SOCIUS

Original Article

Socius: Sociological Research for
a Dynamic World
Volume 8: 1–16
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23780231221124556
srd.sagepub.com

How Norms Emerge from Conventions (and Change)

Wojtek Przepiorka[1], Aron Szekely[2,3], Giulia Andrighetto[3,4,5], Andreas Diekmann[6,7], and Luca Tummolini[3,4]

Abstract

Social norms regulate our behavior in a variety of mundane and far-reaching contexts, from tipping at the restaurant to social distancing during a pandemic. However, how social norms emerge, persist, and change is still poorly understood. Here the authors investigate experimentally whether spontaneously emerging behavioral regularities (i.e., conventions) gain normativity over time and, if so, whether their normative underpinning makes them resistant to changes in economic incentives. To track the coevolution of behavior and normativity, the authors use a set of measures to elicit participants' first- and second-order normative beliefs and their (dis)approval of other participants' behaviors. The authors find that even in the limited duration of their lab experiment, conventions gain normativity that makes these conventions resistant to change, especially if they promote egalitarian outcomes and the change in economic incentives is relatively small. These findings advance our understanding of how cognitive, social and economic mechanisms interact in bringing about social change.

Keywords

convention, social norm, volunteer's dilemma, normative expectations, experimental sociology

Most of our everyday life is governed by social norms, the unwritten rules that make up for much of social order. From choosing the proper way to greet a stranger to judging what counts as a fair proposal in bargaining, social norms are often based on tacit understandings and bear a widespread influence on our decisions. Social norms can be defined as rules governing behavior in social interactions that are sustained by shared expectations of compliance and sanctioning (Bicchieri 2006; Cialdini and Trost 1998; Coleman 1990; Fehr and Schurtenberger 2018; Horne and Mollborn 2020; Pettit 1990).

Because of their role in motivating behavior that may be in conflict with individuals' self-interest, social norms have gained attention as potential solutions to pressing social dilemmas (Nyborg et al. 2016), from supporting the governance of local commons (Ostrom 2000) to containing the spread of pandemic viruses (van Bavel et al. 2020). However, beside their potential to orient individual behavior toward socially beneficial outcomes, social norms can also support and legitimize unpopular, inequitable or even dysfunctional social outcomes. Examples are the maintenance of falsehood (Asch 1951; Willer, Kuwabara, and Macy 2009), persistence of the gender pay-gap (Alesina, Giuliano, and Nunn 2013; Yamaguchi 2019), blood feuds (Grutzpalk 2002), hate speech (Álvarez-Benjume and Winter 2018), compliance with protection rackets (Lipari and Andrighetto 2021), or female genital mutilation (Efferson et al. 2015).

Despite the role of social norms in creating and maintaining social order (Horne and Mollborn 2020), we still have relatively little empirical knowledge about the forces that drive their emergence and change.

One way in which social norms form and spread is through the intentional adaptation and transfer of preexisting normative expectations and sanctioning mechanisms to new social

[1]Utrecht University, Utrecht, the Netherlands
[2]Collegio Carlo Alberto, Turin, Italy
[3]Institute of Cognitive Sciences and Technologies, Italian National Research Council, Rome, Italy
[4]Institute for Futures Studies, Stockholm, Sweden
[5]Mälardalen University, Västerås, Sweden
[6]ETH Zurich, Zurich, Switzerland
[7]Leipzig University, Leipzig, Germany

Corresponding Author:
Wojtek Przepiorka, Utrecht University, Department of Sociology/ICS, Padualaan 14, Utrecht, 3584 CH, the Netherlands
Email: w.przepiorka@uu.nl

contexts. For example, norms can originate from direct communication and explicit bargaining between actors who share similar normative attitudes (Coleman 1990; Horne 2001) or from the deliberative creation of sanctioning institutions that can induce and enforce collectively beneficial behaviors (Diekmann et al. 2014; Fehr and Gächter 2002; Horne 2009; Piskorski and Gorbatâi 2017).

In a different and more fundamental trajectory, social norms can result from spontaneously emerging behavioral regularities that acquire normative force gradually over time (Hawkins, Goodman, and Goldstone 2019). An extensive interdisciplinary literature has explored how conventions, defined as behavioral regularities that serve as arbitrary solutions to coordination problems (Lewis 1969; Sugden 1986), can emerge tacitly as unintended consequences of repeated social interactions (Centola and Baronchelli 2015; Sugden 1986; Young 1993). Importantly, conventions do not require normative force to originate and spread because, once in place, actors have little incentives to deviate from them. Yet it has been repeatedly suggested that well-established behavioral regularities such as conventions can acquire normative force over time (Bicchieri 2006; Lewis 1969; Sugden 1986; Thibaut and Kelley 1959) possibly because increased familiarity with and predictability of a behavior induces actors to perceive a certain behavioral pattern as something that *should* be carried out (Berger and Luckmann 1966; Opp 2004; Sugden 1998; Theriault, Young, and Barrett 2021; Tummolini et al. 2013; Tummolini and Pezzulo 2021; Wrong 1994).

Supporting this view, research in psychology has provided evidence that people tend to infer what is appropriate ("ought") from what is common ("is"), also known as descriptive-to-prescriptive tendency (Roberts, Gelman, and Ho 2017) or common-is-moral association (Eriksson, Strimling, and Coultas 2015). This seems to be a widespread bias that emerges quite early in development; even 2- and 3-year-olds already expect that members of a community know and act in accordance with conventions (Diesendruck and Markson 2011) and are inclined to protest and sanction behaviors violating of conventional, arbitrary rules (Rakoczy, Warneken, and Tomasello 2008). Interestingly, young children seem to perceive even novel actions as normatively prescribed, provided that the adult marks them as examples of a familiar and conventional type (Schmidt, Rakoczy, and Tomasello 2011).

Notwithstanding the pervasiveness of such a link between the descriptive and the normative, this tendency has been so far mostly explored at the individual level as a cognitive bias shaping the attitudes and interpretations of an observer from a third-person perspective (Lindström et al. 2018; Tworek and Cimpian 2016), but without uncovering the consequences of these emerging normative expectations on the alignment of normative views across individuals and at the group level (although see Horne, Tinkler, and Przepiorka 2018).

Building on this literature, we investigate how novel, spontaneously emerging behavioral regularities in small groups not only shape group members' individual expectations about what *will* and *should* happen in a particular situation but also how normative expectations become consistent across group members. As a measure of group consensus, the latter is an important cue to infer that a social norm has formed (Bicchieri 2006; Bicchieri, Lindemans, and Jiang 2014).

The formation of social norms from conventions has important but often overlooked consequences; even socially harmful practices may acquire some level of normative force, which might contribute to make these behaviors more difficult to eradicate through standard approaches such as changes in economic incentives (Bowles and Polanía-Reyes 2012; Gneezy, Meier, and Rey-Biel 2011; Mackie 1996). Despite the high degree of scientific interest and relevant policy implications of this issue, previous studies have approached the effect of emerging normativity on resistance to change only indirectly (Guala and Mittone 2010); a direct empirical test of how social norms emerge from the persistence of social practices is still lacking.

Here we present evidence from a laboratory experiment designed to address two questions: (1) Do conventions that spontaneously emerge in repeated interactions turn into social norms? (2) Do social norms that arise from conventions make behavior resistant to changes in economic incentives?

We use the repeated volunteer's dilemma game (VOD) as the vehicle for the spontaneous emergence of conventions. The VOD is a binary choice, multiperson game in which a single person's act of volunteering is necessary *and* sufficient to produce the collective good for the entire group (Diekmann 1985, 1993). The VOD combines a coordination and a cooperation problem. Although one agent has to bear the cost of producing the collective good, agents have no means to coordinate on the one agent that would do it. As everyone would prefer someone else to volunteer, the collective good may not be produced because of a diffusion of responsibility effect (Przepiorka and Diekmann 2018). However, two series of behavioral experiments conducted in different labs with more than 720 participants have consistently shown that once the VOD is repeated, people are able to find emergent solutions to the dilemma (Diekmann and Przepiorka 2016; Przepiorka, Bouman, and de Kwaadsteniet 2021).

In the symmetric VOD, in which the cost of volunteering is the same for all group members, people are able to tacitly converge on an equitable turn-taking convention, whereas in an asymmetric VOD in which one person has a lower cost of volunteering, an inequitable solitary-volunteering convention emerges. Whereas in turn-taking each person sequentially incurs the cost of volunteering, in solitary volunteering the person with the lowest cost (i.e., the "strong" group member) takes a disproportionately larger share of volunteering. Although both turn-taking and solitary volunteering are social practices found outside of the laboratory, the measures we use in our experiment enable us to identify how conventions develop and turn into norms.

Hypotheses

Regarding our first research question, we expect the conventions that spontaneously emerge from repeated interactions in the experiment to turn into social norms (*normativity hypothesis*). That is, we expect that (1) the consistency among participants' beliefs on both *what ought to be done* and *what others think ought to be done* (i.e., respectively, first-order and second-order normative beliefs) converge gradually over time and that (2) these participants will approve behavior in accordance with the prevailing convention and disapprove of deviations from it. Conventions that spontaneously emerge from repeated interactions can be sustained by participants' beliefs on *what will be done* alone (empirical expectations), without the need for shared expectations on what ought to be done (normative expectations) or systematic patterns of approval of conformity and disapproval of nonconformity. These conventions, however, would not be social norms.

Regarding our second research question, we expect that the stronger the norm underlying a convention is in a group, the longer will it take the group to abandon the convention after a change in economic incentives (the *stickiness hypothesis*).

Finally, research has shown that equitable solutions to social dilemmas exhibit strong focality and are commonly found across different societies (Baumard, André, and Sperber 2013; Engel 2011; Henrich et al. 2005). We therefore expect the turn-taking convention to be stickier (i.e., more resistant to change in economic incentives) than the solitary-volunteering convention (the *asymmetry hypothesis*). However, consistently with previous literature (Gneezy et al. 2011), we also expect that larger changes in economic incentives also induce larger behavioral changes (the *incentive hypothesis*).

Methods

To test our hypotheses, we use a three-part experimental design in which participants interact repeatedly in a VOD with the same group members (Figure 1). Groups are randomly assigned to start in either the symmetric or an asymmetric VOD and then move to the other VOD in the second part. To investigate the effect of economic incentives on behavioral change, we implement two asymmetric VOD conditions: one in which the volunteering costs of the strong group member are 40 percent lower than the volunteering costs of the other group members and one in which these costs are 80 percent lower. To understand if behavioral regularities also acquire normative force, we adopt a set of measures to detect the presence of social norms and their change over time (Andrighetto, Grieco, and Tummolini 2015; Bicchieri et al. 2014; Otten et al. 2020; Szekely et al. 2021). Finally, in the third part, we assess how much approval and disapproval is ascribed to, respectively, conformity with and deviations from the conventions that emerge in the experiment.

