When norm change hurts

Charles Efferson[1,*], Sönke Ehret[1], Lukas von Flüe[1], and Sonja Vogt[1,*]

[1]Faculty of Business and Economics, University of Lausanne, Switzerland

[*]Address correspondence to *charles.efferson@unil.ch* and *sonja.vogt@unil.ch*.

Word count: c. 9500 (excluding abstract, references, and captions)

Key words: social tipping, behaviour change, coordination, conformity, social norms

Abstract

Applied cultural evolution includes any effort to mobilise social learning and cultural evolution to promote behaviour change. Social tipping is one version of this idea based on conformity and coordination. Conformity and coordination can reinforce a harmful social norm, but they can also accelerate change from a harmful norm to a beneficial alternative. Perhaps unfortunately, the link between the size of an intervention and social tipping is complex in heterogeneous populations. A small intervention targeted at one segment of society can induce tipping better than a large intervention targeted at a different segment. We develop and examine two models showing that the link between social tipping and social welfare is also complex in heterogeneous populations. An intervention strategy that creates persistent miscoordination, exactly the opposite of tipping, can lead to higher social welfare than another strategy that leads to tipping. We show that the potential benefits of miscoordination often hinge specifically on the preferences of people most resistant to behaviour change. Altogether, ordinary forms of heterogeneity complicate applied cultural evolution considerably. Heterogeneity weakens both the link between the size of a social planner's intervention and behaviour change and the link between behaviour change and the well-being of society.

1 Introduction

Applied cultural evolution is, to shun euphemism, an attempt to engineer culture. A social planner wants people to behave differently, and she intervenes in society in pursuit of this objective. The interesting twist is that, once people exposed to the intervention start to change behaviour, endogenous cultural evolutionary processes can take effect. Some people change behaviour because they have direct experience with the intervention. Some people change behaviour because they observe others doing so. If the social planner knows how this second process works, she can implement her intervention in a way that maximises the sum of both the direct effect and the associated indirect cultural evolutionary effect. In particular, the indirect effect might far outstrip the direct effect, in which case cultural evolutionary dynamics dramatically amplify the intervention's consequences. This idea is the essence of applied cultural evolution as an attempt to engineer culture [1].

The indirect cultural evolutionary effect occurs because we influence, teach, and learn from each other, and we do not do so randomly [2]. We pay attention to some people and ignore others [3, 4]. Sometimes we follow the majority, and sometimes we do not [5, 6]. Some people provide examples of how to behave, and some people provide examples of how not to behave [7–9]. Some behaviours we simply like, and others we do not [10]. Whatever the details, the necessary result is some kind of cultural evolution at the aggregate level [1, 11, 12]. If we want to know what kind of cultural evolution, the details are crucial [1, 11, 13, 14]. If we discriminate when we learn from each other in one way, cultural evolution unfolds accordingly. If we discriminate in some other way, we can expect cultural evolution to unfold quite differently.

Engineering culture sounds unpleasant, even imperialistic, and sometimes it is [15–17]. However, because most people live in a society, any attempt to modify anyone's behaviour comes with the potential to induce a secondary cultural evolutionary effect. Every policy change, persuasion campaign, marketing push, therapy session, and passing advice for a friend is an intervention that can affect the individuals directly exposed. Because others are watching, it may also activate subsequent cultural evolution. Thus, the crucial question is not a matter of whether we want applied cultural evolution. We have it, and we will keep it. Rather, the crucial question is, do we have the insight and wherewithal to manage cultural evolutionary processes for the benefit of society? This paper develops a model focused on a particular version of this second question. We examine the link between behaviour change and social welfare. We show that in heterogeneous populations, even though everyone faces clear incentives to behave like others, behaving like others is not always best. As a consequence, an intervention that triggers large-scale norm change can actually be worse than an intervention that generates chronic disagreement.

We assume that everyone somehow wants to behave like others because of some mix of conformity and coordination incentives. Conformity and coordination incentives can create multiple equilibria. Social norms, by which we mean a shared understanding of how people should behave and how people do behave, help people collectively pick a specific equilibrium. In the simplest case with two behaviours, one locally stable steady state has everyone choosing one behaviour, and another locally stable steady state has everyone choosing the other behaviour. The population has converged on one of these equilibria, but the two states may not be equally good for society. One can be relatively harmful and the other relatively beneficial. Because both are locally stable, the population can get stuck in the harmful equilibrium. Happily, however, the same conformity and coordination incentives that trap the population in the harmful equilibrium can create the potential for a rapid transition to the beneficial equilibrium. A sufficiently large shock, a social planner's intervention for example, can dislodge the population from the harmful equilibrium and tip it into the basin of attraction for the beneficial alternative. Once this happens, cultural evolutionary forces finish the job.

Conformity and coordination incentives ensure that the population completes the transition to the a new socially beneficial norm without further inputs from the social planner.

This is the basic model of cultural change based on social tipping [9, 18, 19]. It is an extremely influential model of how social planners can recruit cultural evolutionary processes to promote behaviour change. The idea has appeared, in one form or another, across a range of policy-relevant domains [1] related to gender-based violence [16, 20–25] and other forms of gender bias [26–28], natural resource use [29], health [30, 31], species conservation [32], and climate change [33–37].

If everyone is the same, the social planner's task is relatively straightforward. She needs to know how big the initial shock must be. Put differently, she needs to know what proportion of people must change for conformity and coordination to switch from reinforcing the status quo norm to reinforcing the social planner's preferred alternative. The trouble is that people are usually not all the same [38], and ordinary forms of heterogeneity introduce a number of challenges and complexities [1, 9, 14, 39–42]. Bare minimum, the social planner must ask both how big her intervention should be and which segment of the population to target with the intervention. Interestingly, the best answer to this second question can vary, but targeting the individuals most amenable to change is often the worst strategy if the objective is to maximise behaviour change [1, 14]. Here we ask a related but different question. Namely, if the social planner is considering two different intervention targets, which one maximises social welfare? Surprisingly, this question can lead to very different conclusions. Targeting the most amenable segment of the population can actually limit behaviour change, with frequent miscoordination the outcome, but it can lead to the greatest social welfare. As we will see, this paradoxical outcome readily occurs in situations, perhaps typical, where some people want society to transition to a new norm, but others do not.

2 Model and results

Assume an infinitely large population of individuals, $i \in I$, where I is some uncountably infinite indexing set. Everyone is playing a game with two possible choices, SQ and Alt. Individuals pair off randomly to play the game in a periodic fashion with random rematching every period. For reasons explained below, we think of SQ as the "status quo", namely the behaviour everyone chooses before intervention. The social planner does not like everyone playing SQ. She would prefer that everyone switch to choosing Alt, and thus at some point the social planner implements an intervention that promotes Alt as an "alternative" to the status quo. Importantly, we might normally imagine that the social planner wants people to choose Alt because Alt is somehow socially beneficial. While a natural interpretation, we do not assume or insist on this idea. The social planner might instead simply have her own preferences that differ from those of the people. This possibility has proven important in discussions about certain cultural traditions like female genital cutting and early marriage [15–17, 21]. One view of programmes promoting the abandonment of cutting, for example, is that they help the people in a cutting society help themselves by shifting them towards a non-cutting equilibrium that respects human rights and improves outcomes for all involved. Another view is that such programmes are a form of cultural imperialism, with Europeans and their proxies once again imposing European values on the rest of the world [1, 14]. We will not venture a resolution of this kind of dispute. We do, however, respect the validity of such disputes by allowing the possibility that the social planner's desired outcome may not be socially beneficial. This possibility, in fact, lies at the centre of our claim that a social planner's intervention may induce norm change that hurts.

In any case, before intervention, the game is a strict coordination game for everyone, and individuals have heterogeneous preferences (Table 1, Pre-intervention (All)). The game is a strict coordination game for everyone because we assume that, for each i, $a + x_i > a$, and $b + x_i < d$, where $b < d$. Intuitively, everyone faces incentives to match the choices of their partners. This shared interest in coordinating, however, mixes with heterogeneous preferences in the sense that each individual has her own idiosyncratic x_i value. Across individuals, x_i values are somehow distributed on the interval $(0, d - b)$ according to the density function f and its associated cumulative probability function F.

Because everyone faces incentives to coordinate, a focal player's beliefs about what her next partner will play are important, and we can summarise an individual's preferences as an indifference point defined in terms of beliefs. Let $\tilde{q}_i$ be i's belief that her next randomly selected partner will play Alt. The expected payoff (Table 1, Pre-intervention (All)) from choosing SQ, $E[\Pi_i(\text{SQ})]$, and the expected payoff from choosing Alt, $E[\Pi_i(\text{Alt})]$, are the following,

$$\begin{aligned} E[\Pi_i(\text{SQ})] &= (1 - \tilde{q}_i)(a + x_i) + \tilde{q}_i(b + x_i) \\ E[\Pi_i(\text{Alt})] &= (1 - \tilde{q}_i)(a) + \tilde{q}_i(d). \end{aligned} \tag{1}$$

The individual is indifferent between the two choice options if $\tilde{q}_i = x_i/(d-b)$. If $\tilde{q}_i > x_i/(d-b)$, the individual prefers to choose Alt. If $\tilde{q}_i < x_i/(d - b)$, she prefers SQ.

As a kind of reference model, imagine that, for each i, the belief in the current period is simply the actual distribution of choices from the previous period, and each i chooses the behaviour with the highest expected payoff given this belief. Imagine further that individuals choose Alt when indifferent. This model is sometimes called the "threshold" model [14, 39, 43]. The model supports at least two interpretations. First, individuals form beliefs myopically, namely by simply extrapolating from the recent past, and they choose the best option given these beliefs [44]. Second, all individuals are conformists in the sense specified by Boyd and Richerson [11], but conformist social learning mixes with content biases that are heterogeneous across individuals [1]. We motivate our analysis here with the coordination game interpretation, but this is not essential. The two models are isomorphic [1], and our analysis would remain the same were we to rely on the conformity interpretation. Regardless of interpretation, the model is extraordinarily good at predicting behaviour in experimental studies with coordination games [19, 44]. Thus, although we do not limit our analyses by assuming that social dynamics unfold according to this model, we sometimes highlight the steady states of the model as a point of reference.

Notice that, if $x_i < (d-b)/2$, the set of beliefs for which i chooses Alt is larger than the set of beliefs for which i chooses SQ. This is one way of saying that i prefers Alt over SQ. More technically, we will say that, for such an individual, coordinating on Alt risk dominates coordinating on SQ [45]. If $x_i > (d-b)/2$, the opposite holds, and coordinating on SQ risk dominates coordinating on Alt.

Empirical research has shown that, without special countervailing mechanisms in place, risk dominance exerts an extraordinary pull on cultural evolutionary dynamics [46–52]. What would this mean in a heterogeneous population? As others have argued [27], we assume the population is most likely to converge on the equilibrium that a majority of individuals view as risk-dominant. By extension, the situation of interest for our analysis is one in which $F((d-b)/2) < 0.5$. Specifically, $F((d-b)/2) < 0.5 \Rightarrow 1 - F((d-b)/2) > 0.5$, where the second condition means that before intervention a majority of individuals view coordinating on SQ as risk-dominant. In this case, we expect the population to converge on SQ. If the distribution of preferences had been otherwise, the population would have probably converged on Alt, and the social planner would have had no need to intervene in the first place [1, 27]. If most individuals view coordinating on SQ as risk-dominant, the distribution of x_i values should be left-skewed because left skew ensures that $F((d-b)/2) < 0.5$.

Crucially, although we argue that left-skewed x_i distributions represent the situations of interest, this claim is silent about the welfare consequences of coordinating on SQ versus coordinating on Alt. Specifically, we have made no claims so far about the relation between the $a + x_i$ and d, where the former are the payoffs players get when coordinating on SQ and the latter the payoff they get when coordinating on Alt. Many possibilities exist. At one extreme, for each i, $a + x_i > d$, which would mean that coordinating on SQ is socially beneficial in the precise sense that coordinating on SQ is strictly better for everyone than coordinating on Alt. A social planner who intervenes in this situation is simply promoting her own agenda, as discussed above, to the detriment of the people. At the other extreme, for each i, $a + x_i < d$, which would mean that coordinating on SQ is socially harmful because coordinating on SQ is strictly worse for everyone that coordinating on Alt. Both of these extremes are consistent with saying that coordinating on SQ is risk-dominant for a majority of individuals. Risk dominance depends on the relation between x_i and $d - b$, not the relation between $a + x_i$ and d. Situations between the two extremes are also possible, and such situations figure prominently in our analyses below.

At some point, the social planner rolls out an intervention to promote behaviour change. She targets (T) some proportion, $\phi \in (0, 1)$, of the population and incentivises these people to switch from SQ to Alt (Table 1, Post-intervention (T)). The intervention is unequivocally effective in the sense that $h > g$, and thus post-intervention all targeted individuals always choose Alt. This is a strong assumption. It implies that all targeted individuals, regardless of their initial preferences, effectively acquire new preferences because of their experience with the intervention. One natural interpretation is that the intervention in question is an extremely effective persuasion campaign that instils new preferences based on strong personal values. Imagine an individual who becomes extremely well-informed about climate change. Green choices like riding a bike and avoiding beef become intrinsically valuable to this individual (h), and brown choices like driving an SUV and eating steaks become intrinsically painful (g). These intrinsic values dominate decision making in the sense that they are far more important than whether or not the individual manages to coordinate with others. Later, we relax this assumption with simulation models that assume individuals are less likely to respond to the intervention in this way as they become more resistant to change pre-intervention [14, 24].

