Policy to activate cultural change to amplify policy

Charles Efferson[a,1]

This manuscript was compiled on April 22, 2021

Social tipping | Behavior change | Social norms | Coordination | Cultural evolution

If you happened to be driving down the road in Sweden at 04:50 on 3 September 1967, the Swedish government required you to stop. You then had to move slowly from the left to the right side of the road, and at 05:00 you could continue on your way. Although Sweden invested heavily in preparing for this pivotal ten minutes, the transition from left to right created some inevitable confusion (1). Nonetheless, the transition to a new equilibrium was fast. Traffic accidents and insurance claims actually declined immediately after the change, presumably because of extra caution behind the wheel, but they soon returned to normal (2). With a one-time government initiative, Swedes tipped from driving on the left to driving on the right, where they have remained ever since. The rest of us gained a compelling metaphor, arguably too compelling, for how social tipping can support society-wide changes in culture consistent with policy goals.

I say "arguably too compelling" because choosing a side of the road is a special kind of coordination problem maximally suited to rapid change. The question is, when does the potential for rapid social tipping extend to other coordination problems that are similar in some ways but different in others? More broadly, can we predict and even control tipping in settings that are typical precisely because they are more complex than choosing the left or right side of the road? In a companion article, Andreoni et al. (3) examine exactly these questions with a theoretical and experimental approach. Apart from basic scientific interest, the questions are relevant across an impressive array of policy domains where social norms, applied cultural evolution, and tipping appear as related mechanisms for behavior change (4, 5). Example domains range from equality, social justice, and health (6, 7) to resource conservation (8, 9) and climate change (10, 11).

Choosing a side of the road is a special problem for at least three reasons. Simple preferences to coordinate with people nearby do not mix with other motives. Moreover, these preferences are the same for everyone, and they are stable through time. Intuitively, from an ex ante perspective before a society has chosen left or right, everyone agrees that either side is and will remain just as good as the other. The one and only concern is that everyone makes the same choice. Language is similar. "Der Hund" and "le chien" both work fine and will continue to do so; we just need to agree (12, 13). Step outside these two domains, however, and many coordination problems involve a number of additional complexities.

Andreoni et al. (3) add important and realistic complexity by abandoning exactly the three characteristics that make driving and language special problems. They examine a setting in which individuals are randomly paired to play a game. Each player must choose blue or green, and everyone faces incentives to coordinate their choices with their partners. Players play, receive a payoff, update their beliefs about how others play, and then pair off and play again. So far, this sounds like driving, but the similarities end there. Specifically, each player has a ranking over the equilibria of the game, which means the player either prefers coordinating on blue over coordinating on green or vice versa. Players also differ from each other in terms of their rankings, and player rankings change through time.

Andreoni et al. (3) emphasize the evolution of social norms as an organizing principle. A norm is a common behavior together with the widespread belief that the behavior is and should remain common. A norm helps people pick a specific behavior when everyone values choosing the same behavior, which is a problem with multiple solutions. This pressure to behave like others is also why norm evolution can exhibit tipping. If a status quo norm becomes unstable, the pressure to conform can lead the population to coalesce quickly around a new norm.

To develop a modeling framework for how norms evolve, Andreoni et al. (3) decompose preferences into three parts. First, each player faces a basic material incentive that favors either coordinating on blue over coordinating on green or vice versa. Second, each player faces material incentives that are relevant when two players choose different options. Specifically, in addition to the opportunity costs of miscoordination, each player in a miscoordinating pair pays a cost that increases as the player's choice becomes more unusual. We can interpret this cost as punishment. These first two components of the incentive structure are material in the sense that they were monetized in Andreoni et al.'s experiment. More broadly, they represent the public features of decision making that would be readily available for policy intervention. A policy maker, for example, can subsidize the behavior she prefers, tax

[a]Faculty of Business and Economics, University of Lausanne, 1015 Lausanne, Switzerland.
CE wrote the paper.
[1]To whom correspondence should be addressed. E-mail: charles.efferson@unil.ch.
The author declares no competing interests.

the behaviors she does not prefer, and punish deviants. The third component of preferences is an idiosyncratic psychological quantity that appears in the predictive model of Andreoni et al. but was not monetized in their experiment. Among other interpretations, variation in this quantity represents the fact that some people are more open to new experiences than others, a form of ordinary heterogeneity that can affect the spread of innovations in a population (14, 15).