To test our hypothesis that conventions beget social norms, we test whether (1) a *convention* emerges from participants' repeated interactions, (2) these participants develop normative and empirical *consensus* around the convention, and (3) adherence to and deviations from the convention elicit *approval* and *disapproval*, respectively. Note that even though conventions can be sustained without shared normative expectations, there is no reason to assume that a convention has to be in place before it can gain normative underpinning; behaviors and normative expectations can coevolve in a feedback process. In the following paragraphs we explain in detail how we measure these constructs. For convenience, Table 1 contains the most important definitions of terms and acronyms used throughout the article.

Conventions

The latent norm index ($\mathrm{LNI}_{k,m}$) captures the proportion of rounds in which a behavioral pattern (i.e., convention) k is observed in a group of size m for at least m consecutive rounds. Because $m = 3$ in our experiment, we restrict the LNI to capturing solitary volunteering ($k = 1$; e.g., AAA...), strictly consecutive turn-taking among the same two group members ($k = 2$; e.g., ABAB...), and strictly consecutive turn-taking among three group members ($k = 3$; e.g., ABCABC...). To avoid capturing random sequences of behavior, the LNI starts capturing these behaviors only if they occur in at least m consecutive rounds in each of which the collective good is produced by one person only. Consider a group of three participants that have interacted in the VOD for 10 rounds in an indiscernible way but then take turns among the three of them for the remaining 40 rounds of a 50-round sequence. In this case $\mathrm{LNI}_{3,3} = 0.8$. If instead, in rounds 11 and 12, person A and person B volunteer, respectively, and person C fails to volunteer in round 13 (or more than one person volunteers), then these three rounds do not count toward the LNI and $\mathrm{LNI}_{3,3} = 0.74$. The LNI does not capture less regular patterns (e.g., ABBABB...), and the pattern ABCCBAABCCBA... would be recorded as $\mathrm{LNI}_{3,3}$ because within the minimum of m rounds, turn-taking among all three group members is strictly consecutive. However, the occurrence of such irregular patterns is rare because coordinating on them tacitly is difficult. Moreover, LNI is a group-level metric and thus records behavioral patterns irrespective of which subsets of group members engage in them. For example, ABABABCACACA would be recorded as 100 percent $\mathrm{LNI}_{2,3}$ (and 25 percent $\mathrm{LNI}_{3,3}$). However, the $\mathrm{LNI}_{1,3}$ (solitary volunteering) reported for the asymmetric VODs in this article refer exclusively to the behavior of the strong player.

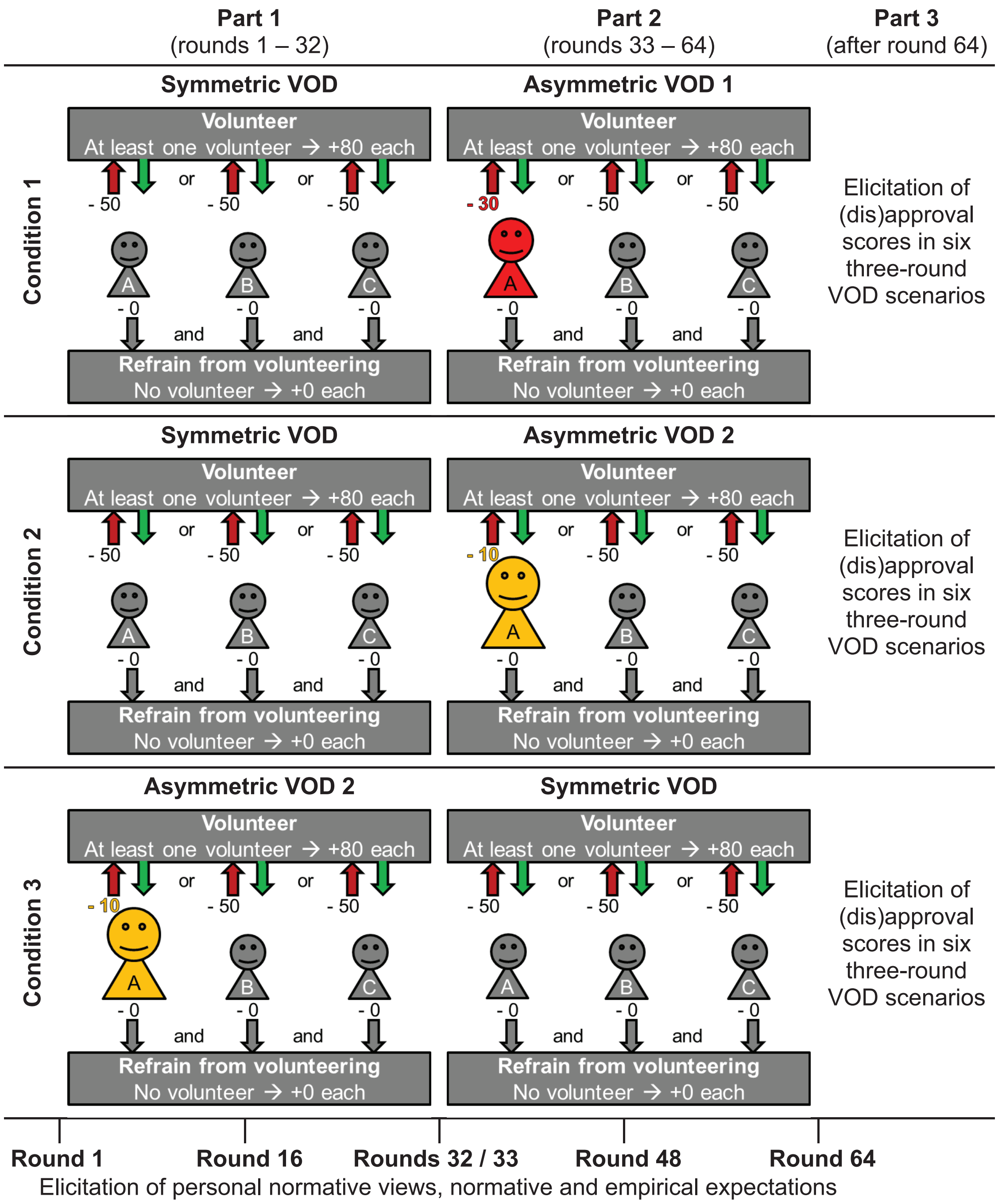


Figure 1. Our three-part experimental design capturing norm emergence and change. First, participants are randomly assigned to one of three experimental conditions. Next, participants are randomly grouped in triads, in which they interact with the same group members in fixed roles (A, B, or C) for 2 × 32 rounds. In part 1 (rounds 1–32), they interact in the symmetric volunteer's dilemma game (VOD) (conditions 1 and 2), in which all members have the same costs from volunteering ($K = 50$), or the asymmetric VOD 2 (condition 3), in which member A has an 80 percent lower cost from volunteering ($K_A = 10$) than members B and C ($K_{B,C} = 50$). In part 2 (rounds 33–64), participants in condition 1 switch to the asymmetric VOD 1, in which member A has a 40 percent lower cost from volunteering ($K_A = 30$) than members B and C ($K_{B,C} = 50$); participants in condition 2 switch to the asymmetric VOD 2, and participants in condition 3 switch to the symmetric VOD. In all conditions and rounds, the benefit from producing the collective good is the same for all group members ($U_{A,B,C} = 80$). At the beginning of rounds 1, 16, 32, 33, 48, and 64, we elicit participants' personal normative views (not incentivized), normative expectations, and empirical expectations (both incentivized) regarding the actions to be taken in the corresponding round. At the end of round 64, in part 3, we present participants individually with six hypothetical three-round, three-person VOD scenarios in a consecutive but random order and let them rate how much they (dis)approve of each group members' actions in these scenarios.

Table 1. Definitions and Acronyms.

VOD (volunteer's dilemma game): a step-level collective good game in which the voluntary and costly act of one player is necessary and sufficient to produce the collective good for all $m \geq 2$ group members.

TT (turn-taking): a behavioral pattern (i.e., convention) in a repeated VOD in which the same $1 < l \leq m$ group members take turns in volunteering after every round while all other $m - l$ group members abstain from volunteering in every round.

SV (solitary volunteering): a behavioral pattern (i.e., convention) in a repeated VOD in which one and the same group member volunteers in every round while all other $m - 1$ group members abstain from volunteering in every round.

LNI (latent norm index): the proportion of rounds in a repeated VOD in which a group of m players coordinates on a specific behavioral pattern (i.e., convention) for at least m consecutive rounds.

NTC (normative tacit consensus): the proportion of group members' overlapping normative expectations (i.e., guesses of other group members' personal normative beliefs) in a particular round of the repeated VOD.

ETC (empirical tacit consensus): the proportion of group members' overlapping guesses of other group members' empirical expectations in a particular round of the repeated VOD.

