# The long-term benefits of following fairness norms: a game-theoretic analysis

Emiliano Lorini, Roland Muehlenbernd

# The Long-Term Benefits of Following Fairness Norms: A Game-Theoretic Analysis

Emiliano Lorini[1(✉)] and Roland Mühlenbernd[2]

[1] IRIT-CNRS, Toulouse University, Toulouse, France
lorini@irit.fr
[2] University of Tübingen, Tübingen, Germany

**Abstract.** In this study we present a game-theoretic model of guilt in relation to sensitivity to norms of fairness. We focus on a specific kind of fairness norm à la Rawls according to which a fair society should be organized so as to admit economic inequalities to the extent that they are beneficial to the less advantaged agents. We analyze the impact of the sensitivity to this fairness norm on the behavior of agents who play a repeated Prisoner's Dilemma and learn via fictitious play. Our results reveal that a great sensitivity to the fairness norm is beneficial in the long term when agents have the time to converge to mutual cooperation.

## 1 Introduction

Prototypical human and artificial societies (*e.g.*, a community, an organization) are populated by agents who have repeated encounters and can decide either to collaborate with the others thereby acting cooperatively, or to exploit the work of the others thereby acting selfishly. In a game-theoretic setting this kind of situations can be represented as an iterated Prisoner's Dilemma (PD) in which agents in the population have repeated one-to-one interactions with others (*i.e.*, at each round two agents in the population meet and play one-shot PD).

The aim of this work is to study how fairness norms tend to emerge in this kind of societies in which agents are assumed (i) to be rational in the sense of being expected utility maximizers, and (ii) to learn from their past experiences. In the paper, we focus on a special kind of fairness norm à la Rawls [20] according to which a fair society should be organized so as to admit economic inequalities to the extent that they are beneficial to the less advantaged agents.

Our analysis is based on the general assumption that agents in the society are heterogenous in the sense of being more or less sensitive to the fairness norm, where an agent's degree of norm sensitivity captures the extent to which the fairness norm has been internalized by the agent. Norm internalization is a concept

> Nobody argues that the art of navigation is not founded on astronomy because sailors cannot wait to calculate the Nautical Almanac. Being rational creatures they go to sea with it already calculated; and all rational creatures go out upon the sea of life with their minds made up on the common questions of right and wrong, as well as on many of the far more difficult questions of wise and foolish.
> *J.S. Mill, Utilitarianism [16, Chap.2]*

DOI:10.1007/978-3-319-25524-8_19

that has been widely discussed in the literature in social sciences and multi-agent systems [2,1,3,10,11]. The idea is that if a given norm is internalized by an agent then there is no need for an external sanction, a reward or punishment to ensure norm compliance. The agent is willing to comply with the norm because, if she does not do this, she will feel (morally) bad. We study the conditions under which an agent's disposition to follow the fairness norm à la Rawls (*i.e.*, the agent's sensitivity to the fairness norm) increases the agent's individual benefit in the long term. In other words, we aim at providing an utilitarian explanation of the internalization of the fairness norm à la Rawls, that is to say, we aim at explaining why rational agents with learning capabilities should become motivated to follow the fairness norm à la Rawls even without external enforcement (*e.g.*, external sanctions, punishment).

