Parochial cooperation and the emergence of signalling norms[†]

Wojtek Przepiorka^{1,*} and Andreas Diekmann^{2,3}

¹ *Department of Sociology / ICS, Utrecht University, Netherlands;*

² *Department of Sociology, University of Leipzig, Germany;*

³ *Department of Humanities, Social and Political Sciences, ETH Zurich, Switzerland*

* Author for correspondence: w.przepiorka@uu.nl

Abstract

Why do people adorn themselves with elaborate body piercings or tattoos, wear obstructing garbs, engage in life-threatening competitions and other wasteful and harmful but socially stipulated practices? Norms of cooperation and coordination, which promote the efficient attainment of collective benefits, can be explained by theories of collective action. However, social norms prescribing wasteful and harmful behaviours have eluded such explanations. We argue that signalling theory constitutes the basis for the understanding of the emergence of such norms, which we call signalling norms. Signalling norms emerge as a result of the uncertainty about who is friend and who is foe. The need to overcoming this uncertainty arises when different groups compete for scarce resources and individuals must be able to identify, trust and cooperate with their fellow group members. After reviewing the mechanisms that explain the emergence of cooperation and coordination norms, we introduce the notion of signalling norms as markers of group distinction. We argue that adherence to signalling norms constitutes a commitment promoting parochial cooperation rather than a quality-revealing signal facilitating partner choice. We formalize our argument in a game-theoretic model that allows us to specify the boundary conditions for the emergence of signalling norms. Our paper concludes with a discussion of potential applications of our model and a comparison of signalling norms with related concepts.

Keywords: Cooperation; Trust; Commitment; Signalling; Group Distinction; Social Norm

[†] To appear in *Philosophical Transactions of the Royal Society B: Biological Sciences*, theme issue on “The language of cooperation: reputation and honest signalling”, edited by Szabolcs Számadó, Daniel Balliet, Francesca Giardini, Eleanor Power and Karoly Takács. We would like to thank Karoly Takács and three anonymous reviewers for their perceptive comments on earlier versions of our paper.

Introduction

Why do people incur costs in terms of time, effort and other resources to create benefits for non-kin others, even in the absence of formal institutions that incentivise such cooperation? Next to psychological mechanisms that explain cooperation by means of, for example, other-regarding preferences (Balliet et al. 2009; Fehr and Schmidt 2006, van Lange 1999), there is a number of social mechanisms that explain cooperation while maintaining the assumption of only self-regarding individuals. It has been argued and shown that people's embeddedness in ongoing social relations can promote cooperation because of expectations of direct and indirect reciprocity (Buskens and Raub 2013, Nowak and Sigmund 2005, Trivers 1971). That is, cooperation emerges when people can expect their deeds to be rewarded and their misdeeds to be punished by either the same interaction partners in repeated encounters (Axelrod 1984, Dal Bó 2005, Fudenberg and Maskin 1986) or by new interaction partners that learn these people's reputations through third-parties (Giardini and Wittek 2019, Przepiorka et al. 2017, Roberts et al. 2021, Sylwester and Roberts 2013, Wu et al. 2016). However, these mechanisms are also in place to enforce collectively inefficient (i.e. wasteful) and even harmful behaviours (Abink et al. 2017, Bliege Bird and Smith 2005, Efferson et al. 2015, Gambetta and Origi 2013, Horne 2009, Opp 2018, Sosis 2005, Willer et al. 2009).

Footbinding, female genital mutilation, extensive body tattoos, lethal initiation rituals, excessive feasts, obstructive dressing codes, and other behaviours that do not seem to serve collective nor individual interests have been enforced through selective incentives. It has been argued that such behaviours exist due to cultural inertia and disappear only with a time-lag in response to changing environmental and social conditions (Mesoudi 2011, Ogburn 1922). This argument suggests, however, that these behaviours were, at some point, functional in serving the collective. Another argument suggests that wasteful and harmful behaviours are part of a repertoire that is invoked by individuals to set group boundaries and facilitate cooperation within groups (i.e. parochial cooperation) (Bernhard et al. 2006, De Dreu et al. 2020, Habyarimana et al. 2009, Handley and Mathew 2020, Sherif et al. 1961). It has been shown that arbitrary markers can evolve to group individuals with similar but unobservable traits that facilitate cooperation between individuals within these groups (Cohen and Haun 2013, Efferson et al. 2008, McNamara 2005). Yet, it remains a puzzle why the distinguishing markers of many social groups can only be acquired through wasteful and harmful behaviours, which bind workforce, mobility, time, and other resources that could be used to create collective benefits.

We argue that signalling theory (Gambetta 2009a, Gintis et al. 2001, Spence 1974, Zahavi 1975, 1977) constitutes the basis for the understanding of the emergence of wasteful and harmful behaviours as markers of group distinction, which we call *signalling norms*. We first review how the emergence of cooperation and coordination norms has been construed in the social and behavioural

sciences. We then introduce the notion of signalling norms and argue that a demand for signalling norms arises because of information asymmetries pertaining to individuals' commitments to a social group (Posner 2000). That is, while cooperation and coordination norms emerge out of a demand for reducing negative externalities of individuals' behaviours, signalling norms emerge out of a demand for reducing uncertainty about who is friend and who is foe.

We devise a game theoretic model which allows identifying the conditions for the emergence and evolution of signalling norms. In particular, we point out that it is the commitment rather than the signalling aspect of the adherence to signalling norms that reduces uncertainty about the cooperative intent of members of a social group. In other words, rather than an unobserved individual trait that can be inferred from norm-abiding behaviour, norm-abiding behaviour impedes cooperative relations with members of other groups and thus increases trust and cooperation within social groups.

We conclude our paper with a discussion of potential applications of our model and a comparison of signalling norms with two related concepts – collective effervescence and credibility-enhancing displays.

Cooperation and coordination norms

Social norms can be defined as rules guiding individual behaviours in social interactions that are sustained by shared expectations and sanctions (Coleman 1990, Hechter and Opp 2001, Horne and Mollborn 2020). Social norms help to overcome social dilemmas that arise when people, by purposefully following their beliefs and preferences, in sum produce outcomes that make them worse off than intended. Norms, such as quiet zones in train compartments, dedicated smoking areas at train stations and airports, signs indicating appropriate walking behaviour on escalators, laws proscribing the use of doping in sports, drinking of alcohol below a certain age, or the production of excessive noise late at night (Coleman 1990, Hechter and Opp 2001, Voss 2001), all serve to reduce the negative externalities that individual actions can impose on others. Such externalities are noise, cigarette smoke, traffic jams, excessive competition, increasing healthcare costs, and so on.¹ Social dilemmas are thus situations in which a gap between individual and collective rationality exists (Dawes 1980, Kollock 1998, Rapoport 1974, van Lange et al. 2013).

