

Moral Norms in a Partly Compliant Society

Sebastian Kranz*

University of Bonn

Revised version August 2009

Abstract

The article investigates social interaction among individuals who differ in their privately known motivation to comply with moral norms that are collectively rational in the sense that they maximize welfare given the distribution of moral motivation in the society. This yields tractable models of rule-consequentialism that can be tested with experimental data. The analysis focuses on two welfare principles: utilitarianism and complier optimality. The latter puts explicit welfare weight only on the type with the highest moral motivation. Already a simple model with two types is in line with a wide range of experimental stylized facts, like conditional cooperation, costly punishment, the role of intentions, or concerns for social efficiency.

Keywords: *moral norms, rule-utilitarianism, social preferences, complier optimality, fairness*

JEL Classifications: A13, C7, D63, D64, D71

*Address: Department of Economics, University of Bonn, Adenauerallee 24-26, 53113 Bonn, Germany; e-mail: skranz@uni-bonn.de. For very valuable discussions and suggestions I am especially grateful to Paul Heidhues. Further, I would like to thank an Associate Editor and two Referees, as well as, Lukas Buchheim, Armin Falk, Urs Fischbacher, Martin Hellwig, Fabian Herweg, Thomas Gall, Werner Güth, Eugen Kovac, Johannes Münster, Georg Nöldeke, Heather Partner, Susanne Ohlendorf, Frank Riedel, Larry Samuelson, Karl Schlag, Reinhard Selten, Avner Shaked, Tymon Tatur and many others for very helpful comments. I thank the Deutsche Forschungsgemeinschaft for financial support through SFB/TR 15.

1 Introduction

The traditional economic postulate of a society inhabited only by selfish individuals has been severely challenged over the last decades by the insights of experimental economics. Although a substantial fraction of people shows behavior that is consistent with selfish behavior, a large fraction of people shows systematic deviations. For example, in one-shot interactions people cooperate in public good games or punish unfair behavior when punishment is costly (see e.g. Fehr and Gächter, 2000a).

We propose a model which can explain a wide range of stylized facts from economic experiments as the outcome of interaction between individuals with different privately known motivation to perform ethically rational actions. The model is deeply inspired by John Harsanyi's work on ethics (e.g. 1977, 1985, 1992). For Harsanyi, ethical (or moral) behavior is based on a notion of *collective rationality* that goes beyond playing individual best-replies given some fixed unselfish preferences:

"The theory of rational behavior in a social setting can be divided into *game theory* and *ethics*. Game theory deals with two or more individuals often having very different interests who try to maximize *their own* (selfish or unselfish) interests in a rational manner against all the other individuals, who likewise try to maximize their own (selfish or unselfish) interests in a rational manner. In contrast, ethics deals with two or more individuals often having very different personal interests, yet trying to promote the *common interests of their society* in a rational manner." (Harsanyi, 1992, p. 672)

Concretely, Harsanyi argues that rational morally motivated individuals should follow rules that maximize welfare in a society given that it becomes commonly known that morally motivated individuals follow these rules. In addition, he argues that society should adapt an utilitarian welfare criterion, i.e. to maximize the expected average utility over all individuals. The combination of both ideas constitutes the philosophical principle of *rule-utilitarianism*. This is a special case of *rule-consequentialism*, which allows also for alternative welfare criteria.¹ Harsanyi (1992, p. 692-694) sketches a formal model of a rule-utilitarian society, which remains, however, at a quite general level that is not suited to make clear

¹For a good survey on the philosophical literature on rule-consequentialism see Hooker (2008).

predictions across economic experiments; nor does it work out implications of partial moral motivation and private information in detail.

The main contributions of this paper are the following: First, we propose and characterize a model of rule-consequentialism that accounts for limited moral motivation and private information of types and can be directly applied to experimental observations. Second, we propose and investigate a welfare principle called ‘complier optimality’ that has several appealing properties. Third, we apply our model to a series of games that were widely studied in the literature on fairness and show that complier optimal behavior matches a large range of empirical stylized facts.²

In our model, there is a commonly known *norm* that specifies for each information set of a given underlying game which actions are permitted or forbidden. For example, if the underlying game is an ultimatum game, one norm would be to forbid player 1 to offer less than 40% of the cake, to forbid player 2 to accept offers below 30% of the cake, and to permit all other actions.

We call a norm *rule-consequentialistic* for a given welfare criterion if no other norm could induce behavior that leads to higher welfare. The exact form of rule-consequentialistic norms depends on the welfare criterion and on how a norm influences players’ behavior. In the basic model, presented in Section 2, we assume that there are only two types of players: selfish types who do not care about breaking norms, and morally motivated *compliers*, whose utility is reduced by a fixed amount when deviating from the norm. Types are private knowledge, but the probability that a player is a complier, which is identical for each player, is commonly known.

For general norms, the label *complier* could be misleading, since a complier does not automatically comply with a norm but deviates whenever the gains from a deviation are sufficiently large. However, under weak regularity conditions, we can establish a *compliance principle*: rule-consequentialistic norms can always be constructed such that they permit only one strategy profile that, in equilibrium, is exactly followed by compliers.

While we give some general characterization of rule-consequentialistic norms, this paper focuses on two welfare principles: utilitarianism (leading to *rule-utilitarian norms*), and a novel principle that we call ‘complier optimality’. A *complier optimal norm* maximizes expected utility of compliers, i.e. no explicit welfare weight is put on the utility of selfish types.

²Our examples include public goods games with and without punishment option, a sequential prisoners’ dilemma or simple trust game, dictator games, and intentional and non-intentional versions of an ultimatum game.

At first thought, one might worry that complier optimal norms could lead to severe negative discrimination against selfish types. However, a rather opposite result holds true in our set-up with private information: selfish types will always benefit from the presence of compliers, but compliers could be severely exploited if norms are not complier optimal. Private information of types implies that selfish players that act in the same way as compliers will also be treated in the same way, i.e. individuals cannot be discriminated for what they are but only for what they do.³ Since selfish types always have the option to mimic compliers, their expected utility is always weakly higher than compliers' expected utility. This implies that complier optimal norms maximize the expected utility of the worst-off type. Rule-utilitarian norms can force compliers to sacrifice so much of their own payoff for selfish types that compliers are worse off than inhabitants of a completely selfish society (this is often the case if the fraction of compliers is rather small, Section 3 offers many examples). In contrast, under complier optimal norms expected payoff of *every type* is weakly higher than in a completely selfish society (this holds true because one candidate for a complier optimal norm is a norm that allows compliers to act selfishly).

Harsanyi offered a famous justification for a utilitarian welfare criterion: welfare judgments should be impartially formed behind a fictitious veil of ignorance where each position in life is considered as equally likely; expected utility maximization behind such a veil of ignorance directly implies an utilitarian welfare criterion. Complier optimal norms also prescribe to be impartial with respect to players' positions; they do not prescribe, however, to account for additional payoff gains that selfish types could achieve under norms that allow to free-ride more effectively than under complier optimal norms.

While it is likely to remain a matter of subjective opinion whether complier optimal norms are deemed normatively more or less appealing than rule-utilitarian norms, we consider it worthwhile to investigate and exemplify the concrete differences between these two related principles.

Another argument to investigate complier optimal norms is given in Appendix C. In a nutshell, we consider a voting-by-feet model where social interactions can take place in different communities that can have different norms. A community's norm can be interpreted as a collective agreement of how to behave in social interactions within the community. Compliers feel guilty when breaking

³This point will be formalized in Appendix B. There we also discuss the case that types are not private information.

this agreement, while selfish types don't. Individuals can freely migrate between communities, i.e. compliers can decide which norms (agreements) they want to follow, but they cannot prevent selfish types from joining their community. We show that complier optimal norms generally emerge in this model.

Section 3, the main part of the paper, analyzes the predictions under complier optimal norms and rule-utilitarian norms for a series of games that have been widely analyzed in the experimental literature on fairness. Already under the assumption that every player's material utility is linear in money, the model with complier optimal norms captures many of the experimental stylized facts. Examples are conditional cooperation, the use of costly punishment, the role of intentions, the observation that reciprocal subjects tend to trust more, or concerns for social efficiency. For zero-sum games, complier optimal norms generally prescribe to act selfishly rational. Altruistic acts in games that are zero-sum in monetary payoffs —like giving in dictator games— can be complier optimal if preferences over material outcomes are extended for factors like risk-aversion, loss-aversion or envy.

Predicted behavior under rule-utilitarian norms is qualitative often similar to behavior under complier optimal norms, but can also substantially differ in some situations. Roughly summarized, both types of norms have in common that they prescribe to punish non-cooperative behavior (or reward cooperative behavior), whenever common knowledge of such a norm can effectively induce selfish players to act more cooperatively. Rule-utilitarian and complier optimal norms usually differ if no norm can effectively induce cooperative behavior by selfish types. This is for example the case when there is no or only a weak punishment option or norms cannot substantially change selfish players' incentives because there are too few compliers. In such cases, rule-utilitarian norms essentially prescribe to act altruistic towards selfish players. An example is to cooperate as a second mover in a sequential prisoners' dilemma game even if the first-mover has defected — a prediction that is almost never observed in experiments (see Section 3.3). In contrast, complier optimal norms essentially prescribe in those cases to act selfishly towards selfish types. Altruistic acts are then only conducted if the probability that the other player is a complier is sufficiently large.

Section 4 extends the model to allow for arbitrary many types with different moral motivation. Main results of the basic model carry over: types with smaller moral motivation are always weakly better off than types with higher moral motivation and a modified version of the compliance principle holds. Furthermore, we establish that under quite general conditions welfare under rule-consequentialistic

norms weakly increases whenever the type distribution shifts towards higher levels of moral motivation. In Appendix A, we analyze a different extension, in which the willingness to comply depends on the size of the welfare gains from collective norm compliance.

Quite surprisingly, Harsanyi's basic ideas about rational ethics have been ignored in large parts of the economic literature on fairness and social preferences, which has become a very popular field in recent years.⁴ Except for the model by Feddersen and Sandroni (2006), who analyze voting behavior in a society with rule-utilitarian individuals, we are not aware of any model of social preferences that cites Harsanyi or discusses whether a model with rule-consequentialistic types could be in line with experimental observations. A brief comparison of our model with related models of social preferences is given in Section 5. Section 6 illustrates possible extensions and concludes.

2 Basic Model

2.1 Basic set-up

We first model the interaction between selfish players and compliers for any arbitrary norm. Afterwards, we define rule-consequentialistic norms and the special cases of complier optimal and rule-utilitarian norms.

Social interaction in absence of compliers shall be described by an *underlying game* of perfect recall with extensive form representation Γ . $N = \{1, \dots, n\}$ denotes the set of players, z an outcome (terminal node) of the underlying game and $u(z) = (u_1(z), \dots, u_n(z))$ the *underlying payoffs*. Σ_i denotes the set of player i 's strategies, and $\Sigma = \Sigma_1 \times \dots \times \Sigma_n$ the strategy space. $A|h$ is the set of possible actions at an information set h .

There is a commonly known *norm* r , which is a correspondence that selects for each information set h of the underlying game, a non-empty subset of actions $r(h) \subset A|h$. We say an action $a \in A|h$ is *permitted* by the norm if $a \in r(h)$ and *forbidden* if $a \notin r(h)$. We say player i has *deviated* from the norm in an outcome z of the underlying game if he has played at least once a forbidden action in that outcome.

⁴For models of social preferences see for example, Rabin (1993), Levine (1998), Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Charness and Rabin (2002), Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006), Segal and Sobel (2007), Cox et. al. (2007) or López-Pérez (2008).

There are two types of players: compliers and selfish players. Utility of a selfish player is always given by the underlying payoff. If a compliant player has deviated from the norm, she feels guilty and her underlying payoff is reduced by an amount $g > 0$; otherwise her utility is given by the underlying payoff.⁵

We assume throughout the paper that players' types are private knowledge and independently drawn by an initial move of nature. A player is a complier with probability $\kappa \in [0, 1]$ and selfish with probability $1 - \kappa$. Types are drawn independently for each player. We call κ the *compliers' share* in the population and assume that κ is common knowledge.