Before that, however, we consider the model in which the intervention leads any targeted (T) individual, regardless of her pre-intervention x_i, to change behaviour. The question is, what do the non-targeted individuals (NT) do? This question lies at the centre of applied cultural evolution in general and norm change based on social tipping specifically. If endogenous cultural evolutionary processes lead to behaviour change among people having no direct experience with the intervention, the social planner has activated these processes, whether intentionally or not, to amplify the direct effect of her intervention.

To analyse long-run welfare, we consider two alternative intervention strategies. Although many more possibilities exist, we analyse the two extremes because they bracket the range of possibilities and intuitively capture the trade-offs the social planner faces [14]. At one extreme, the social planner targets the segment of the population most amenable to change. Specifically, $\exists\, x_A^1 \in (0, d-b)$ such that $F(x_A^1) = \phi$, and the social planner targets everyone with an $x_i \leq x_A^1$. At the other extreme, the social planner targets the segment of the population most resistant to change. In this case, $\exists\, x_R^1 \in (0, d-b)$ such that $F(x_R^1) = 1-\phi$, and the social planner targets everyone with an $x_i > x_R^1$.

After intervention, assume the population stabilises in the long-run on some proportion choosing Alt. Under an amenable target, we denote this proportion as $\hat{q}_A$. Under a resistant target, we use $\hat{q}_R$. Because targeted individuals always choose Alt post-intervention, then $\hat{q}_A, \hat{q}_R \geq \phi$. Because of one additional assumption, $\hat{q}_R \geq \hat{q}_A$ must also hold. Specifically, of the non-targeted individuals who change from SQ to Alt, we assume they do so in order from those most amenable towards Alt to those most resistant. Put differently, non-targeted individuals who change from SQ to Alt do so in order from those with the smallest x_i values to those with the largest x_i values. This assumption is consistent with the threshold model [14, 39], but more intuitively it simply means that people who are relatively favourable towards Alt choose Alt at least as early as those who are relatively unfavourable towards Alt. With this assumption in place, as long as targeted individuals respond the same regardless of their pre-existing preferences, one can show that $\hat{q}_R \geq \hat{q}_A$ must hold [14].

The intuition is the following. Under an amenable target, by targeting the most amenable segment of the population, the social planner chooses the easiest possible task for the intervention. This leaves the hardest possible task for subsequent endogenous cultural evolutionary processes because the non-targeted individuals necessarily comprise a proportion $1-\phi$ of the population as resistant to change as possible. Under a resistant target, in contrast, the social planner takes the hardest possible task for the intervention, but this does not matter because we are assuming the intervention is equally effective regardless of the target. A resistant target also leaves the easiest possible task for endogenous cultural evolution because the non-targeted individuals make up a proportion $1-\phi$ as amenable to change as possible. For this reason, the long-run proportion choosing Alt must be at least as large under a resistant target as under an amenable target [14].

The result of the intervention is a partition of the population (Fig. 1) into either two or three categories of player. First, targeted players choose Alt, which we designate with (Alt,T), and this category always exists. Second, some or all non-targeted players may choose Alt, a category denoted as (Alt,NT). Finally, some or all non-targeted players may stick with SQ, a category denoted as (SQ,NT). Our focus is the welfare effects associated with the long-run partition of the population. Crucially, the partition and its welfare consequences depend on the social planner's intervention strategy.

Under an amenable target, targeted individuals constitute a proportion ϕ of the population from the left tail of the x_i distribution. Specifically, $F(x_A^1) = \phi$, and the targeted subset thus consists of all individuals with x_i values in the interval $(0, x_A^1]$. These individuals form the (Alt,T) category. If additional individuals choose Alt, then $\hat{q}_A > \phi$. These individuals form the (Alt,NT) category, which make up a proportion $\hat{q}_a - \phi$ of the population. Because non-targeted individuals choose Alt in order from those most amenable to most resistant, the most amenable individuals in this category come from somewhere in the middle of the x_i distribution. More precisely, $\exists\, x_A^2 > x_A^1$ such that $F(x_A^2) = \hat{q}_A$. The subset of non-targeted individuals who choose Alt thus consists of all individuals with x_i values in the interval $(x_A^1, x_A^2]$ if $x_A^2 < d - b$ and (x_A^1, x_A^2) if $x_A^2 = d - b$. If $\hat{q}_A = \phi$, this category of player does not exist. Finally, if $\hat{q}_A < 1$, some non-targeted individuals persist in choosing SQ. These individuals necessarily come from the right tail of the x_i distribution. Specifically, individuals with x_i values in the interval $(x_A^2, d - b)$ form this (SQ,NT) category. If $\hat{q}_A = 1$, this category does not exist.

Under a resistant target, we start at the right tail of the x_i distribution. Given an x_R^1 such that $F(x_R^1) = 1 - \phi$, the (Alt,T) individuals are those individuals with x_i values in the interval $(x_R^1, d - b)$. If $\hat{q}_R > \phi$, the (Alt,NT) individuals come from the left tail of the distribution. Specifically, $\exists\, x_R^2 > 0$ such that $F(x_R^2) = \hat{q}_R - \phi$. These (Alt,NT) individuals consist of everyone with x_i values in the interval $(0, x_R^2]$. Finally, if $\hat{q}_R < 1$, some individuals stick with SQ in the long run, and these individuals have x_i values in the interval $(x_R^2, x_R^1]$, which is somewhere in the middle of the x_i distribution. Fig. 1 shows an example of how an amenable versus a resistant target might induce two distinct partitions.

To analyse the welfare consequences of our two intervention strategies, we imagine randomly sampling an individual in a period and calculating this individual's expected payoff. Because we simply focus on the expected payoff of a randomly selected individual, we are adopting a utilitarian view of social welfare. Intuitively, we are assuming that first and foremost the social planner has an obligation to maximise the aggregate payoffs in society. If she has other objectives, like reducing inequality for example, she can redistribute ex post to accomplish these objectives. We relax the focus on payoffs when we develop a more elaborate model below.

To assist with the logic behind calculating an individual's expected payoff from one period of play, Table 2 shows all the ways in which two players can pair off to play, the probabilities of the various pairs, and the payoffs generated for each player in a pair. For convenience, define $V_A = (x_A^2, d-b)$, which is the set of x_i values for (SQ,NT) individuals under an amenable target. Analogously, let $V_R = (x_R^2, x_R^1]$ be the set of x_i values for (SQ,NT) individuals under a resistant target.

Under an amenable target, the expected payoff, $E_A[\Pi_i]$, takes the form,

$$\begin{aligned}
E_A[\Pi_i] = & \frac{\phi^2(2h)}{2} + \frac{\phi(\hat{q}_A-\phi)(h+d)}{2} + \frac{\phi(1-\hat{q}_A)(h+b+E[X_i \,|\, x_i \in V_A])}{2} \\
& + \frac{(\hat{q}_A-\phi)\phi(d+h)}{2} + \frac{(\hat{q}_A-\phi)^2(2d)}{2} + \frac{(\hat{q}_A-\phi)(1-\hat{q}_A)(a+b+E[X_i \,|\, x_i \in V_A])}{2} \\
& + \frac{(1-\hat{q}_A)\phi(b+E[X_i \,|\, x_i \in V_A]+h)}{2} + \frac{(1-\hat{q}_A)(\hat{q}_A-\phi)(b+E[X_i \,|\, x_i \in V_A]+a)}{2} \\
& + \frac{(1-\hat{q}_A)^2(2a+2E[X_i \,|\, x_i \in V_A])}{2}.
\end{aligned} \tag{2}$$

This expression simplifies to

$$\begin{aligned}
E_A[\Pi_i] = & \phi h + a(1-\phi)(1-\hat{q}_A) + b\hat{q}_A(1-\hat{q}_A) \\
& + d\hat{q}_A(\hat{q}_A-\phi) + (1-\hat{q}_A)E[X_i \,|\, x_i \in V_A].
\end{aligned} \tag{3}$$

The expected payoff under a resistant target, $E_R[\Pi_i]$, is analogous,

$$\begin{aligned}
E_R[\Pi_i] = & \phi h + a(1-\phi)(1-\hat{q}_R) + b\hat{q}_R(1-\hat{q}_R) \\
& + d\hat{q}_R(\hat{q}_R-\phi) + (1-\hat{q}_R)E[X_i \,|\, x_i \in V_R].
\end{aligned} \tag{4}$$

$E_A[\Pi_i]$ and $E_R[\Pi_i]$ look similar, but recall that $\hat{q}_A$ and $\hat{q}_R$ can be different. More subtly, the terms $E[X_i \,|\, x_i \in V_A]$ and $E[X_i \,|\, x_i \in V_R]$ are conditional expectations over different parts of the x_i distribution. Specifically, with μ as the Lebesgue measure,

$$\begin{aligned}
E[X_i \,|\, x_i \in V_A] &= \frac{\int_{V_A} x_i f(x_i) \mathrm{d}\mu(x_i)}{\int_{V_A} f(x_i) \mathrm{d}\mu(x_i)} \\
E[X_i \,|\, x_i \in V_R] &= \frac{\int_{V_R} x_i f(x_i) \mathrm{d}\mu(x_i)}{\int_{V_R} f(x_i) \mathrm{d}\mu(x_i)}.
\end{aligned} \tag{5}$$

$E[X_i \,|\, x_i \in V_A]$ is the expected x_i value of (SQ,NT) individuals under an amenable target. If this category exists ($\hat{q}_A < 1$), these individuals will come from the right tail of the x_i

distribution and thus be relatively resistant to Alt. For this reason, $E[X_i \,|\, x_i \in V_A]$ will tend to be large. In contrast, $E[X_i \,|\, x_i \in V_R]$ is the expected x_i value of (SQ,NT) individuals under a resistant target. If this category exists ($\hat{q}_R < 1$), these individuals will come from the middle of the x_i distribution, and so they will only be moderately resistant to Alt. Consequently, $E[X_i \,|\, x_i \in V_R]$ will tend take intermediate values.

To see which intervention strategy yields the highest expected payoff, subtract one from the other,

$$\begin{aligned} E_A[\Pi] - E_R[\Pi] = (1 - \hat{q}_A)E[X_i \,|\, x_i \in V_A] - (1 - \hat{q}_R)E[X_i \,|\, x_i \in V_R] \\ - (\hat{q}_R - \hat{q}_A)\left\{b + (\hat{q}_A + \hat{q}_R)(d - b) - (1 - \phi)a - \phi d\right\}. \end{aligned} \tag{6}$$

Note that h and g disappear. This happens because we assume the intervention is equally effective when targeting resistant versus amenable individuals. The welfare differences between the two interventions strategies thus depend exclusively on effects related to non-targeted individuals. To gain some intuition about what equation (6) represents, assume the two different intervention strategies produce outcomes as different from each other as possible. Specifically, let $\hat{q}_A = \phi$ and $\hat{q}_R = 1$. This means that using the intervention to target amenable people does not induce any behaviour change via endogenous cultural evolution; the only effect is the direct effect of the intervention. In contrast, targeting people resistant to change induces the maximum possible change; everyone eventually chooses Alt. Moreover, the rate of miscoordination is relatively high under the amenable target at $2\phi(1 - \phi)$, but miscoordination never occurs under the resistant target. For this special case, $E_A[\Pi] - E_R[\Pi] > 0$ if and only if the following holds,

$$(1 - \phi)(a + E[X_i \,|\, x_i \in V_A]) + \phi(b + E[X_i \,|\, x_i \in V_A]) > d. \tag{7}$$

Condition (7) compares the expected payoffs of non-targeted individuals given the two intervention strategies. Under a resistant target, all non-targeted players choose Alt. They always coordinate, and they always get a payoff of d. Under an amenable target, all non-targeted players choose SQ. A randomly selected player of this type is paired with another non-targeted player with probability $1 - \phi$. They coordinate, and the expected payoff of the focal player is $a + E[X_i \,|\, x_i \in V_A]$. The focal player is paired with a targeted player with probability ϕ. They miscoordinate, and the expected payoff of the focal player is $b + E[X_i \,|\, x_i \in V_A]$.

Because the x_i are distributed on $(0, d - b)$, $b + E[X_i \,|\, x_i \in V_A] < b + d - b = d$ must be true. However, $a + E[X_i \,|\, x_i \in V_A] > d$ is certainly possible, and condition (7) is also possible.

The outcome depends on both a and the distribution of x_i values among the non-targeted individuals given an amenable target. In this sense, $a + E[X_i \,|\, x_i \in V_A]$ is a measure of the alignment or misalignment between the social planner's objectives and the preferences of a key segment of the population, namely the individuals the social planner does not target when she chooses an amenable target.

At one extreme, alignment is high, and condition (7) does not hold regardless of the x_i distribution. Specifically, note that $\sup V_A = d - b$. Substituting this value for $E[X_i \,|\, x_i \in V_A]$ and rearranging reveals that condition (7) does not hold if $a \leq b$. In this special case, the population fixed on SQ before intervention really is stuck in a harmful equilibrium. Behaviour change is unambiguously good in the sense that a resistant target, which maximises behaviour change, produces larger expected payoffs than an amenable target, which minimises behaviour change, whatever form preference heterogeneity takes.

At the other extreme, misalignment is high, and condition (7) holds under extreme conditions. Specifically, when $\hat{q}_A = \phi$, note that $\inf V_A = x_A^1$, where $F(x_A^1) = \phi$. Substituting this value and rearranging shows that condition (7) holds if and only if $a > (d - \phi b - x_A^1)/(1 - \phi)$. Note that this condition involves x_A^1, which depends on the x_i distribution given ϕ. For this reason, we do not isolate a situation in which condition (7) holds regardless of the entire x_i distribution. Instead, in this special case, we focus on an extreme situation in which condition (7) holds even if the distribution of x_i values among non-targeted players minimises $a + E[X_i \,|\, x_i \in V_A]$. In this special case, behaviour change is unambiguously harmful for all non-targeted individuals given an amenable target. Consequently, an amenable target is best precisely because non-targeted individuals do not change behaviour.