We distinguish between cooperation and coordination dilemmas (Ullmann-Margalit 1977, Przepiorka 2021). Cooperation dilemmas refer to situations in which individuals have conflicting interests and, through the best pursuit of their own interests, reach a worse state than they could have reached by following an alternative plan of action. Coordination dilemmas refer to situations in which individuals' interests are (partly) aligned, but the multiple ways of best pursuing these interests

¹ Positive externalities can also create a demand for social norms that prescribe certain behaviours (Coleman 1990).

makes it difficult for these individuals to tacitly coordinate on reaching a better state. Accordingly, a distinction is made between cooperation norms, which help overcome cooperation dilemmas, and coordination norms, which help to overcome coordination dilemmas (Ullmann-Margalit 1977). In the literature, cooperation norms are also called injunctive norms and coordination norms are also called descriptive norms or conventions (Bicchieri 2006, Cialdini et al. 1990, Hawkins et al. 2019, Lewis 1969). However, the terms descriptive norms and conventions imply a lack of normative force guiding individual behaviours, which is often unjustified in the light of existing sanctioning mechanisms that help to overcome coordination dilemmas (Guala and Mitone 2010, Przepiorka et al. 2021). Coordination norms can thus also be injunctive norms. Moreover, the emergence of both cooperation and coordination norms can be explained by means of the same theoretical framework (Figure 1a).

Figure 1: Emergence of social norms

According to Coleman (1990), social norms are collective-level (i.e. macro-level) constructs that emerge when individual behaviours create negative externalities (at the micro-level) to an extent that produces a demand for regulating these behaviours. A particular norm can emerge through bargaining, communication, collective decision making or other mechanisms that we will not discuss further here (Hechter and Opp 2001, Elster 2007). Once in place, these norms are maintained by positive and negative sanctions (i.e. selective incentives).

At this point it is important to make explicit that social norms are directed at focal actions performed by target individuals and benefit the norm beneficiaries (Coleman 1990). For example, a non-smoking norm is directed at the focal action smoking performed by smokers (the targets of the norm) and benefit the non-smokers (the norm beneficiaries). Adherence to the norm benefits the norm beneficiaries directly by reducing the negative externalities created by the targets. However, the targets and beneficiaries of a social norm do not have to be mutually exclusive groups. Even in the previous example one could argue that smokers benefit from non-smoking norms as well. However, the emergence of other types of norms such as dress codes, table manners, jargon, tastes etc., which are often considered as norms of group distinction (Elias 1939, Bourdieu 1979, Veblen 1899), are more difficult to explain with Coleman's argument (Elster 2007).

One could argue that since the focal actions targeted by norms of group distinction are conventional (e.g., veiling as a sign of a woman's religiosity and belonging to the Muslim community), these norms must be construed as coordination norms. In this case, however, it would remain difficult to pinpoint what the negative externalities are that such norms of group distinction serve to reduce. What kind of coordination dilemmas (or cooperation dilemmas) do these norms help to solve? Rather than improving collective outcomes, many of these norms promote collectively wasteful and individually harmful behaviours (e.g., religious garbs that impede mobility and vision).

Although we are not the first to stumble over this puzzle (Abink et al. 2017, Bliege Bird and Smith 2005, Efferson et al. 2015, Gambetta and Origi 2013, Horne 2009, Opp 2018, Sosis 2005, Willer et al. 2009), previous scholarship provides plausible explanations for the maintenance of inefficient norms but is less explicit about their emergence. For example, it has been argued that inefficient norms are adhered to because adherence is positively sanctioned by others, and since people value social relationships and approval, they adhere to these norms and enforce them (Horne 2009, Horne and Mollborn 2020). At the same time, norms of group distinction, by definition, establish group boundaries that prevent the formation of social relations between members of different groups. For example, the demeanour of someone that wears a religious dress covering their entire body (including their head) to indicate their group affiliation is more likely to keep people at bay than establish new relations with members of other groups.

In what follows we propose that introducing a new type of social norms, signalling norms, can help to explain the emergence of wasteful and harmful markers of group distinction and bridge seemingly incompatible strands of literature on social norms (Coleman 1990, Elster 2007).

Signalling norms

We define signalling norms as social norms the adherence to which by person A changes or reinforces person B's beliefs about A's commitment to a social group. The members of this social group are the beneficiaries of the signalling norm. However, unlike for cooperation and coordination norms, the actions that signalling norms are directed at do not necessarily benefit the beneficiaries directly and may even harm other parties or the target of the norm (e.g., signalling norms prescribing illegal behaviour) (Gambetta 2009b). A signalling norm benefits its beneficiaries indirectly by the information adherence to the signalling norm produces. Adherence to a signalling norm produces information about A's commitment to the group of beneficiaries. In other words, signalling norms are social norms the adherence to which is a signal of trustworthiness (i.e. cooperative intent) observed by the beneficiaries of the signalling norm (see also Barclay 2004, Barclay et al. 2021, Fehrler and Przepiorka 2013). The emergence of signalling norms can thus be explained by the prevalence of uncertainty about and the importance of knowing who is friend and who is foe.

Posner (1998, 2000) was the first to suggest that people's adherence to social norms can be conceived as signals of these people's trustworthiness. Posner's main idea is that social norms are the equilibrium outcomes of signalling games (Camerer 2003, Gintis 2009). Such equilibria arise mainly due to a demand for distinction between long-term and short-term types. These types are defined in terms of their discount factors, i.e. their propensity to engage in repeated interactions. Long-term types are more likely to be available for future interactions than short-term types. Since in repeated interactions cooperation can become a self-regarding best response (Axelrod 1984, Buskens and Raub 2013, Dal Bó 2005, Fudenberg and Maskin 1986), long-term types are more trustworthy (Hardin 2002) and thus more desirable interaction partners than short-term types. However, since individuals' discount factors are not directly observable, there is uncertainty about a potential partner's type. In this case, signalling behaviour (e.g., adherence to a social norm) can help to overcome the information gap allowing for repeated, cooperative interactions between long-term types. In order for types to be distinguishable, it must hold that only the long-term types can afford to engage in signalling behaviour.

Posner (2000) applies his theoretical ideas to explain different, seemingly irrational, social behaviours like gift giving, voting, or discrimination. However, Posner conceives of any norm abiding behaviour as a potential signal that helps to distinguish long-term and short-term types (McAdams 2001). By introducing the notion of signalling norms, we aim to specify the conditions under which the emergence and maintenance of social norms can be explained by signalling theory (Gambetta 2009a, Gintis et al. 2001, Spence 1974, Zahavi 1975, 1977).

Like for cooperation and coordination norms, a necessary condition for the emergence of signalling norms is a demand for such norms (Figure 1b). However, the demand for signalling norms does not arise from the need to proscribe actions that produce negative externalities. The demand for

signalling norms arises from trust problems and people's need to belong to a group of similar and likeminded others (Greene 2013).

