Bonn Econ Discussion Papers

Discussion Paper 11/2006

Moral Norms in a Partly Compliant Society

by

Sebastian Kranz

May / June 2006

Bonn Graduate School of Economics
Department of Economics
University of Bonn
Adenauerallee 24 - 42
D-53113 Bonn

The Bonn Graduate School of Economics is
sponsored by the

Deutsche Post World Net
MAIL EXPRESS LOGISTICS FINANCE

Moral Norms in a Partly Compliant Society

Sebastian Kranz*
University of Bonn

June 2006

Abstract

This paper analyses competition of moral norms and institutions in a society where a fixed share of people unconditionally complies with norms and the remaining people act selfishly. Whether a person is a norm-complier or selfish is private knowledge. A model of voting-by-feet shows that those norms and institutions arise that maximize expected utility of norm-compliers, taking into account selfish players' behavior. Such *complier optimal norms* lead to a simple behavioral model that, when combined with preferences for equitable outcomes, is in line with the relevant stylized facts from a wide range of economic experiments, like reciprocal behavior, costly punishment, the role of intentions, giving in dictator games and concerns for social efficiency. The paper contributes to the literature on voting-by-feet, institutional design, ethics and social preferences.

Keywords: *moral norms, social preferences, reciprocity, fairness, rule utilitarianism, voting-by-feet, cultural evolution, golden rule, social norms*

JEL Classifications: A13, C7, D02, D63, D64, D71, D8, Z13

*Address: Bonn Graduate School of Economics, University of Bonn, Adenaueralle 24-26, 53113 Bonn, Germany; e-mail: skranz@uni-bonn.de. For extremely valuable discussions and suggestions I am especially grateful to Paul Heidhues, Georg Nöldeke and Susanne Ohlendorf. Further, I would like to thank many fellow PhD students from the BGSE, as well as Armin Falk, Thomas Gall, Werner Güth, Frank Riedel, Karl Schlag, Reinhard Selten, Urs Schweizer and participants of the BGSE workshop in Bonn and the EDP Jamboree 2005 in Florence for very helpful comments. Special thanks to Heather for improving my grammar, and to my family, friends and neighbours, who endured endless questioning of how they would act in this or that situation.

1 Introduction

All human societies have certain institutions that structure the way people interact with each other. These interactions are usually governed by moral- or social norms, which are explicit or implicit rules describing how members of a community are supposed to act (see Elster 1989 for a survey). The traditional economists' approach claims that people only comply with such norms when it is in their selfish best interest. Experimental evidence shows, however, that not all people always act selfishly: For example, many people perform costly punishment in one-shot interactions, or contribute to a public good in the absence of punishment (see e.g. Fehr and Gaechter, 2000).

We analyse which moral norms and institutions develop in societies inhabited by two types of people: *compliers*, who unconditionally comply with moral norms by intrinsic motivation, and selfish people, who break norms whenever this is individually rational.

In our model, a community (e.g. a religious group or a city) is populated by a continuum of inhabitants who randomly meet and interact with each other. Social interaction is described by a normal- or extensive-form game[1] (the institution), and a community's moral norm is a fixed, commonly known strategy profile of this game. Whether a player is a complier or selfish is private knowledge, but the share of compliers in a community is commonly known. Hence, selfish players maximize expected payoffs taking into account other selfish players' behavior, the norm and the compliers' share in a community. A norm is *complier optimal* if it maximizes expected utility of all compliers, taking into account the induced equilibrium strategies of selfish players (see Section 2).[2]

Complier optimal norms can be analyzed for any preferences people may have over material payoffs, like risk-aversion, loss-aversion, envy, altruism, spitefulness, etc. We illustrate them for linear utility in own monetary payoffs and for a simple envy-augmented utility function. Our clear predictions capture the relevant stylized facts from a wide range of economic experiments, e.g. costly punishment, conditional cooperation, the role of intentions, giving in dictator games and con-

[1] Our examples in Section 3 include simultaneous and sequential prisoners' dilemma games, public goods games with punishment option, ultimatum games (also a version with non-intentional offers), dictator games and a general result for two-player zero-sum games.

[2] The reader may think of a distant resemblance to a Stackelberg model: First a norm is fixed, then selfish players account for it in their Nash equilibrium strategies. Note, however, that there is no strategic value of commitment for an individual complier, since types are private knowledge. Thus, in terms of expected utility, a selfish player is weakly better off than a complier.

cerns for social efficiency, with no more than 2 parameters for the entire model (see Sections 3-4, where we also discuss the relation to models of social-preferences).

In Sections 5-6, we show that complier optimal norms and institutions arise from competition of norms and institutions via voting-by-feet. In our model, a society consists of a set of communities, each with its own institution and norm. Inhabitants vote by feet and migrate to communities that grant higher expected utility. Expected utility depends on an inhabitant's type, the institution and moral norm, as well as the share of compliers in a community. After having migrated, a complier follows the norm of her new community and selfish players adapt their equilibrium strategies to the new situation. No person can be prevented from migration - especially, compliers are not able to exclude selfish migrants from a community. We define an *open migration-proof equilibrium* for this migration process. It is based on stability against farsighted coalitional deviations, in the spirit of concepts by Chwe (1994) and Conley & Konishi (2002).

We show first that (under some regularity conditions) every society in which the entire population is located in a single community that has a complier optimal institution and norm constitutes an open migration-proof equilibrium. Second, in every open migration-proof equilibrium, compliers' expected utility will be the same as in the society above.

Our approach differs from models that analyse development of pro-social behavior and culture from evolutionary perspectives, and from typical learning models (see Ostrom, 2000 or Bergstrom, 2002 for surveys). This is because in our model compliers, although voting-by-feet on norms and institutions, do not face evolutionary selection pressures when having lower payoffs than selfish players, nor do they become selfish when learning that selfish players are better off.

That is in line with recent experimental results by Gürerk et. al. (2006) who allow individuals to vote-by-feet between a public goods game with costly punishment and one without. Virtually the whole population migrates to the community with the institution that allows for punishment. There, over 40% of subjects comply with the norm to punish non-contributors, although those subjects who do not punish get higher payoffs (example 3.3 illustrates that punishment of non-contributors is a complier optimal norm).

Our model takes the share of compliers in a society's total population as exogenously given. One explanation for norm-compliance can be a complier's desire for consistency - after having promised or being expected to follow a community's

moral norm.[1]

The existence of compliers and selfish people is also in line with Lawrence Kohlberg's (e.g. 1984) famous work on the psychology of moral development. He distinguishes three levels (split into six stages) of moral thinking. At the pre-conventional level, people follow norms only when this serves their own interests or avoids punishment - like the selfish inhabitants in our model. At the conventional level, people follow the existing norms in their society by a sense of duty or a desire for social approval - like compliers. At the post-conventional level, actions are guided by universal principles and a genuine interest in the welfare of others. Such people comply only to sensible norms, like complier optimal norms, and may be the ones that initiate change of inappropriate norms and institutions - like an initial group of migrants in our voting-by-feet model.

The question of which moral norms people should follow has of course been long discussed in ethics - see Binmore (1994) for an overview from a game theoretic perspective and, as one famous example, note Rawls' (1973) "Theory of Justice". John Harsanyi (e.g. 1992) strongly advocated the conception of rule utilitarianism.[2] In his own words, Harsanyi (1985, p.44) describes the crux of rule utilitarianism as follows:

"In fact, the very purpose of rule utilitarianism is to identify the moral code that would maximize expected social utility if it became the accepted moral code of society, that is, if

1. *it were followed by all rational and morally motivated people (obviously we cannot expect it to be followed by people committed to an irrational nonutilitarian moral code or by people having no concern for morality at all), and if*

2. *all members of society knew that this was the moral code followed by all rational and morally motivated people."*

Our assumptions are similar to Harsanyi's conditions: Compliers correspond to the morally motivated people and selfish players to people that have no concern for morality at all.[3] As in Harsanyi's second condition, the moral norm is commonly known in our model. Harsanyi postulates that from an ethical perspective one should select a moral code that maximizes expected social utility, i.e. expected

[1] People's desire for consistency has long been analyzed in psycholgy, see e.g. Heider (1946), Newcomb (1953), Festinger (1957) or Cialdini (1993). For recent economic models see e.g. Ellingsen & Johannesson (2004) and Charness & Dufwenberg (2005).

[2] For a recent application of rule utilitarianism see Feddersen's and Sandroni's (2002) model of participation in elections.

[3] Irrational behaviour should be incorporated in the game that describes social interaction, e.g. by allowing that actions are mis-implemented (see example 3.3).

sum of utility of all inhabitants.

In contrast, complier optimal norms, which arise from our voting-by-feet model, correspond to the following ethical principle: "If people are ex-ante identical - except for the fact that some comply with norms and others are selfish - and selfish people are in expectation never worse off than compliers[1], then follow a norm which maximizes compliers' expected utility." This can be expressed more compactly in a form related to the Golden Rule or Kant's categorical imperative:

"Follow a norm that you want to be followed by all norm-compliers, in a society where some people act selfishly."

The remaining paper is structured as follows. In Section 2 we present our framework for a single community with a given norm and institution. We define the concept of a norm equilibrium and complier optimal norms. Section 3 illustrates complier optimal norms for different games and we show that predicted behavior is in line with the experimental findings. In Section 4 we compare our approach with existing models of social preferences. Section 5 describes the model of voting-by-feet for competition of moral norms and presents the results. In Section 6 we extend the model to competition of institutions, and make one further extension. Section 7 briefly concludes.

2 Analysing a single community

2.1 Basic Definitions

A community has a continuum of inhabitants. A share κ of inhabitants are compliers and the others are selfish. Inhabitants randomly meet each other in groups of two or more persons and play a game with normal-form representation $G = (N, S, u)$, which describes the institution for social interaction in the community. Note that G may be the normal-form representation of an extensive-form game. $N = \{1, ..., n\}$ describes the set of players, $S = S_1 \times ... \times S_n$ is the strategy space (containing mixed strategies), and $u : S \rightarrow R^n$ is the tuple of utility functions. Each role in the social interaction is equally likely for each inhabitant, i.e. each inhabitant is equally likely to be drawn as ith player of the game G.

A community has a commonly known moral norm $r = (r_1, ..., r_n) \in S$, which is a strategy-profile of the game G. The norm fixes behavior for compliers, i.e. if

[1] In our model selfish players are not worse off than compliers because types are private information.

player i is a complier, she must play the strategy $r_i \in S_i$.[1] For the inhabitants of a community the norm is fixed and exogenously given (in the voting-by-feet model, see Section 5, people will endogenously select norms via migration).

For selfish players, the presence of compliers changes the game G. We assume that a player's type is private knowledge. Let $\theta = (\theta_1, ..., \theta_n)$ denote the actually drawn vector of players' types, where $\theta_i = 0$ means that player i is selfish and $\theta_i = 1$ means that player i is a complier. Types are independently drawn. The probability to draw a complier is κ and to draw a selfish player is $1 - \kappa$, with κ being common knowledge.