Between these two extremes, the distribution of preferences among non-targeted individuals, given an amenable target, is important. If a is sufficiently large, and if the distribution of x_i values is sufficiently left-skewed, $a + E[X_i \,|\, x_i \in V_A]$ can be large enough to ensure that condition (7) is satisfied. In this case, even if some people clearly benefit from changing behaviours from SQ to Alt, others do not. Indeed, some individuals do best by sticking with SQ and tolerating frequent miscoordination. These individuals are exactly the people who are not targeted under an amenable target, they are exactly the people who do not change behaviour under an amenable target, and they are exactly the people who ensure that an amenable target with frequent miscoordination is better for society than a resistant target with no miscoordination. In this situation, the social planner is at odds with the most resistant segment of the population under her influence. She leaves these people out of her intervention given an amenable target, and they maintain their pre-existing preferences as a result. These preferences favour SQ to such an extent that coordinating on SQ with other non-targeted individuals more than compensates for miscoordinating when paired with targeted individuals. A resistant target could generate genuine norm change, with everyone coordinating on Alt, but it would do real harm relative to an amenable target with chronic disagreement and persistent miscoordination.

More broadly, we can make further progress if we choose specific parameter values and use a graphical approach to the unrestricted condition (Eq. 6). Accordingly, Figs. 2 – 4 summarise which intervention strategy produces the highest expected payoff (Eq. 6) under two different x_i distributions, two different values of a, and three different values of ϕ. Because these figures work with the unrestricted condition (Eq. 6), they show the relative welfare effects of the alternative intervention strategies for any possible outcome subject to $\phi \leq \hat{q}_A \leq \hat{q}_R \leq 1$. In this way, we consider a wide range of steady states, and thus we do not limit attention to a particular dynamical process. To create these figures, we set $b = 0$ and $d = 1$. We then vary $a \in \{0.25, 0.75\}$, $\phi \in \{0.25, 0.5, 0.75\}$, and the skewness of the x_i distribution. The figures reveal that, when the social planner and the people have partially misaligned preferences, an amenable target often yields a society with higher payoffs than a resistant target. More interestingly, this result frequently obtains even though the amenable target leads to less behaviour change and more miscoordination than the resistant target. This happens under an increasingly broad range of conditions as a increases and as the skew of the x_i distribution increases. In other words, it happens as the disconnect between the social planner and the resistant segment of society becomes more pronounced.

To see this result, note that for all figures the red region expands as we move from left (a,c) to right (b,d) and from top (a,b) to bottom (c,d). Red represents a steady state in which an amenable target produces greater welfare than a resistant target. The move from left to right means the left skew of the x_i distribution increases, and the move from top to bottom means a increases. The expansion of the red region is especially important along the top boundaries of each panel because this represents outcomes for which a resistant target leads to full-fledged norm change (i.e. $\hat{q}_R \approx 1$), but an amenable target does not (i.e. $\hat{q}_A < 1$).

Thus far, we have compared the payoffs under two alternative intervention strategies. We have shown that an intervention strategy that produces a complete shift in the population from one equilibrium to another, with little or no miscoordination the result, can actually be worse in terms of social welfare than an alternative strategy that leads to no norm at all after intervention. What if, in contrast, we hold the intervention strategy constant? Specifically, what if we hold the size (i.e. ϕ) and target (i.e. amenable or resistant) constant and allow the steady state to vary? Intuitively, one might imagine that with such comparisons, where all else is equal, more coordination would be better than less coordination. Intuition is once again, however, potentially misleading. Fig. 5 shows example results.

The results reveal that, when at least some individuals prefer to coordinate on SQ as opposed to Alt, expected payoffs can actually increase with miscoordination all else equal. An amenable target is especially prone to this result because an amenable target leaves the most resistant individuals to persist in choosing SQ. These individuals tolerate frequent miscoordination because they get especially high payoffs when they manage to coordinate on SQ, and this effect can drive up the expected payoff for the entire population. As a increases (Fig. 5c, d), and as the skew of the x_i distribution increases (Fig. 5b, d), expected payoffs can rise with miscoordination rates for the same basic reason. Of particular note, amenable versus resistant targets do not necessarily lead to the same patterns of variation in expected payoffs. Expected payoffs can increase as miscoordination rises under an amenable target but decrease as miscoordination rises under a resistant target (Fig. 5c, $\phi \in \{0.5, 0.75\}$; Fig. 5d, $\phi = 0.75$). As a result, even if we assume the two intervention strategies generate the same behaviour change (i.e. $\hat{q}_A = \hat{q}_R$), miscoordination may be good for social welfare under an amenable target but bad for social welfare under an resistant target. Equally challenging for the social planner, the effects of miscoordination are not even reliably monotonic given an intervention strategy. The solid red line of Fig. 5b provides an example. As we move from $\hat{q}_A = 1$ to $\hat{q}_A = 0.5$, expected payoffs first decrease and then increase. This means that, given an amenable target with $\phi = 0.25$, increasing miscoordination is first bad for society, but then it becomes good. This subtlety can be relevant when the social planner has committed to an intervention strategy, but she is uncertain about the final outcome. Because of the uncertainty, she will not be able to say ex ante if social welfare increases or decreases as coordination rises.

3 A stylised beef-eating illustration

To clarify the intuition, imagine a population of beef eaters and a social planner who would like everyone to switch to plant-based alternatives for reasons related to both public health and climate change. Even though everyone is a regular beef eater before intervention, say because beef-eating is the local culture, people's tastes naturally vary. Some people have tastes that favour coordinating on fruits, vegetables, and beans over yet another meal with blood on the plate and a brick in the stomach. We can call these people the "berry lovers". Other people would prefer to coordinate on stew, cheese, and Waygu instead of fruits, vegetables, and beans. We can call these people the "steak lovers". Between these two groups we have the "omnivores", a group of people who enjoy sharing a steak dinner with friends just as much as they enjoy sharing a meal of salad and strawberries. The berry lovers constitute a minority of the population. Frequencies rise as we move into the omnivores. Steak lovers form the majority, which is exactly why the population converged on beef eating before intervention. In terms of the model above, this stylised population could have $d = 1$, $b = 0$, and $a = 0.5$, with a distribution of x_i values that covers the full support, $(0, d-b)$, but is left-skewed. The berry lovers have x_i values noticeably less than 0.5, the steak lovers have x_i values noticeably greater than 0.5, and the omnivores have x_i values close to 0.5.

Now consider a social planner who implements an intervention of size $\phi = 0.5$ in either the amenable tail of the preference distribution or the resistant tail. Imagine further that, under an amenable target, only the targeted individuals switch to a plant-based diet ($\hat{q}_A = \phi$), but the entire population switches under the resistant target ($\hat{q}_R = 1$). How can the former lead to larger expected payoffs than the latter? Because everyone responds to the intervention in the same way when targeted, the answer depends on what happens among the individuals who are not targeted.

Because steak lovers are common, the social planner targeting the amenable tail has to target the berry lovers, the omnivores, and maybe even a few half-hearted steak lovers to accumulate 50% of the population for her intervention. The remaining non-targeted 50% consists of serious steak lovers who simply continue to choose steak. Because this half of the population has extreme preferences, they can tolerate some miscoordination as long as they get to have their preferred food. Sometimes they have a steak at the table with someone enjoying plant-based alternatives, and perhaps their berry-loving companions chastise them along the way, but at least they get to eat steak. In particular, their payoffs are only slightly less than the coordination payoffs they would have received if they had joined their tablemates and chosen a plant-based meal. That said, when they do coordinate on steak with another serious steak lover, payoffs are especially high all around. These high payoffs follow from the extreme preferences of the non-targeted group, given an amenable target, and the high payoffs more than compensate for the small miscoordination costs that sometimes occur. For exactly this reason, non-targeted individuals persist in choosing steak instead of switching to plant-based alternatives. Their decision to do so has important welfare consequences, and in particular their extreme preferences ensure that choosing steak and occasionally miscoordinating may generate much higher expected payoffs than switching to plant-based alternatives.

In contrast, when the social planner targets the resistant tail, she targets the 50% of the population composed of the serious steak lovers. The remaining non-targeted 50% consists of the berry lovers, the omnivores, and few half-hearted steak lovers. When they all switch to coordinating on plant-based alternatives, the berry lovers experience a gain, but they are a small group. The omnivores are, by definition, approximately indifferent over the equilibria, and so they switch from one norm to another norm that is more or less just as good. The half-hearted steak lovers experience a loss, but they are also a small part of the non-targeted group. All in all, many non-targeted individuals have moderate preferences. Coordinating is important; coordinating on a specific behaviour is not. The non-targeted change behaviour for this reason, but their decision to do so has only moderate welfare effects precisely because many of them have moderate preferences.

As a crucial caveat, the example outlined here is deliberately vague about externalities. One interpretation is that externalities are not present; the payoff matrix for each individual captures the full suite of consequences associated with the choices the individual and her partner can make. If true, the analysis above holds without complication. An intervention among the resistant may produce a complete shift to a plant-based norm, but an intervention in the amenable tail would have produced chronic miscoordination with higher average payoffs. If externalities are present, however, the social planner may be justified in choosing an intervention strategy that generates complete norm change even if she knows this strategy will produce lower perceived payoffs than other intervention strategies. The social planner would be justified with such an approach, for example, if beef is underpriced, which is almost certainly true, because the price does not account for all climatic effects associated with raising cattle instead of plants. As another way to think about this, if the price of beef was correct, the value of a would be lower than it is, and the entire distribution of $a + x_i$ values would shift downward. Most people then would actually perceive a complete shift to a plant-based norm as beneficial, and this would be true even for many people who really enjoy steak.

Efforts to shift norms related to cultural traditions like female genital cutting and early marriage [1, 15, 16, 21] also raise critical questions about the extent to which externalities are present. A viewpoint emphasising cultural relativism might argue that families have exactly the preferences they should have within the context of cultural traditions that value female genital cutting or early marriage. An outsider may not understand these preferences, but this is no reason to discount their legitimacy. In this case, norm change can actually hurt relative to alternative social planning strategies. As always, details related to the distribution of $a + x_i$ values and the set of social planning strategies under consideration are crucial. In stark contrast, a viewpoint emphasising universal human rights might argue that cutting and early marriage perpetuate cultural systems that devalue girls and women, with a host of attendant social costs. By this account, externalities are ubiquitous, and the social planner should promote norm change even if she knows families will perceive themselves as worse off than they would under different intervention strategies.

4 Heterogeneous response to intervention

The analytical model above (§ 2) makes the strong assumption that the intervention is equally effective regardless of how amenable or resistant a targeted individual is before intervention. In practice, however, interventions designed to change behaviour often have heterogeneous effects [53]. To account for this possibility, we developed an agent-based simulation (Supplementary Information) that allows heterogeneous responses to the intervention. These simulations also allow us to show transient dynamics, the inequalities an intervention produces, and any stochastic effects that might occur because the population is finite. First, we explain the generic structure of the model. Then we explain the parameter space we used and key state variables we recorded when simulating.

To begin, a simulation creates a population of $N = 1000$ agents by drawing an x_i value for each agent from a left-skewed beta distribution. In time $t = 1$, everyone chooses SQ. Agents pair off randomly and play coordination games based on their pre-intervention preferences (Table 1, Pre-intervention (All)). Because everyone plays SQ, everyone coordinates and receives a payoff of $a+x_i$. Between $t = 1$ and $t = 2$, the simulation implements an intervention of size ϕ in either the amenable (A) or resistant (R) tail of the x_i distribution. Targeted individuals respond to the intervention with a probability that is a decreasing function of their x_i values. Specifically, a targeted agent i responds with probability $s_i = 1 - x_i/(d - b)$. If a targeted i responds to the intervention, she gets a new payoff matrix (Table 1, Post-intervention (T)), and given this new payoff matrix she changes from SQ to Alt because $h > g$. If a targeted i does not respond, she retains her original payoff matrix. Non-targeted agents also retain their original payoff matrices. Agents pair off randomly again and play coordination games. At this point, only targeted individuals who have responded to the intervention choose Alt. Agents receive payoffs based on their individual payoff matrices, their choices, and the choices of their partners.

Agents update their beliefs myopically by treating the proportion of others choosing Alt in $t = 1$ as their beliefs about the probability a randomly selected partner will choose Alt in $t = 2$. Specifically, let q_t be the proportion choosing Alt in t. If i chose SQ in $t = 1$, she believes her partner in $t = 2$ will choose Alt with probability $q_1 N/(N-1)$. If i chose Alt in $t = 1$, her belief is instead $(q_1 N - 1)/(N-1)$. Agents again pair off randomly to play coordination games in $t = 2$. When deciding how to play, each agent best responds given her myopically updated belief. Agents receive payoffs. The algorithm, which consists of random matching and myopic best responding, repeats until $t = 100$.

Altogether, we ran simulations over the following parameter space.

1. For beta-distributed x_i values, we fixed $\beta = 2$ and allowed α to vary according to $\alpha \in \{2.25, 2.5, 2.75, 3, 8\}$. Left skew increases as α increases.

2. We chose intervention sizes based on $\phi \in \{0.25, 0.5, 0.75, 0.9\}$.

3. Interventions targets were relatively amenable (A) or resistant (R).

4. For targeted agents who respond to the intervention (Table 1, Post-intervention (T)), we fixed $g = 0$ and let h vary based on $h \in \{2, 3\}$.