Consensus

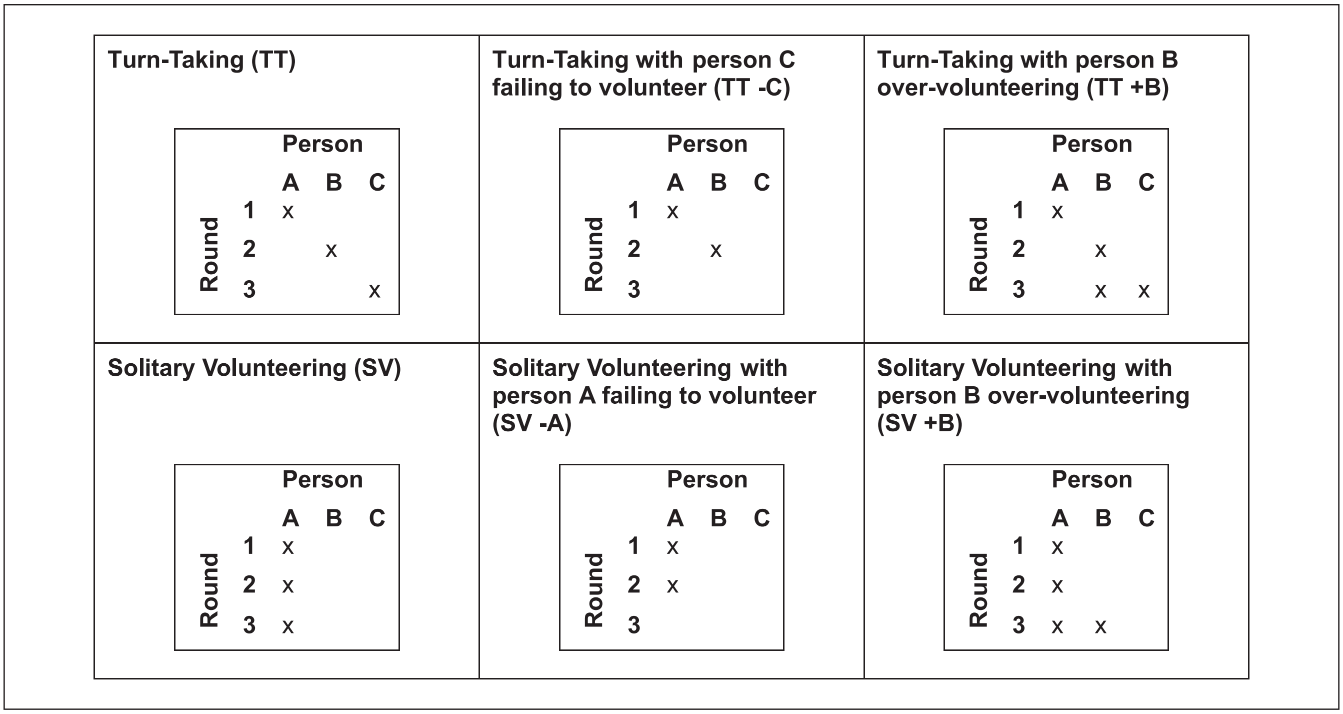
In rounds 1, 16, and 32 (part 1) and rounds 33, 48, and 64 (part 2), we ask participants four questions before they make their decisions in these rounds. First, we ask about participants' normative beliefs: "Who should press 'up' [i.e., volunteer] and who should press 'down' [i.e., not volunteer] in this round?" Participants answer by clicking in the corresponding decision fields of each player (including themselves) in their group (see sample screenshots in the Online Appendix). We also ask participants how confident they are about their answer on a 10-point scale. Participants' personal normative beliefs cannot be compared to an independent measure and therefore their elicitation is not incentivized. Second, we elicit participants' normative expectations. Participants are asked to respond to the question about personal normative beliefs from the perspective of the other two participants in their group. If, for example, person B is the participant responding, they will answer the following two questions: "Suppose you are Person A. Who would Person A think should press 'up' and who should press 'down' in this round?" and "Suppose you are Person C. Who would Person C think should press 'up' and who should press 'down' in this round?" Participants' normative expectations are, in other words, guesses about their two other group members' personal normative beliefs, and participants earn 25c for each correct guess. Each participant can thus earn up to $2 \times 3 \times 25\text{c} = 150\text{c}$ by answering the two questions. Third, we ask for participants' empirical expectations: "Who will press 'up' and who will press 'down' in this round?" Participants' empirical expectations are, in other words, guesses about their two other group members' behavior (whether they will choose "up" or "down" in this round). Participants earn 25c for each correct guess and can thus earn up to $2 \times 25\text{c} = 50\text{c}$ by answering the question.

According to an influential view (Bicchieri 2006; Bicchieri et al. 2014), the formation of a social norm is reflected in the mutual consistency between each member's personal normative beliefs and what each believes that others expect from them. Hence, from participants' normative expectations, we calculate each group's normative tacit consensus (NTC) in the corresponding round. NTC is the proportion of overlapping guesses of other group members' personal normative beliefs. The total number of overlapping guesses is $3 + 3 + 3 = 9$ (i.e., up to three overlaps per dyad about the third group member's personal normative beliefs). The number of overlapping guesses is then divided by 9 to obtain NTC. From participants' empirical expectations, we calculate each group's empirical tacit consensus (ETC). ETC is the proportion of overlapping guesses of the three group members' empirical expectations. The total number of overlapping guesses is $1 + 1 + 1 = 3$ (i.e., up to one overlap per dyad about what the third group member will do). The number of overlapping guesses is then divided by 3 to obtain ETC.

(Dis)approval

After round 64 (in part 3), participants are presented with six scenarios displaying how the three members of another, hypothetical group behaved in the experiment and they are asked to rate, on a five-point scale, how appropriately each of these group members behaved. Figure 2 shows a schematic representation of the six scenarios. The scenarios are presented to participants sequentially in a random order. A scenario comprises three consecutive rounds and the decisions each group member made in these rounds (see sample screenshots in Online Appendix). Three of the six scenarios are related to turn-taking and the other three to solitary volunteering. The same six scenarios are presented to all participants. That is, the solitary-volunteering scenarios were also presented to participants who, in part 2 of the experiment, interacted in the symmetric VOD, and the turn-taking scenarios were also presented to participants who interacted in an asymmetric VOD in part 2.

The turn-taking (TT) and solitary volunteering (SV) scenarios show norm-following behavior: turn-taking among all group members and solitary volunteering by person A (the "strong" group member), respectively. Turn-taking with person C failing to volunteer (TT –C) and solitary volunteering with person A failing to volunteer (SV –A) display norm-breaking behavior through which the collective good is not produced in one round. Turn-taking with person B volunteering in excess (TT +B) and solitary volunteering with person B volunteering in excess (SV +B) show norm-breaking behavior that is inconsistent with the respective behavioral



Turn-Taking (TT)

Round	A	B	C
1	x		
2		x	
3			x

Turn-Taking with person C failing to volunteer (TT -C)

Round	A	B	C
1	x		
2		x	
3			

Turn-Taking with person B over-volunteering (TT +B)

Round	A	B	C
1	x		
2		x	
3		x	x

Solitary Volunteering (SV)

Round	A	B	C
1	x		
2	x		
3	x		

Solitary Volunteering with person A failing to volunteer (SV -A)

Round	A	B	C
1	x		
2	x		
3			

Solitary Volunteering with person B over-volunteering (SV +B)

Round	A	B	C
1	x		
2	x		
3	x	x	

Figure 2. Schematic representation of behavioral scenarios used for elicitation of disapproval scores. The first column displays behavior that is in line with the turn-taking (first row) and solitary volunteering (second row). Columns two and three show deviations from these patterns.

pattern but does not lead to an underproduction of the collective good.

Procedures

We conducted five sessions at the Decision Science Laboratory of ETH Zurich, with 36 participants in each session (60 participants per experimental condition). Hence, 180 participants made 64 decisions resulting in 11,520 decisions in the VOD in total. Sessions lasted approximately 105 minutes, and participants earned on average 48 Swiss francs in the experiment. Participants were 22.83 years old on average (*SD* = 3.19 years), 56 percent were men (101 of 180), and 98 percent were students (176 of 180). Participants were recruited via e-mail from the participant pool of the lab. The experiment was programmed and run in z-Tree (Fischbacher 2007). Instructions were presented to participants on paper and supplemented by instructions on their computer screens. As participants were allocated to experimental conditions within sessions, the instructions were not read aloud, but the experimenter was available to answer questions. Before participating in the experiment, participants answered control questions to ensure they understood the instructions (see experimental instructions in the Online Appendix). Our experimental protocol was reviewed and approved by the ethics review board of ETH Zurich.

Data Analysis Strategy

All test statistics are based on regression model estimations. Statistical significance is set at the 5 percent level (i.e., $\alpha = .05$) for two-sided tests, and we account for the repeated measures obtained on the same groups by estimating cluster-robust standard errors, where every group of three participants forms a cluster (clustering at participant level was used in the analysis of disapproval scores). To test the statistical significance of the differences between regression coefficients, we use linear combinations of these coefficients after estimation. Upon publication, the data and code with which the results can be reproduced will be made available in a public repository.

Results

Normativity Hypothesis

Figure 3 shows average LNIs at six different time points across experimental conditions (the behavioral patterns that emerged in each group in our experiment are shown in the Online Appendix). By inspection, it is clear that in the two experimental conditions that start with the symmetric VOD (conditions 1 and 2), groups predominantly converge on the behavioral regularity of taking turns among all three group members in the creation of the collective good. These groups reach an average $LNI_{3,3}$ of 59 percent in condition 1

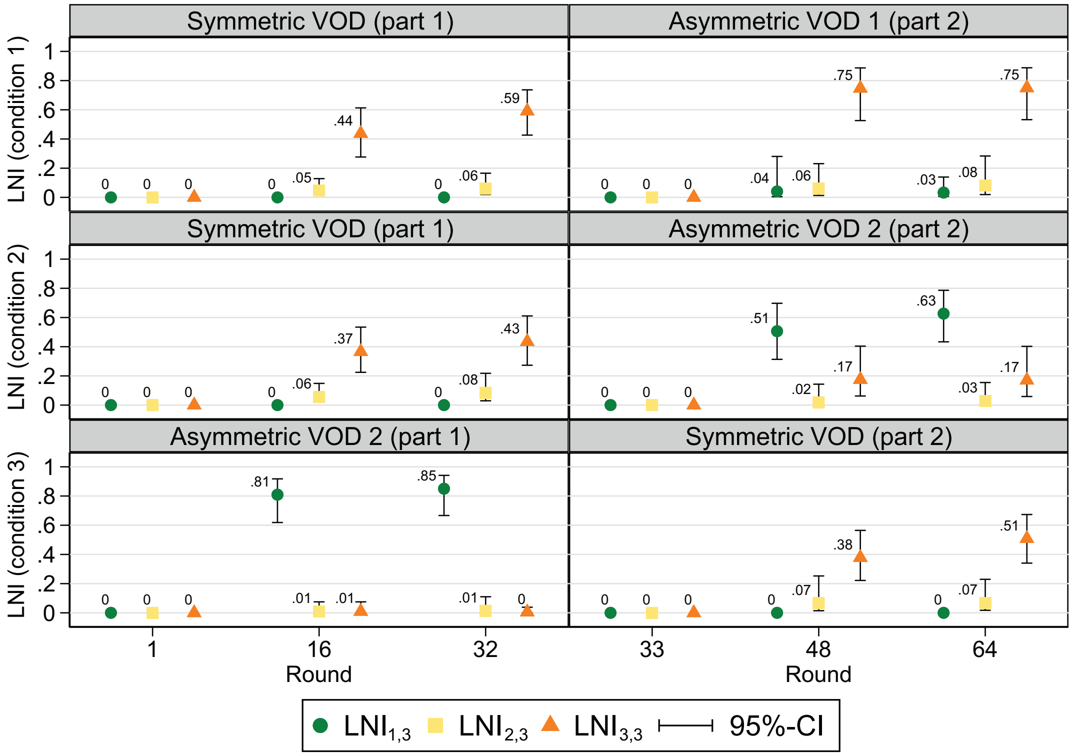


Figure 3. Cumulative proportions of three types of conventions across experimental conditions and time. The figure shows point estimates from ordinary least squares regression models and logit-transformed 95 percent confidence intervals (CIs), which always lie between 0 and 1. In the symmetric VOD condition, in both parts of the experiment, turn-taking among all three group members (LNI$_{3,3}$) emerges as the main convention. In the asymmetric VOD 2 condition, also in both parts, solitary volunteering by the "strong" group member (LNI$_{1,3}$) emerges as the main convention. In the asymmetric VOD 1 condition, which follows the symmetric VOD condition in part 2, turn-taking among all three group members remains the main convention. This indicates that the equitable turn-taking convention is more resistant to change (i.e., "sticky") if economic incentives are too small. Turn-taking among two of the three group members (LNI$_{2,3}$) hardly emerges in any of the conditions.
Note: LNI = latent norm index.

and 43 percent in condition 2 by the end of the first part. The differences in LNI$_{3,3}$ across these two conditions are statistically insignificant (round 16: $b = 0.07$, $t = 0.64$, $p = .524$; round 32: $b = 0.16$, $t = 1.39$, $p = .170$). In contrast, in the experimental condition that starts with the asymmetric VOD 2 (condition 3), we observe an average LNI$_{1,3}$ of 85 percent by the end of part 1 whereby the collective good is repeatedly created by the same volunteer. Taken together these results demonstrate that we were able to induce the emergence of different conventions contingent on VOD payoffs alone. However, although LNI captures the behavioral component of social norm emergence in the repeated VOD (Guala and Mittone 2010), it is not sufficient to make the case that a social norm exists; there must also be consensus among participants that the convention *should* be followed.