Let $s = (s_1, ..., s_n)$ denote the strategy profile played by selfish types in this game of incomplete information and define $s^\theta(s, r)$ by

$$s_i^\theta(s, r) := \begin{cases} s_i & \text{if } \theta_i = 0 \\ r_i & \text{if } \theta_i = 1 \end{cases} . \tag{1}$$

Thus, s^θ describes the strategies that are actually played, when the vector of selfish and compliant types θ is drawn. Let θ_{-i} and $s_{-i}^\theta(s, r)$ denote the types and played strategies of all players except for player i. For a given norm r and compliers' share κ, expected payoff of a *selfish player* i is then given by

$$u_i^{\kappa,r}(s) := \sum_{\theta_{-i}} \Pr(\theta_{-i}|\kappa) u_i(s_i, s_{-i}^\theta(s, r)) \tag{2}$$

where $\Pr(\theta_{-i}|\kappa) = \prod_{j \neq i} \kappa^{\theta_j}(1-\kappa)^{1-\theta_j}$. Since selfish players act individually rational they must play a Nash equilibrium of the induced game with payoff function $u^{\kappa,r}(s)$. We denote this induced game by $G^{\kappa,r} = (N, S, u^{\kappa,r})$ and formally define

Definition 1 *A triple (κ, r, s) of compliers' share, norm and selfish strategy profile is a **norm equilibrium** for a game G if s is a Nash equilibrium of the induced game $G^{\kappa,r}$.*[2]

We will refer to the strategy profile s, which is played by selfish players in a norm equilibrium, as the *selfish (Nash) equilibrium.*

[1] One can specify a guilt-augmented utility function that rationalizes compliers' behaviour. The utility loss from feelings of guilt when violating a norm must be sufficiently big, such that compliers always prefer to follow the norm.

[2] Note that the norm r is *not* endogenously determined in a norm equilibrium. Endogenous in a norm equilibrium are only the selfish equilibrium strategies, which are induced by the given norm and compliers' share.

2.2 Existence of norm equilibria

Can we be sure that for a given game a norm equilibrium exists for all κ and r? Under fairly general conditions the answer is yes, as the following proposition states:

Proposition 1 *If $G = (N, S, u)$ fulfills the following three conditions*

1. *S_i is nonempty, compact and convex,*
2. *$u_i(s)$ is continuous in $(s_1, ..., s_n)$,*
3. *$u_i(s)$ is concave in s_i*

then for every $\kappa \in [0, 1]$ and every $r \in S$ a norm equilibrium exists.

Proof. We need to show that for every $r \in S$ and every $\kappa \in [0, 1]$ the game $G^{\kappa,r}$ has a Nash equilibrium. We note that when G fulfills the three stated conditions, $G^{\kappa,r}$ also fulfills these conditions. For condition 1 this is clear, since $G^{\kappa,r}$ has the same strategy space as G. For conditions 2 and 3 this holds true, because the payoff function of $G^{\kappa,r}$, i.e. $u^{\kappa,r}(s)$, is a linear combination of payoffs described by $u(s)$ and thus continuity / concavity of u implies continuity / concavity of $u^{\kappa,r}(s)$. The last step is to note that conditions 1-3 are sufficient conditions for existence of a Nash equilibrium using the standard Nash-existence proof (see e.g. Mas-Colell et. al. 1995, p. 260-261). ■

In the usual Nash existence proof, only quasi-concavity of $u_i(s)$ in s_i instead of concavity is necessary.[1] Although concavity is a slightly stronger condition, it is nevertheless usually fulfilled, e.g. if S is the set of mixed strategies over a finite action space.

2.3 Equilibrium selection and refinements

The game $G^{\kappa,r}$ may have multiple Nash equilibria, i.e. multiple norm equilibria may exist for a given norm and compliers' share. A selfish equilibrium selection function $\psi : [0; 1] \times S \to S$, which selects for every compliers' share and norm a unique selfish Nash equilibrium of the game $G^{\kappa,r}$, facilitates welfare comparisons between norms.[2] Further, ψ can impose refinements on the class of selected

[1] We need concavity because linear combinations of quasi-concave functions are not necessarily quasi-concave.

[2] To find the rule-utilitarian moral code, Harsanyi (1992 p. 693) proposes a "predictor function", which is similar to our selection function. A selection function guaranties a complete ordering of moral norms with respect to total- or compliers' expected utility.

equilibria. Especially, if G is derived from an extensive form game, selfish Nash equilibria that are not sequentially rational in the corresponding extensive-form game should be ruled out.[1]

In Section 6.1 we show how selfish equilibrium selection can be endogenized by the voting-by-feet process. Until then, we assume that a global equilibrium selection function ψ is given. We denote the selected selfish Nash equilibrium by $s(\kappa, r) \equiv \psi(\kappa, r)$.

2.4 Selfish players' and compliers' expected utility

For a given selfish equilibrium selection function, we can write down the expected payoffs of a selfish player and a complier as a function of κ and r. For a selfish player i expected utility is given by

$$U_i(\kappa, r) := u_i^{\kappa,r}(s(\kappa, r)) = \sum_{\theta_{-i}} \Pr(\theta_{-i}|\kappa) u_i(s_i(\kappa, r), s_{-i}^{\theta}(s(\kappa, r), r)). \tag{3}$$

A compliant player i plays r_i and therefore her expected utility is given by

$$V_i(\kappa, r) := u_i^{\kappa,r}(r_i, s_{-i}(\kappa, r)) = \sum_{\theta_{-i}} \Pr(\theta_{-i}|\kappa) u_i(r_i, s_{-i}^{\theta}(s(\kappa, r), r)). \tag{4}$$

Since we assumed that each position is equally likely for every inhabitant, expected utility from one social interaction is given for selfish inhabitants by $U(\kappa, r) := \frac{1}{n}\sum_{i=1}^{n} U_i(\kappa, r)$ and for compliers by $V(\kappa, r) := \frac{1}{n}\sum_{i=1}^{n} V_i(\kappa, r)$.

Note that in expectations compliers can never be better off than selfish players, i.e.

$$V_i(\kappa, r) \leq U_i(\kappa, r) \ \forall i \text{ and } V(\kappa, r) \leq U(\kappa, r). \tag{5}$$

These inequalities must hold because types are private information and therefore a selfish player i can guarantee himself the expected payoff of a compliant player i by simply playing r_i himself.

2.5 Complier optimal norms and other moral principles

We now formally define complier optimal norms. A norm is complier optimal for compliers' share κ if it maximizes expected utility of a compliant inhabitant. Let $R^o(\kappa)$ denote the set of complier optimal norms for compliers' share κ and let $r^o(\kappa)$ denote an element of $R^o(\kappa)$. Formally:

$$r^o(\kappa) \in R^o(\kappa) := \arg\max_{r \in S} \{V(\kappa, r)\}. \tag{6}$$

[1] Examples 3.2 and 3.5 illustrate how to deal with extensive form games.

One can construct cases where a complier optimal norm does not exist, i.e. $V(\kappa, r)$ has no maximum with respect to r, but only a supremum. Non-existence seems, however, not to be a severe problem, as existence is usually ensured by an appropriate selfish equilibrium selection function (see the examples in Section 3 for illustration). We call a norm equilibrium with a complier optimal norm, i.e. (κ, $r^o(\kappa)$, $s(\kappa, r^o(\kappa))$), a *complier optimal norm equilibrium (CONE).* Recall that selfish players' expected utility can never fall below compliers' expected utility. This provides an ethical justification for complier optimal norms, as pointed out in the introduction.

It is helpful to compare complier optimal norms with alternative moral principles. An utilitarian norm maximizes expected sum of utility (including utility of selfish players), i.e.

$$r^{utilitarian}(\kappa) \in \arg\max_{r \in S} \{\kappa V(\kappa, r) + (1-\kappa)U(\kappa, r)\}. \tag{7}$$

Another principle is to act in a way that would maximize total welfare if everyone acted this way, i.e. if one naively assumed that everyone were a complier. Such norms are given independently of the actual compliers' share by

$$r^{naive} \in \arg\max_{r \in S} V(\kappa = 1, r). \tag{8}$$

3 Examples

3.1 A Prisoners' Dilemma game

Suppose social interaction is described by a two player Prisoners' Dilemma game (PD game), where players can either cooperate C or defect D. We normalize the payoff matrix as follows

	C	D
C	$1, 1$	c, d
D	d, c	$0, 0$

with $c < 0 < 1 < d$ and $c + d < 2$. Note that defection is a strictly dominant strategy and hence selfish players will always defect, i.e. $s(\kappa, r) = D \ \forall \kappa, r$.[1] For a given compliers' share κ, what is the complier optimal norm $r^o(\kappa)$? Compliers' expected utility from cooperation is given by $V(\kappa, C) = \kappa + (1-\kappa)c$ and from

[1] We abbreviate symmetric selfish equilibrium strategies and norms by the strategy of one player, i.e. we write $s(\kappa, r) = D$, instead of $s(\kappa, r) = (D, D)$.

defection by $V(\kappa, D) = 0$. Thus, the complier optimal norm depends on the share of compliers κ in a community. One has,

$$r^o(\kappa) = \begin{cases} C & \text{if } \kappa > \frac{|c|}{1+|c|} \\ D & \text{if } \kappa < \frac{|c|}{1+|c|} \\ C \text{ or } D & \text{if } \kappa = \frac{|c|}{1+|c|} \end{cases} .$$

The intuition behind this result is that when there are only few compliers, i.e. a small κ, compliers are very likely to meet selfish players and would be exploited in the PD game if the norm were cooperation. Thus, only if the probability to meet other compliers is sufficiently high, does the complier optimal norm become cooperation.

What does the utilitarian norm $r^{utilitarian}$ look like? Let us assume for simplicity that the measure of the total population in the community is normalized to 1. If compliers cooperate, expected sum of utility is given by $\kappa V(\kappa, C) + (1-\kappa)U(\kappa, C) = \kappa(\kappa + (1-\kappa)c) + (1-\kappa)(\kappa d)$ and if compliers defect the expected sum of utility is given by 0. This yields

$$r^{utilitarian}(\kappa) = \begin{cases} C & \text{if } \kappa > \frac{|c|-d}{1+|c|-d} \\ D & \text{if } \kappa < \frac{|c|-d}{1+|c|-d} \\ C \text{ or } D & \text{if } \kappa = \frac{|c|-d}{1+|c|-d} \end{cases} .$$

The share of compliers needed to make cooperation the utilitarian norm is smaller than the share needed to make cooperation the complier optimal norm, since $\frac{|c|-d}{1+|c|-d} < \frac{|c|}{1+|c|}$. This is because compliers who cooperate create a positive payoff externality for selfish players. Indeed, the utilitarian norm can be defection only in games with $c + d < 0$, i.e. sum of utility is lower when one player defects and the other cooperates than if both defect. Finally, it is clear that the naive norm is given by the social optimal solution that both players cooperate, i.e.

$$r^{naive} = C.$$

3.2 A sequential Prisoners' Dilemma

We want to illustrate now the effects of a small institutional change in our setting. Modify the PD game such that player 1 moves first and player 2 can condition his decision on the observed behavior of player 1.

We summarize the results for complier optimal norm equilibria. A sequential rational selfish player 2 always defects. For $\kappa < \frac{|c|}{2+|c|}$ there are multiple complier

optimal norm equilibria but on the equilibrium path there is always mutual defection. For $\kappa > \frac{|c|}{2+|c|}$ the unique complier optimal norm is that a compliant player 1 cooperates and a compliant player 2 plays the following Tit-for-Tat strategy: cooperate if player 1 cooperated and defect if player 1 defected. A selfish player 1 cooperates in the resulting norm equilibrium if $\kappa > \frac{|c|}{1+|c|}$ and defects if $\kappa < \frac{|c|}{1+|c|}$.[1]

These predictions are in line with experimental studies of sequential Prisoners' Dilemma games (see Bolle & Ockenfels 1990 or Clark & Sefton 2001), which show that unconditional cooperation is practically never observed.

