Emergence of Latent Norms in a Volunteer's Dilemma

"Take One for the Team!" Individual Heterogeneity and the Emergence of Latent Norms in a Volunteer's Dilemma

Andreas Diekmann, *ETH Zurich*
Wojtek Przepiorka, *Utrecht University and Nuffield College*

The tension between individual and collective interests and the provision of sanctioning mechanisms have been identified as important building blocks of a theory of norm emergence. Correspondingly, most investigations focus on how social norms emerge through explicit bargaining and social exchange to overcome this tension, and how sanctions enforce norm compliance. However, sanctioning presupposes the existence of the behavior at which it is directed, and the question of how behavior worth sanctioning can emerge tacitly if communication is not possible has hitherto received little attention. Here, we argue that game theory offers an ideal framework for theorizing about emergent behavioral regularities and show how latent norms can emerge from actors' recurring encounters in similar social dilemmas. We conduct two experiments in which small groups of subjects interact repeatedly in a volunteer's dilemma. We vary the heterogeneity of group members in terms of their costs of cooperation and the way they encounter each other in subsequent interactions. Our results show that subjects in homogeneous groups take turns at cooperating, whereas in heterogeneous groups mostly the subjects with the lowest costs cooperate. The emergence of solitary cooperation is moderated by the way subjects encounter each other again and their other-regarding preferences.

Introduction

Social norms are a central concept in sociological scholarship. Most sociological scholarship concerned with the impact of social norms on individual behavior and society at large has taken social norms for granted (Coleman 1990; Ellickson

The authors would like to thank Stefan Wehrli from DeSciL, the experimental laboratory at ETH Zurich, for his support in conducting the two experiments. They are grateful to Neli Demireva, Anja Neundorf, Karl-Dieter Opp, Bart Wilson, and two anonymous reviewers for their insightful comments and helpful suggestions. This work was supported by the Swiss National Science Foundation [grant number 100017_124877]. Authors' names are listed in alphabetical order. Authors' contributions are as follows: AD, WP designed the experiments and devised the latent norm index; WP programmed and conducted the experiments; WP analyzed the data; and AD, WP wrote the paper. Please direct correspondence to Wojtek Przepiorka (w.przepiorka@uu.nl).

Social Forces 94(3) 1309–1333, March 2016
doi: 10.1093/sf/sov107
Advance Access publication on 13 October 2015

1991). Only relatively recently did sociologists and other scholars become interested in the social forces that bring social norms about. Theories on the emergence of social norms were put forward by Ullmann-Margalit (1977), Opp (1982), Axelrod (1986), Coleman (1990), Ellickson (1991), Young (1998), Posner (2000), Ostrom (2000), Horne (2001), Bicchieri (2006), and others. In institutional economics, the "property rights tradition" (Demsetz 1967) has contributed to our understanding of the emergence of social norms, and in social psychology, Thibaut and Kelley (1959) are acknowledged forerunners of a theory of norm emergence. Based on this literature, we define a social norm as a rule guiding social behavior, the deviation from (adherence to) which is negatively (positively) sanctioned.

Most theories of norm emergence have in common the idea that actions affecting others (so-called externalities) can give rise to social norms that proscribe or prescribe these actions (Demsetz 1967; Ullmann-Margalit 1977; Axelrod 1986; Coleman 1990; Posner 2000; Ostrom 2000; Horne 2001). Coleman, for instance, posits that externalities can generate a demand for social norms if the acting and affected parties cannot agree on a collectively more favorable outcome. This may be because it is too costly for the two parties to initiate an agreement, or because they are unwilling to accept the redistribution of resources an agreement would imply. Horne (2001), taking Coleman's theory a step further, argues that rather than individual and group interests in isolation, the interaction of both, either met or undermined by individual behavior, would provide a better theoretical framework for the explanation of not only norm demand but also norm content. That is, specifying how the behavior of one actor affects both the actor's outcome and other actors' outcomes will better predict the strength and direction of the norm that could possibly emerge.[1] However, both Coleman and Horne are vague about how behavior constituting a social norm, that is, behavior at which a norm is directed, is instantiated as a social process if actors cannot engage in direct negotiation or social exchange. Wrong (1994) explicates this process as follows:

> In short, interaction generates habits; perceived, they become reciprocal expectations; in addition to their purely predictive and anticipatory nature, sensitivity to them endows them with a constraining or even an obligatory character. This entire process is in no sense willed or even fully foreseen by either party. It is a *sui generis* resultant of their recurrent situated interaction. Whatever the needs, motives, and interests underlying this interaction, its continuation has precipitated mutually binding sets of expectations. Thus do norms grow in unplanned fashion out of ongoing interaction. (Wrong 1994, 48–49)

In contrast to social norms, that is, "rules capable of explicit statement by the actors," Wrong (1994, 48) calls "the expectations that arise concerning habits emerging and crystallizing in the course of repeated interactions" *latent norms*. Accordingly, we define a latent norm as expectations concerning behavioral regularities emerging in the course of repeated interactions. Latent norms can be understood as a precursor of social norms. In this regard, Opp (2004) points out that a possible transition from a behavioral regularity via a latent norm to a social norm should be conceived as a gradual process:

> The more often the members of a group perform a behaviour, B, the stronger is the empirical expectation that B is performed; the stronger this empirical expectation is, the stronger is resentment in case of non-regular behaviours; the stronger this resentment is, the more likely it is that the performance of B becomes a norm. (Opp 2004, 14)

Opp (2004) notes, moreover, that the net utility of the behavioral regularity and the size of the negative externalities produced by non-conformity will moderate the process by which a behavioral regularity can turn into a social norm.

Here, we argue that game theory in general and the game-theoretic analysis of social dilemmas in particular provides an ideal framework for the explanation of behavioral regularities and, as such, should be a first building block in a theory of norm emergence (Ellickson 2001; Voss 2001; Bowles 2004). A social dilemma is defined as a situation of strategic interdependence in which the decisions of individually rational actors lead to an inferior outcome for all or some parties than the decisions of "collectively rational" actors. Collective rationality means that actors, had they an opportunity to communicate and agree on a binding contract, should agree on a combination of actions leading to a welfare-enhancing outcome.[2] We start from the assertion that the gap between individual and collective rationality inherent in social dilemmas (Rapoport 1974) creates a demand for latent norms (Voss 2001), and we investigate experimentally how latent norms can emerge from recurring, structurally similar social dilemmas.

Most studies of norm emergence are quick in invoking sanctions enforcing social norms (e.g., Diekmann and Przepiorka 2015; Fehr and Fischbacher 2004; Horne 2007; Kitts 2006; Przepiorka and Diekmann 2013; Willer, Kuwabara, and Macy 2009; Yamagishi 1986) but rarely ask how behavioral regularities worth sanctioning emerge in the first place. We tune down any attempts to explain norm enforcement for the time being and investigate how latent norms emerge in the repeated volunteer's dilemma. We use the volunteer's dilemma (Diekmann 1985, 1993) as our stage game because it shares important properties with a range of other social dilemmas (Archetti and Scheuring 2011; Bliss and Nalebuff 1984; Bolle 2011; Palfrey and Rosenthal 1984; Weesie 1993, 1994) and has been shown to map many real-world situations relatively well (Barron and Yechiam 2002; Darley and Latané 1968; Eger, Kraft, and Weise 1992; Nelson 1959; Przepiorka and Berger 2015). In our experiments, we vary individual heterogeneity and the way in which actors meet in a similar interaction situation repeatedly. We are interested in the structural conditions under which behavioral regularities and latent norms can emerge because we think this will inform a bottom-up theory of norm emergence in general and establish a better predictor of norm content in particular.

In the next section, we start with identifying three structural dimensions along which social dilemmas should be classified to inform a coherent theory of (latent) norm emergence: cooperation and coordination dilemmas, symmetric and asymmetric dilemmas, and one-shot and repeated dilemmas. In section 3, we introduce the volunteer's dilemma and derive testable hypotheses. We then study the emergence of latent norms in two computerized laboratory experiments. Sections 4 and 5 present the designs and results of our two experiments. Section 6 discusses the implications of our findings for theories of norm emergence and concludes.

Classifying Social Dilemmas

There are many possible ways in which social dilemmas can be classified, and we do not attempt to present a comprehensive classification here, nor are we the first to suggest one to inform a theory of norm emergence. With regard to the latter, Ullmann-Margalit (1977) did the bulk of the work and we just stand on her shoulders, restating what she and others have taught us in a nutshell and giving more emphasis to what we find important. Other classifications of social dilemmas have been suggested, for instance, by Dawes (1980), Messick and Brewer (1983), Kollock (1998), McAdams (2009), and Raub, Buskens, and Corten (2015).