Consensus is measured by means of NTC and ETC. NTC measures a group's consensus on what its members think their fellow members think *should* be done in a particular round (i.e., consensus in normative expectations), and ETC measures a group's consensus on what its members think *will* be done in that round (i.e., consensus in empirical expectations). Both variables can assume values between 0 and 1; values above 0.5 indicate agreement and values below 0.5 indicate disagreement. Only NTC and ETC values above 0.5 together are indicative of a social norm; an ETC value above 0.5 alone would merely be indicative of a descriptive norm (Bicchieri et al. 2014). In Figure 4, we therefore present all groups' NTCs and ETCs in a scatterplot, along with their mean values, across conditions and measurement points.

Figure 4 shows that on average, groups agree on what they think their fellow group members think should be done and what they think will be done in a particular round, and these agreements increase over time. There is also heterogeneity in group consensus, in particular at the first measurement point, and more in ETC than in NTC. However, in both conditions both NTC and ETC increase together, indicating that the prevailing behavioral regularity gains normativity over time. This is first evidence that the conventions that emerge in the first part of our experiment also gain normative

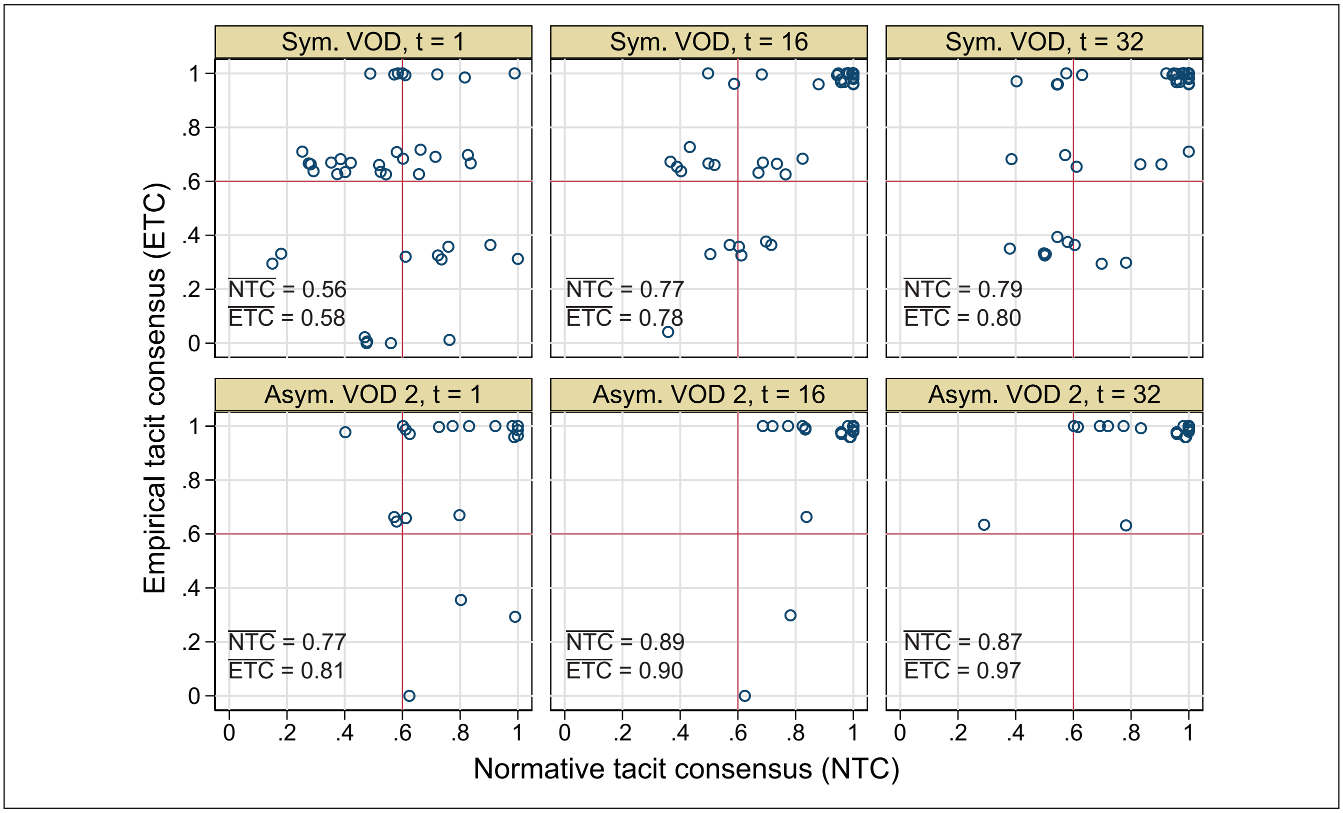


Figure 4. Scatterplots relating normative tacit consensus (NTC) (*x*-axes) and empirical tacit consensus (ETC) (*y*-axes) values. Values shown for each group at the three measurement points in the symmetric volunteer's dilemma game (VOD) conditions (upper row panels; $n = 40$) and asymmetric VOD 2 condition (lower row panels; $n = 20$) in part 1 of the experiment. At all measurement points, mean NTC and ETC lie above 0.5 and are (almost) monotonically increasing over time in both conditions. The upper right quadrant delineated by the red grid lines in each panel is the region in which groups have both NTC and ETC values above 0.6 indicating that these groups have gained substantial normativity on their behavior.

underpinning. Comparing the average NTC and ETC values across the two conventions, it seems that the turn-taking convention takes longer to turn into a social norm than the solitary volunteering convention, which gains high normative force already from the first round. We come back to this point in the part of the article in which we test the asymmetry hypothesis. However, so far we have treated behavior and expectations separately. To show that a certain convention pertains to a certain social norm, we have created two additional measures of accuracy of expectations predicting behavior. Results show that group members' normative and empirical expectations both predict their fellow group members' behavior with high accuracy that increases over time (see Online Appendix, Table S1). Additionally, individual group members' normative and empirical expectations are correlated and these correlations increase over time (see Online Appendix, Table S2)

(Dis)approval

Figure 5 shows average (dis)approval scores pertaining to each of the three persons in the six hypothetical three-round VOD scenarios (also see Figure 2). We present results according to the prevailing behavioral pattern that emerged in part 2 of each treatment. Specifically, turn-taking ($LNI_{3,3}$) dominated at the end of part 2 in the symmetric VOD and asymmetric VOD 1 conditions, so we combine these treatments, whereas solitary volunteering ($LNI_{1,3}$) emerged as the main behavior in the asymmetric VOD 2 condition (see Figure 3). The results separated for each treatment can be found in the Online Appendix (Figure S27) and are entirely consistent with what we display here.

Our expectation is that if turn-taking has become a social norm, in addition to group consensus on normative and empirical expectations, participants should also be willing to approve of compliance with and disapprove of deviations from turn-taking. We expect the corresponding logic to apply if solitary volunteering has turned into a social norm.

Participants from the symmetric and asymmetric VOD 1 treatments (Figure 5, first row) strongly approve of the behavior displayed in the TT scenario ($M = 4.84$, $SD = 0.60$). In TT –C, they strongly approve of person A ($M = 4.57$, $SD = 0.97$) and person B ($M = 4.73$, $SD = 0.70$), who follow turn-taking, and strongly disapprove of person C's inaction ($M = 1.46$,