5. For all agents pre-intervention, non-targeted agents post-intervention, and targeted agents who do not respond post-intervention (Table 1), we fixed $b = 0$, $d = 1$, and let a vary according to $a \in \{0.25, 0.75, 0.9\}$.

When all of these parameter values are fully crossed, the result is 240 unique combinations. For each combination, we simulated 1000 independent populations. These populations differed in terms of the realised distribution of x_i values. For each simulation, in a given t, we calculated the fraction of agents in the population choosing Alt, the frequency with which agents miscoordinated, the average payoff in the population, and the Gini coefficient. The Gini coefficient [54] is a normalised inequality measure that ranges from zero to one. A value of zero indicates perfect equality, while a value of one indicates maximum inequality with a single individual holding all the wealth in the population. Given sorted payoff values in t for a population of size N, i.e. $\pi_{1,t} \leq \pi_{2,t} \leq \dots \pi_{N,t}$, we calculated the Gini as

$$G_t = \frac{2\sum_{i=1}^{N} i\pi_{i,t}}{N\sum_{i=1}^{N} \pi_{i,t}} - \frac{N+1}{N}. \tag{8}$$

Our approach provided four key quantities for characterising any given population in t. We then averaged over our 1000 independently simulated populations to get global averages of our four key quantities in t. We further estimated 95% confidence intervals using a non-parametric bootstrap procedure based on resampling from the set of independently simulated populations (Supplementary Information). We provide our code as Supplementary Information, so the interested reader can repeat the simulation exercise with different parameter values.

Figs. 6 – 9 present some of the key results from our simulations. One key result is that, if the intervention has heterogeneous effects, an amenable target can produce more behaviour change than a resistant target. This cannot happen when the intervention has homogeneous effects and agents myopically best respond [14]. However, if the probability of responding to the intervention declines with resistance to behaviour change, as in our simulations, the social planner can expect to face an important trade-off [1, 14, 42]. An amenable target maximises the direct effect of the intervention, but it minimises the secondary indirect effect. A resistant target minimises the direct effect, but conditional on a direct effect of a given size it maximises the secondary indirect effect.

Because of this trade-off, a resistant target is no longer guaranteed to maximise the sum of the direct and indirect effects. In particular, an amenable target tends to produce more overall behaviour change than a resistant target in situations where neither generates much behaviour change [42], and our results here are consistent with this idea. In Figs. 6 – 8, for example, the initial x_i distributions are only moderately skewed, and total behavioural change ranges from moderate to complete in the sense that the final proportion choosing Alt ranges from a bit more than 0.4 (Fig. 8a, Amenable target) to 1.0 (Fig. 6a, Resistant target). In all of these cases, a resistant target generates more behaviour change than an amenable target. As the skew of the initial x_i distribution becomes more extreme, however, behaviour change becomes very limited in general, with the final proportion choosing Alt remaining below 0.2 Given this limitation, however, an amenable target generates more behaviour change than a resistant target (Fig. 9a).

Crucially, one of our key conclusions from the analytical model above (§ 2) continues to hold in the more complex setting captured by our simulations. Namely, in heterogeneous populations, an amenable target readily leads to more miscoordination but higher average payoffs than a resistant target. In particular, all parameter combinations we considered are consistent with what we might call a non-degenerate form of heterogeneity in the population. By non-degenerate we mean that before intervention some agents view coordinating on SQ as risk-dominant, and some agents view coordinating on Alt as risk-dominant. Similarly, some agents view coordinating on SQ as payoff-dominant, and some agents view coordinating on Alt as payoff-dominant. Given these characteristics of the parameter space we consider, Figs. 6–9 represent a typical set of outcomes.

First, an amenable target always produces at least as much miscoordination as a resistant target and often strictly more. Second, even though amenable targets produce more miscoordination than resistant targets, amenable targets tend to produce average payoffs that are at least as high as resistant targets and often strictly higher. We only identified two cases in which an amenable target produced slightly lower average payoffs than a resistant target. For these two exceptions, h is relatively small ($h = 2$), a is large ($a = 0.9$), the x_i distribution is strongly skewed ($\alpha = 8$), and interventions are relatively small ($\phi \in \{0.25, 0.5\}$). In these two cases, a resistant target creates very little change among the agents targeted, and agents continue to coordinate on SQ at a high rate after intervention. Miscoordination rates are low, but behaviour change is also low, exactly the opposite of what the social planner is trying to accomplish.

Finally, in terms of inequality, an amenable target typically generates higher inequality than a resistant target. This pattern has clear implications for the social planner who wants to limit inequality. If the social planner has a policy tool to redistribute ex post, then she can choose the intervention strategy that maximises average payoffs and then redistribute later, as we assumed for the analytical model (§ 2). If she does not have access to such a policy tool, however, the social planner must somehow resolve the resulting trade-off between competing social objectives. That said, sometimes a resistant target produces the same degree of inequality or even slightly more than an amenable target. We found this pattern in our simulations for 28 combinations of parameter values. Perhaps more important than the associated details, however, the differences in Gini coefficients between the two intervention strategies tend to be small. In our simulations, over all periods and all parameter combinations, the maximum difference in Gini values between an intervention in the amenable tail versus the resistant tail is approximately 0.108. This is about the same as the difference between the United States and Austria in 2018 (`https://data.worldbank.org/indicator/SI.POV.GINI/`). The mean difference in Gini values averaged over all simulation is far smaller than this at 0.036. The tendency for both intervention strategies to produce similar degrees of inequality buttresses our use of expected payoffs as a way to evaluate social welfare.

5 Discussion

As with any model, we have ignored most of what matters. For example, we have ignored other types of interventions that incentivise behaviour change differently. The distinguishing feature of the models above is that payoffs for targeted individuals who respond to the intervention do not depend on the choices of others. Nor do payoffs depend on the pre-intervention preferences of these individuals. In contrast, a social planner could simply subsidise Alt by paying $s > 0$ to targeted individuals who choose Alt, whatever "paying" may mean. With an intervention like this, a targeted individual playing Alt gets $a + s$ or $d + s$. A targeted individual playing SQ gets $a + x_i$ or $b + x_i$, just like a non-targeted player, and for this reason the pre-existing preferences of targeted players affect payoffs so long as targeted players choose SQ after intervention with some positive probability.

Additionally and perhaps most importantly, we have ignored the possibility that non-targeted individuals may change their preferences after they change their behaviour. This possibility would open up an entirely new world of possibilities in terms of social welfare, but it is certainly feasible. One of us, for example, grew up in a run-of-the-mill U.S. city, geographically extensive with no meaningful public transport. When he became a teenager, he really wanted his own vehicle, so he got a job and bought one. The same person has now lived in Switzerland without a car for many years. He is coordinating on an alternative equilibrium. If his preference for his own vehicle had persisted, he would now be suffering relative to the car-based equilibrium of his adolescence and early adulthood. Exactly the opposite has happened. Experiencing the alternative has dramatically crystallised this person's view of how dysfunctional the U.S. equilibrium is, and it has done so in a way that was not possible, bizarrely, back then while sitting in traffic. We can easily imagine an analogous change in preferences, for example, when a family abandons female genital cutting [1]. Regardless, we must be clear about what this mechanism implies. If we believe ex post preference change is the typical mechanism generating welfare improvements after norm change, we are assuming that ex ante the social planner usually knows what is best for people; the people themselves do not.

In spite of the mechanisms we have ignored, we hope to have gained some understanding of how risk dominance, payoff dominance, and behaviour change combine to shape social welfare in heterogeneous populations. We have argued that the relevant situation is one in which a majority of individuals view the status quo tradition (SQ) as the risk-dominant equilibrium ($x_i > (d - b)/2$). If this were not so, the population probably would have never converged on the status quo tradition. With this backdrop in place, we examine welfare effects due to the payoff ranking over equilibria. Some situations are straightforward. Imagine that more or less everyone prefers coordinating on SQ over coordinating on Alt ($a + x_i > d$). In addition to risk dominating Alt, coordinating on SQ also payoff dominates coordinating on Alt. In this case, assuming no externalities, people are doing what they should be doing. A well-intentioned and well-informed social planner would know this and focus on other issues in society.

Alternatively, imagine that more or less everyone prefers coordinating on Alt over coordinating on SQ ($a + x_i < d$). Coordinating on SQ risk dominates coordinating on Alt, but coordinating on Alt clearly payoff dominates coordinating on SQ. People really are stuck in a harmful equilibrium in this case, and individuals cannot afford to deviate as isolated decision makers. A social planner can do real good in this case by engineering a coordinated change in behaviour to the alternative equilibrium. Crucially, however, she should probably not intervene and then just walk away. She should also consider mechanisms [49, 55–57] to counteract the potential fragility of the Alt equilibrium because of any residual tendency to treat coordinating on SQ as risk-dominant. Moreover, we do not know how common situations of this sort actually are. Experimental evidence shows that, although risk dominance does bias choice dynamics, payoff dominance also matters; populations sometimes find a way to the beneficial equilibrium [58].

In both of the situations immediately above, behaviour change and social welfare relate in a simple way. Behaviour change is uniformly bad in the first case and good in the second. Our analysis focuses on situations between these simple cases, situations in which the heterogeneity in the population really matters. We suspect that these cases are not just complicated, but probably also common. In effect, some people view coordinating on Alt as risk-dominant, and some people view coordinating on SQ as risk-dominant. Similarly, some people view coordinating on Alt as payoff-dominant, and some people view coordinating on SQ as payoff-dominant. In scenarios of this sort, the link between behaviour change and social welfare can be varied and counterintuitive. In particular, we have shown that an amenable target can produce a lot of miscoordination but relatively high payoffs because it maximises the chances that the people most resistant to change are exactly the people who do not change behaviour. These individuals have the strongest preferences for coordinating on SQ. They can choose SQ after intervention and tolerate the costs of miscoordination because, when they do coordinate on SQ, they get especially large payoffs. These payoffs can be so large, in fact, that they boost the payoffs of the entire society. If these individuals could pair off to play together at rates above chance, essentially a form of homophily, this effect would be even stronger than it is under the random matching we consider.

More broadly, what is the value of using our knowledge of cultural evolution to engineer beneficial behaviour change? One straightforward answer follows from the idea that beneficial norms can attenuate social welfare loss when markets fail [59, 60]. In today's world, every contract is incomplete, every price is wrong, and externalities are pervasive. Even if we consider a simple transaction like buying milk, the entire global economy is implicated. From the farmer and his cows to the truck driver who drives the milk to the big city, from the company making the vinyl for the seat of the farmer's tractor to the firm extracting the petroleum for the plastic cap on the milk carton, every production step involves dozens of contracts with some piece missing. No one knows the correct price of milk, and thus we have no reason to think that our milk-buying decisions are socially beneficial. Every day, each of us makes countless decisions that affect countless people in ways we do not understand. Activating the cultural evolution of conventions and norms that support socially beneficial choices can attenuate the challenges that follow from pervasive externalities. Our analysis, however, suggests that we still have much to learn about the best way to do so. When people differ in ordinary ways, the task of activating cultural evolution for good can become unexpectedly complex. Recent research has shown that heterogeneity can disrupt any simple monotonic relationship between the size of an intervention and the degree of behaviour change that follows [14, 42]. We have shown here that heterogeneity can also disrupt any simple monotonic relationship between the degree of behaviour change and social welfare. More surprisingly, heterogeneity can even disrupt any simple relationship between the rate of miscoordination and social welfare. Sorting through the complexities, however, will be essential precisely because widespread externalities imply that we need prosocial norms and related informal institutions to fill the welfare gap externalities leave behind.

References

[1] Charles Efferson, Sonja Vogt, and Lukas von Flüe. Activating cultural evolution for good when people differ from each other. In Jeremy Kendal, Rachel Kendal, and Jamie Tehrani, editors, *Oxford Handbook of Cultural Evolution*, chapter TBD. Oxford University Press, 2023.

[2] Rachel L Kendal, Neeltje J Boogert, Luke Rendell, Kevin N Laland, Mike Webster, and Patricia L Jones. Social learning strategies: Bridge-building between fields. *Trends in Cognitive Sciences*, 22(7):651–665, 2018.

[3] Joseph Henrich and Francisco Gil-White. The evolution of prestige: freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, 22(3):165–196, 2001.

[4] Maciej Chudek, Sarah Heller, Susan Birch, and Joseph Henrich. Prestige-biased cul-

tural learning: bystander's differential attention to potential models influences children's learning. *Evolution and Human Behavior*, 33(1):46–56, 2012.

[5] Charles Efferson, Rafael Lalive, Peter J. Richerson, Richard McElreath, and Mark Lubell. Conformists and mavericks: the empirics of frequency-dependent cultural transmission. *Evolution and Human Behavior*, 29(1):56–65, 2008.

[6] Jacob K. Goeree and Leeat Yariv. Conformity in the lab. *Journal of the Economic Science Association*, 1(1):15–28, 2015. ISSN 2199-6776. doi: 10.1007/s40881-015-0001-7. URL http://dx.doi.org/10.1007/s40881-015-0001-7.

[7] Charles Efferson, Rafael Lalive, Maria Paula Cacault, and Deborah Kistler. The evolution of facultative conformity based on similarity. *PLoS One*, 11(12):e0168551, 2016.

[8] Aysha Bellamy, Ryan McKay, Sonja Vogt, and Charles Efferson. What is the extent of a frequency-dependent social learning strategy space? *Evolutionary Human Sciences*, 4, 2022.

[9] Sönke Ehret, Sara Constantino, Elke Weber, Charles Efferson, and Sonja Vogt. Group identities can undermine social tipping after intervention. *Nature Human Behaviour*, Forthcoming:TBD, 2022. doi: https://doi.org/10.1038/s41562-022-01440-5.