We believe that situations of strategic interdependence, such as social dilemmas, can be formalized in game-theoretic terms. Based on a formal model, clear hypotheses can be derived and put to an empirical test. However, not every social dilemma is a prisoner's dilemma (McAdams 2009), and this section aims at indicating the variety of situations in which game-theoretic models of social dilemmas can be used to explain the emergence of social norms. In this regard, this section will also better allow the reader to appreciate the broad applicability of the volunteer's dilemma, the social dilemma under scrutiny in this paper.

Cooperation and Coordination Dilemmas

A first important distinction a theory of norm emergence should make is between cooperation dilemmas, such as the prisoner's dilemma (PD), and coordination dilemmas (Lewis 1969; Ullmann-Margalit 1977). Unlike in cooperation dilemmas, where every actor has an incentive to free-ride on the cooperation of others, in pure coordination dilemmas, the interacting parties' interests fully overlap. Hence, there is no conflict between individual self-interest and collective well-being in pure coordination dilemmas. The gap between individual and collective rationality (i.e., social dilemma) arises from the fact that there are multiple (Nash) equilibria and, without communication, it is difficult for actors to find a tacit agreement on which equilibrium should be selected by all actors.[3]

Equilibrium selection is often arbitrary, path dependent, and the result of a long-lasting diffusion process. Historical events, for instance, can trigger or favor a certain type of behavior and lead to conventions that enable actors to tacitly agree on a non-detrimental course of actions (Lewis 1969). Examples are vehicles driving on the right-hand side of the road (Young 1993), or the emergence of cultural badges (Centola and Baronchelli 2015). By definition, conventions coordinate actors' choices of equilibrium strategies in a coordination game. Thus, conventions create self-fulfilling expectations that the parties to an interaction will act in a certain way, as deviating from these expectations would be self-harming (Lewis 1969; Young 1998). Coordination dilemmas are well suited to explain the emergence of behavioral regularities (i.e., selection of one of the possible equilibria) based on historical events (McAdams 2009). This argument presupposes, however, that actors encounter structurally similar interaction situations repeatedly (we will come back to this point shortly). But even in single encounters, actors may find a tacit agreement on which equilibrium to select, if one equilibrium is more conspicuous than any other. If actors know that they are facing the same coordination dilemma, the nature of the situation may provide "some focal point for each person's expectation of what the other expects him to expect to be expected to do" (Schelling 1980, 57). Not acting on these expectations will lead to an unfavorable outcome for all parties.[4]

Cooperation and coordination dilemmas are not mutually exclusive, and the overlap can be at least twofold. First, cooperation dilemmas often arise because the interacting parties have divergent notions of what cooperative acts are (Winter, Rauhut, and Helbing 2012). Thus, they have first to coordinate on which cooperation dilemma they are mutually facing (McAdams 2009; Garrett and Weingast 1993; Winter 2014). Second, actors' interests can overlap fully or only partly. In the latter case, the coordination dilemma also entails a cooperation dilemma, as choosing a particular equilibrium will benefit one actor more than other actors. Impure coordination dilemmas are thus well suited to describe situations of distributional conflict (Schelling 1980, 58–67; McAdams 2009).[5]

Symmetric and Asymmetric Social Dilemmas

A second distinction should be made between symmetric and asymmetric social dilemmas. Although partly overlapping interests in impure coordination dilemmas imply asymmetry in actors' outcomes, most social dilemmas discussed in the literature are symmetric. They are called symmetric because the decision situation is equivalent from each actor's perspective; symmetry implies individual homogeneity. While symmetry can be a useful simplifying assumption to start a game-theoretic analysis of a social dilemma, it is, of course, one of the more unrealistic ones. Individuals differ in many significant ways. In most naturally occurring social dilemmas, there will be at least one actor who has different preferences, endowments, and/or constraints than their interaction partners.

In his seminal book, "The Logic of Collective Action," Olson (1971[1965]) notes that even in small groups of self-regarding individuals, a public good is likely to be undersupplied unless a "privileged" actor is in the group, who would benefit from providing the public good all by him- or herself. Olson points out that in a situation in which one actor has stronger incentives to cooperate than the other actors, there may arise a "*tendency for the 'exploitation' of the great by the small*" (Olson 1971[1965], 35; italics in original). In a similar vein, Nelson (1959) noted that firms that invest in the collective good of basic research face the risk of being exploited by competing firms, which may eventually patent and market the new products without incurring any R&D costs. The theory of equilibrium selection by Harsanyi and Selten (1988) lends further support to the idea that an equilibrium will be selected in which the most privileged actor cooperates (see also Diekmann [1993] and He, Wang, and Li [2014]). However, perceivable individual differences can be unrelated to the structure of the dilemma and still tacitly single out a focal actor, reinforcing the selection of an equilibrium in which the focal actor cooperates (Schelling 1980).

The above suggests that asymmetry, that is, individual heterogeneity, can be an important determinant of coordinated action and facilitate cooperation and the attainment of mutually beneficial outcomes. Advocates of "Critical Mass Theory" suggest, moreover, that individual heterogeneity may increase the likelihood that a public good will be produced, particularly in the case of "accelerating" production functions (Oliver, Marwell, and Teixeira 1985; Heckathorn 1993; Marwell and Oliver 1993). With accelerating production functions, initial contributions have a relatively small impact on the level of the public good, while later contributions yield a much higher rate of return. Thus, individual heterogeneity may provide for a critical mass of highly interested and resourceful individuals who incur the "start-up costs" and instigate further contributions with ever larger marginal benefits.[6]

One-Shot and Repeated Social Dilemmas

One-shot social dilemmas are an important ingredient in a theory of norm emergence because they define the structure of the immediate interaction (i.e., stage game). However, the notion of latent norms gains momentum once actors are repeatedly confronted with structurally similar social dilemmas. In repeated encounters, actors can form mutual expectations about the future course of action based on their own and their interaction partners' previous actions, and act contingent on these expectations (Opp 2004; Wrong 1994). Also, Ullmann-Margalit's (1977) notion of norms as emergent solutions to social dilemmas is inherently based on the assumption that individuals face similar decision situations repeatedly. For instance, in the one-shot PD, mutual defection is the only Nash equilibrium, whereas in the repeated PD, mutual cooperation too can be part of agents' equilibrium strategies (Axelrod 1984; Voss 2001).

In general, the folk theorem in game theory asserts that a more cooperative equilibrium than mutual defection can be sustained if the likelihood that the agents will meet again is large enough and each agent knows that deviating from the cooperative equilibrium will cause mutual defection forever (Fudenberg and Maskin 1986). Thus, in cooperation dilemmas, repetition allows equilibria in which cooperation is a self-regarding best response. However, the repeated game has many such equilibria, and agents have no rational guidance as to which one to select. This equilibrium selection problem becomes even more severe when the number of agents increases and their actions cannot be perfectly observed by the entire group (Gintis 2009, 185–95).

It has been suggested that, similar to a traffic light system, an exogenous correlating device could orchestrate agents' equilibrium play, and that social norms can be conceived of as such correlating devices (Bowles and Gintis 2011, 89–92). At the same time, these authors acknowledge that "social norms cannot be introduced as a *deus ex machina*, as if laid down by a centralized authority, without violating the objective to provide a 'bottom-up' theory of cooperation that does not presuppose preexisting institutional forms of cooperation" (Bowles and Gintis 2011, 90; italics in original). Thus, the question remains how social norms emerge in the first place.

Here we argue, and we are not the first to do so (see, e.g., Thibaut and Kelley 1959, ch. 8; Opp 1982, 2004; Wrong 1994; Bendor and Swistak 2001; Voss 2001), that social norms can emerge from repeated interactions broadly construed. Social norms do not guide agents' selection of equilibrium strategies in repeated games, as it is unlikely that agents can know a priori whether the game is repeated or one-shot (Delton et al. 2011). It is more plausible that latent norms, once they emerge from repeated interactions, will guide agents' behavior in future encounters by making certain actions in the stage game focal (Posner 2000). The more a latent norm becomes accepted and adhered to, the more non-adherence will cause resentments; once consolidated by a sanctioning mechanism, the latent norm is likely to turn into a social norm (Guala 2013; Horne 2001; Opp 2004). In what follows, we will use the volunteer's dilemma to derive hypotheses regarding the emergence of latent norms, and we will test our hypotheses empirically in two experiments. We give a more comprehensive review of the experimental literature on social norm emergence in the online Supplementary Material, Appendix A.

