show row player payoffs 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. **a**, Payoffs 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, i.e. after the experimental group had settled on a status quo norm. **b**, The intervention encouraged behaviour change by introducing new payoffs that favoured the alternative option among targeted (T) players, regardless of the partner's choice. These payoffs held for the entire post-intervention phase. **c**, Non-targeted (NT) players retained their original payoffs post-intervention.

	(a) Pre-int (all)		**(b)** Post-int (T)		**(c)** Post-int (NT)	
	SQ	Alt	SQ	Alt	SQ	Alt
SQ	200	50	200	50	200	50
Alt	50	200	350	350	50	200

Table 2 | Spillovers by treatment. Spillovers[22] take values in $[-1, 1]$ and provide a normalised measure of long-run behaviour in a population while accounting for the size of the intervention (Methods). Results are from an OLS 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 choice options in the identity treatment resulted in a large reduction in spillovers (Identity).

	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]

The p values are from two-sided z tests.
(Robust standard errors)
[95% confidence intervals]

Table 3 | Participant chooses the alternative behaviour. Linear probability models (Methods) for individual choices in the final periods of the pre- and post-intervention phases. Election is a dummy indicating sessions after 7 November 2020, the day the election was called. Composite dummies are defined jointly over (i) treatment (Neutral vs. Identity), (ii) whether the participant was targeted (T) or not (NT), and (iii) 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. Results are robust to including day fixed effects (Supplementary Table 14), including more periods in the analysis (Supplementary Table 2), and to alternative estimate methods with a random-effects logit model (Supplementary Table 16).

	Choose alternative behaviour	
	Model 1	Model 2
Intercept	0.13 (0.02) p<0.001, [0.10, 0.16]	0.14 (0.02) 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) p=0.19, [-0.08, 0.02]	-0.07 (0.04) p=0.07, [-0.15, 0.01]
(Neutral,NT,Post-int)	0.63 (0.05) p<0.001, [0.53, 0.73]	0.73 (0.06) p<0.001, [0.61, 0.85]
(Neutral,T,Post-int)	0.81 (0.03) p<0.001, [0.76, 0.86]	0.80 (0.04) p<0.001, [0.72, 0.88]
(Identity,NT,Pre-int)	-0.12 (0.02) p<0.001, [-0.15, -0.08]	-0.12 (0.02) p<0.001, [-0.17, -0.07]
(Identity,T,Pre-int)	-0.12 (0.02) p<0.001, [-0.15, -0.08]	-0.13 (0.02) p<0.001, [-0.17, -0.09]
(Identity,NT,Post-int)	0.09 (0.06) p=0.17, [-0.04, 0.21]	0.03 (0.07) p=0.65, [-0.11, 0.17]
(Identity,T,Post-int)	0.53 (0.05) p<0.001, [0.44, 0.63]	0.54 (0.06) 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 p values are from two-sided z-tests.
(Cluster-robust standard errors)
[95% confidence intervals]

Figure 1 | The two images used to label 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).

Figure 2 | Distributions of normalised spillovers by treatment. The spillover[22] is a normalised 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. **a**, The distribution of spillovers in the neutral treatment. **b**, The distribution of spillovers in the identity treatment. The difference in spillovers by treatment is large and highly significant (Table 2).

Figure 3 | Choice dynamics by treatment. 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 participants coordinated on the status quo (blue), coordinated on the alternative (green), or miscoordinated (red) for each period. (**a**) In neutral sessions, participants were relatively slow to converge on the status quo before intervention and relatively fast to converge on the alternative after intervention. (**b**) In identity sessions, participants converged quickly before intervention, requiring fewer overall trials to meet the intervention criteria, but persisted in a state of chronic disagreement after intervention. For reference, under random matching the maximum possible expected rate of miscoordination is 0.5.

Figure 4 | Choice of alternative behaviour by treatment. Effect sizes and 95% confidence intervals from Model 1 in Table 3, $N = 1546$ 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. Curly brackets show results from various linear combinations discussed in the main text, all of which are based on the cluster-robust standard errors from Table 3.The red dashed vertical line represents no effect. To present these linear combinations graphically, we have reordered the effects compared to Table 3.

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

References

[1] Young, H. P. The Evolution of Social Norms. *Annual Review of Economics* **7**, 359–387 (2015). URL https://doi.org/10.1146/annurev-economics-080614-115322.

[2] Rosenfeld, M. J. Moving a Mountain: The Extraordinary Trajectory of Same-Sex Marriage Approval in the United States. *Socius* **3**, 2378023117727658 (2017). URL https://doi.org/10.1177/2378023117727658.

[3] Rode, J. & Weber, A. Does localized imitation drive technology adoption? a case study on rooftop photovoltaic systems in germany. *Journal of Environmental Economics and Management* **78**, 38–48 (2016).

[4] Winkelmann, R. *et al.* Social tipping processes towards climate action: a conceptual framework. *Ecological Economics* **192**, 107242 (2022).