In this sense, signalling norms solve a particular cooperation dilemma, namely the trust dilemma (Dasgupta 1988), and not by the focal actions they prescribe, but by the information these focal actions produce. Like cooperation and coordination norms, once in place, signalling norms are enforced by sanctions. Since individuals adhering to signalling norms are perceived as more trustworthy by the norm beneficiaries (i.e. members of a particular group), these individuals are positively sanctioned by the greater trust that is placed in them. Correspondingly, individuals that do not adhere to signalling norms are negatively sanctioned by not being trusted. The following example will further illustrate the idea.

Muslim women veil as a sign of their religiosity and belonging to the Muslim community. To pass as a Muslim woman requires much more than wearing the right garb, but the sum of signalling norms Muslim women adhere to makes compliance with these norms a very strong signal of being a true member of the Muslim community. Against the often corroborated predictions of modernization and secularization theory, recent evidence suggests that highly religious Muslim women veil more in response to modernizing forces such as access to education, urban living, contacts with non-Muslims etc. (Aksoy and Gambetta 2016). One explanation that Aksoy and Gambetta put forward for their finding is that these women veil more to signal their religiosity and commitment to the Muslim community because the surrounding temptations jeopardize their reputations for modesty (see also Aksoy and Gambetta 2021).

This example illustrates the signalling and commitment aspects of the adherence to signalling norms (see also Posner 2000: 29, Raub 2004). The signalling aspect captures the possibility that by adhering to a signalling norm, people convey information about their otherwise unobservable traits (e.g., beliefs, values, preferences, interdependencies) (Barclay et al. 2021, Fehrler and Przepiorka 2013, Gambetta 2009a, Gintis et al. 2001). The commitment aspect captures the possibility that by adhering to a signalling norm, people constrain the number of alternative interaction partners or make themselves vulnerable to punishment by third-parties (Gambetta and Przepiorka 2019, Hardin 1996, Shelling 1960). However, conceptualising markers of group distinction as signalling norms makes it difficult to explain the emergence of these norms as signals of people's unobservable traits. The main objection is that sanctions enforcing the adherence to signalling norms could incentivise people to adhere to these norms irrespective of their underlying traits (Eriksson 2019). This is all the more true given the behaviours targeted by signalling norms (e.g., veiling) are mostly conventional and therefore unlikely to be causally related to the unobservable trait of interest (e.g., religiosity, trustworthiness). It is thus the commitment rather than the signalling aspect of the adherence to signalling norms that reduces uncertainty about the cooperative intent of (candidate) members of a

social group.² Yet, the cost-benefit differential of adhering to a signalling norm must be large enough for the beneficiaries of the norm to decide whether to establish and maintain or deny (new) relationships with the targets of the norm.

In line with the commitment aspect, religious practices such as veiling foment discrimination from outgroup members, which makes ingroup members less competitive outside of their group. In other words, by subjecting themselves to discrimination, ingroup members reduce opportunities for cooperation outside of their group and thereby increase the costs of leaving their group (Iannaccone 1994, Sosis 2003). Relatedly, Posner (2000: 97-103) emphasizes the importance of group membership for the evolution of cooperation. He argues that agents signal their loyalty (i.e. long-term cooperative intent) with their ingroup by discriminating against members of the outgroup (see also Takács et al. 2021). Again, such signals of group loyalty are reliable since by discriminating against members of the outgroup, agents forgo potential benefits from cooperation with them.³

One important implication of adherence to signalling norms being about commitment rather than signalling is that the type separating behaviour *per se* does not have to be costly (which does not preclude it from being collectively wasteful and harmful for others) (Przepiorka and Berger 2017). By adhering to signalling norms as markers of group distinction, individuals “burn bridges”, that is, they make themselves unavailable for cooperative interactions with members of an outgroup. This lack of outside opportunities, in turn, makes individuals adhering to signalling norms valid and trustworthy interaction partners for members of the ingroup (i.e. the beneficiaries of the signalling norm). What makes adherence to signalling norms type separating is the individual want to earn benefits from repeated interactions with members of a particular (in)group.⁴ The next section devises a game theoretic model that captures the main tenets of signalling norms.

² The distinction between the signalling and commitment aspect of signalling norms is, in principle, equivalent to the distinction between signalling and screening mechanisms in games with asymmetric and incomplete information (Archetti 2011; Archetti et al. 2011; Kübler et al. 2008; Riley 2001). We thus argue that signalling norms are screening devices that emerge out of a group’s demand to identify trustworthy group members. We prefer the term “signalling norms” over “screening norms” because the screening mechanism can be subsumed under the general and more widely known signalling theory framework. We prefer the term “commitment” over “screening” because the former reflects better the bonding aspect of group memberships.

³ Group identity (i.e. social identity) can thus be conceived as a system of signalling norms the adherence to which separates the members of the group from the non-members. This is an important point to make because it links the literature that conceives of social norms as regulating social interactions (Horne and Molborne 2020) with the literature that equates social norms with social identities (Akerlof and Kranton 2000).

⁴ Signalling norms can also be imposed on members of a particular group by members of another group to facilitate discrimination and oppression (e.g., yellow badges Jews were ordered to wear in German-occupied Poland during

Game-theoretic model

Posner's (1998, 2000) idea that social norms are separating equilibrium outcomes of signalling games that help to distinguish long-term types from short-term types has been formalized in a game theoretic model by Przepiorka and Diekmann (2013) (also see Diekmann 2009, Diekmann and Przepiorka 2010, Przepiorka 2009). The following is an edited and extended version of the main part of the model section in Przepiorka and Diekmann (2013). A more detailed game theoretic analysis is provided in the online supplementary material published along with the paper by Przepiorka and Diekmann (2013).

The binary trust game (Figure 2) represents a cooperation dilemma that cannot be overcome by rational and self-regarding players. While it is rational for the truster not to place trust, both the truster and the trustee would be better off if trust was placed and honoured. In this sense is the trust game similar to the prisoner's dilemma game, although the two games differ in three respects: (1) the trust game is sequential (player 1 moves first and player 2 observes player 1's action) whereas the prisoner's dilemma game is simultaneous (both players decide without knowing what the other player does); (2) by defecting (i.e. not placing trust) in the trust game, player 1 precludes player 2 from making a decision; from (1) and (2) it follows that (3) in the trust game only player 2 can exploit a cooperative player 1 whereas in the prisoner's dilemma game also player 1 can exploit a cooperative player 2. In both games, the Nash equilibrium is reached through (anticipated) mutual defection.