Model and Hypotheses

The volunteer's dilemma (VOD) is a step-level public good game where only one actor's cooperation is necessary and sufficient to produce the public good (Diekmann 1985; Palfrey and Rosenthal 1984). Although in the VOD the benefits outweigh the costs of producing the public good, free-riding on another actor's cooperation is even more beneficial. Consequently, everyone may end up earning nothing while waiting for someone else to "volunteer." More formally (see table 1), a public good of value $\sum U_i$ for a group of size $n \geq 2$ is produced by a single actor i choosing C (cooperation) at a cost K_i, where $U_i > K_i > 0 \; \forall \; i$. The public good is not produced if all actors choose D (defection), and there is a welfare loss if more than one actor chooses C.

The VOD is an impure coordination dilemma, where problems of coordination and cooperation are involved simultaneously. First, unlike in the PD, defection is not a dominant strategy, as cooperation guarantees a payoff of $U_i - K_i > 0$. The VOD has n (Pareto-optimal) Nash equilibria in pure strategies, where one actor chooses C and $n - 1$ actors choose D. This can be seen by the fact that in equilibrium no actor has an incentive to change their strategy unilaterally. However, for a group of actors, it is difficult to coordinate on one of the equilibria without communication. Second, actors' interests do not fully overlap. The net benefits from the public good are lower for the volunteer than for a free-rider, which creates potential for distributional conflict.

Table 1. The Volunteer's Dilemma Game

	Number of other actors choosing C				
Actor *i*'s choice	0	1	2	…	$n - 1$
Cooperation (C)	$U_i - K_i$	$U_i - K_i$	$U_i - K_i$	$U_i - K_i$	$U_i - K_i$
Defection (D)	0	U_i	U_i	U_i	U_i

We distinguish between the symmetric and asymmetric VOD. In the symmetric VOD, all actors have the same benefits from and costs of producing the public good (i.e., $U_i = U_j$ and $K_i = K_j \;\forall\; i \neq j$). In an asymmetric VOD (Diekmann 1993), there is at least one actor with different benefits and/or costs than the rest of the group (i.e., $U_i \neq U_j$ and/or $K_i \neq K_j \;\exists\; i \neq j$).

The Symmetric VOD

In the symmetric one-shot VOD, a pure-strategy equilibrium results in "asymmetric" payoffs, where the volunteer earns K less than a free-rider. Therefore, a pure-strategy equilibrium will not be easily attainable without an agreement on who the volunteer should be. Besides the n Nash equilibria in pure strategies, the symmetric one-shot VOD has a further, payoff-symmetric Nash equilibrium in mixed strategies. To see this, denote actor i's probability of choosing D with q_i. Then, the probability that every actor $j \neq i$ will choose D is $\prod q_j$, and the probability that at least one actor $j \neq i$ will choose C is $1 - \prod q_j$. Hence, actor i's expected payoff from choosing D with probability q_i is

$$\pi_i = q_i U_i \left(1 - \prod_{j \neq i}^{n} q_j \right) + (1 - q_i)(U_i - K_i). \tag{1}$$

In a mixed-strategy equilibrium (MSE), every actor i must be indifferent between C and D. That is, the change in actor i's expected payoff from a small change in the probability of choosing D must be zero:

$$\frac{\partial \pi_i}{\partial q_i} = -\,U_i \prod_{j \neq i}^{n} q_j + K_i = 0. \tag{2}$$

Solving the above system of n equations (one for each actor) yields actor i's probability p_i^* $(= 1 - q_i^*)$ of choosing C in the MSE:

$$p_i^* = 1 - \sqrt[n-1]{K_i / U_i}\,. \tag{3}$$

Finally, we can state the MSE probability that the public good will be produced, that is, that at least one actor will choose C (equation 4), and the MSE probability that the public good will be produced efficiently, that is, by one actor only (equation 5):

$$p^* = 1 - \prod_{i=1}^{n} q_i^* \tag{4}$$

$$p_e^* = n p_i^* (1 - p_i^*)^{(n-1)}. \tag{5}$$

The situation changes if the symmetric VOD is repeated an indeterminate number of times. Now, the n actors can coordinate on sharing the costs of producing the public good, for instance, by taking turns in choosing C one after the other. Turn-taking has been observed in experiments (Bornstein, Budescu, and Zamir 1997; Evans, Sibly, and Tisdell 2013; Helbing et al. 2005) and has been shown to be an equilibrium strategy in the infinitely repeated two-person VOD (Lau and Mui 2012). Based on our theoretical argument thus far, we can state our behavioral hypotheses regarding the symmetric VOD:

> **H1:** If the symmetric VOD is repeated for an indeterminate number of times, actors will be more likely to take turns in cooperating than to coordinate on only one actor cooperating throughout, or to cooperate according to the MSE.

The Asymmetric VOD

In the asymmetric one-shot VOD, actor i's probability p_i^* of choosing C in the MSE (i.e., the solution of equation 2) is

$$p_i^* = 1 - \frac{U_i}{K_i}\left(\prod_{j=1}^{n} \frac{K_j}{U_j}\right)^{\left(\frac{1}{n-1}\right)}. \tag{6}$$

A somewhat counterintuitive implication of equation 6 is that actor i's probability of choosing C decreases as U_i increases and/or K_i decreases. In other words, the stronger an actor i is (in terms of benefits from and/or costs of producing the public good), the lower this actor's propensity to choose C. Moreover, for certain combinations of U and K, equation 6 yields negative values (see table 7 in the appendix). This makes the MSE intuitively not very appealing as a model of human behavior in the asymmetric one-shot VOD. For the sake of completeness, equation 4 also gives the MSE probability that the public good will be produced in the asymmetric VOD, and the MSE probability that the public good will be produced efficiently can be generalized for the asymmetric case as follows:

$$p_e^* = \sum_{i=1}^{n} p_i^* \prod_{j \neq i}^{n} (1 - p_j^*). \tag{7}$$

Fortunately, there is an alternative equilibrium concept for the asymmetric one-shot VOD. According to the theory of equilibrium selection by Harsanyi and Selten (1988), for the special case of an asymmetric VOD with one strongest actor, the pure strategy equilibrium will be selected in which the strongest actor chooses C and all other actors choose D (see also Diekmann [1993]). This conjecture has recently received further theoretical (He, Wang, and Li 2014) and empirical support (Cherry, Cotten, and Kroll 2013; Diekmann and Przepiorka 2015; Przepiorka and Diekmann 2013). Moreover, repetition will make tacit coordination on the strongest actor even more likely. We can thus state our behavioral hypotheses regarding the asymmetric VOD:

> **H2:** If the *asymmetric* VOD with one strongest actor is repeated for an indeterminate number of times, actors will be more likely to coordinate

> on only the strongest actor cooperating than on taking turns in cooperating, or to cooperate according to the MSE.

However, an alternative explanation for such behavior could be that being the strongest actor makes that actor focal in Schelling's (1980) sense, and coordination is attained *not because of the asymmetry* of the VOD but because of the "focality" of the strongest actor.

> **H3:** If the *symmetric* VOD with one focal actor is repeated for an indeterminate number of times, actors will be more likely to coordinate on only the focal actor cooperating than on taking turns in cooperating, or to cooperate according to the MSE.

Latent Norm Index (LNI)

In order to test our hypotheses, we develop an index to measure the frequency and stability of behavioral patterns that may emerge in the repeated VOD. Since we are interested primarily in welfare-enhancing outcomes, we restrict our attention to sequences of interactions where the same single actor cooperates repeatedly (solitary volunteering) and where a single actor's cooperation is followed by another single actor's cooperation in the next round (turn-taking).

A sequence of m interactions with n actors is specified by $(x_1, x_2, \dots, x_i, \dots, x_n)_1, \dots, (x_1, x_2, \dots, x_i, \dots, x_n)_j, \dots, (x_1, x_2, \dots, x_i, \dots, x_n)_m$, where $x_i = 0$ if actor i chose D and $x_i = 1$ if actor i chose C. We denote efficient outcomes by the "position" of the single cooperative actor, that is, $(1,0,\dots,0) = 1, (0,1,0,\dots,0) = 2,\dots,(0,\dots0, 1) = n$, and denote inefficient outcomes, where either all actors chose D or more than one actor chose C, by –1. Moreover, with regard to turn-taking, we denote the size of the subset of actors taking turns by $h = 1, 2,\dots, n$. For example, in a group of size $n = 3$, where actors 1 and 3 take turns in volunteering while actor 2 free-rides, actors 1 and 3 are in the subset of turn-taking actors, and thus, $h = 2$. Correspondingly, $h = 1$ denotes solitary volunteering.