Note that compliers' expected utility can never decrease when changing from a simultaneous-move to a sequential PD game but there can be substantial increases, especially when κ is slightly above $\frac{|c|}{1+|c|}$. This is illustrated on the left graph of figure 1 for PD games with $c = -1$. The right graph of figure 1 shows that expected

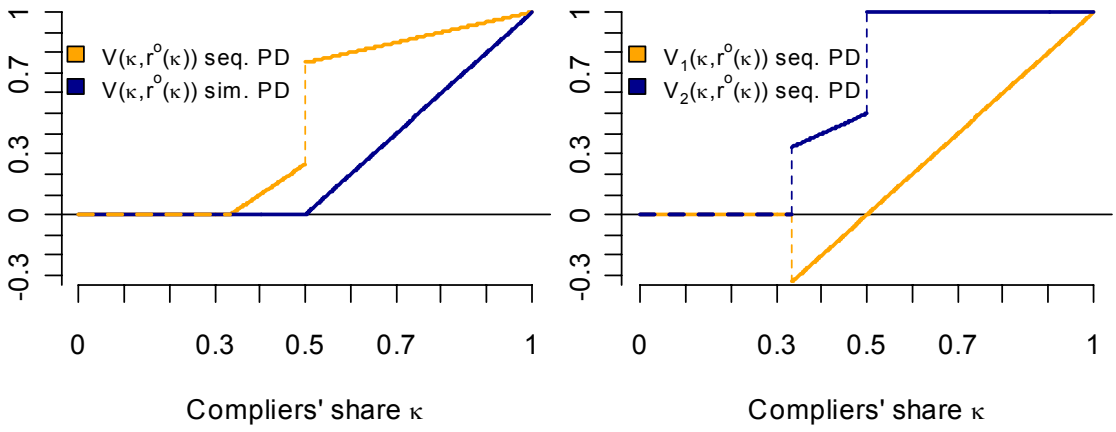

Figure 1: Left: Compliers' expected utility in a sequential PD game compared to a simultaneous-move PD game, with $c = -1$. Right: Expected utility of a compliant player 1 vs a compliant player 2 in the sequential PD game.

utility differs, however, between compliant player 1 and 2. Compliant player 1 gets substantially lower expected utility, which for $\kappa \in (\frac{1}{3}, \frac{1}{2})$ even becomes negative. In this interval, a compliant player 1 accepts excessive exploitation by a selfish player 2, because the decision to cooperate benefits a compliant player 2. Although our model assumes that an inhabitant does not know ex-ante which player she will be, we recommend to augment utility functions over monetary payoffs by some explicit equity concerns, as illustrated in example 3.5 and Section 4.

[1]For $\kappa = \frac{|c|}{1+|c|}$ every mix between C and D is an equilibrium strategy for a selfish player 1. In this case, compliers' expected utility depends on the selfish equilibrium selection function (but the complier optimal norm does not).

3.3 A public goods game with costly and imprecise punishment

This example analyses a two-player public goods game with a quite flexible specification of punishment technologies. There are two players who each own one unit of money. A player can either keep his money for himself, increasing own payoff by 1, or contribute it to a common project, which increases both players' payoff each by $\frac{\beta}{2}$, where β is an exogenous parameter describing the efficiency of the common project. Assume $1 < \beta < 2$, i.e. we have a public goods dilemma situation where it is socially optimal if both players contribute, but individually rational to keep the money (assume, in this example, that utility in monetary payoffs π is simply given by $u_i(\pi_i, \pi_j) = \pi_i$).

Further, players have the opportunity to costly punish non-contributors. A player i decides over a punishment level $x_i \geq 0$ and in the case where the partner did not contribute, the partner's payoff is reduced by x_i. Punishment entails two kinds of costs for the punisher. First, there are direct costs of punishment ϕx_i with $\phi > 0$, which only must be paid if punishment is indeed carried out. Second, there are control costs γx_i with $\gamma \geq 0$, which have to be paid even if punishment is not carried out. To simplify the game, we allow a player to punish only if she contributes to the public project, otherwise she must set $x_i = 0$.

Additionally, we allow for the case that punishment is not completely precise in the sense that with a probability η $(0 \leq \eta \leq 1)$ punishment is erroneously carried out, even when the partner has contributed. This extension illustrates how mistakes can be easily implemented into a game. We think that analysing the effects of possible mistakes is especially important in games of punishment and deterrence.

The reason to analyse control costs as well as imprecise punishment is that imprecise punishment causes negative externalities for other contributors, whereas control costs cause no such externalities. This distinction becomes relevant in the context of Sections 5-6.

To sum up, a players strategy $s_i = (a_i, x_i)$ consists of an action $a_i \in (C, D)$, where C stands for 'contribute' and D for 'do not contribute' and a level of punishment $x_i \in R_0^+$. Final payoffs are summarized in table 1.

If there is some imprecision, i.e. $\eta > 0$ (and $\phi > 0$) or some control costs, i.e. $\gamma > 0$, selfish players will never punish in equilibrium. Therefore, if there are no compliers present, the unique selfish Nash equilibrium is that both players do not contribute and do not punish, i.e. $s_1 = s_2 = (D, 0)$. For compliers, punishment can

Payoffs for player 1 (row)	C (Contribute)	D (Do not contribute)
C (Contribute)	$\beta - \eta(\phi x_1 + x_2) - \gamma x_1$	$\frac{\beta}{2} - (\phi + \gamma)x_1$
D (Do not contribute)	$\frac{2+\beta}{2} - x_2$	1

Table 1: Payoffs for player 1 (row) from the public goods game with costly and imprecise punishment. Payoffs for player 2 (column) are symmetrical.

be beneficial only if the threat of punishment is sufficiently high to make selfish players contribute. We first want to examine the minimal level of punishment x^* compliers must choose to make selfish players contribute. When we assume that a selfish player 2 plays a mixed strategy $s_2 = ((1-\delta, \delta), x = 0)$, i.e. he contributes with probability $(1-\delta)$, and that the norm is given by $r_1 = r_2 = (C, x)$, then a selfish player 1's expected payoff from contributing is given by $\kappa(\beta - \eta x) + (1-\kappa)((1-\delta)\beta + \delta\frac{\beta}{2})$ and from not contributing by $\kappa(\frac{2+\beta}{2} - x) + (1-\kappa)((1-\delta)\frac{2+\beta}{2} + \delta)$. Setting both equal and solving for x gives the minimal level of punishment necessary to induce contribution as

$$x^*(\kappa) = \frac{2-\beta}{2\kappa\,(1-\eta)}.$$

Note that a selfish player 1's decision to contribute is independent of the strategy of selfish player 2, since δ canceled out. We find the same for selfish player 2 and thus have as selfish equilibrium strategies:

$$s_i(\kappa, r = (C, x)) = \begin{cases} (C,0) & \text{if } x > x^*(\kappa) \\ (D,0) & \text{if } x < x^*(\kappa) \\ (C,0)\ ,\ (D,0) \text{ or any mix} & \text{if } x = x^*(\kappa) \end{cases} \quad \text{for } i = 1, 2.$$

In order to have a well defined complier optimal norm, assume that for $x = x^*(\kappa)$ selfish players will contribute. Since punishment is costly, the complier optimal norm postulates either contribution and punishment with $x^*(\kappa)$, contribution without punishment or non-contribution. Compliers' expected utilities from these three norms are given by

$$\begin{aligned} V(\kappa, r = (C, x^*(\kappa)) &= \beta - \eta(\phi + \kappa)x^* - \gamma x^* \\ V(\kappa, r = (C, 0)) &= \kappa\beta + (1-\kappa)\frac{\beta}{2} \\ V(\kappa, r = (D, 0)) &= 1 \end{aligned}$$

If there are no control costs and no imprecision in punishment, i.e. $\gamma = \eta = 0$ then always the complier optimal norm is to punish with at least $x^*(\kappa)$. This

is simply because with exact punishment, there arises no cost from a successful punishment threat, which never has to be carried out, because selfish people will contribute in equilibrium. Hence, as long as players are able to punish up to an unlimited level, even a tiny share of compliers is sufficient to deter selfish players from non-contributing.[1]

Let us assume there are some control costs, i.e. $\gamma > 0$, or some imprecision of punishment, i.e. $\eta > 0$. Then looking at the limit $\kappa \to 0$, shows that for a sufficiently small compliers' share κ the complier optimal norm is to not contribute. On the other hand, by looking at the limit $\kappa \to 1$, we find that for a very high share of compliers, the complier optimal norm is contribution without punishment. This is because it is simply cheaper for compliers to encounter from time to time a selfish player who defects than to suffer the costs of control and damage from imprecision. For intermediate κ, however, a norm with punishment, i.e. $r = (C, x^*(\kappa))$, can be optimal for compliers if punishment is not too expensive and not too imprecise.

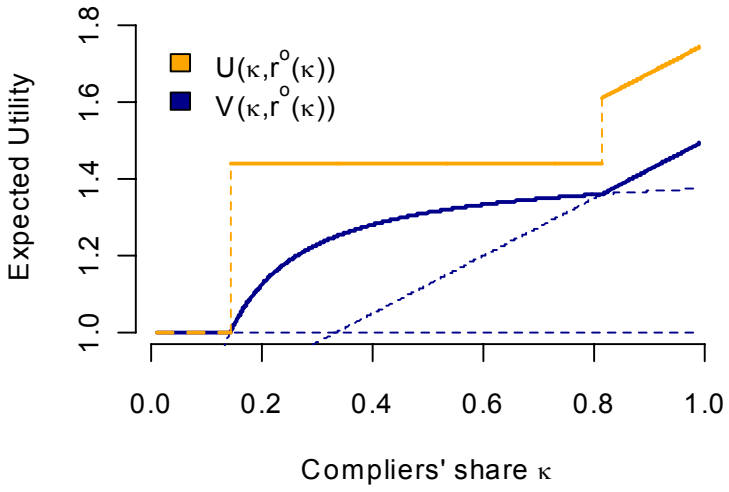

Figure 2: Selfish and compliant players' expected utility under complier optimal norms for a public goods game with imprecise punishment with $\beta = 1.5$, $\phi = 1$, $\gamma = 0$ and $\eta = 0.2$.

Figure 2 illustrates this result for $\beta = 1.5$, $\phi = 1$, $\gamma = 0$ and $\eta = 0.2$. The thick lines show selfish players' and compliers' expected utility when applying

[1]For the case $\gamma = \eta = 0$ it would also be a Nash equilibrium strategy for selfish players to punish at a level high enough to deter from non-contribution. If we have an extensive form game with punishment decision after contribution decision, those strategies are, however, not sequential rational.

the complier optimal norm in a community with compliers' share κ. Note the discontinuous jump in selfish players' expected utility at those levels of κ where complier optimal norm changes from $(D, 0)$ to $(C, x^*(\kappa))$ at $\kappa \approx 0.14$ and then to $(C, 0)$ at $\kappa \approx 0.81$.

3.4 Two player zero-sum games

Consider a two player zero-sum game $G = (\{1, 2\}, S, u)$, with $u_1(s) + u_2(s) = 0$ for all $s \in S$. A general zero-sum theorem states that s^* is a Nash equilibrium of G if and only if s_i^* is a maxmin strategy, i.e. $s_i^* \in \arg\max_{s_i \in S_i} \min_{s_j \in S_j} u_i(s_i, s_j)$. Further, all Nash equilibria give the same expected payoff for player i , denoted by u_i^{maxmin}. We show that also in every complier optimal norm equilibrium expected payoff for both a complier and selfish player i is given by u_i^{maxmin}.

Proposition 2 *Assume $\kappa < 1$. A norm r^o is complier optimal in a two player zero-sum game G if and only if $V_i(\kappa, r^o) = U_i(\kappa, r^o) = u_i^{maxmin}$ $\forall i$.*[1]

Proof. Define $\Delta_i := U_i(\kappa, r) - V_i(\kappa, r)$ and note that $\Delta_i \geq 0$ (see Section 2.4). Expected utility in a zero-sum game is 0, i.e. $\kappa V(\kappa, r) + (1 - \kappa)U(\kappa, r) = 0$. This can be written as $\kappa V(\kappa, r) + (1 - \kappa)\left[V(\kappa, r) + \frac{1}{2}\sum \Delta_i\right] = 0$, yielding $V(\kappa, r) = -(1-\kappa)\frac{1}{2}\sum \Delta_i$. If for some complier $V_j(\kappa, r) \neq u_j^{\text{maxmin}}$ at least one complier i gets lower expected utility than u_i^{maxmin}. Since a selfish player i can guarantee himself expected payoff of at least u_i^{maxmin}, by playing a maxmin strategy, this leads to $\Delta_i > 0$, implying $V(\kappa, r) < 0$. This cannot be complier optimal, since compliers' expected utility is 0 when the norm equals a profile of maxmin strategies. ■

Consider a dictator game where player 1 splits 1 unit of money between him and player 2. Assume both players have identical linear utility in own monetary payoff (using $u_i(\pi_i) = \pi_i - 0.5$ this yields a zero-sum game). By our result, the unique complier optimal norm is that player 1 keeps all money for himself. For intuition, note that with with probability $(1-\kappa)$ transferred money would be given to a selfish player, which would reduce expected monetary payoff of compliers.