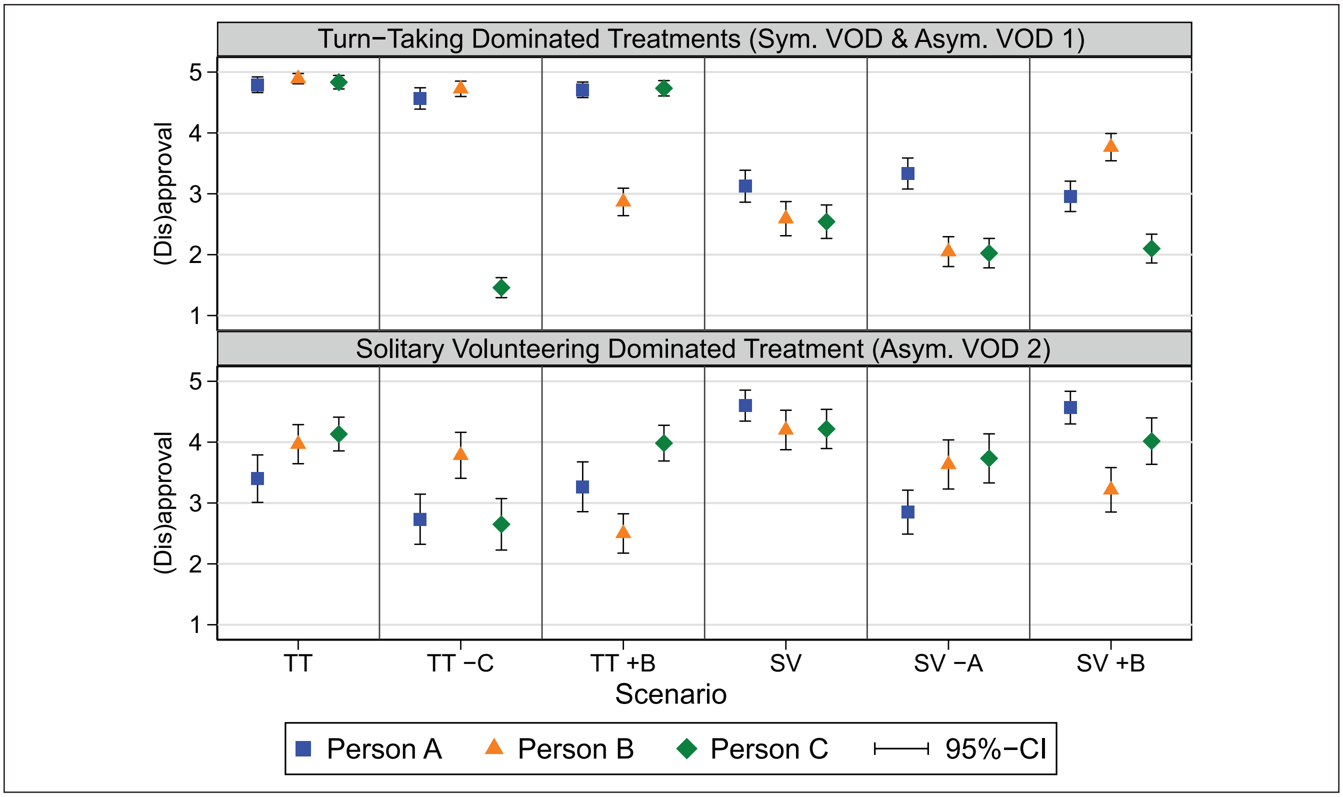


Figure 5. Average (dis)approval scores. The figure shows point estimates from ordinary least squares regression models and 95 percent confidence intervals (CIs). Participants were presented six scenarios, each of which showed a sequence of three volunteer's dilemma game (VOD) rounds that another group may have produced: turn-taking (TT), turn-taking with person C failing to volunteer (TT –C), turn-taking with person B volunteering in excess (TT +B), solitary volunteering by person A (SV), solitary volunteering with person A failing to volunteer once (SV –A), and solitary volunteering with person B volunteering in excess (SV +B). (Dis)approval ratings pertained to each of three persons in the scenarios and went from 1 ("totally disapprove") to 5 ("totally approve"). The VOD payoffs in the scenarios corresponded the VOD payoffs that participants faced in the second part of the experiment (see Figure 1).

SD = 0.91; vs. person A, *b* = 3.11, *p* < .001; vs. person B, *b* = 3.27, *p* < .001). In TT +B, they approve of person A (*M* = 4.71, *SD* = 0.70) and person C (*M* = 4.73, *SD* = 0.69), who follow turn-taking, and disapprove of person B's overvolunteering (*M* = 2.87, *SD* = 1.24; vs. person A, *b* = 1.84, *p* < .001; vs. person C, *b* = 1.87, *p* < .001). In the SV scenario, these participants had intermediate approval of behavior across the three people (*M* = 2.75, *SD* = 1.52). In SV –A, they disapprove of the nonvolunteers, person B (*M* = 2.05, *SD* = 1.35) and person C (*M* = 2.03, *SD* = 1.33), and actually approve, in relative terms, of person A, who fails to volunteer once (*M* = 3.33, *SD* = 1.40; vs. person B, *b* = 1.28, *p* < .001; vs. person C, *b* = 1.30, *p* < .001). In SV +B, they approve the most of person B (*M* = 3.77, *SD* = 1.23), who volunteers in excess, an intermediate amount of person A (*M* = 2.96, *SD* = 1.37; vs. person B, *b* = –0.81, *p* < .001), and least of person C (*M* = 2.10, *SD* = 1.31; vs. person A, *b* = –0.86, *p* < .001), who never volunteers.

Participants from the asymmetric VOD 2 treatment (Figure 5, second row) strongly approve of all persons in SV (*M* = 4.34, *SD* = 1.17) and approve of person B (*M* = 3.63, *SD* = 1.55) and person C (*M* = 3.73, *SD* = 1.55) in SV –A but clearly disapprove of person A, who fails to volunteer (*M* = 2.85, *SD* = 1.39; vs. person B, *b* = –0.78, *p* = .014; vs. person C, *b* = –0.88, *p* = .004) according to the pattern. In SV +B, they approve of person A (*M* = 4.57, *SD* = 1.03) and person C (*M* = 4.02, *SD* = 1.47) and disapprove of the overvolunteering person B (*M* = 3.22, *SD* = 1.40; vs. person A, *b* = –1.35, *p* < .001; vs. person C, *b* = –0.80, *p* = 0.009). In the TT scenario, these participants generally approve of all persons' actions (*M* = 3.83, *SD* = 1.31) but do so to a lesser extent than participants coming from the other two treatments. In TT –C, they approve of person B (*M* = 3.78, *SD* = 1.45) and disapprove of both person A (*M* = 2.73, *SD* = 1.58; vs. person B, *b* = –1.05, *p* < .001) and person C (*M* = 2.65, *SD* = 1.62; *b* = –1.13, *p* = .002). In TT +B, these participants approve of person C's behavior (*M* = 3.98, *SD* = 1.13), have an intermediate approval of person A (*M* = 3.27, *SD* = 1.57), and have the lowest approval for person B, who overvolunteers

(M = 2.50, SD = 1.24; vs. person A, b = –0.77, p = .004; vs. person C, b = –1.48, p < .001).

In summary, we find clear (dis)approval results. Across all scenarios that match the dominant pattern of a treatment (i.e., TT, TT –C and TT +B for symmetric and asymmetric VOD 1, and SV, SV –A and SV +B for asymmetric VOD 2), there is strongly targeted approval for norm followers and disapproval for norm violators. For the scenarios that did not match the dominant pattern of a treatment (i.e., SV, SV –A and SV +B for symmetric and asymmetric VOD 1, and TT, TT –C and TT +B for asymmetric VOD 2), there was more intermediate approval and disapproval indicating less of a normative response to those mismatched patterns. Overall, these findings provide additional support for our first hypothesis that the conventions that emerge in our experiment turn into social norms. Three further findings worth highlighting can be discerned in Figure 5.

First, deviations from turn-taking are disapproved of more than deviations from solitary volunteering. This is most cleanly revealed when comparing the approval of person C in TT –C (M = 1.46, SD = 0.91) in the first row in Figure 5 with the approval of person A in SV –A (M = 2.85, SD = 1.39) in the second row (Δb = –1.39, p < .001). There is also some evidence when comparing the approval of person B in TT +B (M = 2.87, SD = 1.24) in the first row with the approval of person B in SV +B (M = 3.22, SD = 1.40) in the second row (Δb = –0.35, p = .104). Unlike for our measures of consensus (i.e., NTC and ETC), in terms of (dis)approval scores, the normative underpinning of turn-taking is stronger than that of solitary volunteering.

Second, participants disapprove of failures to volunteer more than of overvolunteering when turn-taking is the convention. This can be seen by comparing the approval of person C in TT –C (M = 1.46, SD = 0.91) and approval of person B in TT +B (M = 2.87, SD = 1.24) in the first row in Figure 5 (Δb = 1.40, p < .001), as well as the approval of person A in SV –A (M = 2.85, SD = 1.39) and the approval of person B in SV +B (M = 3.22, SD = 1.40) in the second row (Δb = –0.37, p = .055).

Third, people from the turn-taking dominant treatments (Figure 5, first row) approve of deviations from solitary volunteering. They approve of person A's failure to volunteer in SV –A (M = 3.33, SD = 1.40) more than they approve of person B's (M = 2.05, SD = 1.35; vs. person A, b = 1.28, p < .001) and person C's (M = 2.03, SD = 1.33; vs. person A, b = 1.31, p < .001) inactions. Moreover, they approve of person B's overvolunteering in SV +B (M = 3.77, SD = 1.23) to a greater extent than either person A's actions (M = 2.96, SD = 1.37; vs. person B, b = –0.81, p < .001) or person C's actions (M = 2.10, SD = 1.31; vs. person B, b = –1.67, p < .001). We do not find a similar effect among participants for whom solitary volunteering emerged as main convention (Figure 5, second row); they do not approve of deviations from turn-taking by person C in TT –C (M = 2.65, SD = 1.62; vs. person A, b = –0.08, p = .818; vs. person B, b = –1.13, p = .002) and, on the contrary, disapprove of person B's overvolunteering in TT +B (M = 2.5, SD = 1.24; vs. person A, b = –0.77, p = .004; vs. person B, b = –1.48, p < .001). This asymmetry suggests that turn-taking is more projectible than single volunteering because participants are more willing to interpret (and normatively assess) a different behavioral pattern in light of their dominant convention, thereby implying again a stronger normativity of turn-taking.

Stickiness Hypothesis

According to our second hypothesis, the higher the degree of normativity a group exhibits with regard to a convention, the longer it will take the group to abandon that convention when experiencing a change in economic incentives. To test this hypothesis, we take the following steps. For each group, we (1) determine the primary convention (among $\text{LNI}_{1,3}$, $\text{LNI}_{2,3}$, and $\text{LNI}_{3,3}$) at round 32 and calculate the LNI score associated with that convention in part 1, (2) determine the LNI score of the same convention in part 2, (3) calculate the change in LNI that occurs from part 1 to part 2, and (4) use period 32 NTC and ETC to predict the change in LNI calculated in step 3. The bivariate associations by condition are shown in Figure 6. We expected that groups with high NTC and ETC maintain their dominant LNI, while groups with lower NTC and ETC should abandon their dominant LNI. This implies that positive associations between period 32 NTC or ETC and change in dominant LNI would indicate support for our second hypothesis.

Contrary to our expectations, we find that NTC and ETC do not predict change in main LNI in conditions 1 (symmetric VOD to asymmetric VOD 1) and 2 (symmetric VOD to asymmetric VOD 2). Although they do predict change in LNI in condition 3 (asymmetric VOD 2 to symmetric VOD), the relationship is opposite to what we expected (Figure 6). We find substantially the same result using NTC and ETC averaged across period 16 and 32, using measures of accuracy, or confidence in empirical and normative expectations (see Online Appendix, Figures S28–S30). This implies that our consensus measures, which do strengthen in groups over time (Figure 4), do not explain group-level variation in behavioral stickiness.

We believe the reasons for this finding to be twofold. Either the change in economic incentives is insufficiently strong, in which case behavior and expectations continue along their previous path, or the change is too strong, in which case incentives overrule any stickiness in terms of behavior. Condition 1 (symmetric VOD to asymmetric VOD 1) falls under the first category. As such, consensus cannot increase any more among high consensus groups and only a little in lower consensus groups leading to flat associations. Conditions 2 (symmetric VOD to asymmetric VOD 2) and 3 (asymmetric VOD 2 to symmetric VOD), conversely, fall under the second category. The change in economic incentives is so sharp that most groups abandon their convention not leaving much leverage for consensus to affect behavior.