We define $\lambda_{h,n}$ as the length of a type-h sequence in a group of size n, and the latent norm index as $\mathrm{LNI}_{h,n} = 100 \times \lambda_{h,n}/m$, where an observed sequence registered by the index should be at least as long as the number of actors in the group. This last restriction makes it harder to identify behavioral patterns in larger groups but at the same time helps avoid the measurement of pseudo patterns. In short, the $\mathrm{LNI}_{h,n}$ is the percentage of interactions of a type-h sequence and ranges between 0 and 100.

The following example illustrates how the LNI is applied on two made-up series of $m = 10$ interactions of $n = 3$ actors: 2 1 1 3 2 1 –1 1 2 3 and –1 2 1 2 1 2 2 2 1 3. Recall that –1 denotes an interaction in which either all actors chose D or more than one actor chose C and therefore is not counted as part of a sequence. In the first series, there is one 1-sequence of length two (1 1), and there are six 2-sequences of length two (2 1, 1 3, 3 2, 2 1, 1 2, and 2 3). Since the length of each of these sequences is smaller than n, $\mathrm{LNI}_{1,3}$ and $\mathrm{LNI}_{2,3}$ are both zero. The first series also contains one 3-sequence of length four (1 3 2 1) and another 3-sequence of length three (1 2 3). Since the length of each of these sequences is larger or equal to n, $\mathrm{LNI}_{3,3} = 100 \times (4 + 3)/10 = 70$. The second series contains one 1-sequence of length three (2 2 2), one 2-sequence of length five (2 1 2 1 2), two 2-sequences of length two (2 1 and 1 3), which are not part of a longer 2-sequence, and one 3-sequence of length three (2 1 3). Hence, $\mathrm{LNI}_{1,3} = 100 \times 3/10 = 30$, $\mathrm{LNI}_{2,3}$ is $100 \times 5/10 = 50$, and $\mathrm{LNI}_{3,3} = 100 \times 3/10 = 30$. Note that, as in this example, sequences of different types can overlap but overlaps cannot be longer than n.

Experiment 1

Procedure and Design

Our first experiment comprised 10 experimental sessions with 12 subjects in each session ($N = 120$ subjects in total).[7] Each session consisted of three sequential parts (see table 2). The first part consisted of $m = 56$ rounds, and the second and third parts each consisted of $m = 48$ rounds. At the beginning of each session, the 12 subjects were randomly matched in groups of three, and each group was assigned to one of the four experimental conditions: "symmetric," "asymmetric 1," "asymmetric 2," and "focal point." We will explain the four experimental conditions in detail below.

In each group, the three subjects interacted with one another for the entire number of rounds of one part. At the end of the first part, groups were disbanded and formed anew, such that no subject was in a new group with a subject they had been in a group with before, and no subject was in the same experimental condition as before. The same procedure was applied after the second part. At the end of a session, each subject had consecutively participated in three different experimental conditions, in each of which he or she had interacted with two different subjects.

In every round, each group of three subjects faced a VOD. That is, in every round, subjects had to decide independently whether to choose "up" (i.e., to cooperate) or "down" (i.e., to defect) by clicking on the corresponding area on their decision screen. Choosing "up" earned a subject $U - K$ with certainty.

Table 2. Design of Experiment 1

$n = 3$	Symmetric $U = 80c$ $K = 50c$	Asymmetric 1 $U = 80c$ $K_{1,2} = 50c$ $K_3 = 30c$	Asymmetric 2 $U = 80c$ $K_{1,2} = 50c$ $K_3 = 10c$	Focal Point $U = 80c$ $K = 50c$
Part 1 ($m = 56$)	560	560	560	560
Part 2 ($m = 48$)	480	480	480	480
Part 3 ($m = 48$)	480	480	480	480

Note: The table lists the number of interactions (i.e., rounds) that were recorded in each condition and in each part of experiment 1. *U* denotes the payoff a subject earns if the public good is produced; *K* denotes the individual cost of producing the public good; and 100c correspond to CHF 1.

Choosing "down" earned a subject U, but only if at least one other subject in their group chose "up" in the same round. If all subjects chose "down," they earned nothing. After every round, subjects learned the outcome of the interaction and the decision each subject in their group had made. By the ID number randomly assigned to each subject at the beginning of each part, subjects could follow the decisions of their group members over time.

The four experimental conditions differed in the variant of the VOD subjects faced in each round (see table 2). In the symmetric condition, $U = 80$c and $K = 50$c for all group members. Thus, in the symmetric condition, each subject earned $U - K = 30$c for cooperating. Conditions asymmetric 1 and asymmetric 2 differed from the symmetric condition only in that in each group one subject earned $U - K = 50$c or 70c, respectively, for cooperating. That is, at the beginning of each part, one group member was randomly assigned to be the "strong" subject, who in every round faced a lower cost from cooperation than the other two subjects. Finally, in the focal point condition, the payoff structure of the VOD corresponded to the one in the symmetric condition, but one of the three subjects was singled out by a visual cue. That is, we operationalized the "focal" actor by highlighting one subject's decision area with a different background color. Similar to the strong subject in the asymmetric conditions, the focal subject stayed in his or her role throughout one part of the experiment.

The experiment was conducted at the Decision Science Laboratory of ETH Zurich (DeSciL). Subjects were recruited by e-mail and could register online for one of the available sessions. Subjects were undergraduate students from different departments; 39 percent were female, and they were 21.8 years old on average (*sd* = 2.57). All participants received a show-up fee of CHF 10 and earned CHF 21.5 on average in the experiment. CHF stands for Swiss franc, and at the time the experiment was conducted, CHF 1 corresponded to USD 1.17.

Upon arrival at the lab, subjects were seated in one of the cubicles and were asked to read the experimental instructions they were presented on the computer screen. The instructions described the decision situation and explained how their own decisions and the decisions of the other subjects in their group would affect their payoffs. Several examples illustrated different scenarios. Moreover, they were told that the experimental session comprised three parts, with each part lasting 30 to 60 rounds. They were not told the exact number of rounds of each part, to avoid so-called end-game effects. Subjects were informed that, at the beginning of each part, they would be randomly matched with two other subjects with whom they would stay in one group for the entire part, that their decisions were anonymous, and that they would receive further instructions during the experiment. Then, subjects were asked 10 control questions about the instructions. Questions that elicited at least one wrong answer were read out loud by the experimenter, and the correct answer was explained. Next, the experiment was conducted.

Results and Discussion

Table 3 lists the average $\mathrm{LNI}_{b,3}$ values across the four conditions of experiment 1. In line with hypothesis H1, we find clear evidence for turn-taking in the symmetric

Table 3. LNI$_{h,3}$ Values across Treatments of Experiment 1

	Symmetric			Asymmetric 1			Asymmetric 2			Focal Point		
h	1	2	3	1	2	3	1	2	3	1	2	3
Part 1	3.6	14.1	**20.5**	**23.6**[a]	2.0	15.7	**61.6**	0.5	2.1	0.0[b]	5.5	**35.9**
Part 2	6.5	0.6	**64.0**	**45.6**	11.9	5.2	**69.8**	0.0	7.9	5.6	15.4	**42.3**
Part 3	0.0	9.2	**64.0**	**35.4**	10.6	19.4	**53.8**	4.0	10.0	0.0	7.3	**50.0**
Ø	3.3	8.0	**49.5**	**34.9**	8.2	13.4	**61.7**	1.5	6.7	1.9	9.4	**42.7**

Notes: The average LNI$_{h,3}$ values listed in this table are calculated based on the interaction patterns shown in figures C1 through C12 in the online Supplementary Material, appendix C.
[a]Does not include 9.3 percent of solitary volunteering by a weak subject.
[b]Does not include 8.9 percent of solitary volunteering by a non-focal subject.

condition. On average, 50 percent of the interactions in the symmetric condition were part of a turn-taking 3-sequence, with their proportion increasing over the three parts.

As soon as the stage game becomes asymmetric, the patterns of behavior that emerge are strikingly different from what we find in the symmetric case. In conditions asymmetric 1 and asymmetric 2, turn-taking is identified only in 13 and 7 percent of the interactions, respectively, whereas solitary volunteering by the strongest subject is identified in 35 and 62 percent of the cases, respectively. This is clear evidence in support of hypothesis H2.