However, the assumption that trustees always abuse trust is neither realistic nor useful. If trusters were certain about the trustworthiness of trustees, the notion of trust would be superfluous. The trust problem arises from trusters' uncertainty about trustees' preferences and constraints, which determine these trustees' decisions in an interaction. In the model presented here, we assume two types of trustees that differ in their discount factors and can be characterised as either long-term or short-term (Posner 1998, 2000). A discount factor stands for a trustee's probability of engaging in another interaction with the truster, but it can also be interpreted as the trustee's time preference. Here we adhere to the former interpretation as it better conveys the commitment aspect of the adherence to signalling norms (see previous section).

WWII). In these cases, the beneficiaries of the signalling norm are the members of the dominant group, who also enforce adherence to the signalling norm.

Figure 2: In the binary trust game, the truster (player 1) first decides whether to place trust (t) or not to place trust ($\neg t$). If the truster decides not to place trust, the interaction is terminated and both parties receive payoff P . If, instead, the truster decides to trust, it is the trustee's turn (player 2) to choose whether or not to honour that trust (h or $\neg h$). If the trustee honours the trust, both players receive payoff R . If the trustee does not honour the trust, the trustee receives payoff T while the truster receives S . The payoffs are ordered so that that the trustee abuses trust if the truster places it (i.e. $T > R > P$) and the truster prefers not to place trust rather than find his or her trust abused (i.e. $R > P > S$). The letters T , R , P , and S stand for temptation, reward, punishment and sucker's payoff and are commonly used to denote payoffs in the prisoner's dilemma game (Rapoport and Chammah 1965).

Given a long-term trustee's discount factor (δ_l), he or she strictly prefers to engage in repeated interactions with a truster over a one-time abuse of trust (i.e. $R/(1 - \delta_l) > T > P$).⁵ Given a short-term trustee's discount factor (δ_s), the expected payoff from repeated interactions with a truster is strictly smaller than his or her payoff from a one-time abuse of trust (i.e. $T > R/(1 - \delta_s) > P$). Hence, the long-term trustee and the short-term trustee will differ in what they do in the trust game, as long as their discount factors differ such that:

$$\delta_l > \frac{T - R}{T} > \delta_s \quad (1)$$

Note that an implicit assumption of our model is that an interaction between a truster and a trustee ends if the truster does not place trust or the trustee abuses placed trust. This assumption implies that only the long-term trustee will be deterred from abusing a truster's trust as he or she would otherwise forgo the higher future benefits from repeated cooperative interactions with the truster. The short-term trustee's potential future benefits from repeated interactions with the truster are too small for him or her to resist the temptation of abusing the truster's trust right away. Of

⁵ The present value of the reward from repeated cooperation is $R/(1 - \delta) = R + \delta R + \delta^2 R + \delta^3 R + \dots$ (see, e.g., Osborne 2009)

course, failing to be trusted by a truster (a second time) does not prevent a short-term trustee to try to gain and abuse the trust of another truster elsewhere. However, from the perspective of the trusters that are members of a particular group, a short-term trustee ceases to be part of the equation after one encounter. This is plausible, for example, if one considers that trustees that were not trusted or that abused trust are banned from any further interactions with the trusters that are members of a particular group (Greif 1989, 1993, Takács et al. 2021; also see footnote 6 on this assumption). However, these trusters remain uncertain about the trustworthiness of every new trustee that attempts to interact with one of them.

A truster's uncertainty about a trustee's type can be accounted for in the trust game with incomplete information (Figure 3) (Camerer and Weigelt 1988, Dasgupta 1988, Voss 1998, Bacharach and Gambetta 2001, Raub 2004). Given the probability α to meet a long-term trustee and the payoff structure, a truster only trusts if the expected payoff from trusting is higher than the payoff from not doing so. That is, if:

$$\alpha R \frac{1}{1 - \delta_l} + (1 - \alpha)S > P \quad (2)$$

After solving Equation 2 for α , it can be shown that a truster will abstain from placing trust if α is less than the threshold value α^* , where:

$$\alpha^* = \frac{(P - S)(1 - \delta_l)}{R - S(1 - \delta_l)} \quad (3)$$

Under these conditions (i.e. $\alpha < \alpha^*$), the truster and the long-term trustee could attain a more beneficial outcome if the trustee were able to communicate his or her type credibly.

Figure 3: In the trust game with incomplete information, nature (N) moves first and determines the trustee's type to be long-term (short-term) with probability α ($1-\alpha$). The probability α (as the entire game) is common knowledge and the fact that the truster does not know whether the trustee is long-term or short-term is denoted by the dashed line. If the truster places trust, the long-term trustee honours trust while a short-term trustee does not. In the first case, the truster receives payoff $R/(1-\delta_l)$; in the second case, the truster's payoff is S . The truster prefers placing trust if the trustee is a long-term type over not placing trust at all and is most reluctant to trust a short-term trustee (i.e. $R/(1-\delta_l) > P > S$).

The model can be extended so that the trustee can initially choose whether or not to send a signal at cost c (Figure 4). In order for the truster to interpret the trustee's type, the signal must be type-separating. Then, a separating equilibrium can emerge in which the long-term trustee sends a signal, a short-term trustee does not (i.e. type-separating behaviour), and the truster places trust only if a signal has been sent by the trustee. The signal is type-separating if the long-term trustee can afford to send it while the short-term trustee cannot.⁶ That is, if:

⁶ In our model, we assume that only mutual cooperation can go on forever but not mutual defection. We make this assumption because it makes our model more comprehensible. We obtain similar results under the assumption that in an infinitely repeated trust game the truster and the trustee use a so-called trigger strategy, which starts defecting forever once the other party abstains from cooperation (see, e.g., Osborne 2009). Then, Equation 1 becomes $\delta_l > \frac{T-R}{T-P} > \delta_s$, Equation 3 becomes $\alpha^* = \frac{P_l - S - \delta_s P_s}{R_l - S - \delta_s P_s}$, and Equation 4 becomes $R_l - P_l > c > T - P$, where R_l , P_l and P_s are, respectively, $R/(1-\delta_l)$, $P/(1-\delta_l)$ and $P/(1-\delta_s)$ (see Supplementary Data 1 in Przepiorka and Diekmann 2013).

$$R \frac{1}{1 - \delta_l} - P > c > T - P \quad (4)$$

In our case, sending a signal means adhering to a signalling norm. Note that Equation 4 implies that the costs of adhering to a signalling norm do not have to differ across trustee types for adherence to be type-separating; what matters is the cost-benefit differential (Johnstone 1997; Gambetta 2009). That is, what makes adherence to a signalling norm type-separating is not the signalling costs alone but the long-term trustee's benefit from interacting with the truster repeatedly, which the short-term trustee does not obtain (Przepiorka and Berger 2017). In other words, adherence to signalling norms is rewarded (i.e. positively sanctioned) through cooperation by members of the ingroup. But under what conditions are signalling norms likely to emerge?