The rest of the paper is organized as follows. In Section 2 we present a game-theoretic model of guilt aversion which provides the static foundation of our analysis. The main idea of the model is that agents in a game are motivated both by their personal utilities and by the goal of avoiding guilt feeling. It is assumed that guilt feeling is triggered in case of the violation of an internalized norm. Specifically, the intensity of guilt feeling is proportional to the agent's sensitivity to the norm. In Section 3, we provide a dynamic extension of our model in order to formally specify repeated interactions and learning in a game-theoretic setting. The learning approach we use is the well-known fictitious play [7].[1] Section 4 provides some mathematical results about convergence for fictitious play in the case of iterated PD in which agents are assumed to be more or less sensitive to the fairness norm à la Rawls. Our mathematical analysis of convergence for fictitious play is partial, as it only covers a subset of the set of possible values of norm sensitivity for the agents in the population. Thus, in Section 5, we present some computational results about convergence for fictitious play which complements the analysis of Section 4. Finally, in Section 6, we present some experimental results in the case of iterated PD which highlight the relationship between an agent's degree of sensitivity to the fairness norm à la Rawls and her individual benefit in the long term. Our results reveal that a great sensitivity to this fairness norm is beneficial in the long term when agents have the time to converge to mutual cooperation. As a side note, we would like to remark that a preliminary version of this work by one of the authors has appeared in [9]. One limitation of this previous work is that it was only applied to a specific instance of the Prisoner's Dilemma and not to the entire class. A second limitation is that, differently from the present work, it was not supported by in-depth mathematical analysis of convergence for the fictitious play process. Finally, it did not contain any analysis of the way an agent's sensitivity to the fairness norm influences her benefit in the long term.

[1] We preferred fictitious play over alternative 'learning from the past' models, since it is i) deterministic, thus manageable to be analyzed formally, and ii) well-established in the field.

## 2 Game-Theoretic Model of Guilt Aversion

In this section, we present our game-theoretic model of guilt and of its influence on strategic decision making. We assume that guilt feeling originates from the agent's violation of a certain norm. Specifically, the intensity of an agent's guilt feeling depends on two parameters: (i) how much the agent is responsible for the violation of the norm, and (ii) how much the agent is sensitive to the norm. As emphasized in the introduction, in our model the agent's sensitivity to the norm captures the extent to which the norm is internalized by the agent.

Our model assumes that an agent has two different motivational systems: an endogenous motivational system determined by the agent's desires and an exogenous motivational system determined by the agent's internalized norms. Internalized norms make the agent capable of discerning what from his point of view is *good* (or *right*) from what is *bad* (or *wrong*). If an agent has internalized a certain norm, then she thinks that its realization ought to be promoted because it is *good* in itself. A similar distinction has also been made by philosophers and by social scientists. For instance, Searle [21] has recently proposed a theory of how an agent may want something without desiring it and on the problem of reasons for acting based on moral values and independent from desires. In his theory of morality [13], Harsanyi distinguishes a person's *ethical preferences* from her *personal preferences* and argues that a moral choice is a choice that is based on ethical preferences.

### 2.1 Normative Game and Guilt-dependent Utility

Let us first introduce the standard notion of normal-form game.

**Definition 1 (Normal-form game).** *A normal-form game is a tuple* $G = (N, (S_i)_{i\in N}, U)$ *where:*

- $N = \{1, \ldots, n\}$ *is a finite set of agents or players;*
- *for every* $i \in N$, $S_i$ *is agent* $i$*'s finite set of strategies;*
- $U : N \longrightarrow (\prod_{i\in N} S_i \longrightarrow \mathbb{R})$ *is an utility function, with* $U(i)$ *being agent* $i$*'s personal utility function mapping every strategy profile to a real number (*i.e.*, the personal utility of the strategy profile for agent* $i$*).*

For every $i \in N$, elements of $S_i$ are denoted by $s_i, s_i', \ldots$ Let $2^{Agt*} = 2^N \setminus \{\emptyset\}$ be the set of all non-empty sets of agents (*alias* coalitions). For notational convenience we write $-i$ instead of $N \setminus \{i\}$. For every $J \in 2^{Agt*}$, we define the set of strategies for the coalition $J$ to be $S_J = \prod_{i\in J} S_i$. Elements of $S_J$ are denoted by $s_J, s_J', \ldots$ We write $S$ instead of $S_N$ and we denote elements of $S$ by $s, s', \ldots$ Every strategy $s_J$ of coalition $J$ can be seen as a tuple $(s_i)_{i\in J}$ where agent $i$ chooses the individual strategy $s_i \in S_i$. For notational convenience we write $U_i(s)$ instead of $U(i)(s)$. As usual a mixed strategy for agent $i$ is a probability distribution over $S_i$. Agent $i$'s set of mixed strategies is denoted by $\Sigma_i$ and elements of $\Sigma_i$ are denoted by $\sigma_i, \sigma_i', \ldots$ The set of mixed strategy profiles is defined to be $\Sigma = \Sigma_1 \times \ldots \times \Sigma_n$