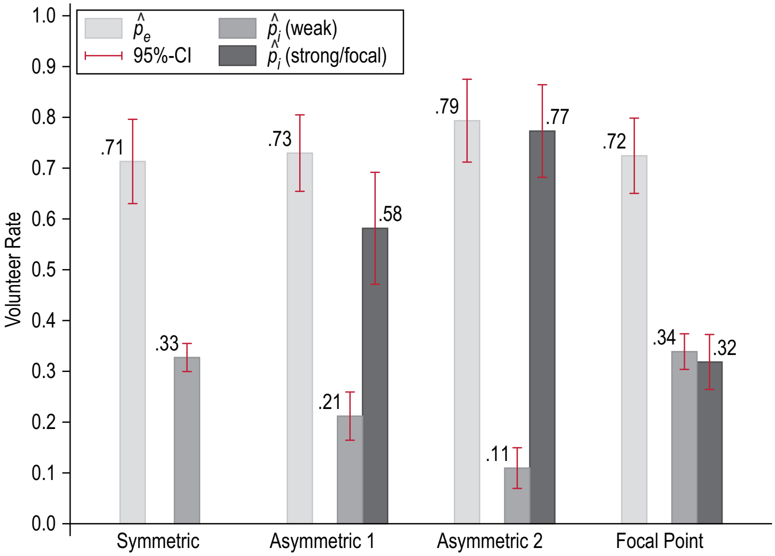
Interestingly, the proportion of solitary volunteering is substantially higher in condition asymmetric 2 than in condition asymmetric 1, a finding our theoretical argument had not allowed us to foresee. This indicates that the degree to which asymmetry facilitates coordinated action depends on the degree of asymmetry itself. In other words, if the strongest actor is not much stronger than the other group members, there seems to be more disagreement with regard to the behavioral pattern subjects should coordinate on than in a condition in which the difference between the strong and the weak subjects is more pronounced.

This last conjecture is supported by the fact that in part three of condition asymmetric 1, turn-taking occurs in over 19 percent of the interactions. Moreover, this finding anticipates our next result. Mere individual distinction is not enough to induce coordination on a single subject to produce the public good throughout. In our focal point treatment, the evidence clearly indicates that subjects take turns (43 percent) rather than coordinate on the focal actor to volunteer solitarily (2 percent). Thus, hypothesis H3 is not supported.

Figure 1 shows the efficiency rates (light-gray bars) and the cooperation rates of weak (medium-gray bars) and strong subjects (dark-gray bars) based on the pooled data from all three parts (figures C13 through C15 in the online Supplementary Material show the same rates for each part separately). Although very different patterns of behavior emerge across the first three experimental conditions, the average efficiency rates are high and statistically indistinguishable from each other ($\chi^2(2) = 2.08, p = 0.353$).

The different behavioral patterns that bring about these high efficiency rates are also reflected in the cooperation rates observed across experimental conditions.

Figure 1. Rates of efficient public good provision and cooperation in experiment 1



In the symmetric condition, where turn-taking dominates, it is not surprising to find a cooperation rate of 33 percent. In the asymmetric conditions, we can distinguish between cooperation rates of weak and strong subjects. The graph clearly shows that it is mostly the strong subject who cooperates, while the weak subjects free-ride. This difference is more pronounced in condition asymmetric 2 than in condition asymmetric 1. The difference in differences is statistically significant ($\chi^2(1) = 15.98$, $p < 0.001$). Finally, in line with turn-taking, cooperation rates are close to 33 percent in the focal point condition for both subject types and do not differ from the cooperation rate observed in the symmetric condition ($\chi^2(2) = 0.50$, $p = 0.778$).

Our first experiment left us wondering why there was a statistically and substantially significant difference in cooperation patterns in the two asymmetric conditions. Our conjecture was that the strong subjects, despite being the ones whose solitary cooperation would produce the public good efficiently, were often reluctant to tacitly agree on them cooperating throughout because this would also lead to them earning less than the other group members, especially in treatment asymmetric 1, where the difference between a strong and weak subject is relatively small. The interaction structure in our second experiment was implemented such that every group member would earn the same (in expectation), even if subjects tacitly agreed that only the strong subject cooperates in every interaction.

Experiment 2

Procedure and Design

Our second experiment differed from our first experiment in that it had only one part comprising 56 rounds, in all conditions the groups of three were disbanded and randomly formed anew after every round,[8] in the two asymmetric conditions the strong subject was determined randomly in every round, and the focal point condition was not implemented (see table 4).

The experiment was conducted at DeSciL, ETH Zurich. Subjects were recruited by e-mail and could register online for one of the available sessions. Subjects were undergraduate students from different departments; 37 percent were female, and they were 21.2 years old on average ($sd = 2.14$). All participants received a show-up fee of CHF 10 and earned CHF 25 on average in the experiment. At the time the experiment was conducted, CHF 1 corresponded to USD 0.98.

We conducted three experimental sessions, with 27 subjects in sessions one and three and 33 subjects in session two ($N = 87$). In each session, subjects were randomized on two of the three experimental conditions and were given treatment-specific instructions on the screen and on paper. The sessions proceeded in a similar way as in experiment 1 (see above). Figures D4 through D6 in the online Supplementary Material show the experimental instructions subjects received as well as the decision and feedback screens they saw during the experiments.

Results and Discussion

Table 5 lists the average $\text{LNI}_{h,3}$ values across the three conditions of experiment 2. The results could not be more clear cut. Unsurprisingly, in the symmetric condition, there is no coordination whatsoever. However, in both asymmetric conditions, in almost every interaction, it is the strongest actor who cooperates while the two other group members free-ride; subjects seem to coordinate on this behavioral pattern almost immediately. Moreover, unlike in experiment 1, there is virtually no difference in solitary volunteering rates across the two asymmetric conditions (94 and 95 percent).

Table 4. Design of Experiment 2

$n = 3$	Symmetric $U = 80c$ $K = 50c$	Asymmetric 1 $U = 80c$ $K_{1,2} = 50c$ $K_3 = 30c$	Asymmetric 2 $U = 80c$ $K_{1,2} = 50c$ $K_3 = 10c$
($m = 56$)	504	616	504

Notes: The table lists the number of interactions that were recorded in each condition of experiment 2. *U* denotes the payoff a subject earns if the public good is produced; *K* denotes the individual cost of producing the public good; and 100c correspond to CHF 1.

Table 5. $\text{LNI}_{h,3}$ Values across Treatments of Experiment 2

	Symmetric			Asymmetric 1			Asymmetric 2		
h	1	2	3	1	2	3	1	2	3
	1.8	2.4	5.2	**94.1**	0.5	0.0	**95.0**	0.0	0.0

Notes: The average $\text{LNI}_{h,3}$ values listed in this table are calculated based on the interaction patterns shown in figures D1 through D3 in the online Supplementary Material, appendix D.

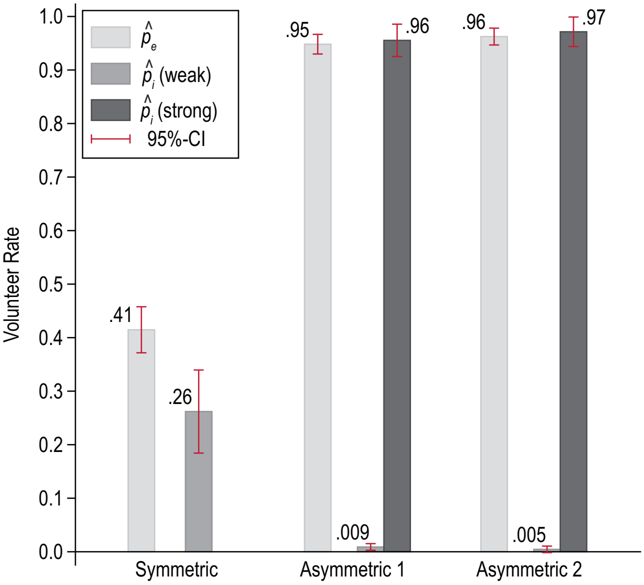
The behavioral patterns that emerged are also clearly reflected in the efficiency and cooperation rates (figure 2). In the symmetric condition, where coordination on turn-taking or solitary volunteering was hardly possible, the average efficiency and cooperation rates are very close to and statistically insignificantly different from the MSE predictions ($\chi^2(1) = 0.98$, $p = 0.323$, and $\chi^2(1) = 1.78$, $p = 0.182$, respectively; see table 7 in the appendix). In the asymmetric conditions, almost perfect efficiency is attained by only the strongest actor cooperating throughout, irrespective of the degree of asymmetry. The difference in differences in cooperation rates across the two asymmetric conditions is, unlike in experiment 1, statistically insignificant ($\chi^2(1) = 0.83$, $p = 0.362$).