For a given underlying game Γ , norm r , compliers share κ , and level of moral motivation g , let $\widehat{\Gamma} = \widehat{\Gamma}(r, \kappa, g, \Gamma)$ denote the extensive form of the resulting game of imperfect information. A strategy-profile of $\widehat{\Gamma}$ is denoted by a pair of strategy profiles of the underlying game $(\sigma^s, \sigma^c) \in \Sigma \times \Sigma$ such that σ_i^s and σ_i^c correspond to the strategies of a selfish and compliant player i , respectively. We say an outcome \widehat{z} of $\widehat{\Gamma}$ *corresponds* to an outcome z of the underlying game if, except for the initial move of nature in \widehat{z} , the histories of moves under \widehat{z} and z are identical. The definition of a norm deviation extends in the obvious way to the game with imperfect information: player i has deviated from the norm in an outcome \widehat{z} if and only if he has deviated from the norm in the corresponding outcome z of the underlying game.

Definition 1 *We say that $(\sigma^s, \sigma^c) \in \Sigma \times \Sigma$ is a **norm equilibrium** for given r, κ, g and Γ and if there exists beliefs μ such that $((\sigma^s, \sigma^c), \mu)$ is a Perfect Bayesian Equilibrium of the resulting game $\widehat{\Gamma}(r, \kappa, g, \Gamma)$.*

If the underlying game has a finite number of outcomes and mixed strategies are allowed then $\widehat{\Gamma}$ always has a Perfect Bayesian Equilibrium, i.e. a norm equilibrium exists for every possible norm. For alternative strategy spaces, however, norm equilibria may fail to exist for some norms.⁶ Generally, let R denote the set of norms for which a norm equilibrium exists for given Γ, κ and g .

⁵For simplicity, we assume that g is an exogenously given parameter. In Appendix B, we illustrate how g can make depended on the resulting welfare under norm compliance. For two different interpretations of norms and the parameter g see the discussion of the extended model in Section 4.

⁶For example, consider an ultimatum game where player 1 must offer an amount $x \in [0, 1]$ and player 2 can either accept or reject the offer, which yields underlying payoffs of $(1 - x, x)$ or $(0, 0)$, respectively. Consider a norm that prescribes player 2 to accept an offer x if and only if $x > \bar{x}$ where $\bar{x} \in (0, 1)$ is some threshold. If $g > \bar{x}$ and $\kappa > \bar{x}$ no norm equilibrium exists. The reason is that there is no optimal offer for a selfish player 1 as $\arg \min\{x | x > \bar{x}\} = \emptyset$. In

It can also be the case that multiple norm equilibria exist for a given norm. We assume that there is an equilibrium selection function ψ that selects for every norm $r \in R$ a unique norm equilibrium.⁷

Let $U_i^s(r, \kappa, g, \Gamma, \psi)$ and $U_i^c(\cdot)$ denote expected utility of a selfish and compliant player i , respectively, in the selected norm equilibrium under a norm $r \in R$. Since we assume that each player has the same probability to be a complier, expected utility of selfish and compliant types is given by $U^s(\cdot) := \frac{1}{n} \sum_{i=1}^n U_i^s(\cdot)$ and $U^c(\cdot) := \frac{1}{n} \sum_{i=1}^n U_i^c(\cdot)$, respectively.

Proposition 1 *In every norm equilibrium, selfish types have weakly higher expected utility than compliers, i.e.*

$$U_i^c(\cdot) \leq U_i^s(\cdot) \quad \forall i \text{ and } U^c(\cdot) \leq U^s(\cdot). \quad (1)$$

Proof. Since types are private knowledge, expected underlying payoffs of a player i only depends on the strategy of the underlying game that he is playing — not on his type directly. Since a selfish player never feels guilty, (σ^s, σ^c) can be a norm equilibrium only if for each player i expected underlying payoff of a selfish type is weakly higher than that of a compliant type; otherwise it is profitable for a selfish type to mimic a compliant type by deviating from σ_i^s to σ_i^c . ■

We say a norm r^o is *complier optimal* for given κ, g and ψ if it maximizes the expected average utility of compliant types, i.e.

$$r^o \in \arg \max_{r \in R} \{U^c(r, \cdot)\}.$$

A *rule-utilitarian norm* r^u maximizes the expected average utility of all players, i.e.

$$r^u \in \arg \max_{r \in R} \{(1 - \kappa)U^s(r, \cdot) + \kappa U^c(r, \cdot)\}.$$

Generally, let $w(\cdot)$ be a real valued function that only depends on expected or resulting underlying payoffs, disutility from guilt and players' types. The function $w(\cdot)$ may be decreasing in underlying payoffs, but we require that $w(\cdot)$ is non-increasing in disutility from guilt. The expected value function $W(\cdot) \equiv Ew(\cdot)$ is

contrast, norm equilibria always exist for norms that prescribe to accept an offer x if and only if $x \geq \bar{x}$. Depending on κ and g , a selfish player 1 will then either offer 0 or \bar{x} (see Section 3.5 for details).

⁷A selection function technically facilitates the analysis, since it guarantees a complete ordering of norms with respect to a given welfare criterion. Robustness of predictions can be checked by considering different equilibrium selection functions.

called a consequentialistic welfare criterion. A norm r^* is called a *rule consequentialistic norm* for a given welfare criterion W if it solves

$$r^* \in \arg \max_{r \in R} \{W(r, \cdot)\}.$$

Clearly, complier optimal and rule utilitarian norms are both special cases of rule consequentialistic norms.

A norm equilibrium (σ^s, σ^c) that is selected for a rule-consequentialistic norm is called a *rule-consequentialistic norm equilibrium*; similarly we define *rule-utilitarian norm equilibria* and *complier optimal norm equilibria*.

2.2 Compliance principle

In the remainder of this section, we establish regularity conditions which ensure that rule-consequentialistic norms exist from which compliers don't deviate in the resulting norm equilibrium. One reason why such norms may not exist is that norms cannot prescribe exact mixing probabilities. In order to manipulate mixing probabilities it can be optimal to forbid an action that compliers still choose with positive probability in the resulting norm equilibrium.⁸ This problem can be circumvented if one allows only pure strategies in the underlying game but includes the randomization devices as explicit moves of nature. For example, one could assume that nature draws a privately observed signal from a standard uniform distribution before each decision node. Since it can depend on the signal whether an action is permitted or forbidden, a norm can then specify exact randomization probabilities.

Another reason why it could be optimal that compliers deviate from the norm in a rule-consequentialistic norm equilibrium, is that the equilibrium selection function ψ may be ill-behaved in the sense that players choose welfare optimal

⁸Consider the following, rather technical, example. There are three players. Player 1 and 2 play the following matching pennies game:

	H	T	
H	1,0	0,1	
T	0,1	1,0	

Player 3 can perform no action, but receives a very high payoff if player 1 chooses H . Assume this payoff is so high that rule-utilitarian norms must maximize the probability with which player 1 chooses H . Assume $\kappa = 1$ and $g < 1$. Then player 1 and 2 will play a completely mixed strategy under every norm. In equilibrium, player 1 mixes such that player 2 is indifferent between choosing H or T . This implies that a rule-utilitarian norm must reduce player 2's payoff from playing T by forbidding him to play T .

strategies only if those strategies are forbidden. For an example, consider a coordination game in pure strategies described by the following payoff matrix:

	A	B
A	2,2	0,0
B	0,0	1,1

Consider the case $g < 1$, in which disutility from guilt is small enough such that both $\sigma^s = \sigma^c = (A, A)$ and $\sigma^s = \sigma^c = (B, B)$ are norm equilibria under every norm.⁹ Consider the following equilibrium selection function ψ : The norm equilibrium $\sigma^s = \sigma^c = (A, A)$ is selected under the norm r' that forbids every player to choose A , while under every other norm the norm equilibrium $\sigma^s = \sigma^c = (B, B)$ is selected. Even though players violate the norm r' in the selected norm equilibrium, r' is still the only rule-utilitarian (and complier optimal) norm under this equilibrium selection function, since the additional payoffs from playing (A, A) instead of (B, B) outweighs compliers' utility loss g from the norm deviation. Hence, in this example compliers will deviate from the only rule-utilitarian norm r' .

Such cases can be ruled out with a weak regularity condition on equilibrium selection. Assume that Γ is in pure strategies. We say a norm r *prescribes* the (pure) strategy-profile of compliers σ^c if at each information set only that action is permitted that compliers choose according to σ^c .

Definition 2 *An equilibrium selection function ψ is **regular** if there do not exist norms r and \tilde{r} with selected norm equilibria (σ^s, σ^c) and $(\tilde{\sigma}^s, \tilde{\sigma}^c)$, respectively, such that*

$$\sigma^c = \tilde{\sigma}^c \text{ but } \sigma^s \neq \tilde{\sigma}^s, \text{ or} \tag{R1}$$

$$\tilde{r} \text{ prescribes } \sigma^c \text{ but } \tilde{\sigma}^c \neq \sigma^c. \tag{R2}$$

Condition (R1) can be stated as follows: *only* if compliers change their behavior across two norms r and \tilde{r} , also selfish player may change their behavior. This condition can always be fulfilled since for given equilibrium strategies of compliers σ^c , the set of equilibrium strategies of selfish players does not depend on the norm. Since compliers are directly affected by norms but selfish players only indirectly, this condition on equilibrium selection seems very natural.

⁹There can exist more norm equilibria. E.g. for $\kappa = \frac{1}{3}$ there exists a norm equilibrium with $\sigma^c = (A, A)$ and $\sigma^s = (B, B)$ under norms that permit every complier to choose A .

Condition (R2) can be stated as follows: if compliers do not play equilibrium strategies σ^c under a norm that prescribes σ^c , then compliers will also not play σ^c under any other norm. This requirement can also always be fulfilled: if under some norm there exists a norm equilibrium in which compliers play σ^c then there clearly must exist a norm equilibrium in which compliers play σ^c under the norm that prescribes σ^c . Given that norms are a natural focal point, it seems strange to assume that compliers coordinate to play σ^c under some norm but do not coordinate to play σ^c under a norm that prescribes to play σ^c . Thus, also this second condition for a regular equilibrium selection function is quite natural.

The equilibrium selection function in the coordination game example above violates condition (R2): compliers do play (B, B) under the norm that prescribes to play (A, A) , but play $\sigma^c = (A, A)$ under the norm that forbids every player to play A . To illustrate why one needs condition (R1), consider the example for the case $\kappa = \frac{1}{3}$. Then, under norms that permit every player to play A , there exists a norm equilibrium with $\sigma^c = (A, A)$ and $\sigma^s = (B, B)$. Consider an equilibrium selection function that selects that norm equilibrium under every norm that permits every player to play A , and otherwise selects the norm equilibrium $\sigma^c = \sigma^s = (A, A)$. If g is sufficiently small, we again find that the rule-utilitarian norm equilibrium is $\sigma^c = \sigma^s = (A, A)$ while every rule-utilitarian norm forbids at least one player to choose A . However, this equilibrium selection function violates the condition (R1): selfish players change their equilibrium strategies under different norms even though compliers do not.

We can now establish the compliance principle:

Proposition 2 (Compliance Principle) *Assume the underlying game Γ is in pure strategies (but possibly has explicit randomization devices) and the equilibrium selection function ψ is regular. Then for every rule-consequentialistic norm equilibrium (σ^s, σ^c) there is a rule-consequentialistic norm that prescribes σ^c .*

Proof. The result follows almost immediately from the definitions. Let r' be a rule-consequentialistic norm and (σ^s, σ^c) be the selected norm equilibrium. Regularity of ψ implies that the norm equilibrium (σ^s, σ^c) is also selected under a norm r , which prescribes σ^c . Since r' and r induce the same norm equilibrium, expected underlying payoffs do not change; furthermore there is no disutility from guilt under r . Since welfare W is assumed to be non-increasing in disutility from guilt, r must also be a rule-consequentialistic norm. ■

For a given equilibrium selection function, we say a norm r *regular* if it prescribes compliers strategies σ^c in the the selected norm equilibrium. The com-

pliance principle implies that, in order to find all rule-consequentialistic norm equilibria, one can neglect without loss of generality all non-regular norms, from which compliers would deviate at some information set. In the examples of Section 3, we make extensive, while usually implicit, use of the compliance principle.

3 Examples

In this section, we illustrate complier optimal and rule-utilitarian norm equilibria for simple games that have been widely studied in the literature on fairness and compare the predictions with the experimental stylized facts. We assume in all examples that there is a strictly positive probability for both compliant and selfish players, i.e. $\kappa \in (0, 1)$.

Furthermore, we assume that all underlying games are in pure strategies and consider a regular equilibrium selection function that is characterized by the following two conditions. First, if a norm equilibrium exists where compliers follow the norm then such a norm equilibrium is selected. Second, from the equilibria that survive the first refinement, we select an equilibrium that maximizes expected average utility of all players. In the examples below, this equilibrium selection essentially boils down to the fact that compliers choose a permitted action if they are indifferent between a permitted and a forbidden action and that selfish players act ‘weakly nice’ in the sense that if they are indifferent between two actions, they choose that action that yields higher payoff for the other players.