Figure 4: In the signalling trust game too, first nature (N) determines the trustee's type, but before the truster decides whether or not to place trust, the trustee decides whether or not to send a signal. The signal is associated with a cost c for the trustee and this cost is incurred first and irrespective of what the truster and the trustee do thereafter. Before the truster decides whether or not to place trust, they observe the trustee's signalling decision.

If $\alpha < \alpha^*$, the separating equilibrium is collectively more beneficial than the equilibrium without a signalling opportunity. That is, if signalling is not possible, trusters will abstain from placing trust and all will receive payoff P . In the case with signalling opportunity, however, trusters' expected payoff is $\alpha R/(1 - \delta_l) + (1 - \alpha)P$ and trustees receive $\alpha[R/(1 - \delta_l) - c] + (1 - \alpha)P$. If $\alpha > \alpha^*$, a

separating equilibrium can also emerge but does not improve collective gains (i.e. the sum of trusters' and trustees' expected payoffs). Without a signalling opportunity, trusters will always place trust and their expected payoff is $\alpha R/(1 - \delta_I) + (1 - \alpha)S$, with trustees receiving $\alpha[R/(1 - \delta_I)] + (1 - \alpha)T$. It can be shown that these collective gains are always larger than the collective gains in the separating equilibrium, if $c > T - P$. In other words, if $\alpha > \alpha^*$, a pooling equilibrium, where both trustee types do not send signals (i.e. type-pooling behaviour) and trusters always place trust, is collectively more beneficial.

Our model thus implies that signalling norms are more likely to emerge if $\alpha < \alpha^*$, than if $\alpha > \alpha^*$. Put differently, if the group of norm beneficiaries (i.e. the trusters) expect the proportion of trustworthy individuals in the target population (i.e. the trustees) to be relatively small, they will be more likely to introduce wasteful markers of group distinction and demand their fellow group members to adopt these markers. However, note that the members of the target population do not have to be untrustworthy in general for a demand for signalling norms to emerge among the group of norm beneficiaries; it is the group of norm beneficiaries that defines the trust problem by determining the benefits that accrue through repeated interactions of their members among each other and the costs that accrue through these members' interactions with outgroup members. How norm beneficiaries define the trust problem that shapes their demand for signalling norms will further depend on their competition with other groups seeking access to scarce resources. Finally, both access to and competition for scarce resources will depend on competing groups' relative power, external shocks such as natural and human-made disasters and technological innovations (see also Romano et al. 2021). Signalling norms must therefore be seen as a side-product of group formation processes that take place in a wider social context that shapes cooperation and conflict between groups (Hess and Hagen 2021, Sherif et al. 1961, Takács et al. 2021, van Tubergen 2020: 277-288).

Discussion

In her seminal book, Ullmann-Margalit (1977) distinguishes between three types of social norms. These are norms of cooperation, norms of coordination, and norms of partiality. While norms of partiality concern the distribution of resources, cooperation and coordination norms emerge spontaneously or by design and may sustain cooperation in social dilemma situations (Coleman 1990). However there are also social norms that do not fit into these three categories (Elster 2007). Examples are norms of etiquette, dress codes, table behaviour and other kinds of seemingly wasteful and even self-harming behaviours. We call these norms “signalling norms” to convey the idea that adherence to these norms signals a person's commitment to a particular group. The emergence of signalling norms can be explained by models stemming from signalling theory developed in biology and economics (Spence 1974, Zahavi 1975, 1977).

By combining insights from signalling theory with rational choice theories of the emergence of social norms from sociology, we highlight how wasteful but socially stipulated behaviour can promote the evolution of parochial cooperation in humans. Signalling norms emerge when a group of trusters is uncertain about the trustworthiness of others they want to interact with in their common interest. In accordance with Posner (1998, 2000), we define trustworthy types by their long-term interests in repeated interactions while untrustworthy types have short-term interests only. Thus, long-term types are expected to reciprocate trustful actions while short-term types defect right from the beginning of an interaction. Note that the short-term and long-term types are not defined through their other-regarding preferences. The model that we outlined here can explain trusters placing trust and trustees behaving trustworthily even if both parties only pursue their own material interests. Put differently, signalling norms facilitate trust and cooperation even among *Homines oeconomici*. Of course, the model can be extended by including other-regarding preferences of honesty (Przepiorka and Berger 2017). Long-term types may develop an intrinsic value of honest behaviour making cooperation more likely even with members of an outgroup.

To be sure, signalling theory and our model cannot predict the specific content of a signalling norm that might evolve in a situation requiring trust. Explanations that, for example, tattoo norms had evolved in delinquent communities for identifying their members are *ex-post*. Any type of behaviour could have evolved for this purpose and a variety of factors contribute to the emergence of particular norms. However, the theory can predict in which situations, determined by the parameters of our model, signalling norms are more likely to emerge. Consider once more the case that the proportion of long-term interaction partners is below the threshold denoted by α^* (Equation 3). Then, the theory predicts that the probability of signalling norms arising and the strictness of compliance with these norms are larger than in a situation where the proportion of long-term types is above the threshold. Moreover, the threshold may vary; it depends on the game payoff parameters R , P , S and the “intensity” of the long-term interest δ_l (the discount factor). These parameters can be systematically varied to derive hypotheses that can be tested in controlled experiments (Przepiorka and Diekmann 2013). Observations from field or survey studies are also valuable for inspecting the theory. For example, Aksoy and Gambetta’s (2016) explanation for why Muslim women veil more in a more secular environment is in accordance with our model predictions. Relatedly, Patel (2012) observes that the proliferation of veiling instigates pious groups to invest more in signals to make a distinction between more and less religious groups. The latter is an example of a signalling norm that is adjusted in response to a reoccurring need for distinguishing true from would-be believers and free-riders.

Although signalling norms are more likely to emerge in situations requiring trust, the trustees must be willing to pay the price for the signal. The price of truthful information has to be large enough to exclude cheaters and free-riders. Consequently, signalling norms are not always the most

efficient solution to trust problems. There are other means to reduce uncertainty about trustees' trustworthiness in social exchange. Contract law and legal sanctions deter cheating but come with transaction and enforcement costs. Legal solutions to trust problems are also not viable in extra-legal contexts. Signalling norms are thus more likely to emerge if the costs of contracts and legal enforcement are too high or formal trust-building mechanism absent altogether. We suppose that this is one reason for why signalling norms are common in delinquent communities. For example, extensive body tattoos used to denote membership in a criminal organization such as the Japanese yakuza or the Russian vory, and these signs of membership were protected through punishment of non-members that adopted them (Gambetta 2009, Varese 2005). In various sub-cultures and in delinquent groups signalling norms promote social order in the absence of state-sanctioned enforcement mechanisms (Ellickson 1991).