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

[6] Mackie, G. Ending Footbinding and Infibulation: A Convention Account. *American Sociological Review* **61**, 999–1017 (1996). URL https://www.jstor.org/stable/2096305.

[7] Nyborg, K. *et al.* Social norms as solutions. *Science* **354**, 42–43 (2016). URL https://science.sciencemag.org/content/354/6308/42. Publisher: American Association for the Advancement of Science Section: Policy Forum.

[8] Christakis, N. A. & Fowler, J. H. The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine* **357**, 370–379 (2007). URL https://doi.org/10.1056/NEJMsa066082.

[9] Arnot, M. *et al.* How evolutionary behavioural sciences can help us understand behaviour in a pandemic. *Evolution, Medicine, and Public Health* **2020**, 264–278 (2020).

[10] Cloward, K. *When Norms Collide: Local Responses to Activism against Female Genital Mutilation and Early Marriage* (Oxford University Press). URL https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780190274917.001.0001/acprof-9780190274917.

[11] Platteau, J.-P., Camilotti, G. & Auriol, E. Eradicating women-hurting customs. *Towards gender equity in development* **319** (2018).

[12] Castilla-Rho, J. C., Rojas, R., Andersen, M. S., Holley, C. & Mariethoz, G. Social tipping points in global groundwater management. *Nature Human Behaviour* **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. *Biological Conservation* **257**, 109092 (2021).

[14] Barrett, S. & Dannenberg, A. Sensitivity of collective action to uncertainty about climate tipping points. *Nature Climate 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). URL `https://science.sciencemag.org/content/364/6436/132`.

[17] Otto, I. M. *et al.* Social tipping dynamics for stabilizing Earth's climate by 2050. *Proceedings of the National Academy of Sciences* **117**, 2354–2365 (2020). URL `https://www.pnas.org/content/117/5/2354`.

[18] Bicchieri, C. & Dimant, E. Nudging with care: The risks and benefits of social information. *Public choice* 1–22 (2019).

[19] Smith, S. R., Christie, I. & Willis, R. Social tipping intervention strategies for rapid decarbonization need to consider how change happens. *Proceedings of the National Academy of Sciences* **117**, 10629–10630 (2020).

[20] Efferson, C. Policy to activate cultural change to amplify policy. *Proceedings of the National Academy of Sciences* **118** (2021).

[21] Smaldino, P. E., Janssen, M. A., Hillis, V. & Bednar, J. Adoption as a social marker: Innovation diffusion with outgroup aversion. *The Journal of Mathematical Sociology* **41**, 26–45 (2017).

[22] Efferson, C., Vogt, S. & Fehr, E. The promise and the peril of using social influence to reverse harmful traditions. *Nature Human Behaviour* **4**, 55–68 (2020). URL https://www.nature.com/articles/s41562-019-0768-2. Number: 1 Publisher: Nature Publishing Group.

[23] Smaldino, P. E. & Jones, J. H. Coupled dynamics of behaviour and disease contagion among antagonistic groups. *Evolutionary Human Sciences* **3** (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. *American Anthropologist* **103**, 992–1013 (2001). URL https://anthrosource.onlinelibrary.wiley.com/doi/abs/10.1525/aa.2001.103.4.992.

[25] Young, H. P. & Burke, M. A. Competition and custom in economic contracts: a case study of illinois agriculture. *American Economic Review* **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? *The Economic Journal* **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. *Journal of the European Economic Association* **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). URL https://science.sciencemag.org/content/360/6393/1116. Publisher: American Association for the Advancement of Science Section: Report.

[30] Bellemare, M. F., Novak, L. & Steinmetz, T. L. All in the family: Explaining the persistence of female genital cutting in West Africa. *Journal of Development Economics* **116**, 252–265 (2015).

[31] Muthukrishna, M. Cultural evolutionary public policy. *Nature Human Behaviour* **4**, 12–13 (2020).

[32] Novak, L. Persistent norms and tipping points: The case of female genital cutting. *Journal of Economic Behavior & Organization* **177**, 433–474 (2020).

[33] Kuran, T. Now out of never: The element of surprise in the east european revolution of 1989. *World Politics: A Quarterly Journal of International Relations* 7–48 (1991).

[34] Shell-Duncan, B. & Hernlund, Y. *Female "circumcision" in Africa: culture, controversy, and change* (Lynne Rienner Publishers, 2000).

[35] Christakis, N. A. & Fowler, J. H. The Collective Dynamics of Smoking in a Large Social Network. *New England Journal of Medicine* **358**, 2249–2258 (2008). URL `https://doi.org/10.1056/NEJMsa0706154`.

[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. Economic and social impacts of the media. In *Handbook of Media Economics*, vol. 1, 723–768 (Elsevier, 2015).

[38] La Ferrara, E. Mass media and social change: Can we use television to fight poverty? *Journal of the European Economic Association* **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. *Bulletin of the World Health Organization* **99**, 819 (2021).

[41] Granovetter, M. Threshold models of collective behavior. *American Journal of Sociology* **83**, 1420–1443 (1978).