[1] Usually, this means that r^o must be profile of maxmin strategies itself. For a counterexample consider, however, a matching pennies game with $u(H, H) = u(T, T) = (1, 0)$ and $u(H, T) = u(T, H) = (0, 1)$. The only maxmin strategy is an equal mix between H and T. For $\kappa = 0.5$, there exists, however, a complier optimal norm equilibrium with norm (H, H) and selfish equilibrium strategies (T, T), i.e. randomization takes place via a player's type.

3.5 Envy in dictator, ultimatum and other games

Note that the concept of complier optimal norms says nothing about players' utility function over monetary payoffs. Players can be expected payoff maximizers, be risk- or loss-averse, or may feel envious when other players have higher payoffs.

We illustrate the interplay between complier optimal norms and emotions using the following utility function over monetary payoffs

$$u_i(\pi_i, \pi_j) = \pi_i - \alpha \max\{\pi_j - \pi_i, 0\} \text{ with } \alpha > 0.$$

The interpretation is that a player feels envy when he has a lower monetary payoff than the other player. The degree of envy α is assumed to be equal for all players, irrespective of whether a player is selfish or a complier. This utility function is a simplified version of inequity aversion by Fehr & Schmidt (1999) - simplified, because we do not incorporate a term that explicitly models "guilt", felt by a player who has a higher payoff than the other. For a rough approximation of experimental evidence, we suggest $\alpha \approx \frac{1}{3}$ (or a bit lower) and $\kappa \approx 0.6$.

Dictator game

In a dictator game player 1 splits an amount of money between him and player 2. Let the total amount of money be normalized to 1 and let x denote the share offered to player 2. Clearly, a selfish player 1 will give nothing to player 2. When a compliant player 1 gives an amount $x^o \leq 0.5$ to player 2, compliers' expected utility is given by $V(\kappa, x^o) = \frac{1}{2}((1 - x^o) + \kappa(x^o - \alpha(1 - x^o - x^o))$. Maximization of this term implies that under a complier optimal norm compliers offer

$$x^o \in \begin{cases} 0 & \text{if } \kappa < \frac{1}{1+2\alpha} \\ [0, 0.5] & \text{if } \kappa = \frac{1}{1+2\alpha} \\ 0.5 & \text{if } \kappa > \frac{1}{1+2\alpha} \end{cases} .$$

The condition $\kappa \lessgtr \frac{1}{1+2\alpha}$ illustrates two factors that determine how much a compliant player 1 should give to player 2. On the one hand, an equal split is beneficial because it reduces envy of a compliant player 2. On the other hand, transferring money has a negative effect because with probability $1 - \kappa$ it is given to a selfish player 2. For $\kappa = \frac{1}{1+2\alpha}$ both effects balance out. For example, if $\alpha = \frac{1}{3}$ an equal split would be complier optimal in a dictator game whenever $\kappa \geq 0.6$. To model offers in between 0 and 0.5 we need a non-linear formulation of envy (see Fehr & Schmidt, 1999, p. 847-848, for discussion of a similar problem).

Andreoni and Miller (2002) performed dictator experiments where transfers were multiplied by an efficiency factor f, i.e. monetary payoffs are given by $(1-x, fx)$. They show that average transfers increase in the efficiency factor. As should be intuitively clear, our model matches this stylized fact.[1]

Ultimatum game

In an ultimatum game player 2 has the opportunity to reject the offer x by player 1 in which case both get paid zero. We assume that a selfish player 1 chooses the highest offer when he is indifferent between more than one offer and that a selfish player 2 acts sequentially rational and furthermore accepts an offer when he is indifferent between accepting or rejecting.

We summarize the results that hold for all complier optimal norm equilibria (see appendix A for the derivations). A selfish player 2 accepts only offers that are weakly higher than $x^* := \frac{\alpha}{1+2\alpha} < 0.5$ and for a compliant player 2 the acceptance threshold is given by

$$x^o := \min\{\kappa + (1-\kappa)x^*, 0.5\} \geq x^*.$$

Both selfish and compliant players 1 will always offer x^o. Note that

$$\lim_{\alpha \to 0} x^o = \min\{\kappa, 0.5\}.$$

Thus in contrast to a dictator game, an infinitesimally small amount of envy suffices to find substantial offers in an ultimatum game.[2]

The stylized facts from ultimatum experiments (see for example the overviews by Güth, 1995, Camerer and Thaler, 1995 or Roth, 1995), can be summarized as follows: The vast majority of offers lie between 0.4 and 0.5, virtually no offer exceeds 0.5 and offers below 0.2 are very rare. Offers near 0.5 are practically never rejected, whereas the rejection rate for offers below 0.2 is very high. Can our model match these stylized facts? For $\alpha = \frac{1}{3}$ we find $x^* = 0.2$. Thus in line with the stylized facts, all offers below 0.2 are rejected. Further we find $x^o = \min\{0.2 + 0.8\kappa, 0.5\}$. This means that already for $\kappa \geq \frac{1}{4}$ observed offers x^o should lie between 0.4 and 0.5.

[1] In the set-up of Andreoni and Miller, compliers should offer $x^o = 0$ if $\kappa < \frac{1}{f+(1+f)\alpha}$, $x^o = 0.5$ if $\frac{1}{f+(1+f)\alpha} < \kappa < \frac{1+(1+f)\alpha}{f}$ and $x^o = 1$ if $\kappa > \frac{1+(1+f)\alpha}{f}$.

[2] The same result can be derived if players are infinitesimal loss-averse with reference level 0.5 or risk averse under equal initial wealth. When players are monetary payoff maximizers, offers between 0 and κ can be found in different complier optimal norm equilibria.

Ultimatum game with non-intentional offers

Blount (1995) performed an experimental treatment where the offer was not selected by the proposer but randomly chosen by a computer. She showed that minimal acceptance levels are significantly lower when the offer was randomly selected, but that some offers still were rejected. Blount's finding can be explained by our model (see appendix A for details). For $\alpha < \kappa$, a compliant player 2 will accept all random offers (for $\alpha > \kappa$ very unequal offers may be rejected). A compliant player 2 still feels envy, but weighs the monetary payoff of a compliant player 1 higher than her envy. The difference to the intentional treatment arises because for random offers a norm has no strategic impact on the behavior of player 1. An envious selfish player 2, however, still rejects every offer below x^*, since it does not matter for him how the offers were selected.

Other games

Our model also captures the stylized facts from other experiments, like mutual giving in trust games (e.g. Berg et. al., 1995) or behavior in best-shot games (see Harrison & Hirshleifer 1989, Prasnikar & Roth 1992 and also Falk et. al., 2003).[1]

Including envy into the prisoners' dilemma games and the public goods game with punishment (examples 3.1-3.3) does not qualitatively change our results. Quantitatively, the identified thresholds for κ increase, which means that compliers are more reluctant to cooperate and more likely to defect, when envy is included.

4 Comparison to social preferences and more

4.1 Comparison to existing models of social-preferences

In order to explain the behavior that is observed in economic experiments, a number of theories on social-preferences have been developed (see Sobel, 2005 for a detailed survey).

For a comparison of some of these theories consider the ultimatum game from example 3.5. In inequity aversion theories, as Fehr & Schmidt (1999) or Bolton & Ockenfels (2000), players do not like inequalities in monetary outcomes, which can explain rejection of low offers. These models have the advantage of being analytically very convenient, but do not account for the role of intentions, e.g.

[1] A formal analysis of these games is available on request from the author.

they do not explain the change in rejection behavior in the treatment with non-intentional offers by Blount (1995).

In Gul & Pesendorfer (2005) (a generalization of Levine, 1998) perceived goodness of others' types matters for own payoff. Rejection can occur in the ultimatum game when own payoff decreases in the payoff of "unkind" types. This model is able to explain Blount's findings. It predicts, however, that unkind types reject more often than kind types. This is somewhat at odds with other experimental results showing that nice types more strongly punish unfair behavior, see e.g. Falk et. al. (2005).

Another set of social-preference models, starting with Rabin's (1993) fairness theory for normal form games, builds on a psychological game framework (Geanakoplos et. al., 1989) where players can get utility from beliefs. The reciprocity models by Dufwenberg & Kirchsteiger (2004) or Falk & Fischbacher (2006) extend the framework to extensive form games. Unequal offers are considered as a sign of unkind intentions and punishing these actions gives emotional satisfaction. These two models can explain reciprocal behavior and the role of intentions very well. They neglect, however, concerns for social efficiency, e.g. they cannot explain the above mentioned findings in the dictator experiments by Andreoni and Miller (2002). Another reciprocity model is given in the appendix of Charness & Rabin (2002). It can account for the role of intentions and social-efficiency concerns, but is quite complex.[1]

Note the principle difference between complier optimal norms and the models described above. In these models of social-preferences offers are rejected solely because player 2 gets emotional satisfaction when punishing "unfair" behavior or when changing "unfair" outcomes. The fact that a commonly known norm of rejecting low offers has a strategic impact on player 1 and thereby increases social utility, is not, however, explicitly considered by player 2. Complier optimal norms, on the other hand, prescribe rejection of offers below x^o, exactly *because* a selfish player 1 can thereby be induced to offer x^o (selfish players' rejections of offers below x^* are, however, still just due to negative emotions).

Models of social preferences often assume a distribution of more than two types, each type being described by one or more parameters. A finer type space allows

[1]For alternative approaches see Cox et. al. (2005) or Segal & Sobel (2006). There is also a recent norm-based approach by López-Pérez (2005), which, however, substantially differs from our model. His E-norms resemble our naive norms (see Section 2.5) combined with equity concerns. In his model, punishment is not part of a norm but arises from anger against norm-violators.

better capturing of behavioral heterogeneity, but comes at the cost of increased analytical complexity. The combination of complier optimal norms with a single envy parameter explains the stylized facts from a wide range of economic experiments with only two parameters for the entire model (κ and α). We offer therefore a novel, tractable and empirically consistent approach to incorporate non-selfish behavior into economic models.

4.2 More types, more emotions

Although a simple model with constant envy for all persons matches the stylized facts of economic experiments quite well, complier optimal norms can be combined with other emotions, like kindness, anger, guilt, shame, etc. This allows for interesting formal statements concerning the interplay between emotions and moral norms. For example, players could feel kindness towards all, including selfish, players, but nevertheless punish selfish actions under a complier optimal norm.

In addition to the distinction of being selfish or compliant, players can be heterogeneous in their emotional attitudes and in other aspects of preferences. In our framework this can be formalized by letting a game start with a move of nature that determines these additional aspects of a player's type. Complier optimal norms then maximize compliers' ex-ante expected utility over this type distribution. Note that norms then may allow compliers to condition their action on their complete type, e.g. very altruistic compliers may be allowed to punish less than very envious compliers.