In the same vein, it has been argued that in pre-state societies, religions contributed to solving cooperation problems arising in increasingly large human groups (van Tubergen 2020: 462-464). Human groups, by defining markers of group distinction related to religious beliefs and practices, established group identities through which members of different groups could be discerned. Since being a member of a religious group often came with considerable benefits, either provided directly by the collective or indirectly through group reputations (Weber 2002 [1920]), religious groups had to protect their identifying markers against free-riders, who would adopt these markers to earn the benefits without contributing to the collective good. Already Durkheim (2001 [1912]) suggested that religious groups, to overcome the free-rider problem, established various kinds of collective activities (e.g., communal praying) that would allow them to monitor the commitment of their members to the group. Durkheim called these religious activities *collective effervescence*. However, it has been objected that collective effervescence, although effective in promoting in-group solidarity (Power 2018), do not signal religious identities beyond the group in which they are practiced. Henrich (2009) argued that members of religious communities could signal their commitment more broadly by engaging in so-called credibility-enhancing displays (CREDs) (see also Norenzayan 2015, Power 2017, Dumas et al. 2021). The more costly it is to perform CREDs, the stronger is the signal that the person performing them is committed to the beliefs and values of the group. CREDs can range from cheap (e.g., wearing a certain garb) to costly (e.g., praying five times a day) to very costly (e.g., self-mutilation) and are performed by individuals in public but not necessarily as part of a collective ritual.

Both collective effervescence and CREDs are related to signalling norms in that they signal individuals' commitments to a group. However, collective effervescence and CREDs are limited to religious identities and the performance of CREDs is mostly voluntary. Signalling norms thus offer

explanations for the widespread use of markers of group distinction by individuals also in other than religious contexts (e.g., social classes, criminal organizations).

References

- Abbink, K., Gangadharan, L., Handfield, T., & Thrasher, J. (2017). Peer punishment promotes enforcement of bad social norms. *Nature Communications*, 8, 609. doi:10.1038/s41467-017-00731-0
- Akerlof, G. A., & Kranton, R. E. (2000). Economics and Identity. *The Quarterly Journal of Economics*, 115(3), 715-753.
- Aksoy, O., & Gambetta, D. (2016). Behind the Veil: The Strategic Use of Religious Garb. *European Sociological Review*, 6(1), 792-806. doi:10.1093/esr/jcw035
- Aksoy, O., & Gambetta, D. (2021). The Politics behind the Veil. *European Sociological Review*, 37(1), 67-88. doi:10.1093/esr/jcaa035
- Archetti, M. (2011). Contract theory for the evolution of cooperation: The right incentives attract the right partners. *Journal of Theoretical Biology*, 269(1), 201-207. doi:10.1016/j.jtbi.2010.10.025
- Archetti, M., Scheuring, I., Hoffman, M., Frederickson, M. E., Pierce, N. E., & Yu, D. W. (2011). Economic game theory for mutualism and cooperation. *Ecology Letters*, 14(12), 1300-1312. doi:10.1111/j.1461-0248.2011.01697.x
- Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.
- Bacharach, M., & Gambetta, D. (2001). Trust in Signs. In K. S. Cook (Ed.), *Trust in Society* (pp. 148-184). New York: Russell Sage Foundation.
- Balliet, D., Parks, C., & Joireman, J. (2009). Social Value Orientation and Cooperation in Social Dilemmas: A Meta-Analysis. *Group Processes & Intergroup Relations*, 12(4), 533-547. doi:10.1177/1368430209105040
- Barclay, P. (2004). Trustworthiness and competitive altruism can also solve the “tragedy of the commons”. *Evolution and Human Behavior*, 25(4), 209-220.
- Barclay, P., Bliege Bird, R., Roberts, G., & Számadó, S. (2021). Cooperating to show that you care: costly helping as an honest signal of fitness interdependence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, forthcoming.
- Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, 442(7105), 912-915. doi:10.1038/nature04981

- Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.
- Bliege Bird, R., & Smith, E. A. (2005). Signaling Theory, Strategic Interaction, and Symbolic Capital. *Current Anthropology*, 46(2), 221-248.
- Bourdieu, P. (1979). *La distinction: Critique sociale du jugement*. Paris: Les Éditions du Minuit.
- Buskens, V., & Raub, W. (2013). Rational Choice Research on Social Dilemmas: Embeddedness Effects on Trust. In R. Wittek, T. Snijders, & V. Nee (Eds.), *The Handbook of Rational Choice Social Research* (pp. 113-150). Stanford, CA: Stanford University Press.
- Camerer, C., & Weigelt, K. (1988). Experimental Test of a Sequential Equilibrium Reputation Model. *Econometrica*, 56(1), 1-36.
- Camerer, C. F. (2003). *Behavioral Game Theory*. Princeton (NJ): Princeton University Press.
- Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places. *Journal of Personality and Social Psychology*, 58(6), 1015-1026. doi:10.1037/0022-3514.58.6.1015
- Cohen, E., & Haun, D. (2013). The development of tag-based cooperation via a socially acquired trait. *Evolution and Human Behavior*, 34(3), 230-235. doi:10.1016/j.evolhumbehav.2013.02.001
- Coleman, J. S. (1990). *Foundations of Social Theory*. Cambridge (MA): The Belknap Press of Harvard University Press.
- Dal Bó, P. (2005). Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games. *American Economic Review*, 95(5), 1591-1604.
- Dasgupta, P. (1988). Trust as a Commodity. In D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relations* (pp. 49-72). Oxford: Basil Blackwell.
- Dawes, R. M. (1980). Social Dilemmas. *Annual Review of Psychology*, 31, 169-193. doi:10.1146/annurev.ps.31.020180.001125
- De Dreu, C. K. W., Gross, J., Fariña, A., & Ma, Y. (2020). Group Cooperation, Carrying-Capacity Stress, and Intergroup Conflict. *TRENDS in Cognitive Sciences*, 24(9), 760-776. doi:10.1016/j.tics.2020.06.005
- Diekmann, A. (2009). *Spieltheorie: Einführung, Beispiele, Experimente*. Reinbek: Rowohlt.
- Diekmann, A., & Przepiorka, W. (2010). Soziale Normen als Signale. Der Beitrag der Signaling-Theorie. In G. Albert & S. Sigmund (Eds.), *Soziologische Theorie kontrovers. Sonderheft der Kölner Zeitschrift für Soziologie und Sozialpsychologie* (pp. 220-237). Wiesbaden: VS Verlag.