3.1 A public goods game

Consider a public goods game with $n \geq 2$ players. Each player i chooses simultaneously a monetary contribution $c_i \in \mathbb{R}_0^+$ to a public good. We assume that underlying payoffs are given for every player i by

$$u_i = \gamma \sum_{j=1}^n c_j - c_i \text{ with } \frac{1}{n} < \gamma < 1. \quad (2)$$

The parameter γ denotes the marginal per capita return of contributions. Since $\frac{1}{n} < \gamma < 1$, total payoffs increase in contributions, but it is strictly dominant for a selfish player to contribute nothing.

If a norm prescribes a compliant player i to contribute \tilde{c}_i , she will follow the norm if and only if the cost of contribution $\tilde{c}_i(1 - \gamma)$ does not exceed the disutility g from violating the norm. This implies that a complier will contribute at most

$\bar{c}_g := \frac{g}{1-\gamma}$. It is straightforward that in the unique rule-utilitarian norm equilibrium all compliers always contribute \bar{c}_g .

Complier optimal norms solve the following trade-off: on the one hand, compliers should contribute, because other compliers can benefit from the positive externalities; on the other hand, contributions imply an expected transfer of resources from compliers to selfish players (who themselves never contribute). For every extra unit that a compliant player i contributes, her own payoff is reduced by $1 - \gamma$, but the payoff of each of the other $n - 1$ players increases by γ . A compliant player i knows for sure that she herself is a complier, but each other player is a complier only with probability κ . Thus, if a compliant player increases her contributions, compliers' expected utility is strictly increased if and only if $\kappa(n - 1)\gamma > 1 - \gamma$, which is equivalent to $\kappa > \frac{1-\gamma}{n\gamma-\gamma}$. Hence, whenever this condition is fulfilled, it is complier optimal to contribute \bar{c}_g . In the case $\kappa < \frac{1-\gamma}{n\gamma-\gamma}$, compliers' expected utility decreases in contributions and it is therefore complier optimal to contribute nothing. Note that under rule-utilitarian norms, compliers' expected utility is negative whenever $\kappa < \frac{1-\gamma}{n\gamma-\gamma}$. Proposition 3 summarizes the results:

Proposition 3 *In the unique rule-utilitarian norm equilibrium selfish players contribute nothing and compliers contribute $\bar{c}_g := \frac{g}{1-\gamma}$. If $\kappa > \frac{1-\gamma}{n\gamma-\gamma}$, this is also the unique complier optimal norm equilibrium; if $\kappa < \frac{1-\gamma}{n\gamma-\gamma}$, there are no contributions under any complier optimal norm, while compliers' expected utility under rule-utilitarian norms is negative.¹⁰*

The level of contributions increases in the marginal per capita return γ , which is in line with the stylized facts from most public good experiments (see e.g. Isaac and Walker, 1988, or the survey of Ledyard, 1995). Furthermore, under complier optimal norms, compliers are more likely to contribute if, ceteris paribus, the number of subjects n increases. This is also consistent with the results of Isaac and Walker (1988), who find that average contributions weakly increase in group size if marginal per capita return γ is kept constant.

¹⁰The results extend straightforwardly to a public goods game where maximal contributions are technologically bounded by a level $\bar{c} > 0$. One then simply has to redefine $\bar{c}_g := \min\{\bar{c}, \frac{g}{1-\gamma}\}$.

3.2 A public goods game with a costly punishment technology

The experimental literature has established that contributions in public goods games can substantially increase if players have the opportunity to punish non-contributors, even if punishment is costly (see e.g. Fehr and Gächter, 2000b). We show that these results are predicted by complier optimal and rule-utilitarian norms.

Consider the two player version of the public goods game from Section 3.1 with a second stage where each player i has the option to spend ϕp_i monetary units to reduce the other player's payoff by p_i units. The level of punishment $p_i \geq 0$ can be freely chosen, and $\phi > 0$ is a parameter that describes the constant marginal cost of punishment. Final payoffs after contributions (c_1, c_2) and punishment levels (p_1, p_2) are given by

$$u_i = \gamma(c_1 + c_2) - c_i - \phi p_i - p_j \text{ for } i, j \in \{1, 2\} \text{ and } j \neq i. \quad (3)$$

Clearly, sequentially rational selfish players will never punish. Compliers are at most willing to pay g for punishment or, equivalently, to punish at most with a level of $\bar{p}_g := \frac{g}{\phi}$. To establish an upper bound on possible contributions by selfish players, consider norms that prescribe maximal punishment of \bar{p}_g towards players who contribute less than a threshold \tilde{c} and no punishment otherwise. If all compliers comply with such a norm, a selfish player contributes \tilde{c} if and only if $\kappa \bar{p}_g \geq (1 - \gamma)\tilde{c}$; otherwise he contributes 0. Thus, selfish players can never be induced to contribute more than $\bar{c}_s := \frac{\kappa \bar{p}_g}{1 - \gamma}$. No norm can induce compliers to extend contributions by more than $\bar{c}_g = \frac{g}{1 - \gamma}$ units above the selfish rational level, i.e. compliers will never contribute more than $\bar{c}_s + \bar{c}_g$.

In every rule-utilitarian norm equilibrium compliers will indeed contribute $\bar{c}_s + \bar{c}_g$ and selfish players contribute \bar{c}_s and there is no punishment on the equilibrium path. It is easy to check that this outcome can be sustained by multiple regular norms: all prescribe to punish a contribution $c < \bar{c}_s$ with some level $p(c) \in [(\bar{c}_s - c)\frac{1 - \gamma}{\kappa}, \bar{p}_g]$ and not to punish contributions of \bar{c}_s and $\bar{c}_s + \bar{c}_g$.

By similar reasoning than in the example without punishment technology, one finds that these norms are also complier optimal if $\kappa \geq \frac{1 - \gamma}{\gamma}$; otherwise it is complier optimal that compliers contribute the same amount \bar{c}_s as selfish players. Proposition 4 summarizes these results:

Proposition 4 *The punishment technology increases the contributions of both*

selfish and compliant types by $\bar{c}_s = \frac{\kappa}{1-\gamma} \frac{g}{\phi}$ in every complier optimal (rule-utilitarian) norm equilibrium.

The comparative statics of \bar{c}_s are quite intuitive: \bar{c}_s increases in the marginal per capita return γ , compliers share κ and level of guilt g and decrease in the marginal cost of punishment ϕ .

That punishment is never conducted on the equilibrium path is due to the strong assumptions that all parameters and the norm are common knowledge and that all players act perfectly rational. It is instructive to slightly relax these assumptions by assuming that with a small probability ε a player irrationally trembles and chooses randomly a contribution level from some distribution with full support on all contribution levels.¹¹ If ε is sufficiently small, rule-utilitarian and complier optimal norms will still induce rational selfish players to contribute \bar{c}_s by punishing lower contributions. Since punishments are now conducted with positive probability, it is never optimal to punish stronger than is necessary for deterrence. This implies that the unique regular complier optimal (rule utilitarian) norm prescribes not to punish any contribution $c \geq \bar{c}_s$ and to punish a contribution of $c < \bar{c}_s$ with $p(c) = (\bar{c}_s - c) \frac{1-\gamma}{\kappa}$. This means punishment is proportional to the negative gap between the actual contribution and the contribution of rational selfish players \bar{c}_s . This prediction is in line with the experimental results by Fehr and Gächter (2000b), who show that punishment is stronger the more a contribution falls behind average contributions.

3.3 A sequential prisoners' dilemma

We now illustrate a sequential prisoners' dilemma game, which can also be interpreted as a simple trust game. There are two players who sequentially decide whether to contribute exactly one unit or nothing to a public good, with payoffs given as in the public good game in Section 3.1. Since there are only 8 different pure strategy-profiles, the following intuitive results are straightforward to verify.

Obviously, a selfish player 2 will never contribute. If $g < 1-\gamma$, also no compliant player 2 can be induced to contribute and by backward induction one finds that also no compliant or selfish player 1 will ever contribute.

Let us consider the case $g \geq 1-\gamma$. A selfish player 1 may strictly prefer to contribute if the norm prescribes player 2 to conditionally cooperate, i.e. to

¹¹ Furthermore, assume that for some $\delta > 0$ the probability density of the trembles has a finite upper bound on the interval $(\bar{c}_s - \delta, \bar{c}_s)$.

contribute if and only if player 1 has contributed. Expected payoff for a selfish player 1 from contributing is then given by $\kappa\gamma + \gamma - 1$. Since his payoff from not contributing is zero, he strictly prefers to contribute whenever $\kappa\gamma + \gamma - 1 > 0$, which is equivalent to $\kappa > \frac{1-\gamma}{\gamma}$. If this condition holds, the unique regular complier optimal and rule-utilitarian norm are identical and prescribe conditional cooperation for player 2 and contribution for player 1.

If the compliers' share is below the threshold $\frac{1-\gamma}{\gamma}$, a selfish player 1 cannot be influenced by any norm and will never contribute. Complier optimal and rule-utilitarian norms differ in this case.

The unique rule-utilitarian norm then prescribes contribution for player 1 and *unconditional* contribution by player 2. This means that, in order to increase total welfare, a player 2 shall contribute, even if he observes that player 1 has not contributed.

In contrast, conditional cooperation by player 2 and contribution by player 1 remains complier optimal norm in the range $\frac{1-\gamma}{3\gamma-1} < \kappa \leq \frac{1-\gamma}{\gamma}$. Given that a compliant player 1 contributes and therefore perfectly separates from a selfish player 1, it is clear that conditional cooperation by player 2 must be complier optimal. Note, however, that a compliant player 1 has a negative expected payoff from contributing in this range. The reason that contributing is still complier optimal is that it creates a positive externality for a compliant player 2. A general interpretation of this result is that it can be complier optimal to trust other people to a larger extent than is individually rational. For $\kappa < \frac{1-\gamma}{3\gamma-1}$, contribution becomes too costly for a compliant player 1 and in no complier optimal norm equilibrium players contribute on the equilibrium path.

The prediction that, as second movers, compliers act reciprocal while, as first movers, they trust more than selfish players is in line with recent experimental results by Altmann et. al. (2008), who find that reciprocal subjects trust significantly more than selfish subjects.¹²

That player 2 either conditionally cooperates or does never cooperate is in line with experimental studies of sequential Prisoners' Dilemma games (see Bolle and Ockenfels, 1990, and Clark and Sefton, 2001, or Berg et. al., 1995, for trust games), where unconditional cooperation by player 2 is almost never observed. One may

¹²The result that complier optimal norms may even prescribe a compliant player 1 to accept negative expected monetary payoff relies on the simplifying assumption that underlying utility is linear in money. This result can vanish if underlying preferences account for risk-, loss- or inequity-aversion (see Section 3.4). The result that a compliant player 1 is more willing to trust than a selfish player 1 will still prevail, however.

therefore consider outcomes under complier optimal norms to be more in line with experimental observations than outcomes under rule-utilitarian norms, which can prescribe unconditional cooperation.¹³

3.4 Dictator games

In a dictator game player 1 splits a given amount of money between him and player 2. First consider the case that every player's underlying utility function is identical, and a linear function of own monetary payoffs only. A dictator game is then a zero-sum game and its unique complier optimal norm equilibrium is that player 1 acts like a selfish player and keeps all money for himself, while every norm from which compliers do not deviate is a rule-utilitarian norm.¹⁴

The dictator game ceases to be a zero-sum game, however, if players are risk or loss averse or feel envious. More equitable splits then yield a higher level of average utility. While in utilitarian welfare economics traditionally risk aversion is considered the main argument for more equitable distributions, loss aversion with respect to a local reference level seems to be more important in situations of small and medium stakes. For example, Rabin (2000) shows that people act much too cautious in lotteries with small or medium stakes for behavior to be consistent with sensible risk preferences over large stakes. There is much evidence that in social settings, the payoffs of the other subject(s) in one's matching group play an important role for reference levels and experienced loss (see, e.g. Fliessbach et. al., 2007, for recent neuro-economic evidence).