[10] Olivier Morin. *How Traditions Live and Die*. Oxford University Press, 2016.

[11] Robert Boyd and Peter J. Richerson. *Culture and the Evolutionary Process*. Chicago: University of Chicago Press, 1985.

[12] Peter J. Richerson and Richerson Boyd. *Not By Genes Alone: How Culture Transformed the Evolutionary Process*. Chicago: University of Chicago Press, 2005.

[13] T. J. H. Morgan and K. N. Laland. The biological bases of conformity. *Frontiers in Neuroscience*, 6(87), 2012. doi: 10.3389/fnins.2012.00087.

[14] Charles Efferson, Sonja Vogt, and Ernst Fehr. The promise and the peril of using social influence to reverse harmful traditions. *Nature Human Behaviour*, 4:55–68, 2020.

[15] Ellen Gruenbaum. *The Female Circumcision Controversy: An Anthropological Perspective*. Philadelphia: University of Pennsylvania Press, 2001.

[16] Karisa Cloward. *When Norms Collide: Local Responses to Activism Against Female Genital Mutilation and Early Marriage*. Oxford University Press, 2016.

[17] David Lawson and Mhairi Gibson. Evolutionary approaches to population health: Insights on polygynous marriage, ‘child marriage’ and female genital mutiliation/cutting. In O. Burger, R. Lee, and R. Sear, editors, *Human Evolutionary Demography*, page TBD. 2023.

[18] Damon Centola, Joshua Becker, Devon Brackbill, and Andrea Baronchelli. Experimental evidence for tipping points in social convention. *Science*, 360(6393):1116–1119, 2018.

[19] James Andreoni, Nikos Nikiforakis, and Simon Siegenthaler. Predicting social tipping and norm change in controlled experiments. *Proceedings of the National Academy of Sciences*, 118(16), 2021.

[20] Gerry Mackie. Ending Footbinding and Infibulation: A Convention Account. *American Sociological Review*, 61:999–1017, 1996.

[21] Bettina Shell-Duncan and Ylva Hernlund. Female “Circumcision” in Africa: Dimensions of the Practice and Debates. In Bettina Shell-Duncan and Ylva Hernlund, editors, *Female “Circumcision” in Africa: Culture, Controversy, and Change*, pages 1–40. Boulder, CO: Lynne Rienner, 2000.

[22] Marc F. Bellemare, Lindsey Novak, and Tara L. Steinmetz. All in the family: Explaining the persistence of female genital cutting in West Africa. *Journal of Development Economics*, 116:252 – 265, 2015. ISSN 0304-3878. doi: http://dx.doi.org/10.1016/j.jdeveco.2015.06.001. URL `http://www.sciencedirect.com/science/article/pii/S0304387815000620`.

[23] Charles Efferson, Sonja Vogt, Amy Elhadi, Hilal El Fadil Ahmed, and Ernt Fehr. Female genital cutting is not a social coordination norm. *Science*, 349(6255):1446–1447, 2015. doi: 10.1126/science.aaa7978.

[24] Sonja Vogt, Nadia Ahmed Mohmmed Zaid, Hilal El Fadil Ahmed, Ernst Fehr, and Charles Efferson. Changing cultural attitudes towards female genital cutting. *Nature*, 538:506–509, 2016.

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

[26] Susan Lee-Rife, Anju Malhotra, Ann Warner, and Allison McGonagle Glinski. What works to prevent child marriage: A review of the evidence. *Studies in Family Planning*,

43(4):287–303, 2012. ISSN 1728-4465. doi: 10.1111/j.1728-4465.2012.00327.x. URL `http://dx.doi.org/10.1111/j.1728-4465.2012.00327.x`.

[27] Jean-Philippe Platteau, Giulia Camilotti, and Emmanuelle Auriol. Eradicating women-hurting customs. In Siwan Anderson, Lori Beaman, and Jean-Philippe Platteau, editors, *Towards Gender Equity in Development*, pages 319–356. Oxford University Press, 2018.

[28] Matthias Schief, Sonja Vogt, and Charles Efferson. Investigating the structure of son bias in Armenia with novel measures of individual preferences. *Demography*, 58:1737–1764, 2021.

[29] Juan Carlos Castilla-Rho, Rodrigo Rojas, Martin S Andersen, Cameron Holley, and Gregoire Mariethoz. Social tipping points in global groundwater management. *Nature Human Behaviour*, 1(9):640–649, 2017.

[30] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 2007(357):370–379, 2007.

[31] Nicholas A Christakis and James H Fowler. The collective dynamics of smoking in a large social network. *New England Journal of Medicine*, 358(21):2249–2258, 2008.

[32] Henry Travers, James Walsh, Sonja Vogt, Tom Clements, and EJ Milner-Gulland. Delivering behavioural change at scale: What conservation can learn from other fields. *Biological Conservation*, 257:109092, 2021.

[33] Karine Nyborg, John M Anderies, Astrid Dannenberg, Therese Lindahl, Caroline Schill, Maja Schlüter, W Neil Adger, Kenneth J Arrow, Scott Barrett, Stephen Carpenter, et al. Social norms as solutions. *Science*, 354(6308):42–43, 2016.

[34] J Doyne Farmer, Cameron Hepburn, Matthew C Ives, T Hale, Thomas Wetzer, Penny Mealy, Ryan Rafaty, Sugandha Srivastav, and Rupert Way. Sensitive intervention points in the post-carbon transition. *Science*, 364(6436):132–134, 2019.

[35] Marwil J Dávila-Fernández and Serena Sordi. Attitudes towards climate policies in a macrodynamic model of the economy. *Ecological Economics*, 169:106319, 2020.

[36] Ilona M Otto, Jonathan F Donges, Roger Cremades, Avit Bhowmik, Richard J Hewitt, Wolfgang Lucht, Johan Rockström, Franziska Allerberger, Mark McCaffrey, Sylvanus SP Doe, et al. Social tipping dynamics for stabilizing earth's climate by 2050. *Proceedings of the National Academy of Sciences*, 117(5):2354–2365, 2020.

[37] Sara M Constantino, Gregg Sparkman, Gordon T Kraft-Todd, Cristina Bicchieri, Damon Centola, Bettina Shell-Duncan, Sonja Vogt, and Elke U Weber. Scaling up change: A critical review and practical guide to harnessing social norms for climate action. *Psychological Science in the Public Interest*, 23(2):50–97, 2022.

[38] Charles Efferson. Policy to activate cultural change to amplify policy. *Proceedings of the National Academy of Sciences*, 118(23), 2021.

[39] Mark Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[40] H Peyton Young. Innovation diffusion in heterogeneous populations: contagion, social influence, and social learning. *American Economic Review*, 99(5):1899–1924, 2009.

[41] Matthew O Jackson and Dunia López-Pintado. Diffusion and contagion in networks with heterogeneous agents and homophily. *Network Science*, 1(1):49–67, 2013.

[42] Robin Schimmelpfennig, Sonja Vogt, Sönke Ehret, and Charles Efferson. Promotion of behavioural change for health in a heterogeneous population. *Bulletin of the World Health Organization*, 99(11):819, 2021.

[43] Duncan J. Watts and Peter Dodds. Threshold models of social influence. *The Oxford Handbook of Analytical Sociology*, pages 475–497, 2009.

[44] Michael Mäs and Heinrich H Nax. A behavioral study of "noise" in coordination games. *Journal of Economic Theory*, 162:195–208, 2016.

[45] John C. Harsanyi and Reinhard Selten. *A General Theory of Equilibrium Selection in Games*. Cambridge: The MIT Press, 1988.

[46] John B Van Huyck, Raymond C Battalio, and Richard O Beil. Tacit coordination games, strategic uncertainty, and coordination failure. *The American Economic Review*, 80(1): 234–248, 1990.

[47] Marc Knez and Colin Camerer. Creating expectational assets in the laboratory: Coordination in 'weakest-link' games. *Strategic Management Journal*, 15(S1):101–119, 1994.

[48] Paul G Straub. Risk dominance and coordination failures in static games. *The Quarterly Review of Economics and Finance*, 35(4):339–363, 1995.

[49] Kenneth Clark, Stephen Kay, and Martin Sefton. When are nash equilibria self-enforcing? an experimental analysis. *International Journal of Game Theory*, 29:495–515, 2001.

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

[51] Charles Efferson, Ryan McKay, and Ernst Fehr. The evolution of distorted beliefs vs. mistaken choices under asymmetric error costs. *Evolutionary Human Sciences*, 2:e27, 2020. doi: 10.1017/ehs.2020.25.

[52] Jasmina Arifovic, Cars Hommes, Anita Kopányi-Peuker, and Isabelle Salle. Ten isn't large! group size and coordination in a large-scale experiment. *American Economic Journal: Microeconomics*, 15(1):580–617, 2023.

[53] Eva Vivalt. Heterogeneous treatment effects in impact evaluation. *American Economic Review*, 105(5):467–70, 2015.

[54] Ulrich Schmidt and Philipp C Wichardt. Inequity aversion, welfare measurement and the Gini index. *Social Choice and Welfare*, 52:585–588, 2019.

[55] Russell Cooper, Douglas V DeJong, Robert Forsythe, and Thomas W Ross. Communication in coordination games. *The Quarterly Journal of Economics*, 107(2):739–771, 1992.

[56] Andreas Blume and Andreas Ortmann. The effects of costless pre-play communication: Experimental evidence from games with pareto-ranked equilibria. *Journal of Economic Theory*, 132(1):274–290, 2007.

[57] Catherine C Eckel and Rick K Wilson. Social learning in coordination games: does status matter? *Experimental Economics*, 10:317–329, 2007.

[58] Raymond Battalio, Larry Samuelson, and John Van Huyck. Optimization incentives and coordination failure in laboratory stag hunt games. *Econometrica*, 69(3):749–764, 2001.

[59] Samuel Bowles. *Microeconomics: Behavior, Institutions, and Evolution.* New York: Russell Sage, 2004.

[60] Sergey Gavrilets and Peter J Richerson. Collective action and the evolution of social norm internalization. *Proceedings of the National Academy of Sciences*, page 201703857, 2017.

Acknowledgements and funding

CE and SV would like the thank the Swiss National Science Foundation (Nr. 100018_185417) for financial support.

Author contributions

CE, SE, and SV developed the initial idea. CE developed and analysed the analytical model. LvF developed and analysed the simulation model with input from SE and CE. CE wrote the paper with feedback from SE, LvF, and SV.

Competing interests

We have no competing interests.

Table 1: **Payoff matrices before and after intervention.** The matrices show row player payoffs. Before intervention, everyone plays a coordination game, but individuals also have heterogeneous preferences (Pre-intervention (All)). Specifically, the x_i are distributed somehow on $(0, d-b)$, where $b < d$. This ensures that, for each i, $a + x_i > a$, and $b + x_i < d$. The social planner then targets a subset of the population with an intervention and incentivises these individuals to choose Alt (Post-intervention (T)) with a new payoff matrix, where $h > g$. Individuals who are not targeted retain their original payoff matrices (Post-intervention (NT)).

Pre-intervention (All)			Post-intervention (T)			Post-intervention (NT)		
	SQ	Alt		SQ	Alt		SQ	Alt
SQ	$a + x_i$	$b + x_i$	SQ	g	g	SQ	$a + x_i$	$b + x_i$
Alt	a	d	Alt	h	h	Alt	a	d

Table 2: **Player matches.** When the population stabilises after intervention, each individuals falls into one of three categories of player. Targeted individuals play Alt (Alt,T). Some or all non-targeted players may also play Alt (Alt,NT). Some or all non-targeted players may play SQ (SQ,NT). The table shows all ways in which two players can pair off to play the game (P_1, P_2), the probabilities associated with these pairings ($P(P_1, P_2)$), and the payoffs that result (π_1, π_2). The table uses a generic $\hat{q}_z$, where $z \in \{A, R\}$, because all terms here have the same structure regardless of whether the target is amenable ($z = A$) or resistant ($z = R$). We use a prime to distinguish between the preferences of Player 1 and Player 2 when the x_i appear.