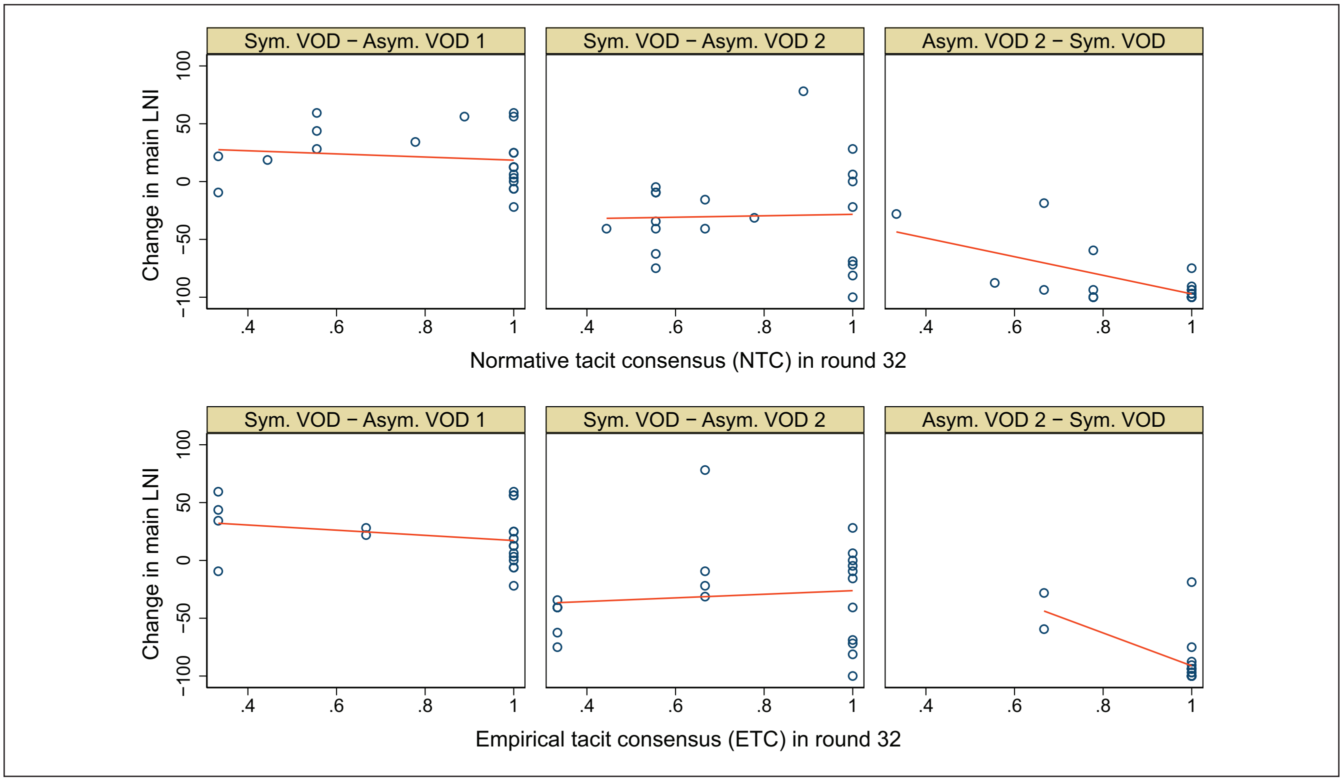


Figure 6. Change in main convention contingent on experimental condition, normative tacit consensus (NTC) and empirical tacit consensus (ETC). The main convention is determined for each group by the mode latent norm index (LNI). In part 1 of the experiment, most groups' main convention was $LNI_{3,3}$ in conditions 1 and 2 and $LNI_{1,3}$ in condition 3 (Figure 3). The change in the main convention is the difference in part 2 LNI and part 1 LNI of a group's main convention in part 1. Slope coefficients showing the relation between change in the main convention and NTC or ETC are based on simple ordinary least squares regressions at the group level (n = 20 per condition).

Asymmetry Hypothesis

According to our third hypothesis, we expect turn-taking to be more resistant to changes in economic incentives than solitary volunteering. To test this hypothesis, we compare the changes in corresponding LNIs from part 1 to part 2 in conditions 2 and 3. In condition 2, groups switch from the symmetric VOD to the asymmetric VOD 2 after round 32. As from round 33, the strong group member has an 80 percent lower cost of volunteering than the other two group members. In condition 3, it is the other way around (Figure 1).

We observe a clear change in behavior due to the change in payoffs in both conditions (Figure 3). In condition 2, turn-taking reaches an LNI of 43 percent at the end of part 1 and is replaced by solitary volunteering, which reaches an LNI of 63 percent at the end of part 2. In condition 3, solitary volunteering reaches an LNI of 85 percent at the end of part 1 and is replaced by turn-taking with an LNI of 51 percent at the end of part 2. However, whereas in part 2 of condition 2 some groups continue taking turns and turn-taking reaches an average LNI of 17 percent by the end of part 2, solitary volunteering ceases to exist entirely in condition 3. Consistent with our asymmetry hypothesis, the difference in differences is statistically significant ([condition 2: $LNI_{3,3}$ in part 2 – $LNI_{3,3}$ in part 1 = –0.26] – [condition 3: $LNI_{1,3}$ in part 2 – $LNI_{1,3}$ in part 1 = –0.85] = 0.59, t = 5.78, $p < .001$). How can we explain this difference in stability between turn-taking and solitary volunteering?

One possibility is that normative (and empirical) consensus is stronger for groups engaged in turn-taking than for groups engaged in solitary volunteering, and this provides the extra stability. Yet NTC and ETC are both higher for solitary volunteering than they are for turn-taking (Figure 4). However, comparisons of NTC and ETC across different conventions may be less meaningful because solitary volunteering and turn-taking have different "complexity." Even assuming that both conventions are uniquely salient in their respective environments, the alternation mechanism of turn-taking requires participants to first coordinate on the order of volunteers (Przepiorka and Diekmann 2018). Variation in groups' abilities to solve this problem will be reflected in both NTC and ETC. In line with this conjecture, participants report lower confidence levels regarding their normative and empirical expectations in rounds 1 and 16 in the two symmetric VOD conditions than in the asymmetric VOD 2 condition (see the Online Appendix for details). Another possibility is that turn-taking is more approved of (and deviation from turn-taking more disapproved of) than solitary volunteering. We find support for this conjecture in the results presented in Figure 5 (see above). The normative underpinning of turn-taking is indeed stronger in terms of (dis) approval scores than that of solitary volunteering.

Incentive Hypothesis

Finally, by comparing the change in behavioral regularities that occur in conditions 1 and 2, we can test our fourth hypothesis that the breakdown of turn-taking is slower if the change in economic incentives is smaller. We find clear support for our hypothesis. In condition 1, turn-taking does not break down at all in part 2 but instead increases (condition 1: $\mathrm{LNI}_{3,3}$ part 2 – $\mathrm{LNI}_{3,3}$ part 1 = 0.16, t = 2.95, p = .005), whereas in condition 2, turn-taking decreases to a large extent (condition 2: $\mathrm{LNI}_{3,3}$ part 2 – $\mathrm{LNI}_{3,3}$ part 1 = –0.26, t = –3.32, p = .002). The difference in differences is statistically significant (b = 0.42, t = 4.41, p < .001).

Discussion

Theoretical analyses, simulation studies, and experimental work have explored how actors tacitly coordinate behavior in repeated interactions leading to the spontaneous emergence of conventions (Hawkins et al. 2019). Conventions emerge in response to coordination problems, in which groups of actors would like to select one of multiple Nash equilibria even if, as a result, payoffs are unequally distributed among them. It is well known that actors' recurrent experience with such situations of strategic interdependence, under some conditions, will give rise to behavioral regularities such as that of turn-taking. This spontaneous convergence on a behavioral regularity has been shown in experiments not only with adult participants but also with children and nonhuman primates (Centola and Baronchelli 2015; Grueneisen and Tomasello 2017; Helbing et al. 2005; Leo 2017; Martin, Biro, and Matsuzawa 2017; Melis et al. 2016). Our study contributes to this literature by addressing two questions: (1) Do spontaneously emerging conventions turn also into social norms, and if so, (2) are these social norms stable in the face of changes in economic incentives? By turning into norms, conventions become "sticky" and only slowly adapt to new environments. This adjustment process will depend not only on how pronounced a change in incentives is but also on the specific content of the convention in question.

The question of the normativity of conventions has been raised in several disciplines (Berger and Luckmann 1966; Lewis 1969; Thibaut and Kelley 1959; Tummolini et al. 2013; Wrong 1994) but rarely empirically examined (Guala and Mittone 2010; Opp 2004). In this study, we have provided what is, to our knowledge, the first empirical evidence that conventions may turn into social norms and, at the same time, how norms contribute to the persistence of behavioral patterns. In particular, we have tested the normativity of conventions hypothesis with state-of-the-art measures of personal normative beliefs of "what ought to be done," normative expectations of "what others think ought to be done," empirical expectations of "what will be done" and approval and disapproval judgments. Moreover, most previous studies on coordination games focused on symmetric situations of complete preference alignment which give rise to conventional regularities such as driving on the left or right side of the road (Guala and Mittone 2010). In contrast, we have focused on impure coordination situations, and in particular on the symmetric and the asymmetric VOD, because these games have been shown to generate different behavioral regularities as a result of small differences in economic incentives (Diekmann and Przepiorka 2016; Przepiorka et al. 2021). In the VOD, a single actor is necessary and sufficient to produce the collective good to the benefit of the group. Although a repeated symmetric VOD is conducive to the emergence of turn-taking, the asymmetric VOD mostly leads to a single actor's persistent volunteering (i.e., solitary volunteering).

To test how conventions emerge spontaneously and turn into social norms under varying incentive regimes, we devised a novel experimental design consisting of three conditions. In condition 1, a series of repeated interactions in the symmetric VOD was followed by interactions in a mildly asymmetric game. Condition 2 increased the degree of asymmetry in the VOD compared to condition 1. Condition 3 reversed the order of the two game sequences in condition 2 by starting with the more asymmetric game followed by the symmetric game. Groups' consensuses on normative and empirical expectations were measured in the course of the experiment to capture the coevolution of conventions, norms and sanctioning intentions in terms of (dis)approval scores that were elicited at the end. This experimental setting has allowed us to examine whether different incentive regimes lead to the emergence of different types of conventions then turning into different types of social norms.

Our results provide clear support for the normativity of conventions hypothesis (Lewis 1969) and a nuanced picture of how the normative underpinning of conventions depends on economic incentives that shape behavioral regularities. In our experiment, groups tacitly achieve consensus not only on what "will" be done but also on what group members think "should" be done, and consensus increases over time in both groups with turn-taking and solitary volunteering conventions. We further observe a clear pattern of approval and disapproval of, respectively, adherence to and deviations from both types of conventions. Disapproval judgments, however, are more pronounced following a failure to volunteer than volunteering in excess, especially with regard to turn-taking. Taken together these results indicate that both conventions turn into social norms but the normative underpinning of turn-taking is stronger than that of solitary volunteering.