Such additional aspects of players' types can be distributed differently for selfish and compliant players, e.g. selfish players may on average be more spiteful than compliers. Note, however, that the condition $U(\kappa, r) \geq V(\kappa, r)$ could then be violated and the findings from our voting-by-feet model in Sections 5-6 may not hold.[1]

Emotions can also depend on the norm, e.g. players may feel angry towards norm violators. Further, selfish players may feel to some degree guilty when violating a norm, which allows a relaxation of the strict distinction between players who always comply with a norm and others who feel not committed at all.

[1] In general, the interpretation of our voting-by-feet result becomes more difficult under additional heterogeneity in types, since our model makes the strong assumption that only the moral type (compliant or selfish) is known before moving to a community. In principle, the voting-by-feet model can, however, also be adapted to analyse the case where players know ex-ante more about their type.

Compliers' equity concerns can also be modelled by letting complier optimal norms maximize some transformed version $\widetilde{V}(\kappa, r)$ of compliers' expected utility, like

$$\widetilde{V}(\kappa, r) := \sum_i V_i(\kappa, r)^\beta \text{ with } 0 < \beta < 1.$$

Here, norms leading to more equal expected utility between players are preferred, i.e. players care about equality in chances. One could also consider only the worst-off player by setting

$$\widetilde{V}(\kappa, r) := V + \beta \min_i\{V_i\} \text{ with } \beta > 0.$$

One can also include a factor that compliers dislike when selfish players are on average better off. A linear formulation of such *exploitation aversion* is incorporated by setting

$$\widetilde{V}(\kappa, r) := V - \gamma(U(\kappa, r) - V(\kappa, r)) \text{ with } \gamma > 0.$$

Exploitation aversion can also be defined non-linear and player dependent, like

$$\widetilde{V}_i(\kappa, r) := \begin{cases} V_i(\kappa, r) & \text{if } U_i(\kappa, r) - V_i(\kappa, r) < \Delta_{\max} \\ -\infty & \text{otherwise} \end{cases} \quad \text{and } \widetilde{V}(\kappa, r) := \frac{1}{n} \sum \widetilde{V}_i(\kappa, r).$$

This is a convenient formulation to capture the idea that a compliant player i is in principle willing to follow the norm r_i, but only if the temptation to act selfishly (measured by the difference in expected utility) does not exceed some level $\Delta_{\max}$. Note that with this formulation of exploitation aversion, our results from the voting-by-feet model in Section 5-6 will carry over without problems.

4.3 Communication and associations

Ellingsen & Johannesson (2004) and Charness & Dufwenberg (2005) are two recent examples that analyse how communicated promises and threats can create commitment effects.[1] Moral norms can be seen as a promise made by all members of a community to act in a certain way. Thus, complier optimal norms offer a theory that says which promises the committed compliers should make in the presence of a certain share of uncommitted selfish players. Analysing the actual process of how people discuss and agree about norms offers a lot of interesting connected research questions - especially because selfish players may prefer different norms

[1] See also Gneezy (2005) and the surveys by Kerr & Kaufman-Gilliand (1994) and Sally (1995), which summarize earlier work on the relation between communication, commitment and cooperation.

than compliers and because of the fact that behavior in pre-game communication and "norm-bargaining" may allow inferences about a player's type. A further question is, what are the effects of discussing norms in an interim-state, where subjects already know their role in the experiment, compared to a symmetric ex-ante state.

Finally, we want to briefly discuss the possibility that - especially in experiments without pre-game communication - experimental subjects in the lab may simply apply known norms from associated real world situations.

For example, Hoffman et. al. (1994) showed that labeling an ultimatum game as a buyer-seller interaction, i.e. the seller sets a price and the buyer can either buy or not buy at this price, reduces the offers by almost 10% with no increase in rejection rate, compared to a traditional labeling of the ultimatum game. A simple request to split some exogenously given prize may remind subjects of equal-split norms, as e.g. known from childhood experience when parents demand equal sharing of sweets between siblings. Market interactions, however, may be less associated with equal-split norms.

Another example is cross-cultural differences. Henrich et. al. (2001) found substantial differences in behavior in the ultimatum game by comparing 15 small scale societies, which substantially differed in their set of norms and customs. One explanation for the different behavior is that participants apply a norm that is used for typical interactions in their society to the ultimatum game.

Although complier optimal norms do not describe a theory of mental associations, a norm-based approach, in general, is likely to facilitate formalizations of such labeling and cultural effects.

5 Competition of norms via voting by feet

5.1 Overview

In this section we illustrate that complier optimal norms arise from a model of voting-by-feet. A society consists of communities with different norms. There is free migration, and inhabitants can move to communities that offer higher expected utility. Voting-by-feet is much analysed in a branch of literature emerging from Tiebout (1956), who analysed local provision of public goods. Different equilibrium concepts have been proposed. Under Nash concepts, e.g. Westhoff (1977), usually too many equilibria exist, some of which trap players in suboptimal states that could be Pareto-improved if coalitional deviations were allowed. Alternative

models consider stability against all possible coalitional deviations, e.g. Greenberg and Weber (1986). Here, the problem is that equilibria often do not exist. Conley and Konishi (2002) discuss these problems and resolve some of them by defining a "migration-proof Tiebout equilibrium", which requires stability only against those coalitional deviations that can be successful when accounting for possibly induced future migration. Concepts of far-sighted stability, e.g. Chwe (1994), are based on a related idea. None of these solution concepts can be directly applied to our model, which has a continuum of inhabitants with privately known types, but we use the basic idea to propose our *open migration-proof equilibrium* as a reasonable stability concept.

This section proceeds with a formal definition of a society. Afterwards we define a *Nash-stable equilibrium* as a society where no individual inhabitant wants to migrate to another populated community. We then account for stability against farsighted-successful coalitional deviations. This leads to the concept of a *migration-proof equilibrium*. To illustrate robustness of our uniqueness result, we furthermore define a *weak migration-proof equilibrium*. Finally, a (weak) *open* migration-proof equilibrium, allows arbitrary empty communities with new norms to enter the society. We then discuss five regularity conditions, before we formally derive the results. In Section 6.1 we extend the results to joint competition of norms and institutions.[1]

5.2 Formal definition of a society

A society consists of a finite number of communities $\{C^j\}_{j\in J}$ indexed by a set J. Each community C^j is characterized by its norm r^j. A society's total population is given by a continuum denoted by $\mu^t = (\mu_c^t, \mu_s^t)$, where μ_c^t is the measure of compliers and μ_s^t the measure of selfish inhabitants. The compliers' share in a society's total population is thus given by $\kappa^t := \frac{\mu_c^t}{\mu_c^t+\mu_s^t}$. An allocation $\alpha = \{(\mu_s^j, \mu_c^j)\}_{j\in J}$ describes how the total population is distributed over the different communities. Relevant characteristics of an allocation are, first, which communities are populated at all and, second, the compliers' share κ^j in a populated community C^j.

In this section we take the game G and the selfish equilibrium selection function ψ to be given and equal in all communities. A society is thus completely described by the collection of the norms in its communities and the allocation of the total population over the different communities: $(\{r^j\}_{j\in J}, \alpha, \mu^t)$.

[1]Note that there are other concepts of institutions. For example in Caplin's and Nalebuff's (1997) model of competition among institutions, institutions are not defined as a game.

Note that a community's norm is a fixed strategy profile, which is not allowed to change with the share of compliers in that community.[1] A framework where norms are allowed to change with the compliers' share in a community will be explored in section 6.2.

5.3 Nash-stable equilibrium

Our first requirement for a stable society, i.e. a society without migration pressures, is that there are no two populated communities where inhabitants of the same type get different expected utility. We say a single selfish / compliant inhabitant *prefers to move* from his origin community C^o to a populated destination community C^d if selfish / compliant inhabitants' expected utility in C^d is *strictly* higher than in C^o. We formally define:

Definition 2 *A society* $(\{r^j\}_{j\in J}, \alpha, \mu^t)$ *constitutes a* ***Nash-stable equilibrium*** *if no inhabitant prefers to move to another populated community.*

Note that every society with a single populated community constitutes a Nash-stable equilibrium, since individuals cannot deviate to empty communities.

5.4 Migration-proof equilibrium

Our concept of a migration-proof equilibrium allows for migration by coalitions. A coalitional migration is described by a finite collection $m = \{(\mu_k, \theta_k, C_k^o, C_k^d)\}_{k=1}^K$ where one entry $(\mu_k, \theta_k, C_k^o, C_k^d)$ means: a group with measure μ_k ($\mu_k > 0$) of inhabitants of type θ_k migrates from origin community C_k^o to a destination community C_k^d. Important for migration-proof stability will be the concept of uncoordinated migration:

Definition 3 *A coalition performs* ***uncoordinated migration*** *if every individual member of the coalition prefers to move from her origin community to the (previously non-empty) destination community (evaluating expected utilities under the pre-migration allocation).*

Uncoordinated migration resembles monotone dynamics in evolutionary models. As long as a society does not constitute a Nash-stable equilibrium, inhabitants migrate to communities that offer higher expected utility for their type.

[1] The notation $r^o(\kappa')$, which we will often use below, thus denotes a norm that is complier optimal for a fixed share of compliers κ', but it does not mean that a norm is a function, which is allowed to change with the compliers' share in a community.

The idea of a farsighted-successful coalitional migration is that the process of uncoordinated migration is anticipated and all members of the farsighted coalition see a chance to end up in a Nash-stable society where they are strictly better off than initially. Formally:

Definition 4 *Migration by a coalition m is* ***farsighted-successful*** *if there exists a (possibly empty) sequence of uncoordinated migrations that starts after the migration of m and ends in a* Nash-stable equilibrium *where all members of m are strictly better off than initially.*

This directly leads to the definition of a migration-proof equilibrium:

Definition 5 *A society $(\{r^j\}_{j\in J}, \alpha, \mu^t)$ constitutes a* ***migration-proof equilibrium*** *if it is a Nash-stable equilibrium and there exists no farsighted-successful coalitional migration.*

Let us reflect for a moment on the informational assumptions behind a farsighted-successful migration. Is it necessary for our results that a farsighted coalition is able to exclude members of certain types? The answer is no. If some inhabitants prefer to be in some destination community of a farsighted coalition, they can just immediately follow to this community by uncoordinated migration. This leads to the same allocation as if they were members of the coalition. Thus, in principle it suffices that a farsighted migration is publicly announced and members then sort themselves into the coalition. To keep the model simple, we do not, however, explicitly model such a public announcement process.

For an illustration of farsighted migration, let G be a PD game and consider a society with two communities: C^C, with 'cooperation' as norm, i.e. $r^C = (C, C)$, and C^D with $r^D = (D, D)$. Let the entire population initially live in C^D. First assume $r^o(\kappa^t) = (C, C)$. Then, there exist many farsighted-successful migrations from C^D to C^C, e.g. a coalition of all inhabitants, or a coalition of some compliers, followed by uncoordinated migration of the remaining inhabitants of C^D. Now assume $r^o(\kappa^t) = (D, D)$. Then no farsighted-successful migration exists. Although a coalition of some compliers could immediately benefit from moving to C^C, this situation is not stable, since as long as some selfish inhabitants are left over in C^D they prefer to move to C^C. When all selfish inhabitants migrate to C^D compliers are, however, worse off than initially.

5.5 Weak migration-proof equilibrium

One may criticize migration-proof equilibria by arguing that too many coalitional deviations are allowed. Shouldn't coalitional deviations also be immediately beneficial? Shouldn't one finally end up in a migration-proof equilibrium instead of only a Nash-stable equilibrium? We formalize these concerns by defining strongly-successful migrations and a corresponding weak migration-proof equilibrium:

Definition 6 *Migration by a coalition m is **strongly successful** if it is immediately beneficial and there exists a (possibly empty) sequence of uncoordinated migrations that starts after the migration of m and ends in a* migration-proof equilibrium *where all members of m are strictly better off than initially.*

Definition 7 *A society $(\{r^j\}_{j\in J}, \alpha, \mu^t)$ constitutes a **weak migration-proof equilibrium** if it is a Nash-stable equilibrium and there exists no* strongly successful *coalitional migration.*

Obviously, every migration-proof equilibrium is a weak migration-proof equilibrium but the opposite may not hold. Our results are very robust in the sense that they hold for both open migration-proof equilibria and weak open migration-proof equilibria.