[42] Efferson, C., Lalive, R. & Fehr, E. The Coevolution of Cultural Groups and Ingroup Favoritism. *Science* **321**, 1844–1849 (2008). URL `https://science.sciencemag.org/content/321/5897/1844`.

[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 in Cognitive Sciences* (2020).

[45] Young, H. P. Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *American Economic Review* **99**, 1899–1924 (2009).

[46] Jackson, M. O. & López-Pintado, D. Diffusion and contagion in networks with heterogeneous agents and homophily. *Network Science* **1**, 49–67 (2013).

[47] Gavrilets, S. The dynamics of injunctive social norms. *Evolutionary Human Sciences* **2** (2020).

[48] Berger, J., Efferson, C. & Vogt, S. Tipping pro-environmental norm diffusion at scale: opportunities and limitations. *Behavioural Public Policy* 1–26 (2021).

[49] Boyd, R. & Richerson, P. J. The evolution of ethnic markers. *Cultural Anthropology* **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. *Nature Communications* **11**, 1–9 (2020).

[52] Thomas, L. M. 'Ngaitana (I will circumcise myself)': Lessons from colonial campaigns to ban excision in Meru, Kenya (2000).

[53] Smaldino, P. E. & Turner, M. A. Covert signaling is an adaptive communication strategy in diverse populations. *Psychological review* (2021).

[54] Iyengar, S., Sood, G. & Lelkes, Y. Affect, Not Ideology: A Social Identity Perspective on Polarization. *Public Opinion Quarterly* **76**, 405–431 (2012). URL `https://doi.org/10.1093/poq/nfs038`.

[55] Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N. & Westwood, S. J. The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science* **22**, 129–146 (2019). URL `https://doi.org/10.1146/annurev-polisci-051117-073034`. _eprint: https://doi.org/10.1146/annurev-polisci-051117-073034.

[56] McConnell, C., Margalit, Y., Malhotra, N. & Levendusky, M. The Economic Consequences of Partisanship in a Polarized Era. *American Journal of Political Science* **62**, 5–18 (2018). URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12330`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12330.

[57] Finkel, E. J. *et al.* Political sectarianism in America. *Science* **370**, 533–536 (2020). URL https://science.sciencemag.org/content/370/6516/533. Publisher: American Association for the Advancement of Science Section: Policy Forum.

[58] Bowles, S. *Microeconomics: Behavior, Institutions, and Evolution* (Princeton: Princeton University Press, 2004).

[59] Henrich, J. *et al.* "economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences* **28**, 795–815 (2005).

[60] Cooper, D. J. & Kagel, J. H. Other-regarding preferences. *The Handbook of Experimental Economics* **2**, 217–289 (2016).

[61] Falk, A. *et al.* Global Evidence on Economic Preferences*. *The Quarterly Journal of Economics* **133**, 1645–1692 (2018). URL https://academic.oup.com/qje/article/133/4/1645/5025666.

[62] Fehr, E. & Schmidt, K. M. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics* **114**, 817–868 (1999).

[63] Fairlie, R. W. & Robinson, J. Experimental evidence on the effects of home computers on academic achievement among schoolchildren. *American Economic Journal: Applied Economics* **5**, 211–40 (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. *American Economic Journal: Economic Policy* **8**, 80–114 (2016).

[65] Schelling, T. C. *The Strategy of Conflict* (Harvard University 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. *American Economic Review* **98**, 1443–58 (2008).

[67] Goeree, J. K. & Holt, C. A. Asymmetric inequality aversion and noisy behavior in alternating-offer bargaining games. *European Economic Review* **44**, 1079–1089 (2000).

[68] Blanco, M., Engelmann, D. & Normann, H. T. A within-subject analysis of other-regarding preferences. *Games and Economic Behavior* **72**, 321–338 (2011).

[69] Beranek, B., Cubitt, R. & Gächter, S. Stated and revealed inequality aversion in three subject pools. *Journal of the Economic Science Association* **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. *Nature Human Behaviour* **5**, 28–38 (2021). URL `https://www.nature.com/articles/s41562-020-01012-5`. Number: 1 Publisher: Nature Publishing Group.

[71] Angrist, J. D. & Pischke, J.-S. *Mostly harmless econometrics* (Princeton University Press, 2009).

[72] Fiorina, M. P. *Unstable Majorities: Polarization, Party Sorting, and Political Stalemate* (Hoover press, 2017).

[73] Stroud, N. J. & Lee, J. K. Perceptions of cable news credibility. *Mass Communication and Society* **16**, 67–88 (2013).

[74] Hughes, L. Alternative rites of passage: Faith, rights, and performance in FGM/C abandonment campaigns in Kenya. *African Studies* **77**, 274–292 (2018).

[75] Gulesci, S. *et al.* A stepping stone approach to understanding harmful norms (2021).

[76] Hänni, S. & Lichand, G. Harming to signal: child marriage vs. public donations in malawi. *University of Zurich, Department of Economics, Working Paper* (2021).