In experiment 2, the role of the strongest actor is assigned randomly in every round, which leads to subjects' payoffs being equalized in the long run, eliminating any potential for distributional conflict. This corroborates that distributional conflict was a driving force in the repeated asymmetric VOD in experiment 1 that may have prevented the emergence of even more efficient behavioral patterns. It seems that without distributional conflict, any degree of asymmetry can be an almost perfect coordinating device.

General Discussion and Conclusions

In this study, we start from the assertion that the gap between individual and collective rationality in social dilemmas creates a demand for latent norms, and we argue that in situations in which actors cannot engage in direct negotiation or social exchange, latent norms can emerge tacitly. The type and content of a latent norm will crucially depend on the structure of the social dilemma and the way the social dilemma is repeatedly encountered by actors. We argue, moreover, that

Figure 2. Rates of efficient public good provision and cooperation in experiment 2



game theory provides a powerful framework to describe and analyze the social dilemma of interest and make predictions about the latent norms that could possibly emerge. We therefore suggest categorizing social dilemmas along at least three dimensions. A distinction should be made between (1) cooperation and coordination dilemmas, (2) symmetric and asymmetric dilemmas, and (3) one-shot and repeated dilemmas. The first two dimensions define the structure of the (one-shot) social dilemma, and the third dimension describes the way the social dilemma is repeatedly encountered by actors. It is only in repeated encounters (possibly of various types) that latent norms can emerge tacitly.[9]

The more a latent norm becomes accepted and adhered to, the more non-adherences will cause resentments; once consolidated by a sanctioning mechanism that punishes non-adherences and rewards adherence, the latent norm is likely to turn into a social norm (Guala 2013; Horne 2001; Opp 2004). Thus, without considering the structure of the social dilemma from which a latent norm could emerge, it is difficult to make predictions about the direction and content of social norms. Clearly, if we conceive latent norms in terms of equilibria in repeated games, we run into the equilibrium selection problem that game theory faces and which it has tried to solve by invoking a notion of social norms (Gintis 2009, ch. 7). But for obvious reasons, this cannot be the way to establish a bottom-up theory of norm emergence (Bowles and Gintis 2011, 90). Thus, at this point it might be sensible to peek outside the theorist's box and have a look at what can be observed empirically. This is what we do in this paper.

We conducted two laboratory experiments with the repeated volunteer's dilemma (VOD). We chose the volunteer's dilemma as our stage game because it shares important properties with a range of social dilemmas and has been shown to map many real-world social dilemmas relatively well. In our experiments, we vary the asymmetry of the VOD and the way in which subjects encounter the same interaction situation repeatedly. Our findings seem quite definitive with regard to our parameterization of these two variables. If the symmetric VOD is repeated within the same group of subjects (partner matching), we observe clear patterns of turn-taking behavior emerging over time. That is, after some time, many groups of three subjects start taking turns at cooperating, which leads to the public good being efficiently produced in every interaction. Moreover, turn-taking leads to equal earnings for all group members. If, however, the groups of subjects are disbanded and randomly formed anew after every interaction (random matching), turn-taking cannot emerge. In this case, we observe cooperation rates very close to what we would expect in a mixed-strategy equilibrium.

The differences in the latent norms that emerge in the partner and random matching conditions are more subtle if group members are heterogeneous (i.e., in the asymmetric VOD). In both matching conditions, the modal latent norm is solitary volunteering by the "strongest" group member (i.e., the group member with the lowest cost of cooperation). However, while in the random matching condition, the overall rate of solitary volunteering is above 94 percent—irrespective of the degree of asymmetry, the rate is substantially lower in the partner matching condition and also differs substantially with the degree of asymmetry. In condition asymmetric 1, where the strong subject's costs of cooperation are not much lower than for the other group members, solitary volunteering is observed in 35 percent of interactions, and even some attempts at turn-taking can be observed. In condition asymmetric 2, where the strong subject's costs of cooperation are much lower than for the other group members, solitary volunteering is at 62 percent of all interactions, but it is still far from the 94 percent observed in the random matching condition. These results are summarized in table 6.

Why are there such big differences in the rates of solitary volunteering across the various asymmetric conditions? Our contention is that these differences result from a three-way interaction of the structure of the social dilemma, the way in which subjects meet each other repeatedly, and other-regarding preferences. If we assume that a significant part of our subjects are inequity averse (e.g., Fehr and Schmidt 1999) and dislike earning less (or more) than the other subjects in their group, then we can plausibly explain part of these differences.

Recall first that in the random matching condition also the role of the strong subject is assigned randomly in every interaction. Thus, on average, every subject is the strong one for the same number of interactions, and if it is always (and only) the strong subject who cooperates, all subjects will end up earning the same, namely, 1/3 × 50c + 2/3 × 80c = 70c and 1/3 × 70c + 2/3 × 80c = 77c in condition asymmetric 1 and condition asymmetric 2, respectively. In the partner matching condition, the role of the strong subject is also assigned randomly, but the strong subject stays in their role for the entire series of interactions with the same group members. Therefore, were these groups to agree on only the strong subjects cooperating throughout, subjects' average payoffs would differ within every group. The weak subjects would earn 80c per round and the strong subject would earn 50c or 70c in conditions asymmetric 1 and asymmetric 2, respectively. However, if a strong subject is inequity averse, they might decide to interrupt the sequence of cooperative moves at some point. Likewise, if a weak subject is inequity averse, they might try to take turns with the strong subject. However, such attempts would necessarily lead to discoordination and would occur more frequently in condition asymmetric 1 than in asymmetric 2. This is what we observe.

These results corroborate Olson's (1971[1965]) conjecture that in the presence of privileged (i.e., strong) actors, who have an increased interest in providing the public good by themselves, there may arise a tendency of the weak actors to exploit the strong actors. At the same time, strong actors are more reluctant to be the ones who provide the public good all the time, the more such a course of action leads to unequal cumulative payoffs. As one reviewer pointed out to us,

Table 6. Summary of Main Results

	Partner Matching	Random Matching
Symmetric VOD	(++) Turn-taking	(0) No pattern
Asymmetric 1 VOD	(+) Solitary volunteering	(+++) Solitary volunteering
Asymmetric 2 VOD	(++) Solitary volunteering	(+++) Solitary volunteering

these results also draw interesting boundary conditions on the predictions derived from critical mass theory (Oliver, Marwell, and Teixeira 1985). While strong actors may indeed instigate the production of public goods by making an initial investment that reduces subsequent production costs, their willingness to do so may fade with them being repeatedly engaged in this role. However, it has been shown that strong actors may be compensated for their repeated contributions to public goods in terms of higher status (Willer 2009) and trust in social exchange (Barclay 2004; Fehrler and Przepiorka 2013). This discussion confirms that an explanation of cooperation in human groups requires careful consideration of the interplay between actors' preferences and the structure of the situation in which they interact (Simpson and Willer 2015).

There have been interesting attempts to investigate how the interplay between actors' preferences and their embeddedness in different types of social networks affect the emergence and enforcement of social norms (e.g., Centola, Willer, and Macy 2005; Helbing et al. 2014). Computational approaches are well suited to simulate large populations of agents interacting in complex social networks. However, experimental studies investigating the emergence of social norms among human subjects are still rare (e.g., Centola and Baronchelli 2015; Corten and Buskens 2010; Winter 2014), and the combination of agent-based simulations with laboratory experiments seems particularly promising (Andrighetto et al. 2013). Laboratory experiments and the random assignment of subjects to experimental conditions maximize the internal validity of empirical results. The exclusion of potential confounders is especially important if the starting point of one's investigation is a formal theoretical model, which makes precise predictions about human behavior. However, once the robustness of a particular lab finding has been established, it is equally important to extend its validity and generalizability by showing that similar results can be obtained in non-lab situations that resemble the setups created in the lab (Przepiorka and Berger 2015).

There are different processes by which social norms emerge. Social norms can emerge through direct communication and bargaining between actors (Coleman 1990; Horne 2001), diffusion processes triggered by so-called norm entrepreneurs (Ellickson 2001), changes in relative prices due to taxes levied by the state or technological innovations (Ellickson 2001; Posner 2000), but also as unintended consequences of individuals' behavior (Opp 2004). Although game-theoretic considerations can inform theorizing about all these mechanisms, the approach we advocate in this paper is aimed primarily at mechanisms of tacit emergence of latent norms. What should be clear from the two experiments reported in this paper is that a comprehensive theory of norm emergence should incorporate theorizing about the emergence of latent norms. Latent norms guide behavior in many domains of social life, and once they reach a certain level of acceptance and are enforced by sanctions, they can become social norms.