5.6 (Weak) open migration-proof equilibrium

For a fixed set of communities there is no possibility that new norms arise, which could challenge the existing ones. For example, every society with only one community constitutes a migration-proof equilibrium. Hence, we need an opportunity for new empty communities to enter the society in order to analyse which norms will develop.

Consider an original society $(\{r^j\}_{j\in J}, \alpha, \mu^t)$ and let $\{r^e\}_{e\in E}$ be a collection of norms of new empty communities that enter the society. We call the society $(\{r^j\}_{j\in J} \times \{r^e\}_{e\in E}, \alpha', \mu^t)$ an augmented society, and define by $\alpha' = (\alpha, 0)$ an allocation of the augmented society where the newly entered communities are empty and the allocation within the original communities is the same as in the original society. We define a (weak) open migration-proof equilibrium by

Definition 8 *A society $(\{r^j\}_{j\in J}, \alpha, \mu^t)$ constitutes a **(weak) open migration-proof equilibrium**, if for every possible collection of entering empty communities $\{r^e\}_{e\in E}$ the augmented society $(\{r^j\}_{j\in J} \times \{r^e\}_{e\in E}, (\alpha, 0), \mu^t)$ is a (weak) migration-proof equilibrium.*

5.7 Regularity conditions

Some regularity conditions are required for our results.

Condition 1 (C1) *A complier optimal norm $r^o(\kappa)$ exists for all κ.*[1]

Condition 2 (C2) *There are at least some compliers in the society, i.e. $\kappa^t > 0$.*

Condition 2 is relevant because we can obviously say nothing interesting about moral norms when it is common knowledge that there are no compliers at all.

For the next condition, let us define the highest payoff that selfish inhabitants can achieve, under the given selfish equilibrium selection function, when no compliers are present by

$$U_{\kappa=0} := \max_{r \in S} U(0, r). \tag{9}$$

Condition 3 (C3) *For every κ compliers can be least as well off as inhabitants of a purely selfish community, i.e. $V(\kappa, r^o(\kappa)) \geq U_{\kappa=0}$ $\forall \kappa$.*

Condition 4 (C4) *Compliers cannot achieve higher expected utility for some $\kappa < \kappa^t$ than they can maximally achieve for κ^t, i.e. $V(\kappa, r^o(\kappa)) \leq V(\kappa^t, r^o(\kappa^t))$ for $\kappa < \kappa^t$.*

Conditions C3 and C4 could only be violated when there are multiple selfish Nash equilibria. To illustrate a violation consider a game G with payoff-matrix

	A	B
A	1,1	0,1
B	1,0	0,0

The unique complier optimal norm is (A,A), but every strategy-profile is a selfish Nash equilibrium. Problems arise, for example, with a selfish equilibrium selection function that selects (B,B) when in a community $\kappa \geq \kappa^t$ and (A,A) when $\kappa < \kappa^t$. Then conditions C3 and C4 are violated and one can easily show that no open migration-proof equilibrium exists where a community has complier share κ^t, i.e. Proposition 3 (below) could not hold.

Conditions C3 and C4 are, however, always fulfilled for "well-behaved" selfish equilibrium selection functions, e.g. when those selfish Nash equilibria are selected that are best for compliers. Conditions C1-C4 are fulfilled for all examples in Section 3.

[1] Condition 1 can be relaxed such that a complier optimal norm has to exist just for κ^t. This requires a more complicated formulation of conditions C3, C4 in this section and of ρ^o and $C5$ in section 6.2.

Condition 5 (C5) *There exists a* $r^o(\kappa^t) \in R^o(\kappa^t)$ *such that for all* $\kappa > \kappa^t$ *it holds true that* $V(\kappa, r^o(\kappa^t)) \geq V(\kappa^t, r^o(\kappa^t))$.

Condition C5 says that at least for some complier optimal norm $r^o(\kappa^t)$ compliers' expected utility is not reduced when the compliers' share is higher than κ^t. This is important for our uniqueness-result where the proof is based on a coalition of compliers that migrates to a community that applies $r^o(\kappa^t)$. In example 3.3 (the public goods game with costly and imprecise punishment) C5 holds if and only if $r^o(\kappa^t)$ includes no punishment or $\eta = 0$ (which means punishment is precise). For all other examples C5 always holds. In Section 6.2 we allow the selected norm to depend on the actual compliers' share in a community and can thus replace condition C5 by a weaker condition, which also holds in all cases of example 3.3.

5.8 Characterization of equilibria

We now formally characterize open migration-proof equilibria. We first show that a society $(\{r^k\}_k, \alpha, \mu^t)$ in which the entire population lives in a single community that applies a complier optimal norm $r^o(\kappa^t)$ is an open migration-proof equilibrium. Then we show that in every (weak) open migration-proof equilibrium, compliers' expected utility is given by $V(\kappa^t, r^o(\kappa^t))$, since otherwise, compliers could perform a strongly-successful migration to an empty community with norm $r^o(\kappa^t)$.

We start with a helpful lemma that characterizes Nash-stable equilibria:

Lemma 1 *There can be no Nash-stable equilibrium that gives some compliers a strictly higher expected utility than* $V(\kappa^t, r^o(\kappa^t))$.

Proof. In every society there is at least one populated community C^a with a compliers' share $\kappa^a \leq \kappa^t$. By C4 compliers' expected utility in community C^a cannot exceed $V(\kappa^t, r^o(\kappa^t))$. In a Nash-stable equilibrium compliers' expected utility must be equal in every populated community and therefore no complier can have expected utility higher than $V(\kappa^t, r^o(\kappa^t))$. ■

We can proceed to our first result:

Proposition 3 *Every society* $(\{r^k\}_k, \alpha, \mu^t)$ *in which the entire population lives in a single community* C^o *that applies a complier optimal norm* $r^o(\kappa^t)$ *that satisfies C5 constitutes an open migration-proof equilibrium.*

Proof. It directly follows from the definition of an open migration-proof equilibrium that if *every* society with the entire population in C^o constitutes a migration-proof equilibrium, then also every such society constitutes an open migration-proof equilibrium. A society with the entire population in a single community is a Nash-stable equilibrium. It remains to show that there exists no farsighted-successful coalitional deviation.

By Lemma 1, compliers' expected utility in a Nash-stable equilibrium is bounded from above by $V(\kappa^t, r^o(\kappa^t))$. Therefore, no coalitional deviation exists that can be farsighted-successful for compliers. When a coalition consisting only of selfish inhabitants deviates by moving to some previously empty communities, in those communities selfish inhabitants' and compliers' expected utility is bounded from above by $U_{\kappa=0}$. Further, compliers' share in C^o cannot decrease when only selfish inhabitants move away and therefore, by C5, expected utility in C^o can not fall below $V(\kappa^t, r^o(\kappa^t))$, which is weakly higher than $U_{\kappa=0}$ by C3. Thus no complier prefers to leave C^o and selfish inhabitants have no chance to attain expected utility higher than in the initial situation, because $U_{\kappa=0} \leq V(\kappa^t, r^o(\kappa^t)) \leq U(\kappa^t, r^o(\kappa^t))$. Hence, no farsighted-successful coalitional deviation exists. ■

We continue with the second result, which says that all weak open migration-proof equilibria (and thus also all open migration-proof equilibria) must be complier optimal:

Proposition 4 *In every weak open migration-proof equilibrium compliers' expected utility equals* $V(\kappa^t, r^o(\kappa^t))$ *in all populated communities.*

Proof. By Lemma 1 there exists no Nash-stable equilibrium and thus also no weak open migration-proof equilibrium where compliers' expected utility exceeds $V(\kappa^t, r^o(\kappa^t))$. Thus it remains to check that there can exist no weak open migration-proof equilibrium where compliers' expected utility is smaller than $V(\kappa^t, r^o(\kappa^t))$.

Denote compliers' expected utility in the original society by V_{orig}. Suppose that this society is a weak open migration-proof equilibrium with $V_{orig} < V(\kappa^t, r^o(\kappa^t))$. By C5 there exists a complier optimal norm $r^o(\kappa^t)$ with $V(\kappa, r^o(\kappa^t)) \geq V(\kappa^t, r^o(\kappa^t)) > V_{orig}$ for all $\kappa \geq \kappa^t$. Augment the society by a community C^o applying such a complier optimal norm $r^o(\kappa^t)$. We show that the immediately beneficial coalitional migration of all compliers to C^o is also always strongly successful, which

contradicts the assumption that the original society is a weak open migration-proof equilibrium.

To show this, we must distinguish two cases:

Case 1: Assume $U(1, r^o(\kappa^t)) > U_{\kappa=0}$. Then a coalition of all selfish inhabitants follows to C^o by uncoordinated migration. The society with the entire population in C^o is an open migration-proof equilibrium by Proposition 3.

Case 2: Assume $U(1, r^o(\kappa^t)) = U_{\kappa=0}$. (Note that by C3 and C5 this implies that $V(\kappa, r^o(\kappa^t)) = U(\kappa, r^o(\kappa^t)) = U_{\kappa=0}$ for all $\kappa \geq \kappa^t$). Let a coalition of all selfish inhabitants that live in communities with expected utility lower than $U_{\kappa=0}$ move to C^o by uncoordinated migration. In the case where all selfish inhabitants have moved to C^o, we have an open migration-proof equilibrium by Proposition 3.

Otherwise, we have a Nash-stable society where some selfish inhabitants live outside C^o and every member of the society has expected utility of $U_{\kappa=0}$. We confirm that such a society constitutes a migration-proof equilibrium. Since compliers can never get expected utility higher than $V(\kappa^t, r^o(\kappa^t)) = U_{\kappa=0}$, they will not participate in a farsighted migration. By the same reason, compliers will also never emigrate by uncoordinated migration from C^o to a community that is populated only by selfish inhabitants. Selfish inhabitants could get expected utility above $U_{\kappa=0}$, however, only in a community with a positive compliers' share. Since selfish inhabitants can never induce compliers to leave C^o, no coalition with only selfish members can perform a farsighted-successful migration. ■

6 Competition of institutions and more

In this section we present two different extensions of the basic voting-by-feet model. We present each extension separately, but it is also straightforward to combine the two.

6.1 Competition of institutions

We first present an extension of the basic model where communities not only differ in the norm, but also in the game that describes social interaction and additionally by the selected Nash equilibria.