P_1	P_2	$P(P_1, P_2)$	π_1	π_2
(Alt,T)	(Alt,T)	ϕ^2	h	h
(Alt,T)	(Alt,NT)	$\phi(\hat{q}_z - \phi)$	h	d
(Alt,T)	(SQ,NT)	$\phi(1 - \hat{q}_z)$	h	$b + x_{i'}$
(Alt,NT)	(Alt,T)	$(\hat{q}_z - \phi)\phi$	d	h
(Alt,NT)	(Alt,NT)	$(\hat{q}_z - \phi)^2$	d	d
(Alt,NT)	(SQ,NT)	$(\hat{q}_z - \phi)(1 - \hat{q}_z)$	a	$b + x_{i'}$
(SQ,NT)	(Alt,T)	$(1 - \hat{q}_z)\phi$	$b + x_i$	h
(SQ,NT)	(Alt,NT)	$(1 - \hat{q}_z)(\hat{q}_z - \phi)$	$b + x_i$	a
(SQ,NT)	(SQ,NT)	$(1 - \hat{q}_z)^2$	$a + x_i$	$a + x_{i'}$

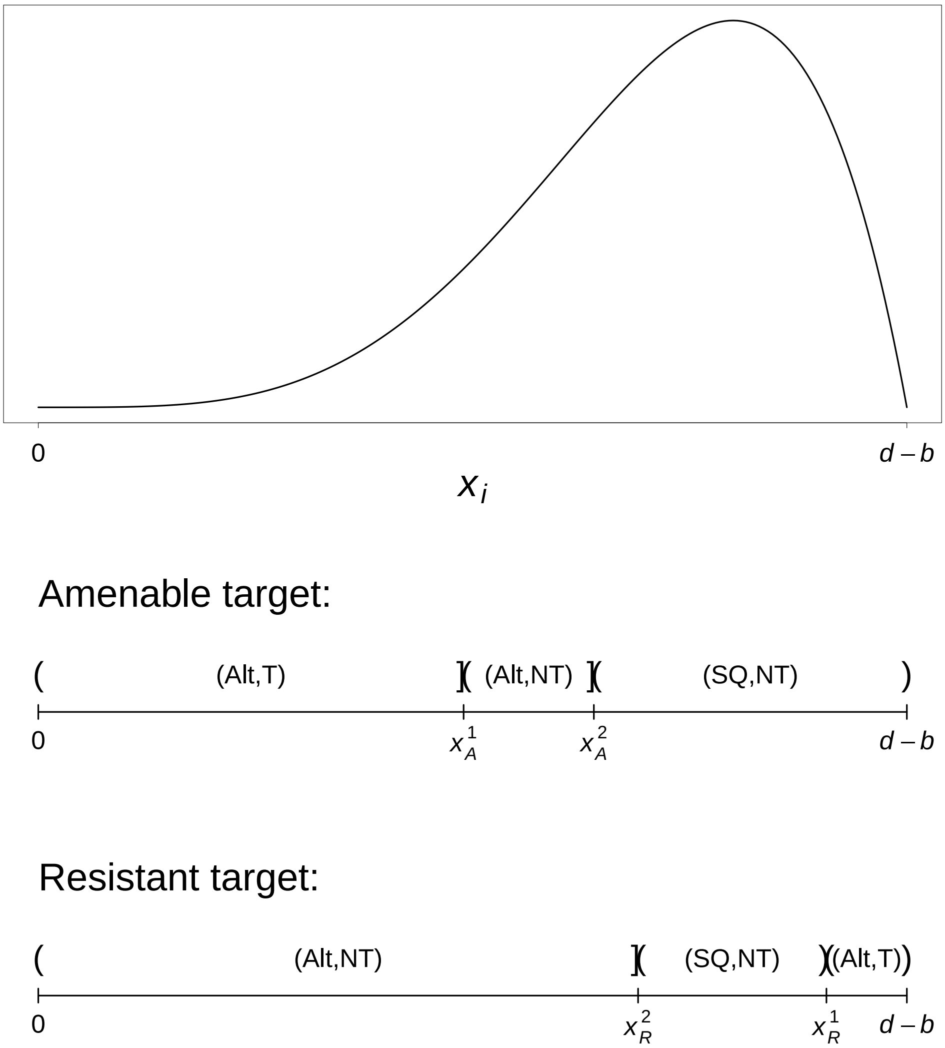

Figure 1: **An example of two different partitions of the population.** The figure shows how an amenable target versus a resistant target might partition the population in different ways given the distribution of x_i values shown. For an amenable target, the example assumes the social planner targets 10% of the population ($\phi = 0.1$). Under an amenable target, the population stabilises post-intervention on 30% choosing Alt ($\hat{q}_A = 0.3$). Specifically, (Alt,T) are the targeted individuals who change behaviour due to direct experience with the intervention, and (Alt,NT) are the non-targeted individuals who change behaviour due to coordination incentives post-intervention. The remaining 70% of the population are the non-targeted individuals who continue choosing the status quo behaviour, denoted (SQ,NT). For a resistant target, the example also assumes the social planner targets 10%, but in the long-run 50% end up choosing Alt. The partition is completely different from the amenable case because the (Alt,T) individuals come from the right tail of the x_i distribution, which leaves the left tail for the (Alt,NT) individuals and the middle for the (SQ,NT) individuals. The parentheses and square brackets denote where the intervals associated with the partition are open or closed respectively.

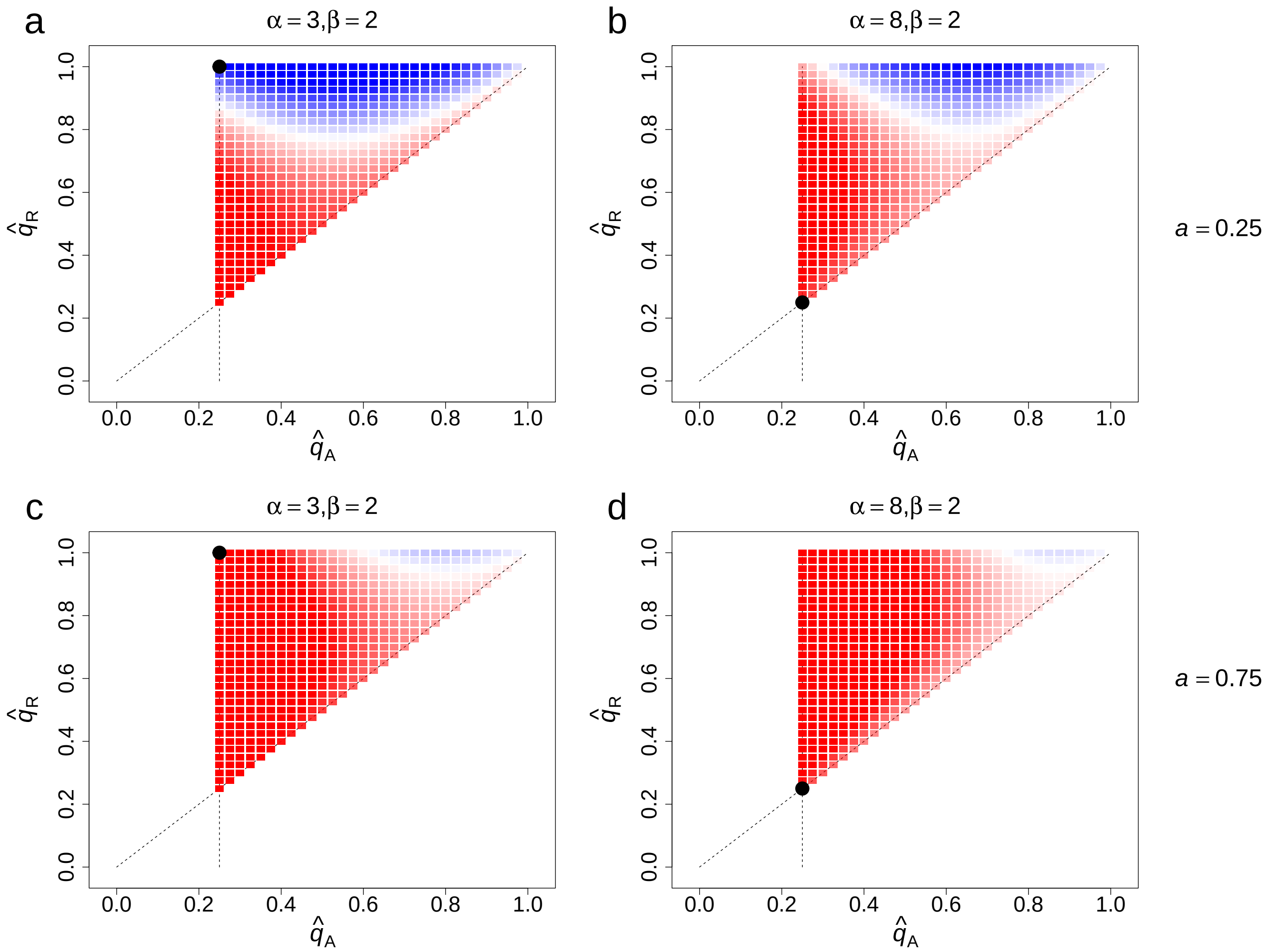

Figure 2: **Comparison of intervention strategies given an intervention of size $\phi = 0.25$.** The graphs represent the values of $E_A[\Pi] - E_R[\Pi]$ (Eq. 6) for all steady states consistent with $\phi \leq \hat{q}_A \leq \hat{q}_R$. Red indicates that $E_A[\Pi] > E_R[\Pi]$, and thus an amenable target is better in terms of social welfare than a resistant target. Blue indicates the opposite. Colour intensity reflects the magnitude of the differences. The black dot shows the steady state under the threshold model [14, 39]. To generate these graphs, $b = 0$, and $d = 1$. The parameter a takes either a moderately low value (**a**, **b**) or a moderately high value (**c**, **d**). The x_i values are beta distributed with shape parameters α and β on the interval $(0, 1)$. The distribution is either weakly left-skewed (**a**, **c**) or somewhat strongly left-skewed (**b**, **d**).

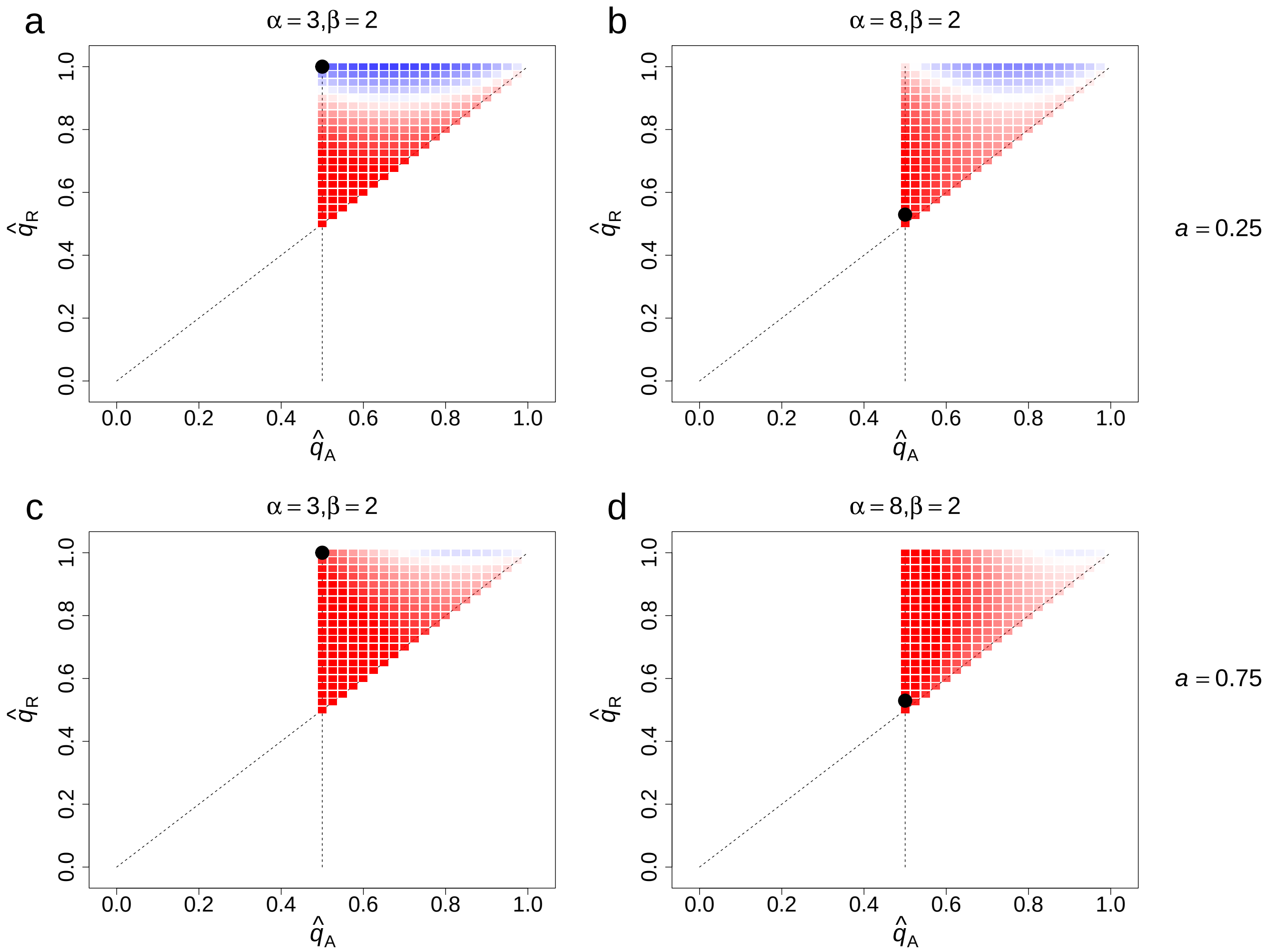

Figure 3: **Comparison of intervention strategies given an intervention of size $\phi = 0.5$.** The graphs represent the values of $E_A[\Pi] - E_R[\Pi]$ (Eq. 6) for all steady states consistent with $\phi \leq \hat{q}_A \leq \hat{q}_R$. Red indicates that $E_A[\Pi] > E_R[\Pi]$, and thus an amenable target is better in terms of social welfare than a resistant target. Blue indicates the opposite. Colour intensity reflects the magnitude of the differences. The black dot shows the steady state under the threshold model [14, 39]. To generate these graphs, $b = 0$, and $d = 1$. The parameter a takes either a moderately low value (**a**, **b**) or a moderately high value (**c**, **d**). The x_i values are beta distributed with shape parameters α and β on the interval $(0, 1)$. The distribution is either weakly left-skewed (**a**, **c**) or somewhat strongly left-skewed (**b**, **d**).