There are good reasons for why the normative underpinning of turn-taking is stronger than of solitary volunteering. Turn-taking is a Nash equilibrium in the repeated, symmetric VOD that fulfills the criteria of Pareto efficiency and equality; turn-taking is thus in line with universal moral principles (Baumard and Sperber 2013; Engel 2011; Henrich et al. 2005). Anthropologists and sociologists have long emphasized the importance of norms of reciprocity, which also imply turn-taking. In a classic work, Gouldner (1960) described the "norm of reciprocity" and the "shadow of indebtedness" that motivates reciprocal action (see Diekmann 2004). Social psychologists Thibaut and Kelley (1959) predicted the emergence of turn-taking norms in the conflict situation named "battle of sex." Psychological studies find strong evidence that children learn turn-taking from the age of five (Melis et al. 2016). Overall, research shows that turn-taking can be considered a fundamental principle of human cooperation in the accomplishment of tasks. It is therefore no surprise that turn-taking has a strong normative anchoring. In contrast, solitary volunteering, although Pareto efficient, generates inequitable outcomes. The single volunteer alone contributes to the collective good and persistently pays a price receiving less than their fellow group members. The strong normativity of turn-taking also explains the phenomenon of the inertia of the norm when faced with changes in the environment.

Although we have shown that conventions may become normative to a different extent, and in ways that may affect the persistence of the underlying social practices, our study cannot answer the question whether there are, or not, conventional regularities that entirely lack such normative dimension. Under what conditions do conventional regularities *fail to* turn into social norms? Linguistic conventions, for instance, that establish an arbitrary connection between a meaning and the label that is used to communicate it are readily recognized as arbitrary by speakers (Hawkins et al. 2019; Lewis 1969) and may amount to a prototypical convention which does not turn into a social norm. Although these conventions may still persist in being followed even when a better alternative is present (Brennan and Clark 1996), this resistance to change may be due more to the cognitive costs of creating a new empirical consensus than to the pull of the old normative consensus. Still, the never-ending debate between descriptivists and prescriptivists in linguistics suggests that there are intuitions about the "right" ways in word usage that we readily pick up and are prepared to enforce (Curzan 2014). More specifically, we suggest that it could be possible that conventional regularities (even linguistic ones), insofar they are widely practiced, are always associated with some level of normativity, however weak. The intrinsic link between predictability and the "sense of should" (Theriault et al. 2021) and the ubiquity of the corresponding descriptive-to-prescriptive tendency (Rogers et al. 2017) suggest that all empirical regularities which become well established and predominant will inevitably also become prescriptive. This might be true not only of arbitrary ones but also of social practices that over time turn out to be socially harmful: honor norms prescribing violent vendettas persist in contemporary societies despite the change of the relevant economic conditions (Nisbett and Cohen 1996). Still, although manipulating incentives in symmetric and asymmetric versions of the VODs successfully varies which behavioral regularities we observe and lets us explore their differential normative strength, all groups in our treatments always face a sequence of cooperation problems whose solution ultimately benefits all the involved parties. It is therefore possible that a baseline level of "exogenous" normativity that participants bring into the lab from the outside is always present. To identify the lower bounds of conventional normativity a different paradigm in which conventional regularities persist in contexts in which they lack any useful social role or are even harmful would be needed (e.g., Abbink et al. 2017).

Conclusions

Social norms regulate our living together in an informal and often subtle way. The ways we dress, great, talk, dine, lead, help, and so on, depend on what we think others expect us to do in a situation. Although we have known that our perceptions of others expectations drive our behaviors in daily life, we understand less how our collective agreement on these expectations comes about. Our research corroborates the validity of a long-standing theoretical idea, namely, that the properties of the situations in which we meet and interact with one another repeatedly, shape behavioral regularities to which we attach the same sense of oughtness over time (Berger and Luckmann 1966; Lewis 1969). Mapping these processes of norm emergence and change advances our understanding of the roles we assume vis-à-vis others, across social contexts and time. The idea that social norms lag behind a change in structural conditions goes back to Thorstein Veblen ([1899] 1994) and William Ogburn (1922). Ogburn proposed the "cultural lag hypothesis," implying that social norms, laws, and institutions often lag behind material change. Interestingly, he chose the problem of overexploitation of natural resources in forestry as an example to illustrate his proposition. Here, we were able to demonstrate that under certain conditions, cultural lags can be produced in a controlled lab experiment.

Acknowledgment

We thank Kasper Otten, Christine Horne, and the participants of the Social Research and Public Policy Seminar at NYU Abu Dhabi, the Sixth International Meeting on Experimental and Behavioral Social Sciences at Utrecht University, the Sociology Colloquium at the School of Business and Economics at Nuremberg University, and the Seminar of the Network for the Advancement of Social and Political Studies at the University of Milan for their helpful comments on earlier versions of the article. We also thank Oliver Brägger and Stefan Wehrli from the Decision Science Laboratory at ETH Zurich for their support with conducting the experiment.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: G.A. and A.D. gratefully acknowledge financial support from, respectively, the Knut & Alice Wallenberg Foundation (KAW 2016.0167) and ETH Zurich research funds.

ORCID iDs

Wojtek Przepiorka https://orcid.org/0000-0001-9432-8696
Aron Szekely https://orcid.org/0000-0001-5651-4711

Supplemental Material

Supplemental material for this article is available online.

References

Abbink, Klaus, Lata Gangadharan, Toby Handfield, and John Thrasher. 2017 "Peer Punishment Promotes Enforcement of Bad Social Norms." *Nature Communications* 8(1):Article 609.

Alesina, Alberto, Paola Giuliano, and Nathan Nunn. 2013. "On the Origins of Gender Roles: Women and the Plough." *Quarterly Journal of Economics* 128(2):469–530.

Álvarez-Benjume, Amalia, and Fabian Winter. 2018. "Normative Change and Culture of Hate: An Experiment in Online Environments." *European Sociological Review* 34(3):223–37.

Andrighetto, Giulia, Daniela Grieco, and Luca Tummolini. 2015. "Perceived Legitimacy of Normative Expectations Motivates Compliance with Social Norms When Nobody Is Watching." *Frontiers in Psychology* 6:1413.

Asch, Solomon E. 1951. "Effects of Group Pressure upon the Modification and Distortion of Judgments." In *Groups, Leadership, and Men*, edited by H. Guetzkow. Pittsburgh, PA: Carnegie Mellon University.

Baumard, Nicolas, Jean-Baptiste André, and Dan Sperber. 2013. "A Mutualistic Approach to Morality: The Evolution of Fairness by Partner Choice." *Behavioral and Brain Sciences* 36(1):59–78.

Berger, Peter L., and Thomas Luckmann. 1966. *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. New York: Anchor.

Bicchieri, Cristina. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge, UK: Cambridge University Press.

Bicchieri, Cristina, Jan W. Lindemans, and Ting Jiang. 2014. "A Structured Approach to a Diagnostic of Collective Practices." *Frontiers in Psychology* 5:1418.

Bowles, Samuel, and Sandra Polanía-Reyes. 2012. "Economic Incentives and Social Preferences: Substitutes or Complements?" *Journal of Economic Literature* 50(2):368–425.

Brennan, Susan E., and Herbert H. Clark. 1996. "Conceptual Pacts and Lexical Choice in Conversation." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22(6):1482–93.

Centola, Damon, and Andrea Baronchelli. 2015. "The Spontaneous Emergence of Conventions: An Experimental Study of Cultural Evolution." *Proceedings of the National Academy of Sciences* 112(7):1989–94.

Cialdini, Robert B., and Melanie R. Trost. 1998. "Social Influence: Social Norms, Conformity and Compliance." In *The Handbook of Social Psychology*, edited by D. T. Gilbert, S. T. Fiske, and G. Lindzey. New York: McGraw-Hill.

Coleman, James S. 1990. *Foundations of Social Theory*. Cambridge, MA: Belknap.

Curzan, Anne. 2014. *Fixing English: Prescriptivism and Language History*. Cambridge, UK: Cambridge University Press.

Diekmann, Andreas. 1985. "Volunteer's Dilemma." *Journal of Conflict Resolution* 29(4):605–10.

Diekmann, Andreas. 1993. "Cooperation in an Asymmetric Volunteer's Dilemma Game: Theory and Experimental Evidence." *International Journal of Game Theory* 22(1):75–85.

Diekmann, Andreas. 2004. "The Power of Reciprocity: Fairness, Reciprocity, and Stakes in Variants of the Dictator Game." *Journal of Conflict Resolution* 48(4):487–505.

Diekmann, Andreas, Ben Jann, Wojtek Przepiorka, and Stefan Wehrli. 2014. "Reputation Formation and the Evolution of Cooperation in Anonymous Online Markets." *American Sociological Review* 79(1):65–85.

Diekmann, Andreas, and Wojtek Przepiorka. 2016. "'Take One for the Team!' Individual Heterogeneity and the Emergence of Latent Norms in a Volunteer's Dilemma." *Social Forces* 93(3):1309–33.

Diesendruck, Gil, and Lori Markson. 2011. "Children's Assumption of the Conventionality of Culture." *Child Development Perspectives* 5(3):189-195.

Efferson, Charles, Sonja Vogt, Amy Elhadi, Hilal El Fadil Ahmed, and Ernst Fehr. 2015. "Female Genital Cutting Is Not a Social Coordination Norm." *Science* 349(6255):1446–47.

Engel, Christoph. 2011. "Dictator Games: A Meta Study." *Experimental Economics* 14:583–610.

Eriksson, Kimmo, Pontus Strimling, and Julie C. Coultas. 2015. "Bidirectional Associations between Descriptive and Injunctive Norms." *Organizational Behavior and Human Decision Processes* 129:59–69.

Fehr, Ernst, and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature* 415(6868):137–40.

Fehr, Ernst, and Ivo Schurtenberger. 2018. "Normative Foundations of Human Cooperation." *Nature Human Behaviour* 2:458–68.

Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics* 10(2):171–78.

Gneezy, Uri, Stephan Meier, and Pedro Rey-Biel. 2011. "When and Why Incentives (Don't) Work to Modify Behavior." *Journal of Economic Perspectives* 25(4):191–209.

Gouldner, Alwin E. 1960. "The Norm of Reciprocity. A Preliminary Statement." *American Sociological Review* 25:161–78.

Grueneisen, Sebastian, and Michael Tomasello. 2017. "Children Coordinate in a Recurrent Social Dilemma by Taking Turns and along Dominance Asymmetries." *Developmental Psychology* 53(2):265–73.