Let there be a set Γ of possible games[1] and let each community C^j select a

[1]Except for requiring that the modified versions of conditions C1-C5 (see in the text below and Section 5) hold, we need no additional restriction on the set of games.

game out of Γ with normal form representation $G^j = (N^j, u^j, S^j)$ that describes the institution of social interaction in the community. Further, each community has a norm r^j which is a strategy profile of G^j. Finally, each community has a selfish equilibrium selection function $\psi^j : [0;1] \times S^j \to S^j$ that selects for every compliers' share and norm a single selfish Nash equilibrium strategy profile of the induced game $G^{j,\kappa,r}$. By defining a set $\Psi(G)$ of allowed equilibrium selection functions for a game G and requiring $\psi^j \in \Psi(G^j)$, one can impose refinements on the class of possibly selected selfish Nash equilibria (see Section 2 for details). Let us denote the selected selfish Nash equilibrium conditional on a selection function ψ by $s(\kappa, r, \psi) \equiv \psi(\kappa, r)$. We denote the triple of game, norm, and selfish equilibrium selection function by $\lambda^j = (G^j, r^j, \psi^j)$ and call it a *norm-institution*. Similar to the definition in Section 2.4, expected utilities of selfish and compliant inhabitants in a community with norm-institution λ^j are given by

$$\begin{aligned} U(\kappa^j, \lambda^j) := &\quad \frac{1}{n}\sum_{i=1}^{n}\sum_{\theta_{-i}} \Pr(\theta_{-i}|\kappa^j) u_i^j(s(\kappa^j, r^j, \psi^j), s_{-i}^{\theta}(s(\kappa^j, r^j, \psi^j), r^j)) \\ V(\kappa^j, \lambda^j) := &\quad \frac{1}{n}\sum_{i=1}^{n}\sum_{\theta_{-i}} \Pr(\theta_{-i}|\kappa^j) u_i^j(r_i^j, s_{-i}^{\theta}(s(\kappa^j, r^j, \psi^j), r^j)). \end{aligned} \tag{10}$$

A norm-institution that maximizes compliers' expected payoff in a community with compliers' share κ is defined by

$$\lambda^o(\kappa) \in \Lambda^o(\kappa) := \arg\max_{\lambda} V(\kappa, \lambda). \tag{11}$$

A society is completely characterized by the norm-institutions of its communities and the allocation of the total population, i.e. by $(\{\lambda^j\}_{j\in J}, \alpha, \mu^t)$. It turns out that the same definitions, which we used to model competition of norms, can be used to model competition of norm-institutions and that we get equivalent results. Simply, every norm that appears in a definition, condition, proposition or proof of Section 5 has to be replaced by the corresponding norm-institution λ.

The same results that hold in equilibrium for norms hold for norm-institutions. To sum up, this implies, first that there is always an open migration-proof equilibrium with the entire population in a community C^o that has a complier optimal norm-institution $\lambda^o(\kappa^t)$, and second that in all weak open migration-proof equilibria compliers' utility is given by $V(\kappa^t, \lambda^o(\kappa^t))$. In other words: A complier optimal combination of institution, norm, and selfish equilibrium strategies arises.

6.2 Allowing norms to vary with the compliers' share in a community

In this extension of the basic model we allow communities to adapt their norm to the actual compliers' share. Let $\rho : [0,1] \to S$ be a so called *norm-function* that selects a norm for every compliers' share. A *complier optimal norm-function* ρ^o selects for every possible compliers' share a complier optimal norm $r^o(\kappa)$. Formally,

$$\rho^o \in \{\rho \mid \rho^o(\kappa) \in R^o(\kappa) \text{ for all } \kappa \in [0,1]\} \tag{12}$$

Note that the following identity holds: $V(\kappa, \rho^o(\kappa)) \equiv V(\kappa, r^o(\kappa))$.[1] We assume now that each community has a norm-function instead of a fixed norm. Thus a society is formalized by a collection of norm-functions for each community and the allocation of the population over the communities $(\{\rho^j\}_{j\in J}, \alpha, \mu^t)$. We can easily adapt the framework of Section 5 to this modified approach. The definitions remain unchanged except for the fact that we use the new definition of a society. Conditions C1-C4 will not be changed, but we can relax C5 as follows:

Condition C5 (modified version). $V(\kappa, r^o(\kappa)) \geq V(\kappa^t, r^o(\kappa^t))$ *for* $\kappa > \kappa^t$.

Being symmetric to C4, this version of C5 means that the maximal expected utility compliers can achieve shall not be lower for compliers' shares above κ^t than it is for κ^t. All examples in Section 3 fulfill this modified version of C5. The following modified version of Proposition 3 holds:

Proposition 3 (modified version). *Every society* $(\{\rho^k\}_k, \alpha, \mu^t)$ *in which the entire population lives in a single community* C^o *that applies a complier optimal norm function* ρ^o *constitutes an open migration-proof equilibrium.*

The proof is identical to the original proof of Proposition 3 in Section 5. Further, Proposition 4 carries over with exactly the same wording and only a small modification in the proof.[2] Thus when allowing norms to adapt to the compliers' share in a community, we find also for the public goods game with imprecise punishment that always an open migration-proof equilibrium exists and that in every (weak) open migration-proof equilibrium, compliers' expected utility equals $V(\kappa^t, r^o(\kappa^t))$.

[1] Do not get confused by the notation: $\rho^o(\kappa)$ denotes a function value of the norm function ρ^o, whereas $r^o(\kappa)$ is simply a label for a norm that is complier optimal for a compliers' share κ.

[2] Substitute in the original proof of Proposition 4 sentences 5-6 by: "Augment the society by a community C^o applying a complier optimal norm function ρ^o.".

7 Conclusions

We have explored moral norms and institutions in a society inhabited by a fixed share of norm-compliers and selfish players under the assumption that types are private knowledge. We have shown that in a model of voting-by-feet, where communities with different norms and institutions compete over inhabitants, complier optimal norms and institutions arise.

The concept of complier optimal norms can be combined with various emotions or preferences over material outcomes. We illustrated this for a simple envy-augmented utility function. With a total of just 2 parameters - compliers' share and degree of envy (equal for all players) - this behavioral model makes clear predictions that capture the stylized facts from a wide range of economic experiments, like conditional cooperation, costly punishment, giving in dictator games and concerns for social efficiency, as well as the role of intentions in variations of ultimatum and best-shot games.

Complier optimal norms have an ethical foundation related to rule utilitarianism. Following a complier optimal norm can be interpreted as a form of *social rationality* in societies where inhabitants' desire for honest commitment is private information.

As mentioned before, the idea that some people follow norms by a desire for consistency has long been analysed in social-psychology and also has become the focus of some recent models and experiments in economics. Under this premise, the explicit design of rules of conduct, i.e. norms, becomes an important element of institutional design. The result of our voting-by-feet model suggests complier optimality as a stability criterion in this context.

Related to this, we analyse in our research in progress how traditional methods of mechanism design can be extended to joint *norm-mechanism* design. We show, for example, how this can be done for the revelation principle. The basic idea of a *direct norm-mechanism* is to make truth-telling the norm for compliers and to provide incentives that selfish players truthfully reveal that they are selfish.

Complier optimal norms are derived from a symmetric ex-ante situation, where inhabitants are ignorant about their role in later social interactions (except for knowing the own moral motivation). An exciting topic for future research is to analyse how development of and acceptance to norms is influenced by the amount of knowledge people have about their likely position in later social interactions. This is also connected to the general role of communication in the formation of norms. Communication has become the focus of a number of interesting recent

economic experiments and, concerning these questions, we expect many insightful future studies to come.

References:

Andreoni, J., Miller, J. H., 2002. "Giving according to GARP: An experimental test of the Consistency of Preferences for Altruism", Econometrica 70(2): 737-753.

Berg, J., Dickhaut, J., McCabe K., 1995. "Trust, Reciprocity, and Social History," Games and Economic Behavior 10: 122–142.

Bergstrom, T. C., 2002. "Evolution of Social Behaviour: Individual and Group Selection", Journal of Economic Perspectives 16(2): 67-88.

Binmore, K., 1994. "Game Theory and the Social Contract", Cambridge MIT Press.

Blount, S. 1995. "When social outcomes aren't fair: The effect of causal attributions on preferences." Organisational Behavior & Human Decision Processes 63: 131-144.

Bolle, F., Ockenfels P. 1990. "Prisoner's dilemma as a game with incomplete information," Journal of Economic Psychology 11: 69-84.

Bolton, G., Ockenfels A., 2000. "ERC: A Theory of Equity, Reciprocity, and Competition," American Economic Review 90(1): 166-193

Camerer, C., Thaler, R., 1995. "Ultimatums, dictators, and manners," Journal of Economic Perspectives 9: 209-219.

Caplin, A., Nalebuff, B., 1997. "Competition among Institutions," Journal of Economic Theory 72(2): 306-342.

Charness, G., Dufwenberg, M., 2003. "Promises & Partnership," Research Papers in Economics 2003:3, Stockholm University, Department of Economics.

Charness, G., Rabin, M., 2002. "Understanding social preferences with simple tests," The Quarterly Journal of Economics 117: 817-869

Chwe, M. S.-Y., 1994. "Farsighted Coalitional Stability", Journal of Economic Theory 63: 299-325.

Clark, K., Sefton, M., 2001. "The sequential prisoner's dilemma: Evidence on reciprocation," Economic Journal 111: 51-68.

Cialdini, R. B., 1993. "Influence: Science and Practice", 3rd edition New York: Harper Collins.

Conley, J. P., Konishi, H., 2002. "Migration-proof Tiebout equilibrium: existence and asymptotic efficiency", Journal of Public Economics 86(2): 243-262.

Cox, J. C., Friedman, D., Gjerstad, S., 2005. "A Tractable Model of Reciprocity and Fairness", Discussion paper, UCSD.

Dufwenberg M., Kirchsteiger G., 2004. "A Theory of Sequential Reciprocity", Games and Economic Behaviour 47: 268–298

Ellingsen, T., Johannesson M., 2004. "Promises, Threats, and Fairness", Economic Journal 114: 397-420.

Elster, J., 1989. "Social Norms and Economic Theory", Journal of Economic Perspectives 3(4): 99-117

Falk, A., Fehr, E., Fischbacher, U., 2003. "On the Nature of Fair Behavior," Economic Inquiry, vol. 41(1): 20-26

Falk, A., Fehr, E., Fischbacher, U., 2005. "Driving Forces Behind Informal Sanctions," Econometrica 73(6): 2017-2030

Falk, A., Fischbacher U., 2006. "A Theory of Reciprocity", Games and Economic Behaviour 54: 293-315

Feddersen, T. J., Sandroni A., 2002. "A Theory of Participation in Elections", mimeo, University of Rochester

Fehr, E., Schmidt K. M., 1999. "A Theory Of Fairness, Competition, And Cooperation", The Quarterly Journal of Economics 114(3): 817-868

Fehr, E., Gaechter S., 2000. "Fairness and Retaliation: The Economics of Reciprocity", Journal of Economic Perspectives 14(3): 159-181.

Festinger, L., 1957. "A Theory of Cognitive Dissonance", Stanford: Stanford University Press.

Geanakoplos, J., Pearce, D., Stacchetti, E., 1989. "Psychological Games and Sequential Rationality", Games & Economic Behavior 1: 60–79.

Gneezy, U., 2005. "Deception: The Role of Consequences", American Economic Review 95(1): 384-394.

Greenberg, J. and S. Weber, 1986. "Strong Tiebout equilibrium under restricted preference domain", Journal of Economic Theory 38: 101-111.

Gul, F., Pesendorfer, W., 2005. "The Canonical Type Space for Interdependent Preferences", mimeo.

Gürerk, Ö, Irlenbusch, B., Rockenbach, B., 2006. "The Competitive Advantage of Sanctioning Institutions", Science, 7 April 2006: 108-111.

Güth, W.; 1995. "On ultimatum bargaining experiments - A personal review", Journal of Economic Behavior and Organization 27: 329-344.

Harrsion, G. W., Hirshleifer J., 1989. "An Experimental Evaluation of Weakest Link / Best Shot Models of Public Goods", Journal of Political Economy 97: 201-225.

Harsanyi, J., 1985. "Does Reason Tell Us What Moral Code to Follow, and Indeed, to Follow Any Moral Code at All?", Ethics 96: 42-55.

Harsanyi, J., 1992. "Game and Decision Theory in Ethics", in the Handbook of Game Theory, vol 1, Edited by R. Aumann and S. Hart.

Henrich, J., Boyd, R., Bowles, S., et al., 2001. "In search of homo economicus: Behavioural experiments in 15 small-scale societies", American Economic Review 91: 73–78.

Heider, F., 1946. "Attitudes and Cognitive Organization", Journal of Psychology 21: 107–112.

Hoffman, E., McCabe, K., Shachat, K., Smith, V., 1994. "Preferences, Property Rights, and Anonymity in Bargaining Games," Games and Economic Behavior 7(3): 346-380

Kerr, N.L., Kaufmann-Gilliland, C.M., 1994. "Communication, Commitment, and Cooperation in Social Dilemmas", Journal of Personality and Social Psychology 66(3): 513–529.