Even though reference levels are likely influenced by alternative factors, we suspect that the case of pure envy, in which the reference level is solely determined by the payoffs of the other players, is not a bad approximation in interactive situations. We therefore illustrate our model for the following specification of

¹³In many circumstances, this difference between complier optimal and rule-utilitarian norms vanishes, however. For example, rule-utilitarian norms also prescribe conditional cooperation if contributions can be selected from an interval and be arbitrarily small, since then a selfish first mover can always be induced to contribute at least a bit. Furthermore, risk- loss- or inequity aversion can cause underlying utility to be sufficiently non-linear in money and therefore can cause conditional cooperation. Finally, as we have already shown, conditional cooperations arises if the compliers' share κ is large enough.

¹⁴Since the sum of underlying payoffs is constant it is clear that every norm that does not cause disutility from feeling of guilt is a rule-utilitarian norm. Since compliers can at most get the same expected payoff than selfish players (Proposition 1), every norm equilibrium in which compliers get the same expected payoff as selfish players is complier optimal. This implies that it is always complier optimal to act selfishly in zero-sum games.

underlying utility:

$$u_i(\pi) = \pi_i - \frac{\alpha}{n-1} \sum_{j \neq i} \max\{\pi_j - \pi_i, 0\} \text{ with } 0 < \alpha. \quad (4)$$

The vector $\pi = (\pi_1, \dots, \pi_n)$ denotes players' monetary payoffs and α measures the degree of envy, which is assumed to be equal for all players, irrespective of whether a player is selfish or a complier. These underlying preferences resemble the inequity aversion preferences introduced by Fehr and Schmidt (1999), with the difference that we do not assume that players feel disutility when being materially better off. Neither do we assume that players experience positive status utility from being materially better off than other players.

The total monetary amount that can be distributed in the dictator game shall be normalized to 1, i.e. g measures which fraction of the pie a compliant player could at most be induced to give (at least if $g \leq \frac{1}{2}$; otherwise envy plays a role). Let $x \in [0, 1]$ denote the amount given to player 2.

It is strictly dominant for a selfish dictator to give nothing. For all levels of $\alpha > 0$, the unique regular rule-utilitarian norm prescribes to give $x = \min\{\frac{1}{2}, g\}$, since this reduces overall envy as much as possible. Complier optimal norms balance a similar trade-off as in the public goods game: giving reduces compliers' average utility if the recipient is a selfish type, but it increases compliers' average utility if the recipient is a complier whose envy is reduced. It is straightforward to verify that the unique regular complier optimal norm prescribes to give $x = \min\{\frac{1}{2}, g\}$ if $\alpha > \frac{1-\kappa}{2\kappa}$ and to give nothing if $\alpha < \frac{1-\kappa}{2\kappa}$.

Thus, whether it is complier optimal to give or not to give is a question of how bad the other player feels if he gets nothing. In an empirical study, Krupka and Weber (2008) present different variants of the dictator game and ask people to judge the ethical appropriateness of different behavior. While in almost all variants it is judged very inappropriate if the dictator keeps all money, a strong exception is a variant where player 1 has the option to keep all money (10\$) without letting player 2 know that he participated in a dictator game and that player 1 got money. Keeping silently the money is on average considered more appropriate than letting player 2 know about the game and giving 30% of the cake (or to give 30% in the standard dictator game). One interpretation of this result is that people assume that player 2 feels no or only little envy or loss if he does not know that he could have gotten money; therefore it is not considered inappropriate that player 1 keeps all money. Such reasoning would be roughly in line with our model of complier optimal norms.

Andreoni and Miller (2002) investigate dictator experiments where transfers are multiplied by an efficiency factor f , i.e. monetary payoffs are given by $(1-x, fx)$. They show that for higher efficiency factors, dictators are more willing to make an equal split or even allow the responder to have a higher payoff. As can be easily checked, the model with complier optimal norms can match these stylized facts.¹⁵

3.5 Ultimatum games

An ultimatum game extends a dictator game by giving player 2 (called responder) the opportunity to reject the offer x by player 1 (called proposer), in which case both get paid zero. We consider the same preferences as in the dictator game in Section 3.4. It is straightforward to show that an envious selfish responder accepts only offers that are weakly higher than $x^* := \frac{\alpha}{1+2\alpha}$. Furthermore, we find:

Proposition 5 *In every complier optimal and rule-utilitarian norm equilibrium selfish proposers offer $x^s = \min\{\frac{1}{2}, \kappa + (1-\kappa)x^*, x^* + \frac{g}{1+2\alpha}\}$ and all lower offers are rejected by compliers. Under rule-utilitarian norms compliers offer $\min\{\frac{1}{2}, x^s + g\}$. Under complier optimal norms, compliers offer $\min\{\frac{1}{2}, x^s + g\}$ if $\alpha > \frac{1-\kappa}{2\kappa}$ and x^s if $\alpha < \frac{1-\kappa}{2\kappa}$.*

Proof. To maximize total welfare, rule-utilitarian norms must induce those offers from selfish and compliant types that are as near as possible to an equal split. Across all norm equilibria, let x^s denote the offer of selfish proposer that is nearest to an equal split. It is obvious that $x^* \leq x^s \leq \frac{1}{2}$ and that x^s can be induced by a norm that prescribes compliers to accept an offer x if and only if $x \geq x^s$. Compliers are willing to reject all offers below x^s if and only if $0 \geq x^s - \alpha(1-2x^s) - g \Leftrightarrow x^s \leq \frac{\alpha+g}{1+2\alpha} = x^* + \frac{g}{1+2\alpha}$. A selfish proposer is willing to offer x^s instead of x^* if and only if $1-x^s \geq (1-\kappa)(1-x^*) \Leftrightarrow x^s \leq \kappa + (1-\kappa)x^*$. These results imply $x^s = \min\{\frac{1}{2}, \kappa + (1-\kappa)x^*, x^* + \frac{g}{1+2\alpha}\}$. Compliers can at most be induced to give g more than selfish players, i.e. $\min\{\frac{1}{2}, x^s + g\}$ is the nearest offer to an equal split that compliers can be induced to make.

To find complier optimal norms, we first use Lemma 1 in Appendix D, which implies that in no complier optimal norm equilibrium, compliers give less than selfish players. It is then straightforward that in every complier optimal norm equilibrium, selfish players must contribute x^s . That compliers give more would

¹⁵For an example assume for simplicity that g is large enough to make compliers follow every norm. Complier optimal norms then prescribe to give nothing if $\kappa < \frac{1}{f+(1+f)\alpha}$, equalize final payoffs if $\frac{1}{f+(1+f)\alpha} < \kappa < \frac{1+(1+f)\alpha}{f}$ and give everything if $\kappa > \frac{1+(1+f)\alpha}{f}$.

only be complier optimal if it were optimal to give money in the dictator game, which is strictly the case if and only if $\alpha > \frac{1-\kappa}{2\kappa}$. ■

In contrast to the dictator game, in the ultimatum game an arbitrary small amount of envy suffices to find substantial offers in complier optimal norm equilibria, since

$$\lim_{\alpha \rightarrow 0} x^s = \min\left\{\frac{1}{2}, \kappa, g\right\}.^{16} \quad (5)$$

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 rejection rates for offers below 0.2 are very high. In our model, we find that all offers below 0.2 are rejected if $x^* = 0.2$, which corresponds to $\alpha = \frac{1}{3}$. For this level of α we find that already for $\kappa \geq \frac{1}{4}$ and $g \geq \frac{1}{3}$ all equilibrium offers are weakly above 0.4.

Ultimatum game with non-intentional offers

Blount (1995) performed an experimental treatment where the offer was not intentionally 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. Like other models of social preferences that account for the role of intentions (see Section 5), our model can explain that rejection goes down if first movers act non-intentional. It is straightforward to check that for $\alpha < \min\{\kappa, g\}$, complier optimality prescribes that a compliant responder accepts all offers; for rule-utilitarian norms the condition is $\alpha < \min\{1, g\}$. If envy is larger, very unequal offers may be rejected under both norms, however.¹⁷ A compliant responder still feels envy, but weighs the monetary payoff of a compliant proposer higher than her envy. The difference to the intentional treatment arises because for random offers, a norm has no strategic impact on proposers' behavior. An envious selfish responder, however, still rejects

¹⁶The same result holds if players are slightly loss averse with a fixed reference level of $\frac{1}{2}$.

¹⁷For $\alpha > g$, a compliant responder's disutility from very unequal offers is so large that she would reject offers of $x < \frac{\alpha-g}{2\alpha+1}$ even if the norm prescribes to accept every offer. For $\alpha > \kappa$ (under complier optimality) or $\alpha > 1$ (under rule-utilitarianism) disutility from envy is so large that even total welfare can be improved if very unequal offers are rejected. However, the resulting rejection thresholds are always strictly below the rejection threshold of compliant and selfish types in the intentional ultimatum game. This means our model predicts a difference between intentional and non-intentional ultimatum games for the whole range of feasible parameters.

every offer below x^* , since it does not matter for him how the offers were selected. Our model is also in line with the outcomes from other experiments that explore the role of intentions, like the best-shot games (see Harrison and Hirshleifer 1989, Prasnikar and Roth 1992) or the mini-ultimatum games analyzed in Falk et. al. (2003).

4 Extended model with multiple types

There are different ways to extend the basic two-type model to a model with multiple types that differ in their moral motivation. This section proposes an extension that achieves the following goals: First, it allows for substantial behavioral variety. Second, a modified compliance principle holds. Third, moral motivation is collectively rational in the sense that under quite weak conditions, welfare always (weakly) increases if moral motivation in a population increases.

A type in the extended model shall be described by his moral motivation $g \in G$ with the type space G being some compact subset of \mathbb{R}_0^+ . Types are private information and independently drawn for each player from a commonly known distribution F with support on G .

We change the definition of a norm in the extended model. The reason is the following: If one simply used the same definition of a norm as in the basic model and assumed that utility of a player with type g is reduced by g if he breaks a norm, the enlarged type space would not translate into much behavioral heterogeneity. For an example, consider the public goods game from Section 3.1 and let \tilde{c} denote the lowest contribution that would not violate a given norm. Then every player with a type $g \geq (1 - \gamma)\tilde{c}$ would contribute \tilde{c} , while every other player would contribute nothing. Hence, at most two different contribution levels would be predicted for every given norm and distribution of types.

In the extended model, a norm shall be a correspondence that selects for each information set h of the underlying game, and for each type $g \in G$ a non-empty subset of permitted actions $r(h, g) \subset A|h$ with the requirement that $r(h, g) \subset r(h, g')$ whenever $g > g'$. The requirement means that one cannot forbid an action for a lower type that is allowed for some higher type. Similar to the basic model, the underlying payoff of a player i of type g is reduced by g whenever he has deviated from a norm, i.e. having played at least one action that is not permitted for his type.

The definitions of the extended game $\hat{\Gamma}$, norm equilibria and equilibrium se-

lection functions follow in the obvious way along the lines of the basic model. Strategy profiles of $\widehat{\Gamma}$ and norm equilibria are denoted by $(\sigma^g)_{g \in G} \in \Sigma^G$ and specify for every type g and player i a strategy of the underlying game.

There are two slightly different interpretations of norms and types in this model. First, one could interpret a norm as the general opinion in a society about how appropriate different actions are. The measure of appropriateness could be defined as the highest type for which the action is still permitted: $\arg \max_g \{a | h \in r(h, g)\}$. In other words, one could interpret the actions that are permitted for a perfect complier with $g = \infty$ as an ‘ideal norm’, and consider norms for other types as measures of appropriateness of some non-ideal action.¹⁸ Higher types are more receptive to the social opinion and feel guilty more quickly and more severe when playing an action that is not considered appropriate enough.

Rule-consequentialistic norms, also have an alternative interpretation: The type g measures the degree of a player’s general moral motivation, in the sense that it describes how much of her own payoff she is willing to give up for increasing welfare W . A norm can then be understood as a device to coordinate the individual’s efforts towards higher welfare in a collectively rational way.

Since the two interpretations do not contradict each other, norms and types can well be considered as capturing elements of both interpretations. Main results from the basic model carry over to the extended model.

Proposition 6 *In every norm equilibrium, a players’ expected utility and expected underlying payoffs are weakly decreasing in her type.*

The proof is similar to the proof of Proposition 1 and therefore omitted. The result follows from the fact that players with lower moral motivation always have the option to mimic types with higher moral motivation.

We will now establish a modified compliance principle for the extended model. Assume that Γ is in pure strategies. We say a norm r *permits* a norm equilibrium $(\sigma^g)_{g \in G}$ if at every information set h and for every type $g > 0$, the action that this type selects under σ^g is permitted for this type.