Notes

1. Coleman (1990) notes that a demand for social norms is a necessary but not a sufficient condition for their emergence; a second and sufficient condition requires that

actors establish mechanisms of sanctioning norm violations to meet the demand. In addition to norm demand and norm enforcement, Horne (2001) identifies norm proliferation as a third necessary condition for the emergence of social norms.

2. Here, we consider "conjoint norms" as defined by Coleman (1990). A conjoint norm has the property that "each actor is simultaneously beneficiary and target of the norm" (Coleman 1990, 247). In contrast, "disjoint norms" may emerge in situations where one group of actors is responsible for the externalities while another group suffers (negative externalities) or profits from externalities. A classic example is a plant at a river that pollutes the water of downstream living residents. Disjoint norms redistribute wealth, are dependent on power and hierarchy, and are often not welfare enhancing (see Hechter and Borland [2001]).
3. A Nash equilibrium is, loosely speaking, a combination of strategies in which no single actor gains from changing her strategy provided that all other actors stick to their strategies. Consider the traffic example. If both actors choose "driving on the left," no actor gains and an actor will even lose if she switches unilaterally to "driving on the right." See, for example, Gintis (2009) or Rasmusen (2007) for a more precise and technical definition.
4. Legal expressions (i.e., formal social norms) establish focal points around which interacting agents coordinate (McAdams 2009, 234). That is, legal actors deliberately create "self-fulfilling expectations that the legally obligatory behavior will occur." Examples are property rights, courts settling disputes in favor of one party, and constitutions, but also technical standards.
5. Ullmann-Margalit (1977, 82) points out that there is a fuzzy transition between impure coordination dilemmas and what she calls situations of favoritism or partiality. Both types of dilemma situations entail distributional conflict, but the latter are clearly characterized by role asymmetry of the interacting agents, a distinction we make in the next footnote.
6. Another type of asymmetry arises if the interacting agents occupy different roles. In game theory, such interactions are known as principal-agent interactions and presuppose an unequal distribution of resources, information, and/or power. This type of asymmetry will be less important for our discussion, because it also presupposes that norms have emerged that justify and reinforce this asymmetry. Ullmann-Margalit (1977, 173) calls these norms partiality norms. Here, we are more interested in how an arbitrary initial distribution of resources could lead to the emergence of partiality (and other) norms.
7. Note that the parameterization of the two experiments described in the following were based partly on the results obtained in a pilot experiment. The pilot experiment is described in detail in the online Supplementary Material, appendix B. In the pilot experiment, we used the symmetric VOD only and varied group size and information feedback experimentally. All our experiments were programmed and conducted with the software z-Tree (Fischbacher 2007).
8. Despite the fact that subjects were fully informed about the other group members' decisions at the end of every round, at the beginning of the next round they received no information about what their new interaction partners had decided in their previous rounds.
9. Even if the parties to the interaction had the possibility and were willing to negotiate or engage in social exchange, establishing a social norm that is directed at behavior in a social dilemma that no one expects to encounter again does not make any sense. This is not to say that a single event cannot trigger the establishment of a social norm (through negotiation and social exchange), which then prevents the event from happening again.

Appendix

Table 7. Theoretical Predictions for the 3-Person VOD

	Symmetric $U_i = 80c$ $K_i = 50c$		Asymmetric $U_i = 80c$ $K_{1,2} = 50c$ $K_3 = 30c$		Asymmetric $U_i = 80c$ $K_{1,2} = 50c$ $K_3 = 10c$	
	MSE	TT	MSE	HS	MSE	HS
$p^*_{1,2}$	0.209	0.333	n/a	0	n/a	0
p^*_3	0.209	0.333	n/a	1	n/a	1
p^*	0.506	1	n/a	1	n/a	1
p^*_e	0.393	1	n/a	1	n/a	1

Notes: The table lists equilibrium point predictions of cooperation rates (p_i^*), rates of public good provision (p^*), and rates of efficient public good provision (p_e^*) for the parameterizations of the 3-person volunteer's dilemma used in the two experiments. These predictions are derived from the theoretical considerations in the Model and Hypotheses section. MSE stands for mixed-strategy equilibrium, TT stands for turn-taking, and HS stands for Harsanyi Selten. Note that in both versions of the asymmetric game an MSE does not exist for the chosen combination of parameters (see the Model and Hypotheses section for details).

About the Authors

Andreas Diekmann is Professor of Sociology at the Swiss Federal Institute of Technology, ETH-Zurich. His research interests focus on theories of social cooperation, experimental game theory, research methods and statistics, and environmental and population sociology. His publications have appeared in the *International Journal of Game Theory*, *Proceedings of the Royal Society* B, *Journal of Conflict Resolution*, *Sociological Methods & Research*, and *American Sociological Review*.

Wojtek Przepiorka is Assistant Professor at the Department of Sociology at Utrecht University. His research interests are organizational behavior, analytical and economic sociology, game theory, and quantitative methodology. His recent publications include "Punitive Preferences, Monetary Incentives, and Tacit Coordination in the Punishment of Defectors Promote Cooperation in Humans" (*Scientific Reports* 5, with A. Diekmann), and "Responsibility Attribution for Collective Decision Makers" (*American Journal of Political Science* 59, with R. Duch and R. Stevenson).

Supplementary Material

Supplementary material is available at *Social Forces* online, http://sf.oxfordjournals.org/.

References

Andrighetto, Giulia, Jordi Brandts, Rosaria Conte, Jordi Sabater-Mir, Hector Solaz, and Daniel Villatoro. 2013. "Punish and Voice: Punishment Enhances Cooperation When Combined with Norm-Signaling." *PLoS ONE* 8(6):e64941.

Archetti, Marco, and Istvan Scheuring. 2011. "Coexistence of Cooperation and Defection in Public Goods Games." *Evolution* 65(4):1140–48.

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

_______. 1986. "An Evolutionary Approach to Norms." *American Political Science Review* 80(4):1095–111.

Barclay, Pat. 2004. "Trustworthiness and Competitive Altruism Can Also Solve the "Tragedy of the Commons." *Evolution and Human Behavior* 25(4):209–20.

Barron, Greg, and Eldad Yechiam. 2002. "Private E-Mail Requests and the Diffusion of Responsibility." *Computers in Human Behavior* 18(5):507–20.

Bendor, Jonathan, and Piotr Swistak. 2001. "The Evolution of Norms." *American Journal of Sociology* 106(6):1493–545.

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

Bliss, Christopher, and Barry Nalebuff. 1984. "Dragon-Slaying and Ballroom Dancing: The Private Supply of a Public Good." *Journal of Public Economics* 25(1–2):1–12.

Bolle, Friedel. 2011. "Passing the Buck." European University Viadrina Frankfurt (Oder), Department of Business Administration and Economics, Discussion Paper No. 308.

Bornstein, Gary, David Budescu, and Shmuel Zamir. 1997. "Cooperation in Intergroup, *N*-Person, and Two-Person Games of Chicken." *Journal of Conflict Resolution* 41(3):384–406.

Bowles, Samuel. 2004. *Microeconomics: Behavior, Institutions, and Evolution*. Princeton, NJ: Princeton University Press.

Bowles, Samuel, and Herbert Gintis. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton, NJ: Princeton University Press.

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

Centola, Damon, Robb Willer, and Michael Macy. 2005. "The Emperor's Dilemma: A Computational Model of Self-Enforcing Norms." *American Journal of Sociology* 110(4):1009–40.

Cherry, Todd L., Stephen J. Cotten, and Stephan Kroll. 2013. "Heterogeneity, Coordination, and the Provision of Best-Shot Public Goods." *Experimental Economics* 16(4):497–510.

Coleman, James S. 1990. *Foundations of Social Theory*. Cambridge: Belknap Press of Harvard University Press.

Corten, Rense, and Vincent Buskens. 2010. "Co-Evolution of Conventions and Networks: An Experimental Study." *Social Networks* 32(1):4–15.

Darley, John M., and Bibb Latané. 1968. "Bystander Intervention in Emergencies: Diffusion of Responsibility." *Journal of Personality and Social Psychology* 8(4):377–83.

Dawes, Robyn M. 1980. "Social Dilemmas." *Annual Review of Psychology* 31:169–93.