With all three parts of the theoretical incentive structure in place, each individual has an indifference point. If the proportion of individuals recently choosing green is at least as large as this indifference point, the individual in question chooses green by assumption. The population consists of a distribution of indifference points. This distribution changes through time and in turn influences how behavior and associated norms evolve.

In Andreoni et al.'s (3) experimental sessions, material incentives initially favored coordinating on blue over coordinating on green, and groups immediately adopted a blue norm as a result. With a blue norm in place, material incentives began to change. At a given point in time, for any individual whose material incentives favored blue over green, these incentives would switch the ranking with probability 0.1. As these new incentives trickled into the population, the distribution of indifference points should have become increasingly favorable for green.

Fig. 1 shows a stylized simulation in which this steady trickle leads to tipping. In $t = 1$, no one faces material incentives that favor coordinating on green. All three parts of the incentive structure combine to create a distribution of indifference points that is not favorable for green, and no one chooses green. Material incentives then begin to change, and the distribution of indifference points drifts downward. For a while, behavior change lags behind as everyone continues to conform to the status quo blue norm. At $t = 6$, changes in behavior start to race ahead of the changes in material incentives, and by $t = 9$ the entire population has switched to choosing green. This is social tipping. Coordination and conformity oppose the behavioral effects of changing incentives at first, but then suddenly a new regime appears in which they amplify these effects.

This kind of tipping, however, may not occur, and altogether Andreoni et al. (3) implemented nine experimental treatments to examine a variety of behavioral mechanisms. Four treatments operated directly via material incentives. Andreoni et al. manipulated the material incentives related to coordinating, and they manipulated the material punishment associated with miscoordinating. Their model does an outstanding job of predicting observed tipping (3, Fig. 4). In one especially revealing treatment, Andreoni et al. allowed the participants themselves to set the punishment costs of miscoordinating. This is like a situation in which a policy maker uses a combination of taxes and subsidies to promote a specific behavior, but the punishment of norm violations is an informal affair that citizens handle themselves. In this treatment, participants consistently set punishment costs too high. Doing so saved them the short-run costs of miscoordinating while transitioning to a new norm, but using punishment to block transitions brought substantial opportunity costs in the long-run.

Four additional treatments manipulated the information and expectations participants had about the changes occurring in their groups. In one treatment, participants received immediate feedback about what others were choosing, an approach designed to mimic the speed of modern communications. One can imagine that readily available information would have facilitated tipping, but it did not. Instead, it seems to have made the early prevalence of blue salient, and this treatment had no effect on tipping. In another treatment, Andreoni et al. (3) cut the size of experimental groups from 20 to 10, which increased the relative influence of each decision maker. This significantly increased tipping. Surprisingly, however, when transitions to a green norm occurred, they were long drawn-out affairs with a lot of miscoordination along the way. Average earnings were especially low as a result. This result shows that transitions to a new, socially beneficial equilibrium can actually be socially harmful depending on how long the transition takes.

In the "Public awareness" and "Preference poll" treatments, Andreoni et al. (3) introduced two mechanisms designed to make private information public (6). Under public awareness, the experimenters gave participants a running log of the kinds of changes in material incentives taking place within the group. The preference poll polled group members about their preferred norm after several periods of play and immediately made the poll results public. Both of these treatments revealed information about participants that would have otherwise remained private, and even trivial revelations of this sort can strongly affect cultural evolution (16). The result in both treatments was a significant increase in tipping to the socially beneficial norm.

Finally, Andreoni et al. implemented a treatment that rewarded those who first attempted to instigate norm change, but only when these attempts were successful. This extra reward for agents of change seems to have motivated individuals predisposed to change anyway, but it also ignored people with a status quo bias. As the authors point out, tipping requires behavior change among both types, both those who are ready to lead the way to a new norm and those who are not. The results across groups in this treatment were highly unpredictable, with half of the groups tipping to green and half sticking with blue. Altogether, Andreoni et al. used a convincing policy-inspired mix of treatments to detail several behavioral subtleties related to tipping. At the same time, their study highlights how much we still need to learn about the various scenarios in which a policy maker might want to activate endogenous cultural change.

One important scenario is when the population is sub-divided into groups that have distinct social identities tied to the norms and behaviors in question. For example, imagine a situation in which some people have tied their social identities to their shared decision to wear face masks in a pandemic, while others have based their social identities on rejecting masks (17). In cases like this, the distribution of indifference points will look quite different from those assumed in Andreoni et al. (see also Fig. 1). The distribution will tend to be strongly bimodal, with one mode for the group that likes one behavior and another mode for the group that likes the other behavior. Tipping points may not exist in situations like this, and the most challenging situation of all is when the groups have social identities that are not only distinct, but oppositional (18). Oppositional identities would mean, for example, that the group rejecting masks values this stance precisely because of the difference it creates with respect to the group wearing masks (19). If preferences take this form, the policy maker who sparks a commitment to her preferred norm in one group likely entrenches and adds value to a different norm in the other group (18). The increasingly sectarian nature of U.S. politics (20) suggests that dynamics of this sort could be common in the future.