Kohlberg, L., 1984, "The Psychology of Moral Development", Harper & Row San Francisco

Levine, D. K., 1998. "Modeling Altruism and Spitefulness in Experiments", Review of Economic Dynamics 1(3): 593-622.

López-Pérez, R., 2005. "Emotions Enforce Fairness Norms (a Simple Model of Strong Reciprocity)", mimeo

Mas-Collel, A., Whinston, M., Green, J., 1995. "Microeconomic Theory", Oxford University Press

Newcomb, T., 1953. "An Approach to the Study of Communicative Acts", Psychological Review 60: 393–404.

Ostrom, E., 2000. "Collective Action and the Evolution of Social Norms", Journal of Economic Perspectives 14(3): 137-158.

Prasnikar, V., Roth, A., 1992, "Considerations of Fairness and Strategy: Experimental Data from Sequential Games", Quarterly Journal of Economics 107 (3): 865-888

Rabin, M., 1993. "Incorporating Fairness into Game Theory and Economics", American Economic Review 83 (5): 1281-1302.

Rawls, J., 1971, "A Theory of Justice", Harvard University Press, Cambridge.

Roth, A. E., 1995. "Bargaining experiments", in Kagel, J., Roth, A. (Eds.), Handbook of Experimental Economics. Princeton University Press, Princeton, pp. 253-348.

Sally, D. F., 1995. "Conversation and Cooperation in Social Dilemmas: Experimental Evidence from 1958 to 1992", Rationality and Society 7(1): 58–92.

Sobel, J.,2005. "Interdependent Preferences and Reciprocity," Journal of Economic Literature 43(2): 396–440.

Segal, U., Sobel, J., 2006. "Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings.", University of California, San Diego Discussion Paper, revised version.

Tiebout, C. M., 1956. "A Pure Theory of Local Public Expenditures", Journal of Political Economy 64(5): 416–424.

Westhoff, F., 1977. "Existence of equilibrium in economies with a local public good", Journal of Economic Theory 14: 84-112.

Appendix A

In this appendix we proof the results that we stated in Example 3.4.

Ultimatum Game

Note that the results are based on the selfish equilibrium selection function from Example 3.4, i.e. player 2 accepts an offer when he is indifferent between accepting or rejecting. Further, if player 1 is indifferent between two offers he selects the higher offer. These assumptions imply that selfish players do not play mixed strategies.

Lemma 2 *A selfish player 2 accepts an offer x if and only if $x \geq x^* := \frac{\alpha}{1+2\alpha}$.*

Proof. A selfish player 2 will obviously accept all offers $x \geq 0.5$. When he accepts an offer $x < 0.5$ his utility is given by $x - \alpha((1-x) - x)$ when he rejects the offer his utility is given by 0. This implies that an offer is accepted if and only if $x \geq \frac{\alpha}{1+2\alpha} = x^*$. ■

Proposition 5 *In every complier optimal norm equilibrium player 1 always offers $x^o := \min\{0.5, \kappa + (1-\kappa)x^*\}$ and all offers are accepted on the equilibrium path. (Note that $x^* \leq x^o \leq 0.5$.)*

Proof. Take a norm r^o where a compliant player 1 offers x^o and a compliant player 2 accepts x^o for sure and rejects all other offers.

I. In the first part of the proof we show that in the norm equilibrium with r^o a selfish player 1 also offers x^o. When a selfish player 1 offers x, his expected utility is given by

$$u_1^{\kappa,r^o}(x) = \begin{cases} 0 & \text{if } x < x^* \\ (1-\kappa)(1-x) & \text{if } x^* \leq x < x^o \\ 1-x & \text{if } x^o \leq x \leq 0.5 \\ 1-x-\alpha(2x-1) & \text{if } 0.5 < x \end{cases}.$$

There are only two candidates for maxima: x^* and x^o. The selfish player offers x^o if and only if $(1-\kappa)(1-x^*) \leq 1-x^o \Leftrightarrow x^o \leq \kappa + (1-\kappa)x^*$, which is fulfilled by the definition of x^o.

II. In the second part of the proof we show that every norm equilibrium with a different equilibrium outcome than under r^o yields a strictly lower compliers' expected utility.

We start by discussing some upper bounds on compliers' expected utility. The expected sum of utility of player 1 and 2 is given by $T := 2\left[(1-\kappa)U + \kappa V\right]$. We call T *total utility.* Since $V_1 \leq U_1$ and $V_2 \leq U_2$ has to hold, compliers' expected utility is bounded from above by $\frac{1}{2}T$. If we know that a compliant player 1 has strictly lower expect utility than a selfish player 1, i.e. $\Delta_1 \equiv U_1 - V_1 > 0$, the upper bound decreases to $\frac{1}{2}(T - \Delta_1)$, because $V_2 \leq U_2$ must still hold. For a

given norm equilibrium, let A denote the expected total disutility caused by envy and let R be the expected share of rejected offers. Total utility is then given by $T = 1 - A - R$. This implies $\frac{1}{2}(1 - A - R - \Delta_1)$ as upper bound for compliers' expected utility.

Consider first the case $x^o = 0.5$. Under r^o no player ever feels envious and therefore total utility is given by 1 and compliers' expected utility reaches its upper bound of 0.5. In any other equilibrium outcome there would either be some envy or rejected offers, which would lead to a strictly lower compliers' expected utility.

Consider now the case $x^o < 0.5$. Compliers' expected utility under r^o is then given by $V^o = \frac{1}{2}(1 - A^o)$, where $A^o{=}\alpha((1 - x^o) - x^o) = \alpha(1 - 2x^o)$ is the total disutility by envy (experienced by player 2). In the following 5 steps we show that norms with different behavior on the equilibrium path must lead to strictly lower compliers' expected utility than r^o.

1. Let r' be a norm that differs from r^o such that with positive probability offers below x^o are made and accepted. Since there are lower offers under r', total envy A' under r' is strictly higher than under r^o. This means compliers' utility is bounded from above by $\frac{1}{2}(1 - A') < V^o$.

2. Consider a norm r' that differs from r^o such that with positive probability an offer of x^o is rejected on the equilibrium path. Rejecting an offer of x^o reduces total envy by $\alpha(1 - 2x^o)$ but also reduces total monetary payoff by 1. This reduction in monetary payoff reduces the total utility by more than the reduction of envy increases it. This is most easily seen by observing that $x^o > x^*$ and that therefore a player 2 considers the decrease in monetary payoff (already of his share) to be more severe than the positive effects of the reduction in envy (recall the proof of Lemma 2). Hence, under r' total utility is bounded by a level strictly below V^o.

3. Consider a norm r' that differs from r^o such that with positive probability offers below x^o are made and accepted *and* offers of x^o are rejected on the equilibrium path. It is also straightforward to show that such a norm yields compliers' expected utility strictly below V^o (we omit the steps that are very similar to 1. and 2.).

4. Note that there exists no norm where a selfish player 1 makes offers above x^o (in the actual case with $x^o < 0.5$). To see this, assume compliers want to induce a selfish player to offer some $x^s > x^o$. The best way to achieve this is to accept only offers of x^s and to reject all other offers. Using similar calculations than in the first part of the proof, we find, however, that a selfish player 1 offers x^* instead of x^s whenever $(1 - \kappa)(1 - x^*) > 1 - x^s \Leftrightarrow x^s > \kappa + (1 - \kappa)x^* = x^o$.

5. Finally, we show that there is no complier optimal norm r' where compliers make offers above x^o with positive probability. Note that it cannot be complier optimal to make offers above 0.5, since total envy is minimized by offering 0.5. Further, a compliant player 2 should accept all offers $x \geq x^o$, since it is obviously not complier optimal to reject an offer in between x^o and 0.5.

Let $F(x)$ denote the distribution function of compliant player 1's offers under r'. The difference in expected utility of a selfish player 1 to that of a compliant player 1 is given by $\Delta_1' = \int_{x^o}^{0.5}(x - x^o)dF(x)$. The difference in total envy between r^o and r' is given by $A^o - A' = 2\kappa\alpha \int_{x^o}^{0.5}(x - x^o)dF(x)$. Compliers' expected utility under r' is bounded from above by $\frac{1}{2}(1 - A' - \Delta_1) = \frac{1}{2}(1 - A^o) + \frac{1}{2}(A^o - A') - \Delta_1)$, which, therefore, can be written as $V^o + \frac{1}{2}[2\kappa\alpha - 1] \int_{x^o}^{0.5}(x - x^o)dF(x)$. This upper bound is greater or equal than V^o only if $2\kappa\alpha - 1 \geq 0$, i.e. $\alpha \geq \frac{1}{2\kappa}$. For $\alpha \geq \frac{1}{2\kappa}$ we find, however, $x^* \geq \frac{1}{2(1+\kappa)}$, which implies $\kappa + (1 - \kappa)x^* \geq 0.5 + \frac{\kappa^2}{1+\kappa}$ and thus $x^o = 0.5$. This contradicts the assumption, made for this case, that $x^o < 0.5$. ■

Ultimatum game with non-intentional offers

The following proposition and corollary characterize complier optimal norms in Blount's treatment with non-intentional random offers by player 1.

Proposition 6 *A compliant player 2 will accept an offer x in a complier optimal norm equilibrium if and only if $\frac{\alpha-\kappa}{1+2\alpha-\kappa} \leq x \leq \left|\frac{(1+\alpha)\kappa}{2\kappa\alpha-(1-\kappa)}\right|$.*

Proof. First note that the strategies of player 2 have no influence on how the offers are randomly selected. We can therefore analyse separately for any given offer x, whether this offer should be accepted or rejected in a complier optimal norm equilibrium.

For $x < 0.5$ an offer may be rejected when player 2 feels too envious. Since a compliant player 2 also values the utility of a compliant player 1, she clearly accepts all offers that a selfish player 2 would accept. Hence, we only have to analyse which offers $x < x^*$ should be accepted. If a compliant player 2 accepts an offer $x < x^*$ then compliers' expected utility is given by $\frac{1}{2}[\kappa(1 - x)] + \frac{1}{2}[x - \alpha(1 - 2x)]$ and otherwise it is 0. This implies that a complier rejects offers $x < 0.5$ if and only if $x < \frac{\alpha-\kappa}{1+2\alpha-\kappa}$.

It is interesting that under a huge degree of envy also very high offers could be rejected by player 2. This is because for $x > 0.5$ player 1 feels envy and if this envy is strong enough, a compliant player 2 could decide to reject the offer. This would lead to an outcome where both earned nothing and can therefore not envy each other. When there is an offer $x > 0.5$ and a compliant player 2 would accept it, compliers' expected utility is given by $\frac{1}{2}[(1 - x) - \alpha(2x - 1)] + \frac{1}{2}[x]$. If a complier rejects the offer, compliers' expected utility is given by $\frac{1}{2}[(1 - \kappa)((1 - x) - \alpha(2x - 1))] + 0$. This implies that a complier rejects offers $x > 0.5$ if and only if $x > \left|\frac{(1+\alpha)\kappa}{2\kappa\alpha-(1-\kappa)}\right|$. ■

Corollary 1 *If $\alpha\kappa \leq 1$ no offers above 0.5 and if $\alpha \leq \kappa$ no offers at all will ever be rejected by a compliant player 2 in a complier optimal norm equilibrium.*

Proof. For $\alpha\kappa \leq 1$ or $\alpha \leq \kappa$ we find $\left|\frac{(1+\alpha)\kappa}{2\kappa\alpha-(1-\kappa)}\right| \geq 1$ and for $\alpha \leq \kappa$ we find $\frac{\alpha-\kappa}{1+2\alpha-\kappa} \leq 0$. The rest follows from Proposition 6. ■