In the extended model, we call an equilibrium selection function ψ *regular* if the following condition holds: if a norm equilibrium $(\sigma^g)_{g \in G}$ is selected under some

¹⁸For example, if one would ask somebody what would be the ‘norm’ in a dictator game, one probably would get an ideal norm as answer, like ‘make an equal split’. However, Krupka and Weber (2008) show that when people are asked how appropriate a certain split in the dictator game is considered, that the appropriateness ratings are relatively smoothly decreasing in the amount the dictator gives. Norms in our model can be interpreted as such ratings of appropriateness.

norm then $(\sigma^g)_{g \in G}$ is also selected under every norm that permits it. As in the basic model, one can easily show that a regular ψ always exists.

Proposition 7 (Compliance Principle for the extended model) *Assume the underlying game Γ is in pure strategies (possibly with randomization devices) and ψ is regular. Then every rule-consequentialistic norm equilibrium is permitted by a rule-consequentialistic norm.*

The proof is similar to the proof in the basic model and therefore omitted. The compliance principle ensures that one can neglect w.l.o.g. rule-consequentialistic norms from which some type with $g > 0$ would deviate at some information set.

Let $U_i^g(r, F, \Gamma, \psi)$ denote the expected utility of a player i who is of type g , and $U^g(\cdot) = \frac{1}{n} \sum U_i^g(\cdot)$ the average expected utility of a type g . Rule-utilitarian norms are simply defined as those norms that maximize $\int_{g \in G} U^g(r, \cdot) dF(g)$.

We use the term *complier* for the highest type in G , which shall be denoted by \bar{g} . The label is sensible in so far that compliers are those types that can be easiest induced to comply with some non-selfish strategy. A complier optimal norm shall maximize expected utility $U^{\bar{g}}(r, \cdot)$ of the type with the highest moral motivation. It follows from Proposition 6 that complier optimal norms also maximize expected utility of the worst-off type.

For an example, consider the public goods game from Section 3.1. It is clear that there is a unique rule-utilitarian norm equilibrium: a player of type g contributes exactly $\bar{c}_g = \frac{g}{1-\gamma}$ units. We characterize complier optimal norms for the case that the type distribution F is atomless. Let $c^o \leq \frac{\bar{g}}{1-\gamma}$ denote the amount that compliers are prescribed to contribute under a complier optimal norm. It is clearly complier optimal that every other type $g < \bar{g}$ contributes the highest amount that he can be induced to contribute, which is given by $\min\{c^o, \frac{g}{1-\gamma}\}$. Let $\kappa^o(c^o) := 1 - F((1-\gamma)c^o)$ denote the fraction of types that contribute the same amount c^o as compliers. A marginal increase in c^o increases a complier's expected utility whenever $(n-1)\gamma\kappa^o(c^o) - (1-\gamma) \geq 0$. Therefore, complier optimality requires

$$\kappa^o(c^o) = \frac{1-\gamma}{(n-1)\gamma}.^{19} \tag{6}$$

Comparing with Section 3.1, we find that compliers' contribution c^o is set such that the share of players $\kappa^o(c^o)$ who act like compliers is equal to the minimal complier's share that is required in the basic model to make positive contributions

¹⁹The assumption $\frac{1}{n} < \gamma < 1$ for a public goods game implies that the right hand side of this inequality lies between 0 and 1.

complier optimal. The result that types with $g > (1 - \gamma)c^o$ give not more than c^o , could be interpreted as a form of exploitation aversion: a player dislikes to give much more than everybody else and therefore prefers a sufficiently high probability that other players give equally much.

We finish this section with another general result for the extended model. It requires a short definition: we say an underlying game starts with a randomization device if it begins with a move of nature that draws for each player i a privately observed signal s_i from a uniform distribution that is irrelevant for underlying payoffs.

Proposition 8 *Let Γ be in pure strategies but start with a randomization device, and let W be a welfare criterion that can be written as a weighted sum of expected utilities with weights that are non-decreasing in types.²⁰ Let F and \tilde{F} be two type distributions and assume rule-consequentialistic norms for W are chosen under both distributions. If F weakly first-order stochastically dominates \tilde{F} , then —at least for some equilibrium selection functions— expected welfare under F is weakly higher than under \tilde{F} .*

Proof. We first proof the result for the case that both F and \tilde{F} are atomless, which implies that for every type $g \in G$ there exists a unique type $\tilde{g}(g) \leq g$ implicitly defined by $\tilde{F}(\tilde{g}(g)) = F(g)$. Under \tilde{F} , let the equilibrium selection function be regular and $(\tilde{\sigma}^g)_{g \in G}$ be a rule-consequentialistic norm equilibrium that is permitted by the rule-consequentialistic norm \tilde{r} . We define the norm r by $r(h, g) = \tilde{r}(h, \tilde{g}(g)) \forall h, g$, i.e. an action is permitted for a type g whenever it is permitted for type $\tilde{g}(g)$ under \tilde{r} . It follows straightforwardly that $(\sigma^g)_{g \in G}$ with $\sigma^g := \tilde{\sigma}^{\tilde{g}(g)} \forall g$ must be a permitted norm equilibrium under r and F which yields the same distribution of outcomes of the underlying game. This implies that welfare under $(\sigma^g)_{g \in G}$ and F is weakly higher as under $(\tilde{\sigma}^g)_{g \in G}$ and \tilde{F} .

If F or \tilde{F} have atoms, we can use the signal s_i to define *quasi-types* which have an atomless distribution. Let $\mu_F(g)$ denote the probability mass that the distribution F puts on a type g . We define $g^*(g, s_i) := F(g) - s_i \mu_F(g)$ as the quasi-type of a type g who receives a signal s_i . Similarly, we define quasi types $\tilde{g}^*(\tilde{g}, s_i)$ under \tilde{F} . Let F^* and \tilde{F}^* denote the resulting atomless distributions of

²⁰This condition on W is sufficient but not necessary for our result. It is, for example, fulfilled for an utilitarian or complier optimal welfare criterion. From the structure of the proof of Proposition 8 it becomes evident that similar results can be derived also for alternative consequentialistic welfare criteria.

quasi-types. The rest of the proof follows from similar steps as in the case where F and \tilde{F} are atomless, using quasi-types instead of types. ■

Proposition 8 basically states that, under rather weak regularity conditions, welfare always increases if a positive fraction of the population becomes stronger morally motivated. This result supports an interpretation of our model as a benchmark case of partial, but collectively rational moral motivation.

5 Other models of social preferences

In this section, we compare our model with some well-known models of social preferences (for a detailed survey on social preferences, see Sobel, 2005).

To start the comparison, consider a model with the following altruistic preferences:²¹ $u_i(\pi) = \pi_i + \alpha_i \sum_{j=1}^n \pi_j$ where $\pi = (\pi_1, \dots, \pi_n)$ is a vector of material payoffs and a player's type α_i measures the altruistic weight that player i puts on the sum of all players' material payoffs. It is instructive to show that this formulation of altruism does not capture collectively rational moral motivation, because total welfare can easily decrease if players become more altruistic.

Consider a public goods game with a first stage, in which players can either contribute one unit of money or nothing. There is a second stage, in which players can reduce the payoff of (only) those players who did not contribute by a fixed amount $p > 1$. In contrast to the example in Section 3.2, punishment shall not be costly, but a player who reduces the payoff of another player shall even receive a small positive amount ε of money. Thus, if all players are completely selfish, non-contributors are always punished at stage 2 and in the unique equilibrium outcome everybody contributes. An altruistic player, however, would not punish in stage 2 if $\alpha_i p > \varepsilon$. Hence, if initially all players are selfish and then all become slightly altruistic—just enough to not punish anymore, but not enough to contribute in the absence of punishment—they will stop contributing and total welfare decreases.

As is common practice when observations from one-shot experiments are studied with models of social preferences, we have assumed in this example that altruistic preferences are narrowly bracketed in the sense that individuals only care about the resulting payoffs from the actual interaction; possible long-run effects on a society as a whole are not considered.

One could imagine a larger model, where altruists care about the utility of

²¹A mathematical formulation of altruistic preferences was already given by Edgeworth (1881).

all individuals in society. Then a decision to punish non-contributors may be beneficial for society as a whole if one believes that the punishment makes the punished individual act more cooperatively in future interactions. Even when it is unlikely that there will be substantial effects on own future payoffs, such that repeated game effects for selfish players are negligible, total effects on society may be larger and therefore relevant for altruists. While such a model of broad-bracketed altruism may be a sensible explanation for observed punishment, it has the drawback that it requires a complete model of the world and therefore would seem rather complex.

Our model with rule-utilitarian norms could be interpreted as a rough approximation to such a model of broad bracketed altruism. The idea is that altruistic players follow rule-utilitarian norms, because they assume that long run expected utility in society is maximized if one always complies to these norms and thereby stabilizes the beliefs in society that a sufficient number of people follow these norms. Similarly, the model with complier optimal norms could be considered as an approximation of a model with broad-bracketed conditional altruism, in which conditional altruists only care about own payoffs and the payoffs of other conditional altruists.

Alternative models of social preferences explain costly punishment by assuming that people, on average, have relatively strong negative emotions. In inequity aversion theories, as Fehr and Schmidt (1999) or Bolton and Ockenfels (2000), costly punishment is explained by sufficiently strong feelings of envy towards players that have a higher payoff than oneself. These models have the advantage to be analytically convenient, but also the drawback that they cannot account for the role of intentions.

In the following models, emotions towards other players can depend on those players' choice sets and thereby account for intentions. In Levine (1998), reciprocal types dislike non-altruistic or spiteful types, and, depending on the choice set, an observed action can signal different information about a player's type.

In the reciprocity models by Dufwenberg and Kirchsteiger (2004) or Falk and Fischbacher (2006) another player's action is classified as 'unkind' if he also could have chosen another action that would lead to a higher payoff on the equilibrium path for oneself (Falk and Fischbacher also include equity concerns). Players get emotional satisfaction from punishing 'unkind' actions and rewarding 'kind' actions. Since in general, it can depend on players' beliefs and belief hierarchies whether an action is considered unkind or not, these models are based on psycho-

logical games (Geanakoplos et. al., 1989) which makes them somewhat complex.²² Charness and Rabin (2002) consider a welfare criterion that is a mixture of the sum and minimum of all players' payoffs. Reciprocal players dislike other players whose strategies show that they do not put sufficient weight on the welfare criterion. Their exact solution concept is implicitly also based on psychological games and quite complex.

Most closely related to our model is the norm-based approach by López-Pérez (2008). Norms are defined in the same way as in our basic model, as a collection of permitted and forbidden actions, and there is also one selfish type and a type that feels disutility when violating the norm. The main difference to our model is that López-Pérez does not consider rule-consequentialistic norms, but norms that would be optimal only in a world in which everybody always follows the norm.²³ Punishment of norm violators is not a direct feature of these norms. Instead, norms become irrelevant once they are openly violated by at least one player. Players are then assumed to feel angry and get emotional satisfaction from punishing the deviators.

We do believe that negative emotions, like envy or anger, definitely play a role in punishment decisions. That is why we illustrated for the case of envy, how negative emotions can be included in our model. On the other hand, if a moral motivation to act collectively rational also plays a substantial role, models that do not account for this effect may overestimate the influence and strength of negative emotions.

Since predictions depend crucially on functional form of emotions, it is difficult to draw inferences on the relative importance of negative emotions compared to rational moral concerns directly from the outcomes of economic experiments. For example, except for the inequity aversion models, all models discussed above have linear formulations, which in public goods experiments with linear punishment technology only predict two forms of punishment: either maximum punishment or no punishment. As is shown in Section 3.2, our model instead yields punishment

²²Rabin (1993) was the first to develop a fairness theory based on psychological games. He considers only 2 player normal form games, however.

²³Norms that would be welfare optimal only if everybody complied with the norm, may easily prescribe choices that reduce welfare if selfish types are present. For example, consider a two player game where player 1 can decide between a sure payoff vector of (99, 99) or playing a prisoners' dilemma game with payoffs (100, 100) under mutual cooperation, but a total payoff of 0 if at least one player defects. While choosing the prisoners' dilemma game would be welfare optimal if everybody would afterwards cooperate, the sure payoff vector leads to a higher welfare if there are more than 1% selfish players.

levels that are increasing in the negative deviation from some target level, which is better in line with empirical evidence. One cannot, however, conclude from these results that anger is not important. They only show that tractable linear formulations of anger do not capture all aspects of observed behavior.