- Dumas, M., Power, E. A., & Barker, J. (2021). When does reputation lie? Dynamic feedbacks between costly signals, social capital, and social prominence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, forthcoming.
- Durkheim, E. (2021 [1912]). *The Elementary Forms of Religious Life*. Oxford: Oxford University Press.
- Efferson, C., Lalive, R., & Fehr, E. (2008). The Coevolution of Cultural Groups and Ingroup Favoritism. *Science*, 321(5897), 1844-1849. doi:10.1126/science.1155805
- Efferson, C., Vogt, S., Elhadi, A., Ahmed, H. E. F., & Fehr, E. (2015). Female genital cutting is not a social coordination norm. *Science*, 349(6255), 1446-1447. doi:10.1126/science.aaa7978
- Elias, N. (1994 [1936]). *The Civilizing Process. Sociogenetic and Psychogenetic Investigations*. Oxford: Blackwell.
- Ellickson, R. C. (1991). *Order Without Law: How Neighbors Settle Disputes*. Cambridge, MA: Harvard University Press.
- Elster, J. (2007). *Explaining Social Behavior: More nuts and bolts for the social sciences*. Cambridge: Cambridge University Press.
- Eriksson, L. (2019). Social Norms as Signals. *Social Theory and Practice*, 45(4), 579-599. doi:10.5840/soctheorpract2019123072
- Fehr, E., & Schmidt, K. M. (2006). The Economics of Fairness, Reciprocity and Altruism – Experimental Evidence and New Theories. In S.-C. Kolm & J. M. Ythier (Eds.), *Handbook of the Economics of Giving, Altruism and Reciprocity, Volume 1* (pp. 615-691). Amsterdam: North Holland.
- Fehrler, S., & Przepiorka, W. (2013). Charitable giving as a signal of trustworthiness: Disentangling the signaling benefits of altruistic acts. *Evolution and Human Behavior*, 34(2), 139-145. doi:10.1016/j.evolhumbehav.2012.11.005
- Fudenberg, D., & Maskin, E. (1986). The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. *Econometrica*, 54(3), 533-554.
- Gambetta, D. (2009). *Codes of the Underworld: How Criminals Communicate*. Princeton: Princeton University Press.
- Gambetta, D. (2009). Signaling. In P. Hedström & P. Bearman (Eds.), *The Oxford Handbook of Analytical Sociology* (pp. 168-194). Oxford: Oxford University Press.
- Gambetta, D., & Origgi, G. (2013). The LL game: The curious preference for low quality and its norms. *Politics, Philosophy & Economics*, 12(1), 3-23. doi:10.1177/1470594X11433740

- Gambetta, D., & Przepiorka, W. (2019). Sharing compromising information as a cooperative strategy. *Sociological Science*, 6, 352-379. doi:10.15195/v6.a14
- Giardini, F., & Wittek, R. (Eds.). (2019). *The Oxford Handbook of Gossip and Reputation*. Oxford: Oxford University Press.
- Gintis, H. (2009). *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction (2nd ed.)*. Princeton (NJ): Princeton University Press.
- Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly Signaling and Cooperation. *Journal of Theoretical Biology*, 213, 103-119.
- Greene, J. (2013). *Moral Tribes: Emotion, Reason and the Gap Between Us and Them*. New York: Penguin Books.
- Greif, A. (1989). Reputation and Coalitions in Medieval Trade: Evidence on the Maghribi Traders. *Journal of Economic History*, 49(4), 857-882. doi:10.1017/S0022050700009475
- Greif, A. (1993). Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition. *American Economic Review*, 83(3), 525-548.
- Guala, F., & Mittone, L. (2010). How history and convention create norms: An experimental study. *Journal of Economic Psychology*, 31(4), 749-756. doi:10.1016/j.joep.2010.05.009
- Habyarimana, J., Humphreys, M., Posner, D. N., & Weinstein, J. M. (2009). *Coethnicity: Diversity and the Dilemmas of Collective Action*. New York: Russell Sage Foundation.
- Handley, C., & Mathew, S. (2020). Human large-scale cooperation as a product of competition between cultural groups. *Nature Communications*, 11, 702. doi:10.1038/s41467-020-14416-8
- Hardin, R. (1996). Trustworthiness. *Ethics*, 107(1), 26-42.
- Hardin, R. (2002). *Trust and Trustworthiness*. New York: Russell Sage Foundation.
- Hawkins, R. X. D., Goodman, N. D., & Goldstone, R. L. (2019). The Emergence of Social Norms and Conventions. *TRENDS in Cognitive Sciences*, 23(2), 158-169. doi:10.1016/j.tics.2018.11.003
- Hechter, M., & Opp, K.-D. (Eds.). (2001). *Social Norms*. New York: Russell Sage Foundation.
- Henrich, J. (2009). The evolution of costly displays, cooperation and religion: credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior*, 30(4), 244-260. doi:10.1016/j.evolhumbehav.2009.03.005

- Hess, N., & Hagen, E. (2021). Competitive gossip: The impact of domain, resource value, resource scarcity, and coalitions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, forthcoming.
- Horne, C. (2009). *The Rewards of Punishment: A Relational Theory of Norm Enforcement*. Stanford (CA): Stanford University Press.
- Horne, C., & Mollborn, S. (2020). Norms: An Integrated Framework. *Annual Review of Sociology*, 46, 467-487. doi:10.1146/annurev-soc-121919-054658
- Iannaccone, L. R. (1994). Why Strict Churches Are Strong. *American Journal of Sociology*, 99(5), 1180-1211.
- Johnstone, R. A. (1997). The Evolution of Animal Signals. In J. R. Krebs & N. B. Davies (Eds.), *Behavioral Ecology: An Evolutionary Approach* (pp. 155-178). Oxford: Blackwell.
- Kollock, P. (1998). Social Dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology*, 24, 183-214.
- Lewis, D. (1969). *Convention: a philosophical study*. Cambridge: Harvard University Press.
- McAdams, R. H. (2001). Signaling Discount Rates: Law, Norms, and Economic Methodology. *Yale Law Journal*, 110(4), 625-689.
- McNamara, T. (2005). 21st Century Shibboleth: Language Tests, Identity and Intergroup Conflict. *Language Policy*, 4, 351–370. doi:10.1007/s10993-005-2886-0
- Mesoudi, A. (2011). *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences*. Chicago, IL: University of Chicago Press.
- Norenzayan, A. (2013). *Big Gods: How Religion Transformed Cooperation and Conflict*. Princeton, NJ: Princeton University Press.
- Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437(7063), 1291-1298.
- Ogburn, W. F. (1922). *Social change with respect to culture and original nature*. New York: Huebsch.
- Opp, K.-D. (2018). Externalities, Social Networks, and the Emergence of Norms: A Critical Analysis and Extension of James Coleman's Theory. *Social Research: An International Quarterly*, 85(1), 167-196.
- Osborne, M. J. (2009). *An Introduction to Game Theory, International Edition*. Oxford: Oxford University Press.