A second issue involves the options available to the policy maker. Andreoni et al. implement several treatments that reflect the

kinds of choices a policy maker might consider to try and provoke tipping. Their treatments represent policy initiatives that subsidize the desired behavior, punish the undesired behavior, influence the information people have, and reward those who instigate change. These are all important possibilities. In addition, a policy maker might also want to constrain an intervention to a specific segment of the population. Indeed, much of the policy appeal of tipping follows from the idea that an intervention touches only some people. When these people change their behavior, however, the effect spills over to generate additional change among those never exposed to the intervention. If a policy maker wants a constrained approach of this sort, she must decide whom to target. Some strategies prioritize the effects among those directly exposed to the intervention while minimizing the changes that occur among those not exposed. Other strategies do the opposite, with a range of trade-offs in between the extremes (18).

Tipping has a theatrical quality, with rapid changes that somehow seem both surprising and obvious after they have occurred. Tipping is also tempting as a policy tool because it implies the policy maker can recruit social interactions within a population to point cultural evolution in a specific direction. Empirically, however, people are strikingly heterogeneous in terms of how they learn from and react to the choices of others (21, 22). This suggests that tipping and other cultural evolutionary processes can easily involve a daunting level of complexity. Andreoni et al. (3) have provided an important study of ways to examine and manage some of this complexity.

1 E Flock, Dagen H: The day Sweden switched sides of the road (photo). *Wash. Post* **N/A** (2012).
2 D Bierend, Throwback Thursday: Hilarity ensues as Sweden starts driving on the right side of the road. *Wired* **N/A** (2014).
3 J Andreoni, N Nikiforakis, S Siegenthaler, Predicting social tipping and norm change in controlled experiments. *Proc. Natl. Acad. Sci.* **?**, ? (2021).
4 HP Young, The evolution of social norms. *Annu. Rev. Econ.* **7**, 359–387 (2015).
5 M Muthukrishna, Cultural evolutionary public policy. *Nat. Hum. Behav.* **4**, 12–13 (2020).
6 C Bicchieri, *Norms in the wild: How to diagnose, measure, and change social norms.* (Oxford University Press), (2016).
7 JP Platteau, G Camilotti, E Auriol, Eradicating women-hurting customs. *Towards gender equity development* **319** (2018).
8 JC Castilla-Rho, R Rojas, MS Andersen, C Holley, G Mariethoz, Social tipping points in global groundwater management. *Nat. Hum. Behav.* **1**, 640–649 (2017).
9 H Travers, J Walsh, S Vogt, T Clements, E Milner-Gulland, Delivering behavioural change at scale: What conservation can learn from other fields. *Biol. Conserv.* **257**, 109092 (2021).
10 K Nyborg, et al., Social norms as solutions. *Science* **354**, 42–43 (2016).
11 MJ Dávila-Fernández, S Sordi, Attitudes towards climate policies in a macrodynamic model of the economy. *Ecol. Econ.* **169**, 106319 (2020).
12 D Centola, A Baronchelli, The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proc. Natl. Acad. Sci.* **112**, 1989–1994 (2015).
13 D Centola, J Becker, D Brackbill, A Baronchelli, Experimental evidence for tipping points in social convention. *Science* **360**, 1116–1119 (2018).
14 J Henrich, Cultural transmission and the diffusion of innovations: Adoption dynamics indicate that biased cultural transmission is the predominate force in behavioral change. *Am. Anthropol.* **103**, 992–1013 (2001).
15 EM Rogers, *Diffusion of Innovations.* (Simon and Schuster), (2010).
16 JK Goeree, TR Palfrey, BW Rogers, RD McKelvey, Self-correcting information cascades. *The Rev. Econ. Stud.* **74**, 733–762 (2007).
17 C Moya, , et al., Dynamics of behavior change in the COVID world. *Am. J. Hum. Biol.* **e23485** (2020).
18 C Efferson, S Vogt, E Fehr, The promise and the peril of using social influence to reverse harmful traditions. *Nat. Hum. Behav.* **4**, 55–68 (2020).
19 SM Utych, Messaging mask wearing during the COVID-19 crisis: Ideological differences. *J. Exp. Polit. Sci.* **N/A**, 1–11 (2020).
20 EJ Finkel, et al., Political sectarianism in America. *Science* **370**, 533–536 (2020).
21 A Mesoudi, L Chang, SR Dall, A Thornton, The evolution of individual and cultural variation in social learning. *Trends Ecol. & Evol.* **31**, 215–225 (2016).
22 RL Kendal, et al., Social learning strategies: Bridge-building between fields. *Trends Cogn. Sci.* **22**, 651–665 (2018).