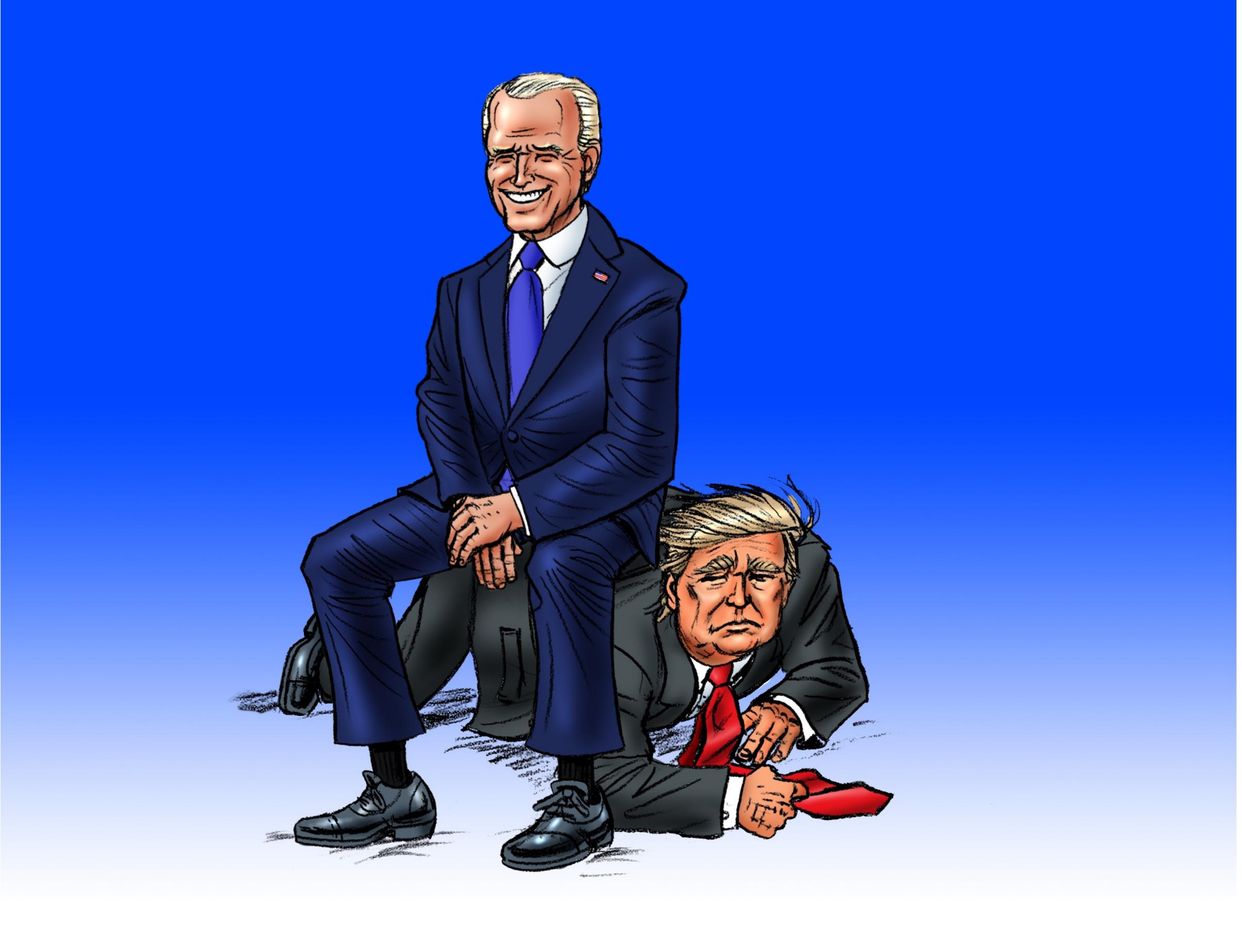
[77] Efferson, C., Vogt, S. & von Flüe, L. Activating cultural evolution for good when people differ from each other. In Kendal, J., Kendal, R. & Tehrani, J. (eds.) *Oxford Handbook of Cultural Evolution* (Oxford University Press, 2023).