- Patel, D. S. (2012). Concealing to reveal: The informational role of Islamic dress. *Rationality and Society*, 24(3), 295-323. doi:10.1177/1043463112440683
- Posner, E. A. (1998). Symbols, Signals, and Social Norms in Politics and the Law. *Journal of Legal Studies*, 27, 765-798.
- Posner, E. A. (2000). *Law and Social Norms*. Cambridge (MA): Harvard University Press.
- Power, E. A. (2017). Discerning devotion: Testing the signaling theory of religion. *Evolution and Human Behavior*, 38(1), 82-91. doi:10.1016/j.evolhumbehav.2016.07.003
- Power, E. A. (2018). Collective ritual and social support networks in rural South India. *Proceedings of the Royal Society B*, 285(1879), 20180023. doi:10.1098/rspb.2018.0023
- Przepiorka, W. (2009). *Reputation and Signals of Trustworthiness in Social Interactions*. Zurich: ETH Zurich, Dissertation Nr. 18649. doi:10.3929/ethz-a-005994644
- Przepiorka, W. (2021). Game-theoretic models. In G. Manzo (Ed.), *Research Handbook on Analytical Sociology* (pp. forthcoming). Cheltenham: Edward Elgar.
- Przepiorka, W., & Berger, J. (2017). Signaling theory evolving: Signals and signs of trustworthiness in social exchange. In B. Jann & W. Przepiorka (Eds.), *Social dilemmas, institutions and the evolution of cooperation* (pp. 373-392). Berlin: De Gruyter Oldenbourg.
- Przepiorka, W., Andrighetto, G., Diekmann, A., Székely, A., & Tummolini, L. (2021). How norms emerge from conventions (and change). Unpublished manuscript, Department of Sociology/ICS, Utrecht University, the Netherlands.
- Przepiorka, W., & Diekmann, A. (2013). Temporal Embeddedness and Signals of Trustworthiness: Experimental Tests of a Game Theoretic Model in the United Kingdom, Russia, and Switzerland. *European Sociological Review*, 29(5), 1010-1023. doi:10.1093/esr/jcs078
- Przepiorka, W., Norbutas, L., & Corten, R. (2017). Order without law: Reputation promotes cooperation in a cryptomarket for illegal drugs. *European Sociological Review*, 33(6), 752-764. doi:10.1093/esr/jcx072
- Rapoport, A. (1974). Prisoner's Dilemma – Recollections and Observations. In A. Rapoport (Ed.), *Game Theory as a Theory of Conflict Resolution* (pp. 17-34). Dordrecht: Reidel.
- Rapoport, A., & Chammah, A. M. (1965). *Prisoner's Dilemma. A Study of Conflict and Cooperation*. Ann Arbor: University of Michigan Press.
- Raub, W. (2004). Hostage Posting as a Mechanism of Trust: Binding, Compensation, and Signaling. *Rationality and Society*, 16(3), 319-365.

- Riley, J. G. (2001). Silver Signals: Twenty-Five Years of Screening and Signaling. *Journal of Economic Literature*, 39(2), 432-478.
- Roberts, G., Raihani, N., Bshary, R., Manrique, H. M., Farina, A., Samu, F., & Barclay, P. (2021). The benefits of being seen to help others: indirect reciprocity and reputation-based partner choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, forthcoming.
- Romano, A., Giardini, F., Columbus, S., De Kwaadsteniet, E., Kisfalusi, D., Triki, Z., . . . Hagel, K. (2021). Reputation and Socio-Ecology in Humans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, forthcoming.
- Schelling, T. C. (1960). *The Strategy of Conflict*. Cambridge (MA): Harvard University Press.
- Sherif, M., Harvey, O. J., White, J. B., & Sherif, C. W. (1961). *Intergroup Conflict and Cooperation: The Robber's Cave Experiment*. Norman, OK: University Book Exchange.
- Sosis, R. (2003). Why aren't we all Hutterites? Costly Signaling Theory and Religious Behavior. *Human Nature*, 14(2), 91-127.
- Sosis, R. (2005). Does Religion Promote Trust? The Role of Signaling, Reputation, and Punishment. *Interdisciplinary Journal of Research on Religion*, 1, 1-30.
- Spence, M. A. (1974). *Market Signaling: Informational Transfer in Hiring and Related Screening Processes*. Cambridge (MA): Harvard University Press.
- Sylwester, K., & Roberts, G. (2013). Reputation-based partner choice is an effective alternative to indirect reciprocity in solving social dilemmas. *Evolution and Human Behavior*, 34(3), 201-206. doi:10.1016/j.evolhumbehav.2012.11.009
- Takács, K., Gross, J., Testori, M., Letina, S., Kenny, A., Power, E., & Wittek, R. (2021). Networks of Reliable Reputations and Cooperation: A Review. *Philosophical Transactions of the Royal Society B: Biological Sciences*, forthcoming.
- Trivers, R. L. (1971). The Evolution of Reciprocal Altruism. *Quarterly Review of Biology*, 46(1), 35-57.
- Ullmann-Margalit, E. (1977). *The Emergence of Norms*. Oxford: Oxford University Press.
- van Lange, P. A. M. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology*, 77(2), 337-349. doi:10.1037/0022-3514.77.2.337
- van Lange, P. A. M., Joireman, J., Parks, C. D., & van Dijk, E. (2013). The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*, 120(2), 125-141. doi:10.1016/j.obhdp.2012.11.003

- van Tubergen, F. (2020). *Introduction to Sociology*. London: Routledge.
- Varese, F. (2005). *The Russian Mafia: Private Protection in a New Market Economy* Oxford: Oxford University Press.
- Veblen, T. (1994 [1899]). *The Theory of the Leisure Class: An Economic Study of Institutions*. New York: Dover Publications.
- Voss, T. (1998). Vertrauen in modernen Gesellschaften. Eine spieltheoretische Analyse. In R. Metze, K. Mühler, & K.-D. Opp (Eds.), *Der Transformationsprozess: Analysen und Befunde aus dem Leipziger Institut für Soziologie* (pp. 91-129). Leipzig: Leipziger Universitätsverlag.
- Voss, T. (2001). Game-Theoretical Perspectives on the Emergence of Social Norms. In M. Hechter & K.-D. Opp (Eds.), *Social Norms* (pp. 105-136). New York: Russell Sage Foundation.
- Weber, M. (2002 [1920]). *The Protestant Ethic and the Spirit of Capitalism*. Oxford: Blackwell.
- Willer, R., Kuwabara, K., & Macy, M. W. (2009). The False Enforcement of Unpopular Norms. *American Journal of Sociology*, 115(2), 451-490.
- Wu, J., Balliet, D., & van Lange, P. A. M. (2016). Gossip Versus Punishment: The Efficiency of Reputation to Promote and Maintain Cooperation. *Scientific Reports*, 6, 23919.
doi:10.1038/srep23919
- Zahavi, A. (1975). Mate selection-A selection for a handicap. *Journal of Theoretical Biology*, 53(1), 205-214.
- Zahavi, A. (1977). The cost of honesty: Further Remarks on the Handicap Principle. *Journal of Theoretical Biology*, 67(3), 603-605.