| | C | D |
|---|---|---|
| C | $R, R$ | $S, T$ |
| D | $T, S$ | $P, P$ |

**Fig. 1.** Prisoner's dilemma (with player 1 being the row player and player 2 being the column player).

and its elements are denoted by $\sigma, \sigma', \ldots$ The utility function $U_i$ reflects agent $i$'s endogenous motivational system, *i.e.*, agent $i$'s desires.

A well-known example of normal-form game is the Prisoner's Dilemma (PD) in which two agents face a social dilemma. The PD is represented in Figure 2.1.

Each agent in the game can decide either to cooperate (action $C$) or to defect (action $D$) and has an incentive to defect. Indeed, it is assumed that, if an agent defects, she gets a reward that is higher than the reward obtained in the case of cooperation, no matter what the other agent decides to do. In other words, cooperation is strongly dominated by defection. The social dilemma lies in the fact that mutual defection, the only Nash equilibrium of the game, ensures a payoff for each agent that is lower than the payoff obtained in the case of mutual cooperation. The Prisoner's Dilemma can be compactly represented as follows.

**Definition 2 (Prisoner's Dilemma).** *A Prisoner's Dilemma (PD) is a normal-form game* $G = (N, (S_i)_{i \in N}, U)$ *such that:*

- $N = \{1, 2\}$;
- *for all* $i \in N$, $S_i = \{C, D\}$;
- $U_1(C, C) = R$, $U_1(D, D) = P$, $U_1(C, D) = S$ *and* $U_1(D, C) = T$;
- $U_2(C, C) = R$, $U_2(D, D) = P$, $U_2(C, D) = T$ *and* $U_2(D, C) = S$;

*and which satisfies the following two conditions:*

***(C1)*** $T > R > P > S$,
***(C2)*** $S = 0$.

Condition (**C1**) is the typical one in the definition of the Prisoner's Dilemma. Condition (**C2**) is an extra *normality* constraint which is not necessarily assumed in the definition of PD. It is assumed here to simplify the analysis of the evolution of fairness norms.

The following definition extends the definition of normal-form game with a *normative* component. Specifically, we assume that every outcome in a game is also evaluated with respect to its ideality degree, *i.e.*, how much an outcome in the game conforms to a certain norm. Moreover, as pointed above, we assume that an agent in the game can be more or less sensitive to the norm, depending on how much the norm is internalized by her.

**Definition 3 (Normative game).** *A normative game is a tuple* $NG = (N, (S_i)_{i \in N}, U, I, \kappa)$ *where:*

- $(N, (S_i)_{i \in N}, U)$ *is a normal-form game;*

– $I : \prod_{i \in N} S_i \longrightarrow \mathbb{R}$ *is a function mapping every strategy profile in* $S$ *to a real number measuring the degree of ideality of the strategy profile;*
– $\kappa : N \longrightarrow \mathbb{R}_{\geq 0}$ *is a function mapping every agent in* $N$ *to a non-negative real number measuring the agent's sensitivity to the norm.*

For notational convenience we write $\kappa_i$ instead of $\kappa(i)$ to denote agent $i$'s sensitivity to the norm.

Following current psychological theories of guilt [12], we conceive guilt as the emotion which arises from an agent's self-attribution of responsibility for the violation of an internalized norm (*i.e.*, a norm to which the agent is sensitive). Specifically, intensity of guilt feeling is defined as *the difference between the ideality of the best alternative state that could have been achieved had the agent chosen a different action and the ideality of the current state, — capturing the agent's degree of responsibility for the violation of the norm —, weighted by the agent's sensitivity to the norm.* The general idea of our model is that the intensity of guilt feeling is a monotonically increasing function of the agent's degree of responsibility for norm violation and the agent's sensitivity to the norm.