Delton, Andrew W., Max M. Krasnow, Leda Cosmides, and John Tooby. 2011. "Evolution of Direct Reciprocity under Uncertainty Can Explain Human Generosity in One-Shot Encounters." *Proceedings of the National Academy of Sciences USA* 108(32):13335–40.

Demsetz, Harold. 1967. "Toward a Theory of Property Rights." *American Economic Review* 57(2):347–59.

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

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

Diekmann, Andreas, and Wojtek Przepiorka. 2015. "Punitive Preferences, Monetary Incentives, and Tacit Coordination in the Punishment of Defectors Promote Cooperation in Humans." *Scientific Reports* 5:10321.

Eger, Thomas, Manfred Kraft, and Peter Weise. 1992. "On the Equilibrium Proportion of Innovation: A Game-Theoretic Approach." *Economics Letters* 38(1):93–7.

Ellickson, Robert C. 1991. *Order Without Law: How Neighbors Settle Disputes*. Cambridge, MA: Harvard University Press.

_______. 2001. "The Evolution of Social Norms: A Perspective from the Legal Academy." In *Social Norms*, edited by Michael, Hechter, and Karl-Dieter Opp, 35–75. New York: Russell Sage Foundation.

Evans, Shane, Hugh Sibly, and John Tisdell. 2013. "Turn-Taking in Finitely Repeated Symmetric Games: Experimental Evidence." School of Economics & Finance, Commerce Building, University of Tasmania, TAS 7001, Australia.

Fehr, Ernst, and Urs Fischbacher. 2004. "Third-Party Punishment and Social Norms." *Evolution and Human Behavior* 25(2):63–87.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114(3):817–68.

Fehrler, Sebastian, and Wojtek Przepiorka. 2013. "Charitable Giving as a Signal of Trustworthiness: Disentangling the Signaling Benefits of Altruistic Acts." *Evolution and Human Behavior* 34(2):139–45.

Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics* 10(2):171–8.

Fudenberg, Drew, and Eric Maskin. 1986. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica* 54(3):533–54.

Garrett, Geoffrey, and Barry R. Weingast. 1993. "Ideas, Interests, and Institutions: Constructing the European Community's Internal Market." In *Ideas and Foreign Policy: Beliefs, Institutions, and Political Change*, edited by Judith, Goldstein, and Robert O. Keohane, 173–206. Ithaca: Cornell University Press.

Gintis, Herbert. 2009. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton, NJ: Princeton University Press.

Guala, Francesco. 2013. "The Normativity of Lewis Conventions." *Synthese* 190(15):3107–22.

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

He, Jun-Zhou, Rui-Wu Wang, and Yao-Tang Li. 2014. "Evolutionary Stability in the Asymmetric Volunteer's Dilemma." *PLoS ONE* 9(8):e103931.

Hechter, Michael, and Elizabeth Borland. 2001. "National Self-Determination: The Emergence of an International Norm." In *Social Norms*, edited by Hechter, Michael, and Karl-Dieter Opp, 186–233. New York: Russell Sage Foundation.

Heckathorn, Douglas D. 1993. "Action and Group Heterogeneity: Voluntary Provision versus Selective Incentives." *American Sociological Review* 58(3):329–50.

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.

Helbing, Dirk, Wenjian Yu, Karl-Dieter Opp, and Heiko Rauhut. 2014. "Conditions for the Emergence of Shared Norms in Populations with Incompatible Preferences." *PLoS ONE* 9(8):e104207.

Horne, Christine. 2001. "Sociological Perspectives on the Emergence of Social Norms." In *Social Norms*, edited by Michael, Hechter, and Karl-Dieter Opp, 3–34. New York: Russell Sage Foundation.

_______. 2007. "Explaining Norm Enforcement." *Rationality and Society* 19(2):139–70.

Kitts, James A. 2006. "Collective Action, Rival Incentives, and the Emergence of Antisocial Norms." *American Sociological Review* 71(2):235–59.

Kollock, Peter. 1998. "Social Dilemmas: The Anatomy of Cooperation." *Annual Review of Sociology* 24:183–214.

Lau, Sau-Him Paul, and Vai-Lam Mui. 2012. "Using Turn Taking to Achieve Intertemporal Cooperation and Symmetry in Infinitely Repeated 2 × 2 Games." *Theory and Decision* 72(2):167–88.

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

Marwell, Gerald, and Pamela Oliver. 1993. *The Critical Mass in Collective Action: A Micro Social Theory*. Cambridge: Cambridge University Press.

McAdams, Richard H. 2009. "Beyond the Prisoners' Dilemma: Coordination, Game Theory, and Law." *Southern California Law Review* 82(2):209–58.

Messick, David M., and Marilynn B. Brewer. 1983. "Solving Social Dilemmas: A Review." In *Review of Personality and Social Psychology*, vol. 4, edited by Ladd, Wheeler, and Phillip R. Shaver, 11–44. Thousand Oaks, CA: Sage.

Nelson, Richard R. 1959. "The Simple Economics of Basic Scientific Research." *Journal of Political Economy* 67(3):297–306.

Oliver, Pamela E., Gerald Marwell, and Ruy Teixeira. 1985. "A Theory of the Critical Mass. I. Interdependence, Group Heterogeneity, and the Production of Collective Action." *American Journal of Sociology* 91(3):522–56.

Olson, Mancur. 1971[1965]. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.

Opp, Karl-Dieter. 1982. "The Evolutionary Emergence of Norms." *British Journal of Social Psychology* 21:139–49.

_______. 2004. "'What 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.

Palfrey, Thomas R., and Howard Rosenthal. 1984. "Participation and the Provision of Discrete Public Goods: A Strategic Analysis." *Journal of Public Economics* 24(2):171–93.

Posner, Eric A. 2000. *Law and Social Norms*. Cambridge, MA: Harvard University Press.

Przepiorka, Wojtek, and Andreas Diekmann. 2013. "Individual Heterogeneity and Costly Punishment: A Volunteer's Dilemma." *Proceedings of the Royal Society B* 280(1759):20130247.

Przepiorka, Wojtek, and Joël Berger. 2015. "The Sanctioning Dilemma: A Quasi-Experiment on Social Norm Enforcement in the Train." Unpublished manuscript, Department of Sociology/ICS, Utrecht University, the Netherlands.

Rapoport, Anatol. 1974. "Prisoner's Dilemma: Recollections and Observations." In *Game Theory as a Theory of Conflict Resolution*, edited by Anatol, Rapoport, 17–34. Dordrecht: Reidel.

Rasmusen, Eric. 2007. *Games and Information: An Introduction to Game Theory*. Malden, MA: Blackwell Publishing.

Raub, Werner, Vincent Buskens, and Rense Corten. 2015. "Social Dilemmas and Cooperation." In *Handbuch Modellbildung und Simulation in den Sozialwissenschaften*, edited by Norman, Braun, and Nicole J. Saam, 597–626. Wiesbaden: Springer VS.

Schelling, Thomas C. 1980. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Simpson, Brent, and Robb Willer. 2015. "Beyond Altruism: Sociological Foundations of Cooperation and Prosocial Behavior." *Annual Review of Sociology* 41:43–63.

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

Ullmann-Margalit, Edna. 1977. *The Emergence of Norms*. Oxford: Oxford University Press.

Voss, Thomas. 2001. "Game-Theoretical Perspectives on the Emergence of Social Norms." In *Social Norms*, edited by Michael, Hechter, and Karl-Dieter Opp, 105–36. New York: Russell Sage Foundation.

Weesie, Jeroen. 1993. "Asymmetry and Timing in the Volunteer's Dilemma." *Journal of Conflict Resolution* 37(3):569–90.

_______. 1994. "Incomplete Information and Timing in the Volunteer's Dilemma: A Comparison of Four Models." *Journal of Conflict Resolution* 38(3):557–85.

Willer, Robb. 2009. "Groups Reward Individual Sacrifice: The Status Solution to the Collective Action Problem." *American Sociological Review* 74(1):23–43.

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

Winter, Fabian. 2014. “Asymmetric Incentives Hinder the Emergence of Norms in the Battle of the Prisoner’s Dilemma.” *Journal of Mathematical Sociology* 38(4):302–20.

Winter, Fabian, Heiko Rauhut, and Dirk Helbing. 2012. “How Norms Can Generate Conflict: An Experiment on the Failure of Cooperative Micro-motives on the Macro-Level.” *Social Forces* 90(3):919–46.

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

Yamagishi, Toshio. 1986. “The Provision of a Sanctioning System as a Public Good.” *Journal of Personality and Social Psychology* 51(1):110–16.

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

_______. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton, NJ: Princeton University Press.