Valuable insights concerning the relative importance of different reasons for costly punishment may be delivered by neuro-economic studies. Knoch et. al. (2006) investigate the question whether subjects that are confronted with a low offer in an ultimatum game have a strong immediate emotional impulse to reject the low offer that can be overwritten by rational concerns to take the money, or whether there is a strong immediate impulse to take the money that can be overwritten by social or moral concerns to reject the low offer. They show that disruption of the dorsolateral prefrontal cortex (DLPFC), a part of the brain that is widely thought as a control instance that can inhibit immediate motivational impulses, leads to substantially higher acceptance rates in ultimatum games. They interpret this result as indication that subjects have an immediate selfish impulse to accept low offers, but that a non-disrupted DLPFC can inhibit such an selfish impulse if it contradicts social or moral norms.

6 Concluding remarks

This paper has analyzed social interactions with types that are partly motivated to act ethically in a collectively rational way. It contributes to both normative and positive analysis.

Normatively, we show how rational limited moral motivation can be concretely modeled for general consequentialistic welfare criteria. For the case that types are private information, we propose complier optimality as a possible normative alternative to rule-utilitarianism, and investigate which behavior these two concepts concretely prescribe in some classical situations of social interaction described in Section 3.

At the same time, the model is quite tractable and its predictions are in line with a large body of stylized facts from economic experiments. Our approach is therefore also valuable for positive analysis that has the goal of predicting likely outcomes under different possible institutions that can govern social interactions.

We focused our analysis on complier optimal and rule-utilitarian norms and illustrated how emotions, like envy, can be incorporated via the underlying preferences. Instead of modifying underlying preferences, one could also account for

emotions indirectly by modifying the welfare criterion. For example, the welfare criterion $W = U^c - \delta(U^s - U^c)^2$ with $\delta > 0$ embodies an explicit form of exploitation aversion in the sense that compliers dislike if selfish types are on average better off. Also equity concerns can be embodied in the welfare criterion. For example, the welfare criterion $W = E(\frac{1}{n} \sum_i u_i + \beta \min_i \{u_i\})$ with $\beta > 0$ leads to modified rule-utilitarian norms that put extra weight on the worst-off player in each outcome; $W = U^c + \beta E(\min_i \{u_i\})$ with $\beta > 0$ leads to a similar modification of complier optimal norms.

An important avenue for future research is to relax the assumption that the distribution of types and the resulting norm are common knowledge. One natural approach is the following. Nature first draws a specific type distribution, and then draws players' types from that distribution. Each player then receives a noisy signal about the actual type distribution and a norm can condition on this signal for determining which actions are permitted or forbidden.

We conjecture that heterogeneity in beliefs will reduce welfare in many situations, even if rule-consequentialistic norms account for the fact that beliefs are heterogeneous. For example, in a public goods game with costly punishment, it could be optimal that players, who believe that the fraction of compliers is relatively high, punish certain low levels of contributions, even though these contributions are optimally chosen by players who believe that the fraction of compliers is low.

Communication can then play an important role: if subjects can exchange messages before playing the public goods game, they may coordinate on a minimal acceptable contribution level and agree to punish only contributions below that level. Modelling the corresponding communication and bargaining processes, given that players' moral motivation and their beliefs over type distributions are both private information, is an interesting challenge for future research.

References:

- Altmann, S., Dohmen, T., Wibral, M., 2008. "Do the Reciprocal Trust Less?," *Economics Letters*, 99(3): 454-457.
- 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.
- Blount, S. 1995. "When Social Outcomes Aren't fair: The Effect of Causal

- Attributions on Preferences,” *Organisational Behavior and Human Decision Processes* 63: 131-144.
- Bolle, F., Ockenfels P. 1990. “Prisoners’ 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.
- Charness, G., Rabin, M., 2002. “Understanding social preferences with simple tests,” *The Quarterly Journal of Economics* 117: 817-869.
- Clark, K., Sefton, M., 2001. “The Sequential Prisoner’s Dilemma: Evidence on Reciprocation,” *Economic Journal* 111: 51-68.
- Chwe, M. S.-Y., 1994. “Farsighted Coalitional Stability,” *Journal of Economic Theory* 63: 299-325.
- 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., 2007. “A Tractable Model of Reciprocity and Fairness,” *Games and Economic Behaviour* 59: 17-45.
- Dufwenberg M., Kirchsteiger G., 2004. “A Theory of Sequential Reciprocity,” *Games and Economic Behaviour* 47: 268-298.
- Edgeworth, F. Y., 1881. “Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences”, London: Kegan Paul.
- Falk, A., Fehr, E., Fischbacher, U., 2003. “On the Nature of Fair Behavior,” *Economic Inquiry*, vol. 41(1): 20-26.
- Falk, A., Fischbacher U., 2006. ”A Theory of Reciprocity,” *Games and Economic Behaviour* 54: 293-315.
- Feddersen, T, Sandroni, A. 2006. ”A Theory of Participation in Elections,” *American Economic Review*, 96(4): 1271-1282.
- 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., 2000a. “Fairness and Retaliation: The Economics of Reciprocity,” *Journal of Economic Perspectives* 14(3): 159-181.
- Fehr, E., Gaechter S., 2000b. “Cooperation and Punishment in Public Goods Experiments,” *American Economic Review*, 90(4): 980-994.
- Fliessbach, B. Weber, P. Trautner, T. Dohmen, U. Sunde, C. E. Elger, A. Falk, 2007. "Social Comparison Affects Reward-Related Brain Activity in the Human Ventral Striatum", *Science* 318 (5854): 1305 - 1308.

- Geanakoplos, J., Pearce, D., Stacchetti, E., 1989. "Psychological Games and Sequential Rationality," *Games and Economic Behavior* 1: 60–79.
- Greenberg, J. and S. Weber, 1986. "Strong Tiebout equilibrium under restricted preference domain," *Journal of Economic Theory* 38: 101-111.
- Greenberg, J., 1990. "The Theory of Social Situations," Cambridge University Press.
- Gürerk, Ö., Irlenbusch, B., Rockenbach, B., 2006. "The Competitive Advantage of Sanctioning Institutions," *Science* 312 (5770): 108-111.
- Güth, W.; 1995. "On Ultimatum Bargaining Experiments - A Personal Review," *Journal of Economic Behavior and Organization* 27: 329-344.
- Harsanyi, 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., 1977. "Rule Utilitarianism and Decision Theory," *Erkenntnis* 11: 25-53.
- 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.
- Hooker, B., 2008 "Rule Consequentialism", *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edited by E. N. Zalta.
- Isaac, R. M., Walker, J. M., 1988., "Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism," *Quarterly Journal of Economics* 103(1): 179-199.
- Knoch, D., Alvaro, P., Meyer, K., Treyer, V., Fehr, E., 2006. "Diminishing Reciprocal Fairness by Disrupting the Right Prefrontal Cortex," *Science* 314 (5800), 829-832.
- Krupka E., Weber, R., 2008. "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?," *IZA Discussion Papers* 3860.
- Ledyard, J. 1995. "Public Goods: A Survey of Experimental Research", in Kagel, J., Roth, A. (Eds.), *Handbook of Experimental Economics*. Princeton University Press.
- Levine, D. K., 1998. "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics* 1(3): 593-622.
- López-Pérez, R., 2008. "Aversion to Norm-Breaking: A Model," *Games and Economic Behavior* 64: 237-267.

- 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.
- Rabin, M., 2000. “Risk Aversion and Expected-utility Theory: A Calibration Theorem,” *Econometrica* 68 (5): 1281-1292.
- Roth, A. E., 1995. “Bargaining Experiments,” in Kagel, J., Roth, A. (Eds.), *Handbook of Experimental Economics*. Princeton University Press.
- Segal, U., Sobel, J., 2007. “Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings,” *Journal of Economic Theory*, 136(1): 197-216.
- Sobel, J., 2005. “Interdependent Preferences and Reciprocity,” *Journal of Economic Literature* 43(2): 396-440.
- 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: Endogenous willingness to comply

Mainly for simplicity, we assumed in the paper that the willingness to comply with a norm g is exogenously given. It seems plausible, however, that individuals are more willing to comply with a norm if (collective) compliance yields a relatively high welfare gain. This appendix illustrates how this idea can be introduced into our framework.

The extension is illustrated for the basic model with two types. We take the underlying game Γ , compliers’ share κ , welfare criterion $W(\cdot)$ and equilibrium selection function ψ as given and assume that for all given levels of $g \geq 0$ the compliance principle holds and that welfare under rule-consequentialistic norms is weakly increasing in g .

For a given norm r , let W_r denote the resulting welfare if all compliers would comply with r and let $\tilde{g}(W_r)$ be a continuous and increasing function that measures compliers’ endogenous willingness to comply as a function of the resulting welfare under compliance. We say a norm r is *sustainable* if in the basic model for $g = \tilde{g}(W_r)$ no compliant player would deviate from r . Let \tilde{R} denote the set of all

sustainable norms. We call $\tilde{r}^* \in \arg \max_{r \in \tilde{R}} W_r$ an *optimal rule-consequentialistic norm* given $\tilde{g}(\cdot)$.

Let $r^*(g)$ denote a regular rule-consequentialistic norm in the basic model for a given level of g . Define $g^* = \max\{g \mid g \leq \tilde{g}(W_{r^*(g)})\}$ as the highest level of moral motivation g such that compliers comply with the norm $r^*(g)$ under the endogenous willingness to comply $\tilde{g}(\cdot)$. It is straightforward to show that $r^*(g^*)$ is an optimal rule-consequentialistic norm given $\tilde{g}(\cdot)$. Hence, one can solve the model with endogenous willingness to comply via two steps. The first step is to characterize rule-consequentialistic norms for the basic model with exogenous g . The second step is to find g^* .

We exemplify the procedure for the public goods game from Section 3.1 and a rule-utilitarian welfare criterion. Let the endogenous willingness to comply be given by

$$\tilde{g}(W) = \sqrt{W}.$$

Recall from Section 3.1 that under a rule-utilitarian norm $r^*(g)$ all compliers contribute $\bar{c}_g = \frac{g}{1-\gamma}$. Straightforward calculation shows that the resulting welfare (i.e. average expected utility) is given by $W_{r^*(g)} := \frac{\kappa(\gamma n - 1)}{1-\gamma}g$. The highest level of g that fulfills $g \leq \tilde{g}(W_{r^*(g)})$ is given by $g^* = \frac{\kappa(\gamma n - 1)}{1-\gamma}$. Resulting contributions are given by

$$\bar{c}_{g^*} = \frac{g^*}{1-\gamma} = \frac{\kappa(\gamma n - 1)}{(1-\gamma)^2}.$$

Contributions are smoothly increasing in κ, n and γ . Compared to the case of an exogenous willingness to comply g , the effect of the marginal per capita return γ on contributions is now magnified.

Appendix B: More on the normative aspects of complier optimality

This appendix addresses some issues on the normative aspects of complier optimality. We first illustrate how complier optimal norms are connected to norms that don't assign welfare weight according to types, but according to players behavior. In the second subsection, we discuss the assumption that types are private knowledge.

B.1 Relations to norms that put equal welfare weight on everybody that acts according to the norm

One aspect of complier optimality that may seem normatively unappealing is that the welfare criterion discriminates players according to their types — not according to their behavior. However, in this appendix we establish formally that complier optimality is equivalent to a welfare principle that puts equal welfare weight on everybody who *acts* according to the norm — independent of his type. Formally, we say that a player i *acts* according to a given norm r , if at every information set his strategy σ_i will always choose an action that is permitted by the norm r .

There are different ways to formalize the idea to maximize welfare of everybody who acts according to the norm. The first idea would be to choose those norms that maximize the *sum* of expected utility of all individuals that act according to the chosen norm. However, this welfare principle has the odd feature that the selected norms depend strongly on how underlying payoffs are normalized. If one transforms underlying payoffs by adding a sufficiently large positive constant to every payoff, this welfare principle will always lead to norms that allow to act selfishly, e.g. to contribute nothing in a public goods game. The reason is as follows: if the added constant is sufficiently large, then the sum of utility is maximized if the number of people who follow the norm is maximized. This means the largest welfare is achieved if also selfish types act according to the norm, which is only the case if the norm permits selfish behavior. In contrast, if one adds a sufficiently large negative constant to underlying payoffs, welfare optimal norms set such high moral standards that nobody (neither selfish nor compliant types) acts according to the norm; this guarantees a non-negative welfare of zero.