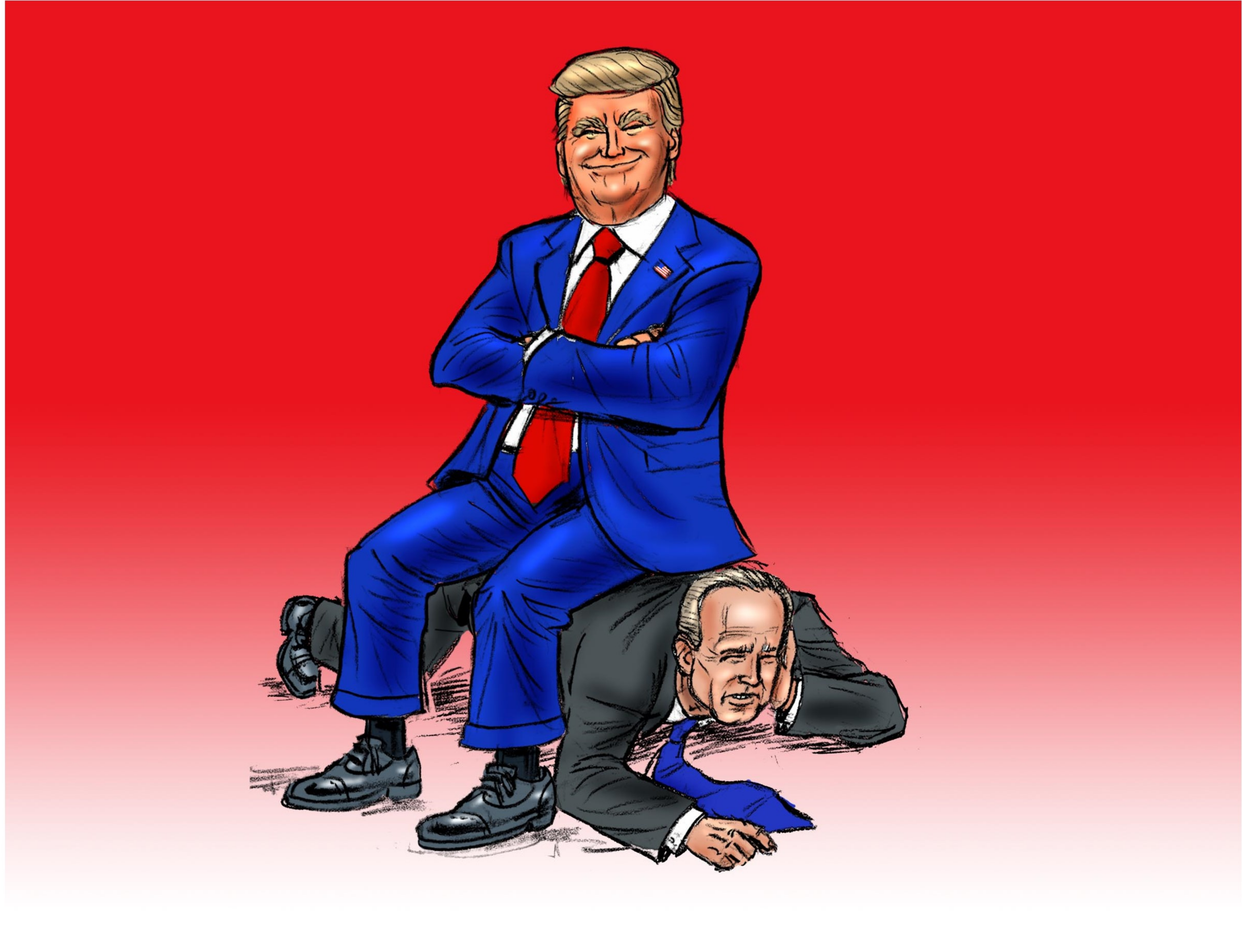
[78] Chen, Y. & Li, S. X. Group identity and social preferences. *American Economic Review* **99**, 431–57 (2009).

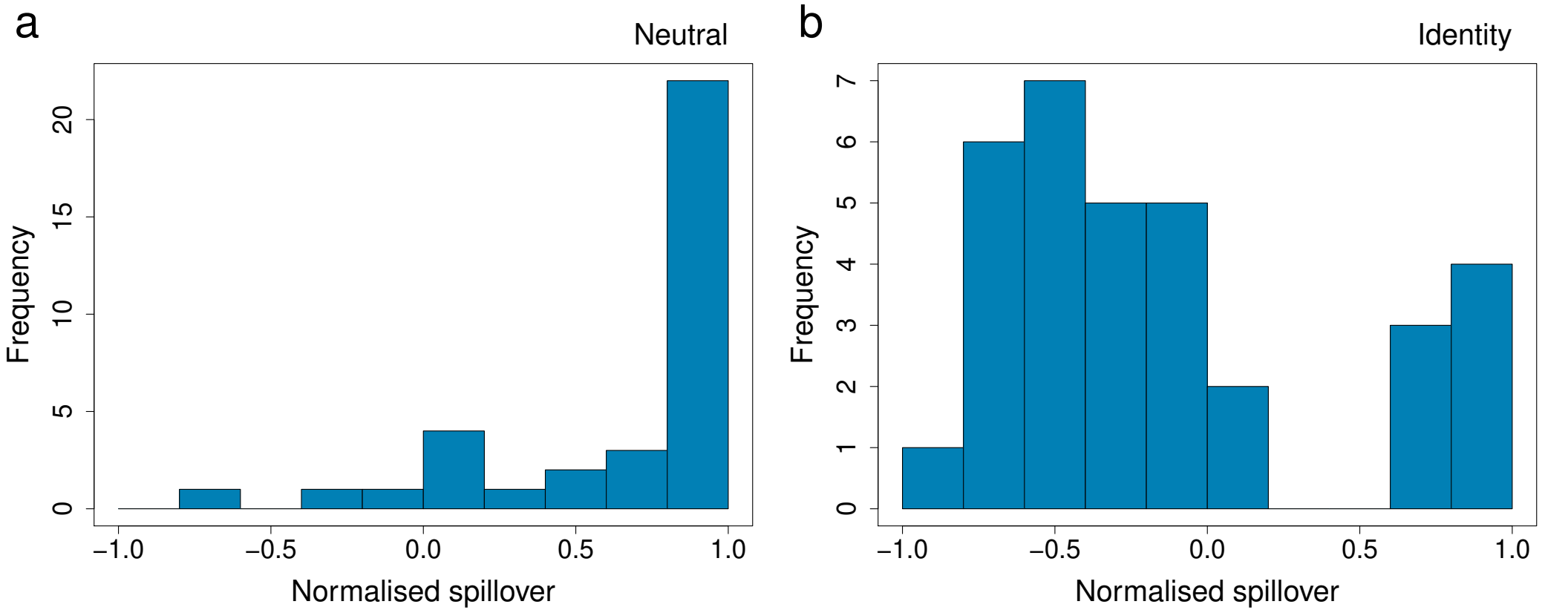
[79] Mason, L. Ideologues without Issues: The Polarizing Consequences of Ideological Identities. *Public Opinion Quarterly* **82**, 866–887 (2018). URL `https://doi.org/10.1093/poq/nfy005`.