Grutzpalk, Jonas. 2002. "Blood Feud and Modernity: Max Weber's and Émile Durkheim's Theories." *Journal of Classical Sociology* 2(2):115–34.

Guala, Francesco, and Luigi Mittone. 2010. "How History and Convention Create Norms: An Experimental Study." *Journal of Economic Psychology* 31(4):749–56.

Hawkins, Robert X. D., Noah D. Goodman, and Robert L. Goldstone. 2019. "The Emergence of Social Norms and Conventions." *Trends in Cognitive Sciences* 23(2):158–69.

Helbing, Dirk, Martin Schönhof, Hans-Ulrich Stark and Janusz A. Hołyst. 2005. "How Individuals Learn to Take Turns: Emergence of Alternating Cooperation in a Congestion Game and the Prisoner's Dilemma." *Advances in Complex Systems* 8(1):87–116.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath, et al. 2005. "'Economic Man' in Cross-Cultural Perspective: Behavioral Experiments in 15 Small-Scale Societies." *Behavioral and Brain Sciences* 28(6):795–855.

Horne, Christine. 2001. "Sociological Perspectives on the Emergence of Social Norms." Pp. 3–34 in *Social Norms*, edited by M. Hechter and K.-D. Opp. New York: Russell Sage.

Horne, Christine. 2009. *The Rewards of Punishment: A Relational Theory of Norm Enforcement*. Stanford, CA: Stanford University Press.

Horne, Christine, and Stefanie Mollborn. 2020. "Norms: An Integrated Framework." *Annual Review of Sociology* 46:467–87.

Horne, Christine, Justine Tinkler, and Wojtek Przepiorka. 2018. "Behavioral Regularities and Norm Stickiness: The Cases of Transracial Adoption and Online Privacy." *Social Research: An International Quarterly* 85(1):93–113.

Leo, Greg. 2017. "Taking Turns." *Games and Economic Behavior* 102:525–47.

Lewis, David. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Lindström, Björn, Simon Jangard, Ida Selbing, and Andreas Olsson. 2018. "The Role of a 'Common Is Moral' Heuristic in the Stability and Change of Moral Norms." *Journal of Experimental Psychology: General* 147(2):228–42.

Lipari, Francesca, and Giulia Andrighetto. 2021. "The Change in Social Norms in the Mafia's Territories: The Anti-racket Movement of Addiopizzo." *Journal of Institutional Economics* 17:227–42.

Mackie, Gerry. 1996. "Ending Footbinding and Infibulation: A Convention Account." *American Sociological Review* 61(6):999–1017.

Martin, Christopher Flynn, Dora Biro, and Tetsuro Matsuzawa. 2017. "Chimpanzees Spontaneously Take Turns in a Shared Serial Ordering Task." *Scientific Reports* 7:14307.

Melis, Alicia P., Patricia Grocke, Josefine Kalbitz, and Michael Tomasello. 2016. "One for You, One for Me: Humans' Unique Turn-Taking Skills." *Psychological Science* 27(7):987–96.

Nisbett, Richard, and Dov Cohen. 1996. *Culture of Honor: The Psychology of Violence in the South*. Boulder, CO: Westview.

Nyborg, Karine, John M. Anderies, Astrid Dannenberg, Therese Lindahl, Caroline Schill, Maja Schlüter, W. Neil Adger, et al. 2016. "Social Norms as Solutions." *Science* 354(6308):42–43.

Ogburn, William F. 1922. *Social Change with Respect to Culture and Original Nature*. New York: Huebsch.

Opp, Karl-Dieter. 2004. "'What Is Is Always Becoming What Ought to Be.' How Political Action Generates a Participation Norm." *European Sociological Review* 20(1):13–29.

Ostrom, Elinor. 2000. "Collective Action and the Evolution of Social Norms." *Journal of Economic Perspectives* 14(3):137–58.

Otten, Kasper, Vincent Buskens, Wojtek Przepiorka, and Naomi Ellemers. 2020. "Heterogeneous Groups Cooperate in Public Good Problems Despite Normative Disagreements About Individual Contribution Levels." *Scientific Reports* 10:16702.

Pettit, Philip. 1990. "Virtus Normativa: Rational Choice Perspectives." *Ethics* 100(4):725–55.

Piskorski, Mikołaj Jan, and Andreea Gorbatâi. 2017. "Testing Coleman's Social-Norm Enforcement Mechanism: Evidence from Wikipedia." *American Journal of Sociology* 122(4):1183–1222.

Przepiorka, Wojtek, and Andreas Diekmann. 2018. "Heterogeneous Groups Overcome the Diffusion of Responsibility Problem in Social Norm Enforcement." *PLoS ONE* 13(11):e0208129.

Przepiorka, Wojtek, Loes Bouman, and Erik W. de Kwaadsteniet. 2021. "The Emergence of Conventions in the Repeated Volunteer's Dilemma: The Role of Social Value Orientation, Payoff Asymmetries and Focal Points." *Social Science Research* 93:102488.

Roberts, Steven O., Susan A. Gelman, and Arnold K. Ho. 2017 "So It Is, So It Shall Be: Group Regularities License Children's Prescriptive Judgments." *Cognitive Science* 41:576–600.

Rakoczy, Hannes, Felix Warneken, and Michael Tomasello. 2008. "The Sources of Normativity: Young Children's Awareness of the Normative Structure of Games." *Developmental Psychology* 44(3):875–81.

Schmidt, Marco F. H., Hannes Rakoczy, and Michael Tomasello. 2011. "Young Children Attribute Normativity to Novel Actions without Pedagogy or Normative Language." *Developmental Science* 14(3):530–39.

Sugden, Robert. 1986. *The Economics of Rights, Co-operation, and Welfare*. Oxford, UK: Basil Blackwell.

Sugden, Robert. 1998. "Normative Expectations: The Simultaneous Evolution of Institutions and Norms." Pp. 73–100 in *Economics, Values, and Organization*, edited by A. Ben-Ner and L. Putterman. Cambridge, UK: Cambridge University Press

Szekely, Aron, Francesca Lipari, Alberto Antonioni, Mario Paolucci, Angel Sánchez, Luca Tummolini, and Giulia Andrighetto (2021). "Evidence from a Long-Term Experiment That Collective Risks Change Social Norms and Promote Cooperation." *Nature Communications* 12(1):1–7.

Theriault, Jordan E., Liane Young, and Lisa Feldman Barrett. 2021. "The Sense of Should: A Biologically-Based Framework for Modeling Social Pressure." *Physics of Life Reviews* 36:100–36.

Thibaut, John W., and Harold H. Kelley. 1959. *The Social Psychology of Groups*. Oxford, UK: Wiley.

Tummolini, Luca, Giulia Andrighetto, Cristiano Castelfranchi, and Rosaria Conte. 2013. "A Convention or (Tacit) Agreement betwixt Us: On Reliance and Its Normative Consequences." *Synthese* 190:585–618.

Tummolini, Luca, and Giovanni Pezzulo. 2021. "The Epistemic Value of Conformity: Comment on 'The Sense of Should: A Biologically-Based Framework for Modeling Social Pressure' by Jordan E. Theriault, Liane Young, and Lisa Feldman Barrett." *Physics of Life Reviews* 36:74–76.

Tworek, Christina M., and Andrei Cimpian. 2016. "Why Do People Tend to Infer 'Ought' from 'Is'? The Role of Biases in Explanation." *Psychological Science* 27(8):1109–22.

van Bavel, Jay J., Katherine Baicker, Paulo S. Boggio, Valerio Capraro, Aleksandra Cichocka, Mina Cikara, Molly J.

Crockett, et al. 2020. “Using Social and Behavioural Science to Support Covid-19 Pandemic Response.” *Nature Human Behaviour* 4(5):460–71.

Veblen, Thorstein. [1899]1994. *The Theory of the Leisure Class: An Economic Study of Institutions*. New York: Dover.

Willer, Robb, Ko Kuwabara, and Michael W. Macy. 2009. “The False Enforcement of Unpopular Norms.” *American Journal of Sociology* 115(2):451–90.

Wrong, Dennis H. 1994. *The Problem of Order: What Unites and Divides Society*. New York: Free Press.

Yamaguchi, Kazuo. 2019. *Gender Inequalities in the Japanese Workplace and Employment: Theories and Empirical Evidence*. Singapore: Springer.

Young, H. Peyton. 1993. “The Evolution of Conventions.” *Econometrica* 61(1):57–84.

Author Biographies

Wojtek Przepiorka is an associate professor in the Department of Sociology at Utrecht University. His research interests are in analytical, economic, and environmental sociology, organizational behavior, and quantitative methodology. He uses a diverse set of quantitative research methods to investigate, among others, how social norms emerge, are enforced and change. His recent publications include “Moderators of Reputation Effects in Peer-to-Peer Online Markets: A Meta-Analytic Model Selection Approach” (*Journal of Computational Social Science*, with R. Jiao and V. Buskens).

Aron Szekely is an assistant professor of sociology at Collegio Carlo Alberto in Turin, Italy. He studies social mechanisms, including reputation, signaling, and social norms and their effects in situations of cooperation and conflict. Much of his research uses experiments to study microlevel individual decision making or agent-based models to explore emergent macrolevel social phenomena.

Giulia Andrighetto is a research director at the Institute of Cognitive Sciences and Technologies of the National Research Council of Italy in Rome, where she is the coordinator of the Laboratory of Agent Based Social Simulation. She is also a senior researcher at the Institute for Future Studies in Stockholm, Sweden. Her research focuses on the emergence, enforcement, change and decay of social norms and their effects on cooperation and conflict. Her research topics include cooperation, altruism, honesty, and bad norms and misinformation. She uses theoretical and computational models, combined with online and laboratory experiments, surveys, and big data to answer these and related questions about norms.

Andreas Diekmann is a senior professor of sociology at the University of Leipzig and professor emeritus at ETH Zurich. His areas of research are social cooperation and experimental game theory, environmental and population sociology, and methods of empirical social research. Current research activities focus on studies of climate policy and experimental research on social dilemmas and the emergence of social norms. His recent publications include “Emergence of and Compliance with New Social Norms: The Example of the COVID Crisis in Germany” (*Rationality and Society*).

Luca Tummolini is a senior researcher at the Institute of Cognitive Sciences and Technologies of the Italian Research Council in Rome and an associated researcher at the Institute for Future Studies in Sweden with a PhD in cognitive science. His research interests are on social interaction and the cognitive mechanisms that enable humans to flexibly coordinate and collaborate with one another: from shared deliberation in small groups to conformity with population-wide regularities such as conventions and social norms. Current projects examine the role of reputation at the group level in the formation of norms promoting aggression, retaliation, and punishment.