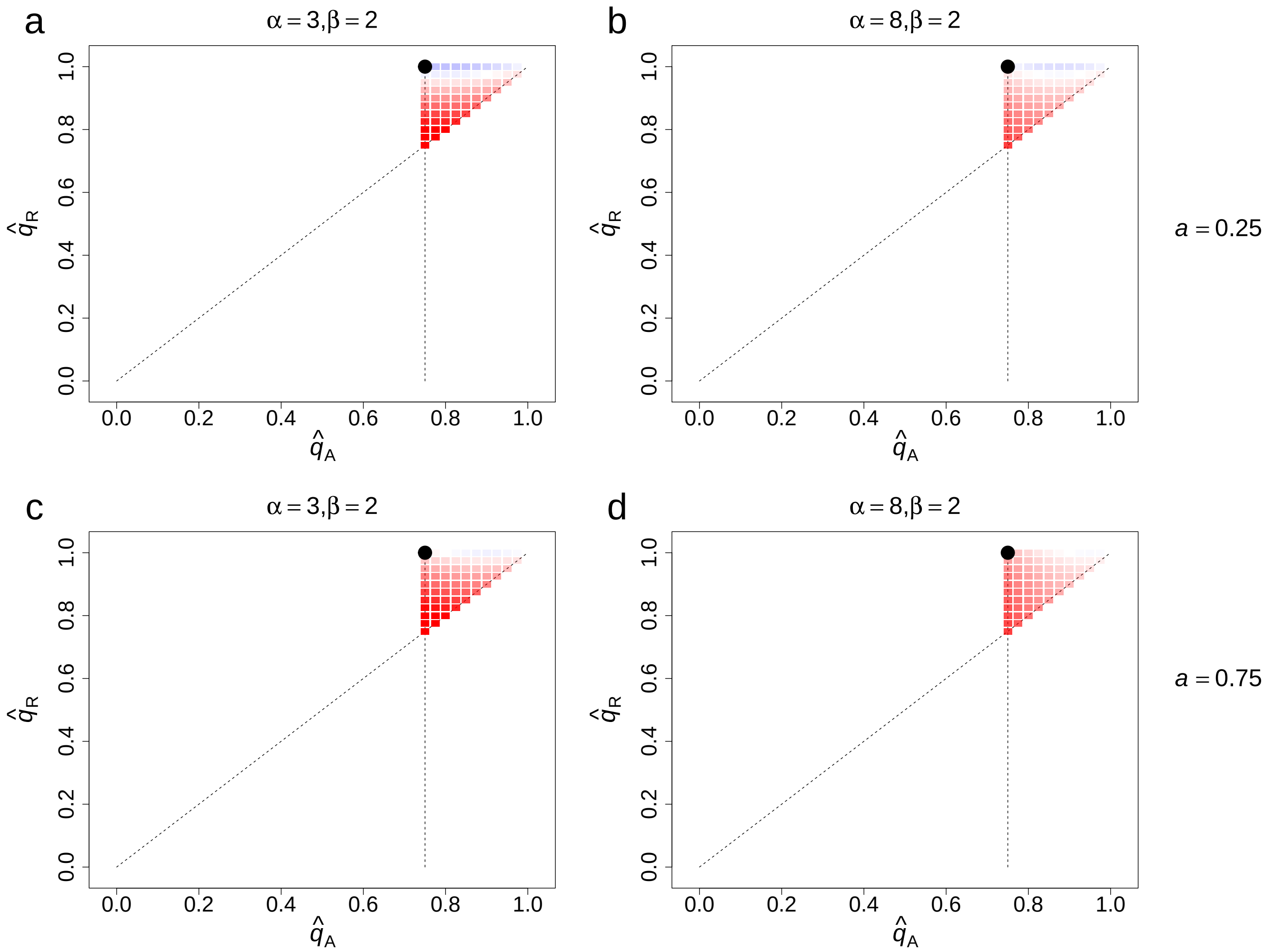

Figure 4: **Comparison of intervention strategies given an intervention of size $\boldsymbol{\phi = 0.75}$.** The graphs represent the values of $E_A[\Pi] - E_R[\Pi]$ (Eq. 6) for all steady states consistent with $\phi \le \hat{q}_A \le \hat{q}_R$. Red indicates that $E_A[\Pi] > E_R[\Pi]$, and thus an amenable target is better in terms of social welfare than a resistant target. Blue indicates the opposite. Colour intensity reflects the magnitude of the differences. The black dot shows steady state values for both intervention strategies under the threshold model [14, 39]. To generate these graphs, $b = 0$, and $d = 1$. The parameter a takes either a moderately low value (**a**, **b**) or a moderately high value (**c**, **d**). The x_i values are beta distributed with shape parameters α and β on the interval $(0, 1)$. The distribution is either weakly left-skewed (**a**, **c**) or somewhat strongly left-skewed (**b**, **d**).

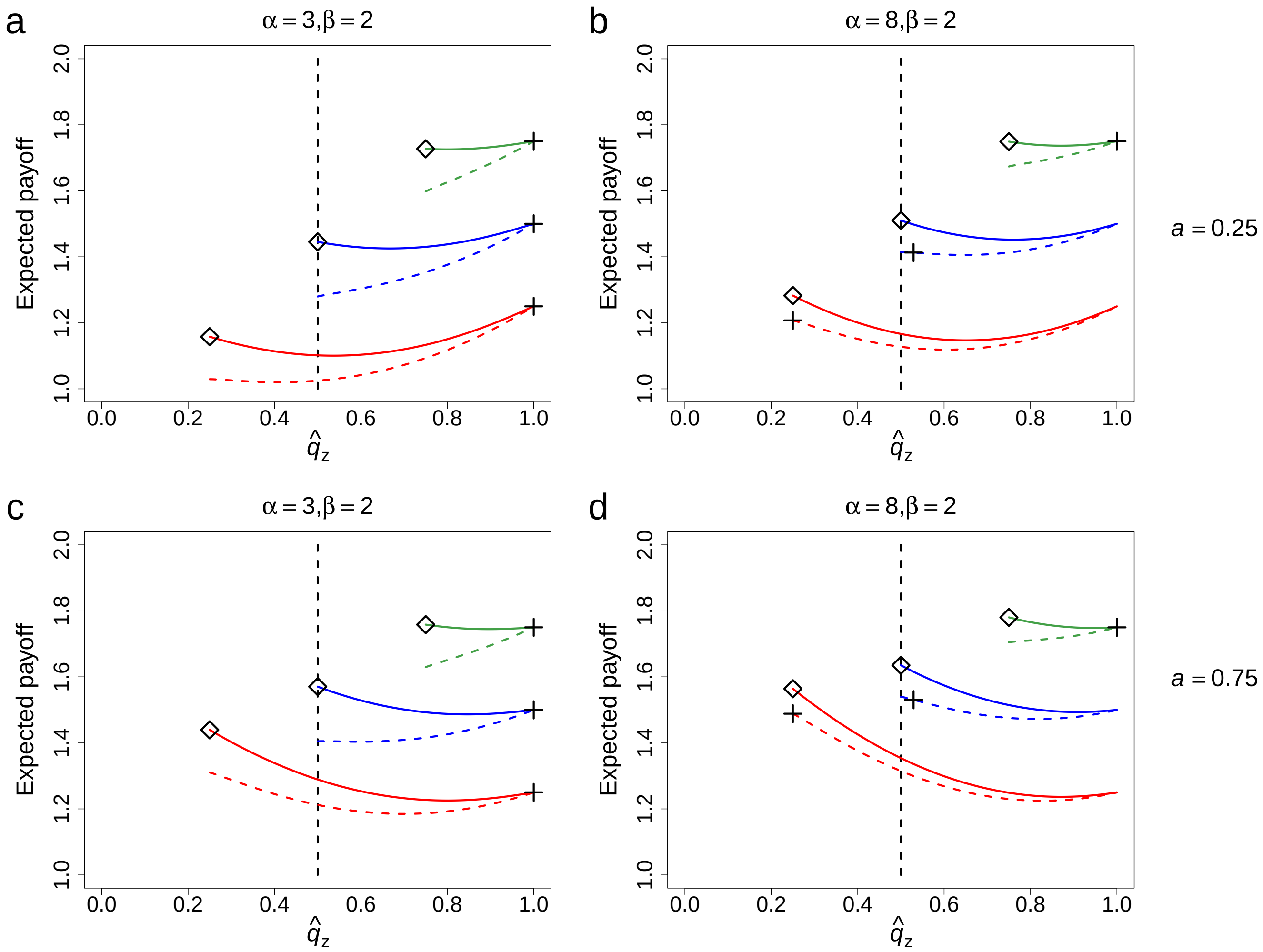

Figure 5: **Social welfare and (mis-)coordination.** The graphs show the expected payoff of a randomly selected individual as a function of the steady state proportion choosing Alt, $\hat{q}_z$, where $z \in \{A, R\}$. Solid lines show the expected payoffs under an amenable target (Eq. 3) and dashed lines under a resistant target (Eq. 4). Intervention sizes are $\phi = 0.25$ in red, $\phi = 0.5$ in blue, and $\phi = 0.75$ in green. The $\diamond$ shows the steady state value under the threshold model [14, 39] given an amenable target and the $+$ given a resistant target. To generate these graphs, $b = 0$, and $d = 1$. The parameter a takes either a moderately low value (**a**, **b**) or a moderately high value (**c**, **d**). The x_i values are beta distributed with shape parameters α and β on the interval $(0, 1)$. The distribution is either weakly left-skewed (**a**, **c**) or somewhat strongly left-skewed (**b**, **d**). Miscoordination reaches the maximum possible rate under random matching when $\hat{q}_z = 0.5$. Thus, any time expected payoffs increase as the steady state approaches this value, social welfare is increasing even though miscoordination is also increasing.

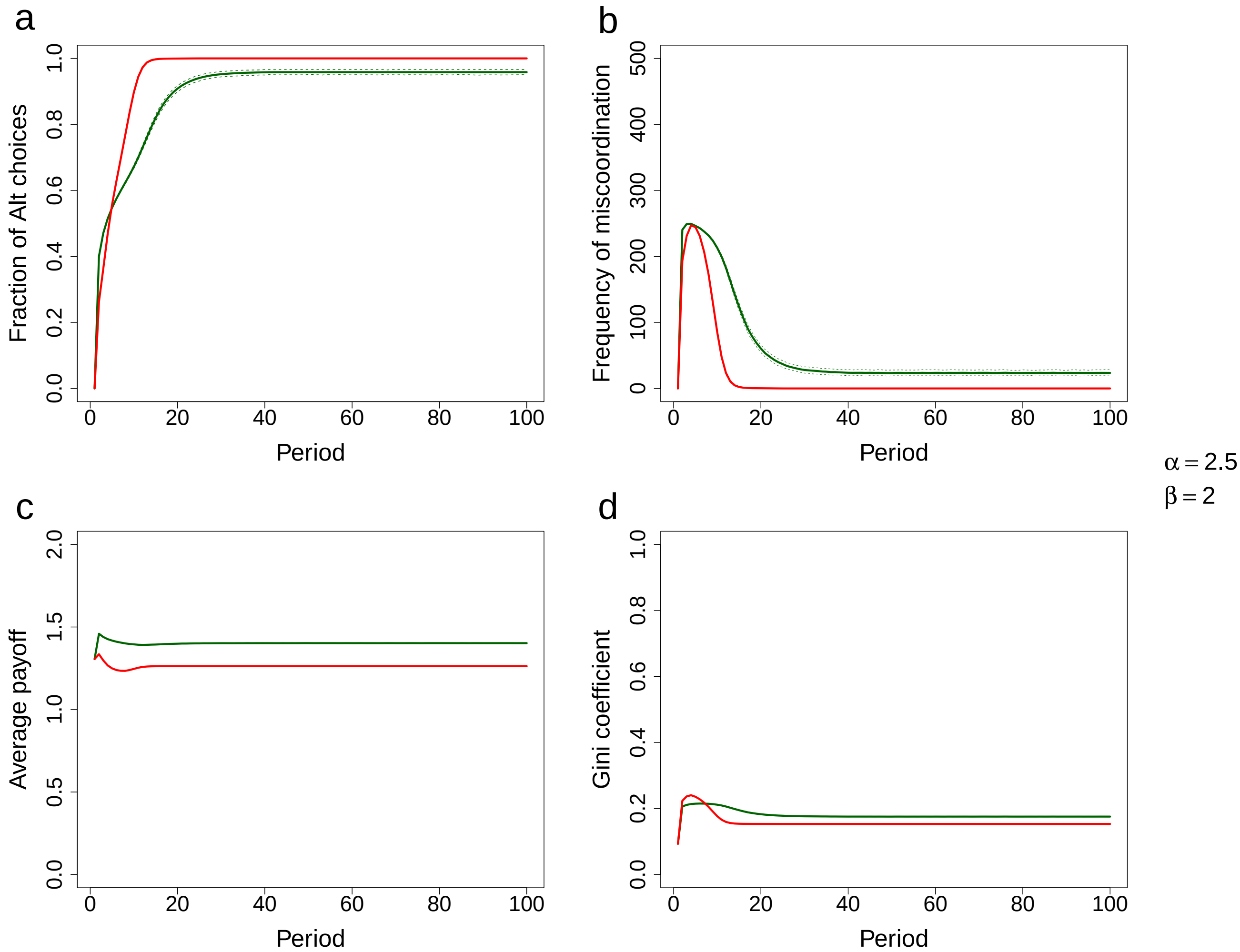

Figure 6: The figure shows a comparison between the dynamics of amenable and resistant target. The solid lines show values averaged over all 1000 simulations. The dashed lines show 95% bootstrapped confidence intervals. Populations with amenable targets are shown in green, and populations with resistant targets are shown in red. Graphs show the faction of agents choosing Alt (**a**), the frequency of miscoordination (**b**), average payoffs (**c**), and the Gini coefficient (**d**). The initial conditions at $t = 1$ are pre-intervention. Thus, a comparison between $t = 1$ and any $t > 1$ shows how populations changed, conditional on an amenable or resistant target, as a result of the intervention. Aside from α and β, parameter values are $\phi = 0.75$, $a = 0.75$, and $h = 2$.