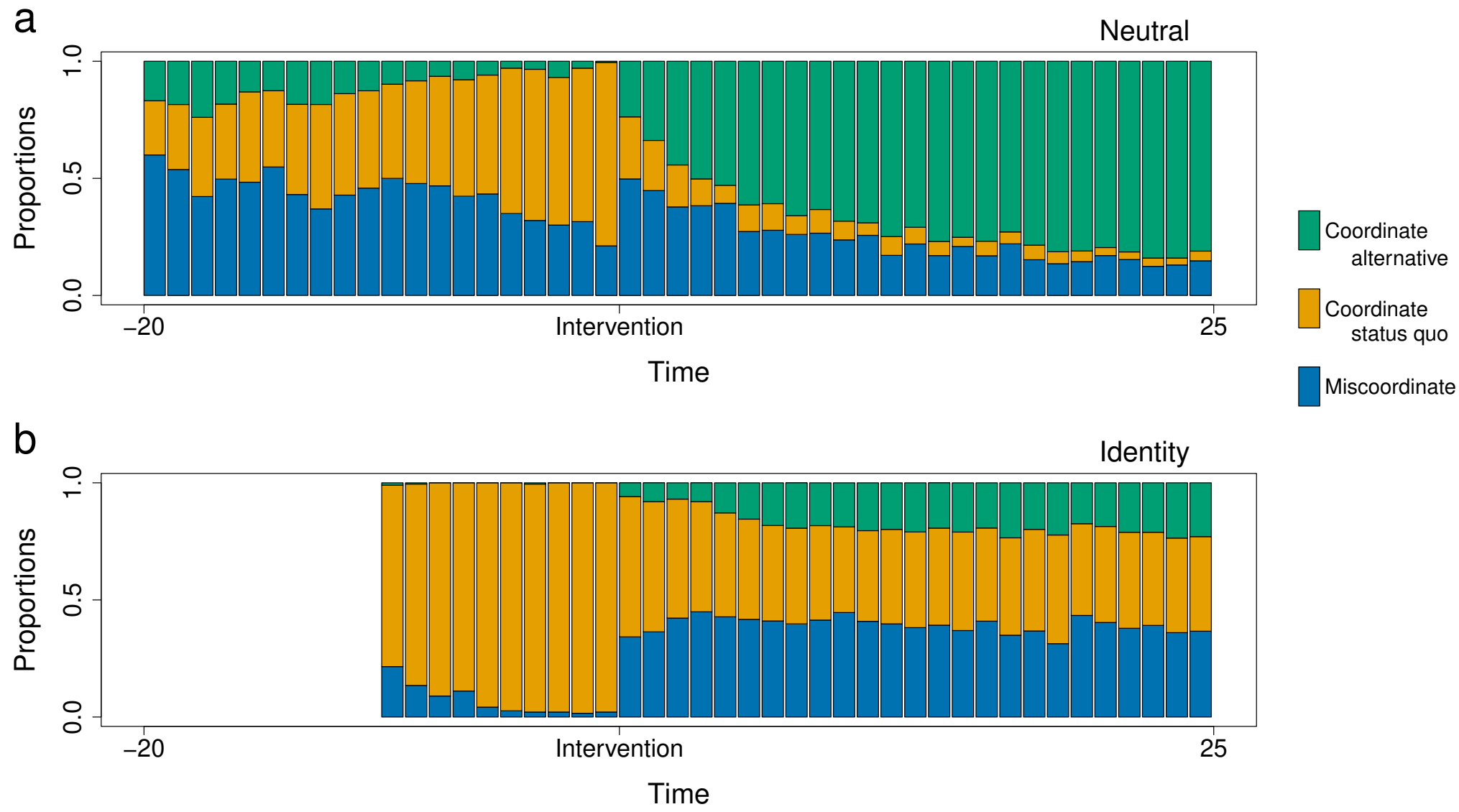
[80] Wooldridge, J. M. *Econometric Analysis of Cross Section and Panel Data* (MIT press, 2010).