To avoid this problem, one could consider norms that maximize the *average* expected utility of all individuals that act according to the chosen norm (and normalize welfare to $-\infty$ if nobody follows the norm). While this principle is robust to normalizations of underlying utility, it still has an unappealing feature: rather odd asymmetries between players can be welfare optimal. To exemplify this point, consider a symmetric two player public goods game where players can contribute either one unit or nothing, with underlying payoffs as in Section 3.1. Assume that g is large enough such that compliers always comply, and let $\kappa = \frac{1}{2}$ and $\gamma = \frac{2}{3}$. It can be easily checked that the average expected utility of those players who follow the norm is largest under an asymmetric norm that prescribes player 1 to contribute one unit but permits player 2 to contribute nothing. The intuitive reason is that under this norm resources are transferred from player 1 to

player 2, which is optimal since at the same time welfare of player 1 weights twice as much as welfare of player 2 under this norm (both compliant and selfish types of player 2 act according to the norm, while only compliant types of player 1 do so).

To rule out such asymmetries, we consider a third way to put equal welfare weight on everybody who acts according to the norm. For a given norm equilibrium, let \tilde{U}_i denote the average utility of those types of player i that act according to the norm and assume that $\tilde{U}_i = -\infty$ if no type of player i acts according to the norm. Let $\tilde{W} = \frac{1}{n} \sum \tilde{U}_i$ be the welfare criterion where for a given player i every type that acts according the norm gets equal welfare weight and —across players— every player gets equal welfare weight. The following result establishes that norms that maximize \tilde{W} are essentially equivalent to complier optimal norms:

Proposition 9 *Let the conditions for the compliance principle be fulfilled and the underlying game have finite payoffs. Then every norm that maximizes \tilde{W} must be complier optimal and every regular complier optimal norm maximizes \tilde{W} .*

Proof. Consider a regular complier optimal norm r^o (recall that regularity of r^o implies that compliers act according to r^o in the selected norm equilibrium). Note that if both a compliant and selfish player i act according to r^o , they have the same expected utility. Hence, under every regular complier optimal norm it holds true that $\tilde{W}(r^o, \cdot) = U^c(r^o, \cdot)$.

Let \tilde{r} be a norm that maximizes \tilde{W} . At least one type of every player must act according to \tilde{r} , since otherwise we would have $\tilde{W} = -\infty$, but a finite payoff can be guaranteed by a norm that simply permits every action. It is straightforward that if a selfish player i acts according to some norm, it must also be the case that a compliant player i will act according to that norm. This implies that all compliers act according to \tilde{r} . This implies that $U^c(\tilde{r}) = \tilde{W}(\tilde{r})$.

Since $\tilde{r} = \arg \max_r \tilde{W}(r)$ and $r^o = \arg \max_r U^c(r^o)$ and the equalities $\tilde{W}(r^o) = U^c(r^o)$ and $U^c(\tilde{r}) = \tilde{W}(\tilde{r})$ hold, we know that $\tilde{W}(r^o) = U^c(r^o) = \tilde{W}(\tilde{r}) = U^c(\tilde{r})$. This means \tilde{r} is complier optimal and r^o maximizes \tilde{W} . ■

B.2 Comments on the case that players' types are not private knowledge

Our result that selfish types are never worse off than compliers, relies, in particular, on the assumption that types are private knowledge. If types are assumed to be

common knowledge,²⁴ it can easily be the case that selfish types are worse off than compliers in rule-consequentialistic norm equilibria.

However, also under rule-utilitarian norms selfish types may be worse off than compliers if types are common knowledge. For example, consider a sequential prisoners' dilemma game in which mutual defection is more efficient than the asymmetric outcome in which one player cooperates and the other defects. It is easy to see that under rule-utilitarian norms a selfish second mover is worse off than a compliant second mover (and that first movers of both types are equally well off).

In other games, selfish types may remain better off than compliers, even under complete information and complier optimal norms. For example, consider a public goods game with a large number of players. In order to benefit the other compliers, complier optimal norms can still prescribe to contribute positive amounts even if some selfish players are known to be present.

Nevertheless, it is true that selfish players are more often worse off than compliers under complier optimal norms than under rule-utilitarian norms. For example, in a dictator game where player 1 can choose between the payoff vectors $(1, -100)$ and $(0, 0)$ and $g > 1$, complier optimal norms would prescribe to pick the first distribution whenever player 2 is selfish, while rule-utilitarian norms always prescribe to pick the more efficient and more equitable second distribution. It may even be the case that selfish types are worse off under complier optimal norms than in a purely selfish society. For example, consider a three person dictator game, where player 1 can choose between the two payoff vectors $(1, 0, 5)$ or $(0, 5, -100)$. A selfish player will always choose the first payoff vector. In contrast, if $g > 1$, a compliant player 1 will pick the second vector in the case that player 2 is compliant and player 3 is selfish.

Given that there is no clear payoff ranking between types, we find it difficult to pick a normatively 'appropriate' welfare criterion for the case that players' types are commonly known. There are always Harsanyi's arguments in favor of an utilitarian welfare criterion, but we do not find it clear that e.g. in a two player public goods game a compliant player should be morally forced to contribute even if he knows for sure that the other player is selfish and does not contribute. In our opinion, a welfare criterion that puts some weight on selfish types and also penalizes payoff differences between different types (and perhaps players) seems to

²⁴If types are common knowledge, a norm would map the vector of drawn types into a strategy-profile of the underlying game. Rule-consequentialistic norms and norm equilibria can be defined similarly as in the case of privately known types.

be a decent choice if one assumes that types are generally common knowledge. We do not think that complier optimality can be well justified on normative grounds, if one assumes that players can directly see other players' types.

However, in most situations we find it reasonable to assume that knowledge of another player's type only becomes available if one has knowledge about that player's behavior in past situations. In that case, one should consider the larger game that includes these past situations. It may then have been a completely rational strategy of a selfish player to reveal his type in an earlier situation, even though this reduces his payoffs in later situations. This means that for the larger game, it would still hold true that rational selfish types have weakly higher expected utility under complier optimal norms than compliers, i.e. the arguments for a normative justification for complier optimality still apply.

On the other hand, acting selfishly in previous situations could also be an irrational decision given the resulting penalties in later interactions. We think the model should then either be augmented for irrational types — who clearly should get some positive welfare weight— or allow for both selfish and compliant types to make mistakes (as in Section 3.2). Then an irrational decision to act selfishly in a previous situation would not reveal that the player is a rational selfish type. A welfare criterion that does not put explicit welfare weight on rational selfish types could then still be defended on normative grounds.

Appendix C: Competition of norms via voting by feet

This appendix provides an alternative justification for complier optimal norms in our two type model. It shows that complier optimal norms arise in a model of competition of norms via voting-by-feet.

The basic idea of the voting-by-feet model is as follows. A society can consist of different communities that can differ in their norm and in their endogenously determined population structure (characterized by the share of compliers in a community's population). Social interaction within a community that has a norm r and a compliers' share κ , shall be described by our basic model of Section 2. This means that members of a community meet randomly in groups of n people and play the game $\widehat{\Gamma}(r, \kappa, g, \Gamma)$. A norm can be interpreted as a collective agreement regulating how inhabitants of the community are supposed to act. Compliant types feel disutility of g if they break this collective agreement, while selfish types

do not feel disutility if they break it.

I assume that all individuals can freely choose in which community they want to live while types are still private information and nobody can be excluded from a community. This means that by choosing their community, compliers freely decide which norm (collective agreement) they accept to follow. Compliers cannot prevent selfish types from joining their community, however.

Voting-by-feet models are much analyzed in a branch of literature emerging from Tiebout (1956), who analyzed local provision of public goods. There is a notorious difficulty, however, in finding an appropriate equilibrium concept. One faces a problem of too many equilibria when using a Nash concept (e.g. Westhoff, 1977) or of non-existence when requiring stability against all immediately beneficial coalitional deviations (see e.g. Greenberg and Weber, 1986). Conley and Konishi (2002) discuss these problems and resolve some of them for a special Tiebout model 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. More generally applicable are the theory of social situations by Greenberg (1990) or the largest consistent set by Chwe (1994), which are based on related ideas.

These concepts, however, are defined only for a finite number of inhabitants. Also, they consider a full information environment where types of all players are known and there are no informational problems in coalition formation. Thus, we cannot directly use these concepts but introduce a modified concept called *migration-proof equilibrium* to extend the basic ideas to our set-up.

C.1 Formal definition of a society

A society is populated by a continuum of compliant and selfish inhabitants of measures $\hat{\mu}_c > 0$ and $\hat{\mu}_s > 0$, respectively. The compliers’ share in the society’s total population is denoted by $\hat{\kappa} := \frac{\hat{\mu}_c}{\hat{\mu}_c + \hat{\mu}_s}$.

A society is described by a finite set of communities $\{C^j\}_{j \in J}$ (indexed by a set J) over which the total population is distributed. Each community is characterized by a tuple $C^j = (\mu_s^j, \mu_c^j, r^j)$, where μ_s^j and μ_c^j are the measures of selfish and compliant inhabitants and r^j is the community’s norm. We assume that every community has a positive measure of inhabitants and let κ^j denote the compliers’ share in community C^j . The underlying game Γ , compliers’ disutility from breaking a norm g , and the equilibrium selection function ψ is identical in each community. Expected utility of a compliant inhabitant of community C^j shall be given by

$U^c(r^j, \kappa^j, \Gamma, g, \psi)$ and of a selfish inhabitant by $U^s(r^j, \kappa^j, \Gamma, g, \psi)$ (see Section 2). Hence, we assume that each position in the game (i.e. being player 1 or player 2 or ...) is equally likely for each inhabitant.²⁵ Subsequently, we denote expected utilities in a community more compactly by $U^c(r^j, \kappa^j)$ and $U^s(r^j, \kappa^j)$. We use the notation $r^o(\kappa)$ for a norm that is complier optimal for compliers' share κ .

C.2 Nash-stable equilibrium

Our first requirement for a stable society is that there are no two 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 3 *A society $\{C^j\}_{j \in J}$ constitutes a **Nash-stable equilibrium** if no inhabitant prefers to move to another existing community.*

Note that every society consisting of a single community is trivially Nash-stable.

C.3 Motivating migration-proof equilibrium

In the concept of migration-proof equilibrium we also to allow for the possibility that inhabitants can collectively migrate to an existing community or collectively create a new community. The concept takes seriously the limits of coalition formation due to private information of types.

We want to motivate the basic ideas with a simple example: Let the underlying game be the public goods game from Section 3.1 and consider a society with initially only one community C^0 , which has a norm that permits to contribute nothing. Consider now the possibility of public announcements like the following: "To all compliers in community C^0 , let us collective migrate and create a new community C^1 , where will all follow the norm to contribute an amount $c \leq \bar{c}_g$ to the public good! You will all be better off there!". Our solution concept will be based on the idea that the addressees of such announcements are willing to migrate if and only if migration is strictly beneficial even in the case where other, non-invited, inhabitants, who prefer to follow to the new community, join the migration.

²⁵Since this assumption is quite strong, the result of our voting-by-feet model has more bite for symmetric underlying games.

In our example, also all selfish players prefer to migrate from C^0 to C^1 , once some compliers follow the announcement and move to C^1 . Since types are private knowledge, we assume that compliers are not able to exclude the selfish inhabitants from their new community C^1 . If all selfish inhabitants follow to C^1 , resulting compliers' share in C^1 will be again $\hat{\kappa}$. Thus, in this example, the joint migration will be beneficial in the long run for the originally addressed compliers only if contributing c in a community with compliers' share $\hat{\kappa}$ gives compliers a positive expected utility.

C.4 Formalizing migration-proof equilibrium

Let a collection $M = (\{m_c^j, m_s^j\}_{j=1}^J, C^d)$, describe a simultaneous migration to a community C^d . An entry $\{m_c^j, m_s^j\}$ means that from community C^j compliers of measure m_c^j and selfish inhabitants of measure m_s^j participate in the migration M . As one building stone of our definition, we need to introduce the notion of uncoordinated migration:

Definition 4 *We say a migration M can occur uncoordinatedly if the destination community C^d already exists and every participant of M prefers to move from her origin community to the destination community (evaluating expected utilities as given before the migration).*

Obviously, uncoordinated migration can occur only in societies that are not Nash-stable. In the spirit of the example above, we also want to allow for announced migrations, where a set of inhabitants is invited to jointly migrate to some new or existing community C^d . If people follow the announcement and the coordinated migration takes place, the society may not be Nash-stable anymore and other inhabitants may follow to C^d via uncoordinated migration. We assume that individuals are skeptical about announcements and only want to participate in the announced migration, if this is strictly beneficial, no matter who follows to C^d via uncoordinated migration. Formally:

Definition 5 *An **announced migration** M to a destination community C^d is **successful** if and only if for every sequence of uncoordinated migrations to C^d that may occur afterwards, the participants of M are strictly better off in C^d than they were initially.*

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

Definition 6 *A society constitutes a **migration-proof equilibrium** if it is Nash-stable and no successful announced migration exists.*

C.5 Conditions

We will now present joint conditions on the game and selfish equilibrium selection function, that are required for our results.