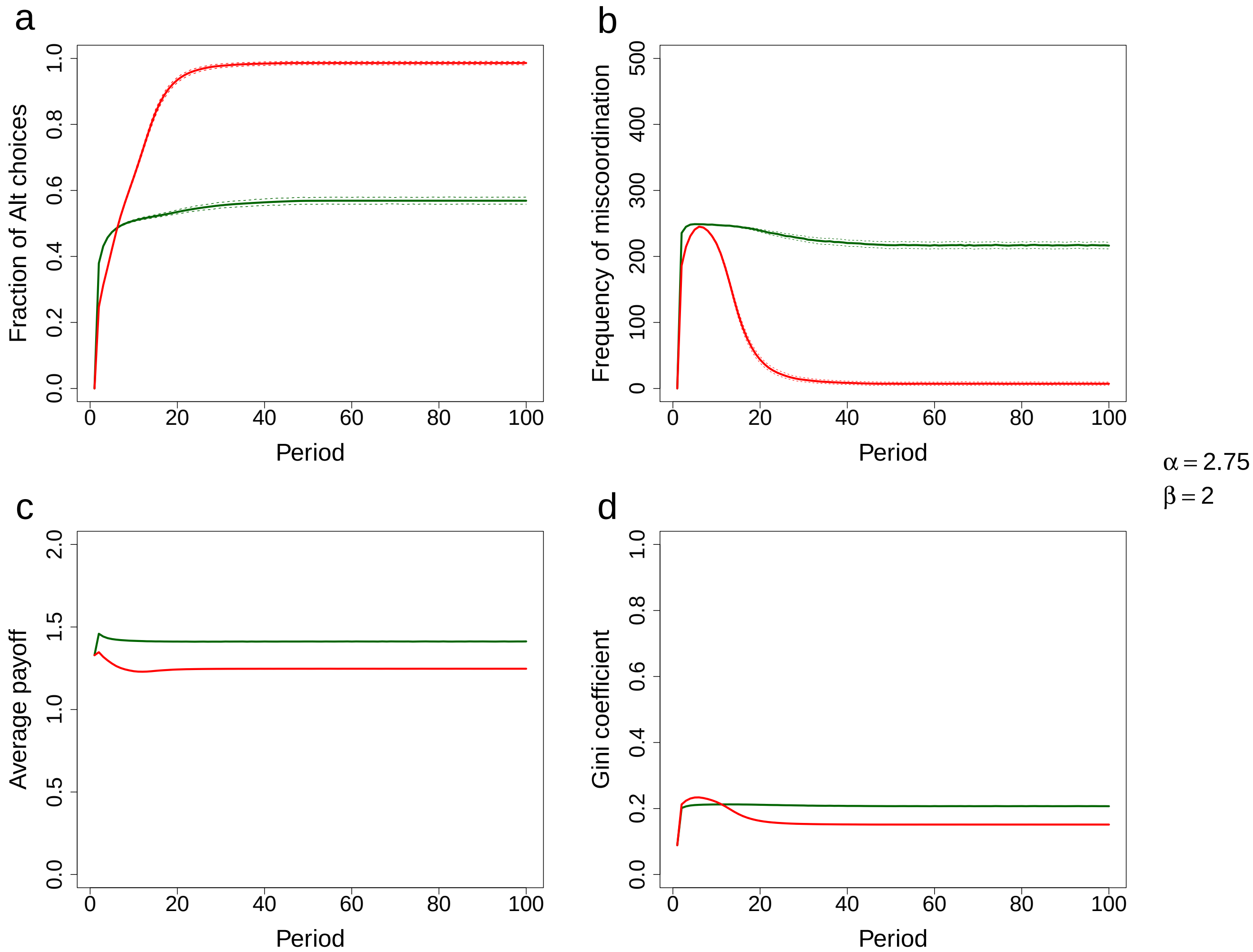

Figure 7: The figure shows a comparison between the dynamics of amenable and resistant target. The solid lines show values averaged over all 1000 simulations. The dashed lines show 95% bootstrapped confidence intervals. Populations with amenable targets are shown in green, and populations with resistant targets are shown in red. Graphs show the faction of agents choosing Alt (**a**), the frequency of miscoordination (**b**), average payoffs (**c**), and the Gini coefficient (**d**). The initial conditions at $t = 1$ are pre-intervention. Thus, a comparison between $t = 1$ and any $t > 1$ shows how populations changed, conditional on an amenable or resistant target, as a result of the intervention. Aside from α and β, parameter values are $\phi = 0.75$, $a = 0.75$, and $h = 2$.

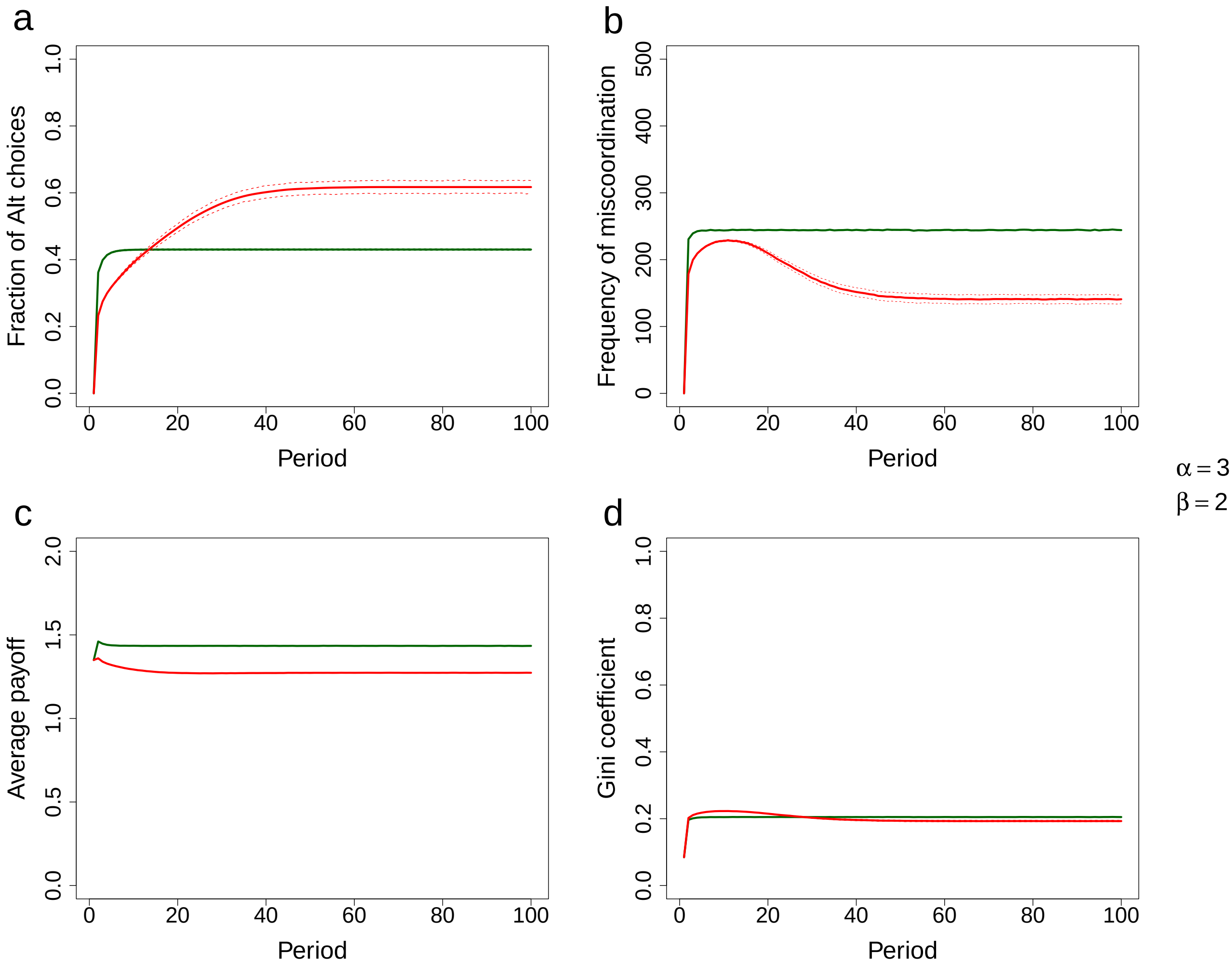

Figure 8: The figure shows a comparison between the dynamics of amenable and resistant target. The solid lines show values averaged over all 1000 simulations. The dashed lines show 95% bootstrapped confidence intervals. Populations with amenable targets are shown in green, and populations with resistant targets are shown in red. Graphs show the faction of agents choosing Alt (**a**), the frequency of miscoordination (**b**), average payoffs (**c**), and the Gini coefficient (**d**). The initial conditions at $t = 1$ are pre-intervention. Thus, a comparison between $t = 1$ and any $t > 1$ shows how populations changed, conditional on an amenable or resistant target, as a result of the intervention. Aside from α and β, parameter values are $\phi = 0.75$, $a = 0.75$, and $h = 2$.

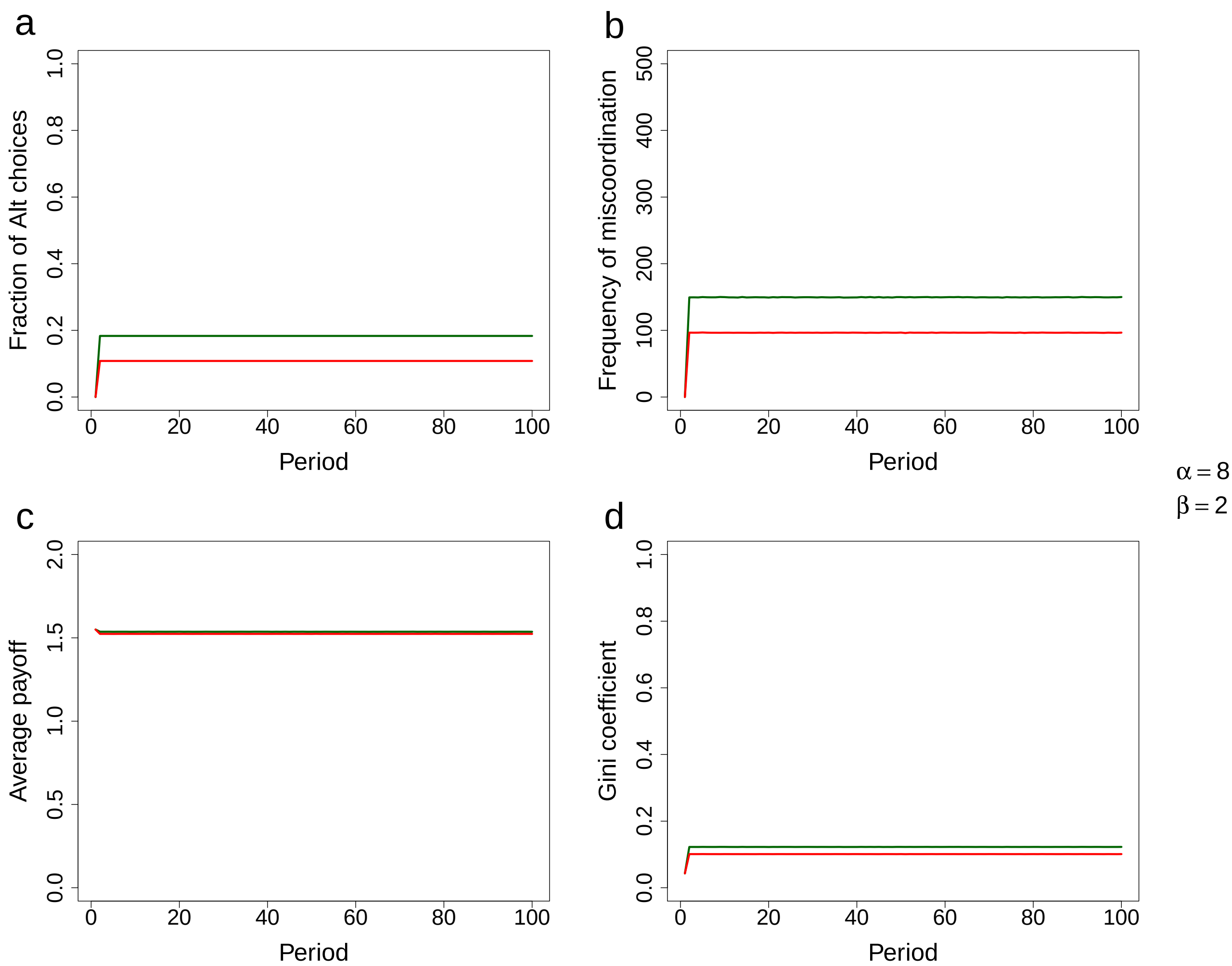

Figure 9: The figure shows a comparison between the dynamics of amenable and resistant target. The solid lines show values averaged over all 1000 simulations. The dashed lines show 95% bootstrapped confidence intervals. Populations with amenable targets are shown in green, and populations with resistant targets are shown in red. Graphs show the faction of agents choosing Alt (**a**), the frequency of miscoordination (**b**), average payoffs (**c**), and the Gini coefficient (**d**). The initial conditions at $t = 1$ are pre-intervention. Thus, a comparison between $t = 1$ and any $t > 1$ shows how populations changed, conditional on an amenable or resistant target, as a result of the intervention. Aside from α and β, parameter values are $\phi = 0.75$, $a = 0.75$, and $h = 2$.