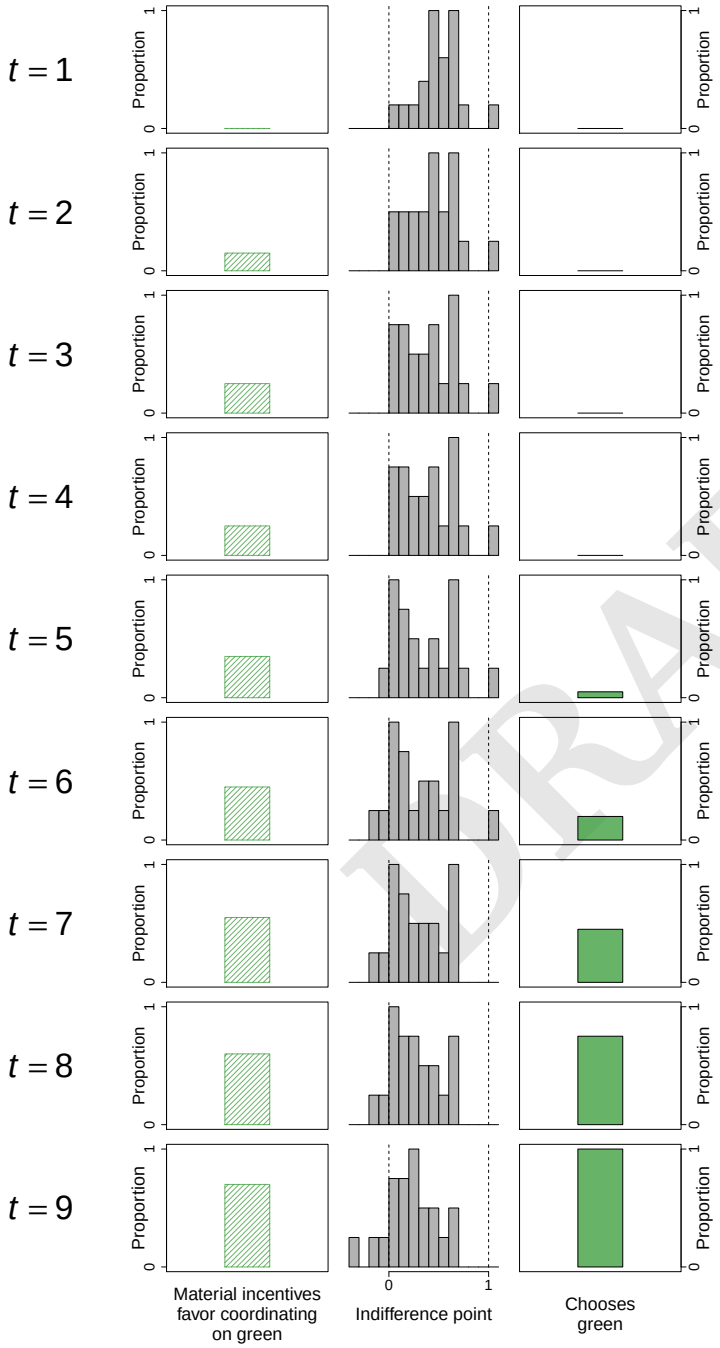

Fig. 1. An example of tipping based on the framework used in Andreoni et al. (3). In $t = 1$, everyone faces material incentives that favor coordinating on blue over coordinating on green (Column 1). The distribution of indifference points is relatively unfavorable for green as a result (Column 2), and everyone chooses blue (Column 3). As time passes, individuals experience changing material incentives. The distribution of thresholds drifts steadily downward, in favor of green, but for a while (e.g. $t \leq 5$) this generates little change in behavior. At some point (e.g. $t \geq 6$), behavior change suddenly accelerates, and the population transitions rapidly to a new norm. Broadly speaking, Andreoni et al. (3) examine when the rapid change in behavior does or does not follow the change in incentives.