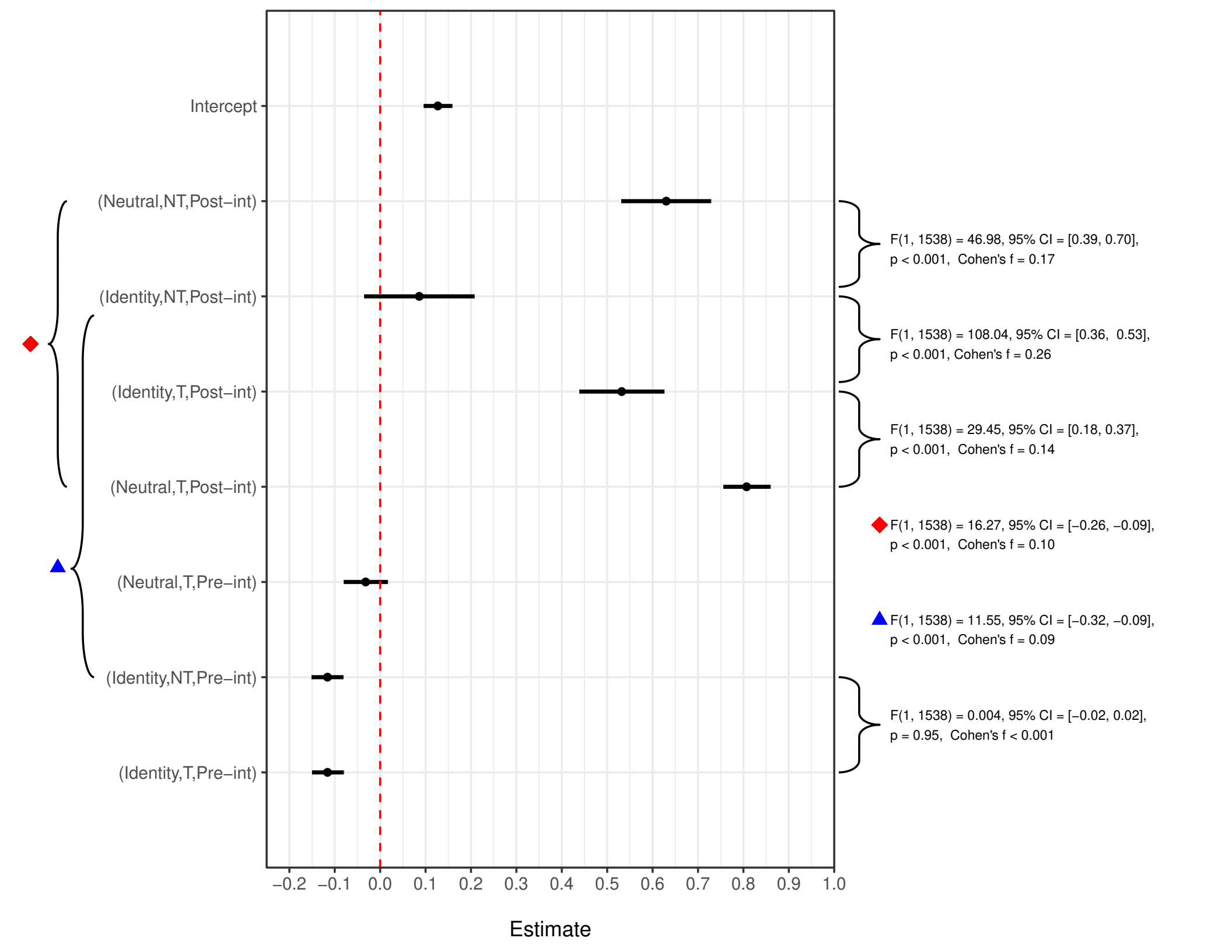
[81] Arai, M. Cluster-robust standard errors using r. *Note available http://people. su. se* (2011).