Condition 1 (C1) *A complier optimal norm r^o exists for all $\kappa \in [0, 1]$.*

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

$$U_{\kappa=0} := \max_{r \in R} U^s(r, \kappa = 0, .). \quad (7)$$

Condition 2 (C2) *For every κ compliers can be least as well off as inhabitants of a purely selfish community, i.e. $U^c(r^o(\kappa), \kappa) \geq U_{\kappa=0} \forall \kappa$.*

Condition 3 (C3) *The highest expected utility that compliers can achieve in a community with the societies' share of compliers $\hat{\kappa}$ is as least as high as the expected utility compliers can achieve in a community with a smaller compliers' share, i.e. $U^c(r^o(\kappa), \kappa) \leq U^c(r^o(\hat{\kappa}), \hat{\kappa})$ for all $\kappa < \hat{\kappa}$.*

Note that the conditions stated in Proposition 8 (see Section 4) guarantee that there is always an equilibrium selection function such that Conditions (C2) and (C3) hold. Conditions C1-C3 are fulfilled for all examples in Section 3 and are sufficient for our existence result (Proposition 10 below). For the uniqueness result (Proposition 11 below), we additionally need the following condition:

Condition 4 (C4) *There exists a complier optimal norm $r^o(\hat{\kappa})$ such that for all $\kappa > \hat{\kappa}$ one has $U^c(r^o(\hat{\kappa}), \kappa) \geq U^c(r^o(\hat{\kappa}), \hat{\kappa})$.*

Condition C4 says that at least for some complier optimal norm $r^o(\hat{\kappa})$, compliers' are not worse off when the fraction of compliers is higher than $\hat{\kappa}$. This condition is also fulfilled for all examples in Section 3.

C.6 Results

The following propositions characterize all migration-proof equilibria, and show that norms arise that are complier optimal for the society's share of compliers $\hat{\kappa}$.

Proposition 10 *Assume conditions C1-C3 hold. A society consisting of a single community that applies a complier optimal norm $r^o(\widehat{\kappa})$ constitutes a migration-proof equilibrium.*

Proof. Since it has only a single community, the described society is Nash-stable. It remains to show that there exists no successful announced migration. First note that no announcement that asks only selfish players to migrate can be successful, since they would get expected utility of $U_{\kappa=0}$ in the new community, if no one followed. By C2 and Proposition 1, $U_{\kappa=0} \leq U^c(r^o(\widehat{\kappa}), \widehat{\kappa}) \leq U^s(r^o(\widehat{\kappa}), \widehat{\kappa})$, i.e. selfish players cannot be strictly better off by such migration.

Consider now the case that the announcement asks some compliers to migrate to a community C^d . This is only successful if compliers are strictly better off in C^d . Condition C3 implies that if there are still compliers outside C^d , they then want to follow to C^d by uncoordinated migration. In the (unlikely) case that $U^c(r^d, \kappa)$ is not weakly increasing in κ , compliers' expected utility in C^d may already drop after this uncoordinated migration below the initial level $U^c(r^o(\widehat{\kappa}), \widehat{\kappa})$. Otherwise, the remaining selfish players outside C^d want to follow to C^d , since (again by C2 and Proposition 1) $U_{\kappa=0} \leq U^c(r^o(\widehat{\kappa}), \widehat{\kappa}) < U^c(r^d, \kappa^d) \leq U^s(r^d, \kappa^d)$. Now all inhabitants are in C^d and compliers of the announced migration cannot be strictly better off than initially. ■

Proposition 11 *Assume conditions C1-C4 hold. In every migration-proof equilibrium compliers' expected utility equals $U^c(r^o(\widehat{\kappa}), \widehat{\kappa})$ in all communities.*

Proof. We now show that in every migration-proof equilibrium compliers' expected utility equals $U^c(r^o(\widehat{\kappa}), \widehat{\kappa})$ in all communities. It follows directly from the definition of a Nash-stable society that compliers' expected utility must be equal in all communities. 1. In no Nash-stable society can compliers have utility higher than $U^c(r^o(\widehat{\kappa}), \widehat{\kappa})$. This is because in every society there is at least one populated community C' with a compliers' share $\kappa' \leq \widehat{\kappa}$ and by C3 compliers' expected utility in community C' cannot exceed $U^c(r^o(\widehat{\kappa}), \widehat{\kappa})$.

2. It remains to check that there can exist no migration-proof equilibrium where compliers' expected utility is smaller than $U^c(r^o(\widehat{\kappa}), \widehat{\kappa})$. Denote compliers' expected utility in the original society by U_{orig}^c . Suppose for a proof by contradiction that this society is a migration-proof equilibrium with $U_{orig}^c < U^c(r^o(\widehat{\kappa}), \widehat{\kappa})$. By C4 there exists a complier optimal norm $r^o(\widehat{\kappa})$ with $U^c(r^o(\widehat{\kappa}), \kappa) \geq U^c(r^o(\widehat{\kappa}), \widehat{\kappa}) > U_{orig}^c$ for all $\kappa \geq \widehat{\kappa}$. Consider an announced migration that asks all compliers to migrate to a new community C^o with this norm $r^o(\widehat{\kappa})$. No matter how many selfish

players follow to C^o , the compliers' share in C^o is always bigger than or equal to $\widehat{\kappa}$. By the inequality stated above, compliers' expected utility in C^o is therefore strictly higher than originally, which means the announced migration is strongly successful. The original society was therefore not a migration-proof equilibrium.

■

The basic intuition behind these results is simply that selfish players want to be where compliers are and compliers (who cannot get rid of selfish players) want to be in a place with a complier optimal norm.

C.7 Simultaneous competition of norms and institutions

It is straightforward to extend the voting-by-feet model, by allowing communities also to differ in the game that describes the enforceable rules of social interaction (this game is labeled as an ‘institution’) and in the way how norm equilibria are selected.

Assume social interactions can be structured in different ways that are characterized by a set of possible underlying games Ω . Let $R(\Gamma)$ denote the set of norms for the underlying game $\Gamma \in \Omega$ for which a norm equilibrium exists for every value of $\kappa \in [0, 1]$. For each game Γ there shall be a set of possible equilibrium selection functions $\Psi(\Gamma)$. An equilibrium selection function $\psi \in \Psi(\Gamma)$ selects for every norm $r \in R(\Gamma)$ and every compliers' share a norm equilibrium of the induced game $\widehat{\Gamma}(r, \kappa, g, \Gamma)$.²⁶ A community shall be characterized by its population as well as a triple of game, norm and equilibrium selection function $\lambda = (\Gamma, r, \psi)$ with $r \in R(\Gamma)$ and $\psi \in \Psi(\Gamma)$. We call λ a *norm-institution* and denote by Λ the set of all possible norm-institutions.

Expected utility of compliers and selfish inhabitants within a given community are denoted by $U^c(\kappa, \lambda)$ and $U^s(\kappa, \lambda)$, respectively. A norm-institution that is complier optimal for complier share κ is defined by

$$\lambda^o(\kappa) \in \Lambda^o(\kappa) := \arg \max_{\lambda \in \Lambda} U^c(\kappa, \lambda). \quad (8)$$

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. To derive this model, we simply have to replace every norm that appears in conditions C1-C4 and in Propositions 8 and 9 and their proofs by the corresponding norm-institution. The proofs of Propositions 8 and 9 carry

²⁶If equilibrium selection shall not be part of the voting-by-feet model, one can define $\Psi(\Gamma)$ to contain only a single equilibrium selection function for each game Γ .

over, because they only make use of conditions C1-C4 and Proposition 1, which implies that in no community selfish players can be worse off than compliers.

To sum up, this implies, first that there is always a migration-proof equilibrium with the entire population in a community C^o that has a complier optimal norm-institution $\lambda^o(\widehat{\kappa})$, and second that in all migration-proof equilibria compliers' utility is given by $U^c(\widehat{\kappa}, \lambda^o(\widehat{\kappa}))$. In other words: A complier optimal combination of institution, norm, and selfish equilibrium strategies arises.

Gürerk et. al. (2006) conducted an economic experiment where subjects could vote-by-feet between two different institutions: a public goods game with a costly punishment option and one without (Section 3.2 illustrates the benefits of a punishment option for compliers). In their experiment, virtually the whole population migrates to the community with the institution that allows for costly punishment, which is in line with our theoretical prediction. There over 40% of subjects punish non-contributors and high levels of contribution can be stabilized.

Appendix D: Simple sequential fairness games

In this appendix, we derive a small result that is helpful to characterize complier optimal and rule-utilitarian norms in the following class of games:

Definition 7 *A simple sequential fairness game is a two player game in pure strategies with the following structure: Player 1 moves first and chooses an action $a_1 \in A_1$, which is observed by player 2 who chooses then an action $a_2 \in A_2$. Furthermore, there is a complete ordering \succeq_{\heartsuit} on A_1 :*

$$a_1 \succeq_{\heartsuit} a'_1 :\Leftrightarrow u_2(a_1, a_2) \geq u_2(a'_1, a_2) \quad \forall a_2 \in A_2$$

We say a_1 is strictly nicer than a'_1 if $a_1 \succeq_{\heartsuit} a'_1$ and $a'_1 \not\prec_{\heartsuit} a_1$.

Examples for simple sequential fairness games are trust games, gift-exchange games, sequential prisoners' dilemma games, or ultimatum games.

Lemma 1 *Consider the basic model and assume $\kappa \in (0, 1)$. In no complier optimal (or rule-utilitarian) norm equilibrium of a simple sequential fairness game does a selfish player 1 choose a strictly nicer action than a compliant player 1.*

Proof. Let a_1^c and a_1^s and $a_2^c(a_1)$ and $a_2^s(a_1)$ denote the pure strategies of compliant and selfish players in a complier optimal (rule-utilitarian) norm equilibrium,

respectively. That compliers choose a_1^c instead of a_1^s can only be prescribed by a rule-utilitarian norm if

$$\begin{aligned} & \kappa u_1(a_1^s, a_2^c(a_1^s)) + (1 - \kappa)u_1(a_1^s, a_2^s(a_1^s)) + & \text{(RU)} \\ & \kappa u_2(a_1^s, a_2^c(a_1^s)) + (1 - \kappa)u_2(a_1^s, a_2^s(a_1^s)) \leq \\ & \kappa u_1(a_1^c, a_2^c(a_1^c)) + (1 - \kappa)u_1(a_1^c, a_2^s(a_1^c)) + \\ & \kappa u_2(a_1^c, a_2^c(a_1^c)) + (1 - \kappa)u_2(a_1^c, a_2^s(a_1^c)). \end{aligned}$$

Similarly, complier optimality requires

$$\begin{aligned} & \kappa u_1(a_1^s, a_2^c(a_1^s)) + (1 - \kappa)u_1(a_1^s, a_2^s(a_1^s)) + \kappa u_2(a_1^s, a_2^c(a_1^s)) \leq & \text{(CO)} \\ & \kappa u_1(a_1^c, a_2^c(a_1^c)) + (1 - \kappa)u_1(a_1^c, a_2^s(a_1^c)) + \kappa u_2(a_1^c, a_2^c(a_1^c)). \end{aligned}$$

Rationality of a selfish player 1 implies

$$\begin{aligned} & \kappa u_1(a_1^s, a_2^c(a_1^s)) + (1 - \kappa)u_1(a_1^s, a_2^s(a_1^s)) \geq & \text{(I1)} \\ & \kappa u_1(a_1^c, a_2^c(a_1^c)) + (1 - \kappa)u_1(a_1^c, a_2^s(a_1^c)) \end{aligned}$$

For a proof by contradiction assume a_1^c is not (weakly) nicer than a_1^s . This implies

$$\kappa u_2(a_1^s, a_2^c(a_1^s)) > \kappa u_2(a_1^c, a_2^c(a_1^c)) \quad \text{(I2)}$$

and

$$(1 - \kappa)u_2(a_1^s, a_2^s(a_1^s)) > (1 - \kappa)u_2(a_1^c, a_2^s(a_1^c)). \quad \text{(I3)}$$

Adding (I1) and (I2) one finds that (CO) is violated and adding additionally (I3) also (RU) is violated. ■