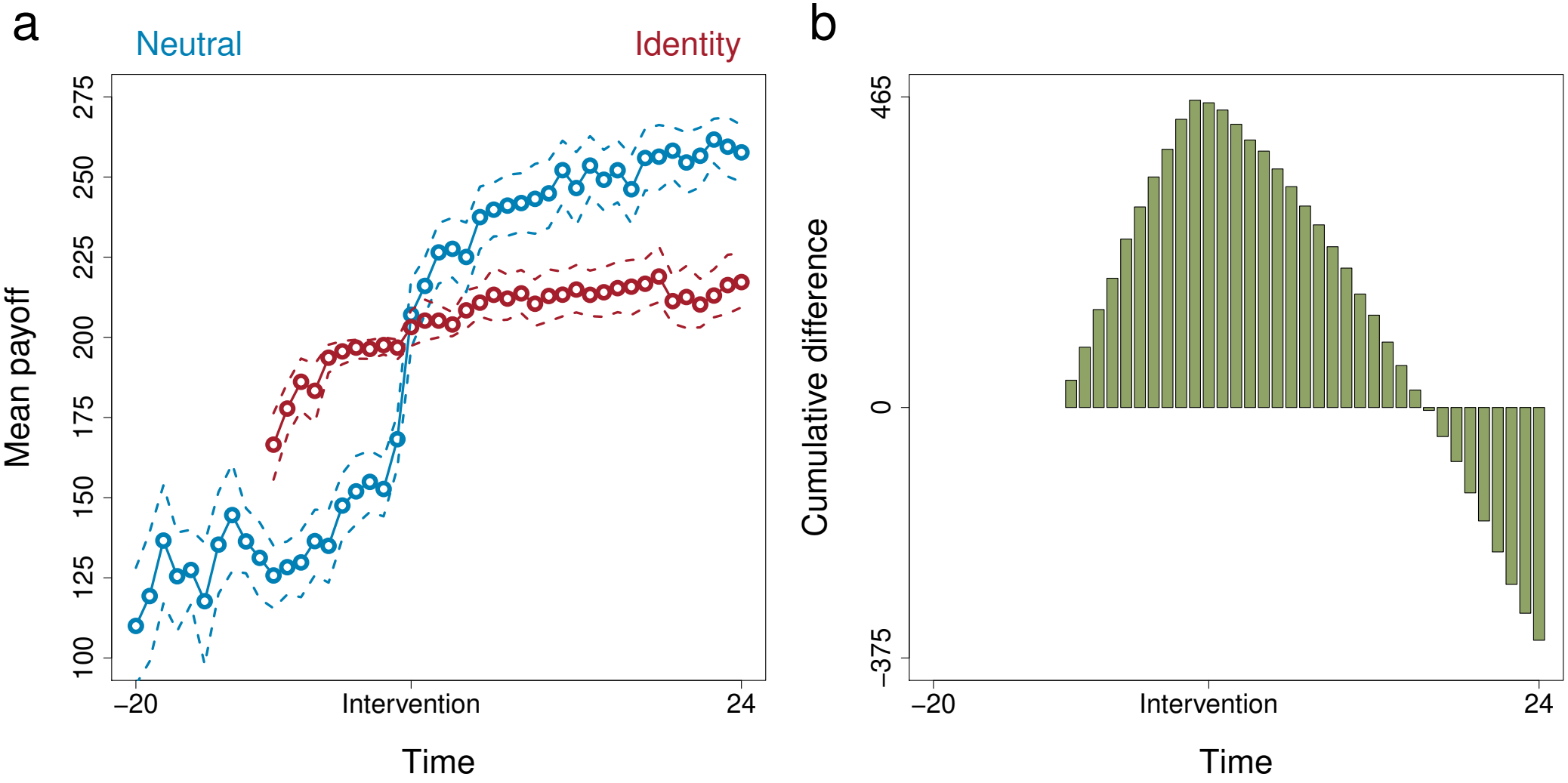





a
Neutral
Frequency
0
5
10
15
20
−1.0
−0.5
0.0
0.5
1.0
Normalised spillover
b
Identity
Frequency
0
1
2
3
4
5
6
7
−1.0
−0.5
0.0
0.5
1.0
Normalised spillover


a
Neutral
Proportions
1.0
0.5
0.0
–20
Intervention
25
Time
Coordinate alternative
Coordinate status quo
Miscoordinate
b
Identity
Proportions
1.0
0.5
0.0
–20
Intervention
25
Time


Intercept
(Neutral,NT,Post–int)
(Identity,NT,Post–int)
(Identity,T,Post–int)
(Neutral,T,Post–int)
(Neutral,T,Pre–int)
(Identity,NT,Pre–int)
(Identity,T,Pre–int)
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Estimate
F(1, 1538) = 46.98, 95% CI = [0.39, 0.70], p < 0.001, Cohen's f = 0.17
F(1, 1538) = 108.04, 95% CI = [0.36, 0.53], p < 0.001, Cohen's f = 0.26
F(1, 1538) = 29.45, 95% CI = [0.18, 0.37], p < 0.001, Cohen's f = 0.14
F(1, 1538) = 16.27, 95% CI = [−0.26, −0.09], p < 0.001, Cohen's f = 0.10
F(1, 1538) = 11.55, 95% CI = [−0.32, −0.09], p < 0.001, Cohen's f = 0.09
F(1, 1538) = 0.004, 95% CI = [−0.02, 0.02], p = 0.95, Cohen's f < 0.001


a
Neutral
Identity
Mean payoff
100
125
150
175
200
225
250
275
–20
Intervention
24
Time
b
Cumulative difference
465
0
–375
–20
Intervention
24
Time