**Definition 4 (Guilt).** *Let* $NG = (N, (S_i)_{i \in N}, U, I, \kappa)$ *be a normative game. Then, the guilt agent* $i$ *will experience after the strategy profile* $s$ *is played, denoted by* $Guilt(i,s)$*, is defined as follows:*

$$Guilt(i,s) = \kappa_i \times (\max_{s'_i \in S_i} I\big(s'_i, s_{-i}\big) - I(s))$$

The following definition describes how an agent's utility function is transformed depending on the agent's feeling of guilt. In particular, the higher the intensity of guilt agent $i$ will experience after the strategy profile $s$ is played, the lower the (transformed) utility of the strategy profile $s$ for agent $i$. Note indeed that the value $Guilt(i,s)$ is either positive or equal to 0. Guilt-dependent utility reflects both agent $i$'s desires and agent $i$'s moral considerations determined by her sensitivity to the norm.

**Definition 5 (Guilt-dependent utility).** *Let* $NG = (N, (S_i)_{i \in N}, U, I, \kappa)$ *be a normative game. Then, the guilt-dependent utility of the strategy profile* $s$ *for agent* $i$ *is defined as follows:*

$$U_i^*(s) = U_i(s) - Guilt(i,s)$$

It is worth noting that the previous definition of guilt-dependent utility is similar to the definition of regret-dependent utility proposed in regret theory [14]. Specifically, similarly to Loomes & Sugden's regret theory, we assume that the utility of a certain outcome for an agent should be trasformed by incorporating the emotion that the agent will experience if the outcome occurs.

### 2.2 Fairness Norms

In the preceding definition of normative game an agent $i$'s utility function $U_i$ and ideality function $I$ are taken as independent. There are different ways of linking the two notions.

For instance, Harsanyi's theory of morality provides support for an utilitarian interpretation of fairness norms which allows us to reduce an agent $i$'s ideality function $I$ to the utility functions of all agents [13]. Specifically, according to the Harsanyi's view, a fairness norm coincides with the goal of maximizing the collective utility represented by the weighted sum of the individual utilities.

**Definition 6 (Normative game with fairness norm à la Harsanyi).** *A normative game with fairness norm à la Harsanyi is a normative game* $NG = (N, (S_i)_{i \in N}, U, I, \kappa)$ *such that for all* $s \in S$*:*

$$I(s) = \sum_{i \in N} U_i(s)$$

An alternative to Harsanyi's utilitarian view of fairness norms is Rawls' view [20]. In response to Harsanyi, Rawls proposed the *maximin* criterion of making the least happy agent as happy as possible: for all alternatives $s$ and $s'$, if the level of well-being in the worst-off position is strictly higher in $s$ than in $s'$, then $s$ is better than $s'$. According to this well-known criterion of distributive justice, a fair society should be organized so as to admit economic inequalities to the extent that they are beneficial to the less advantaged agents.Following Rawls' interpretation, a fairness norm should coincide with the goal of maximizing the collective utility represented by the individual utility of the less advantaged agent.

**Definition 7 (Normative game with fairness norm à la Rawls).** *A normative game with fairness norm à la Rawls is a normative game* $NG = (N, (S_i)_{i \in N}, U, I, \kappa)$ *such that for all* $s \in S$*:*

$$I(s) = \min_{i \in N} U_i(s)$$

In this paper we focus on fairness norm à la Rawls. In particular, we are interested in studying the relationship between the agents' sensitivities to this kind of norm and their behaviors in a repeated game such as the Prisoner's Dilemma in which the agents learn from their past experiences. To this aim, in the next section, we provide a dynamic extension of our model of guilt aversion.

## 3 Dynamic Extension

In the dynamic version of our model, we assume that every agent in a given normative game has probabilistic expectations about the choices of the other agents. These expectations evolve over time. The following concept of history captures this idea.

**Definition 8 (History).** *Let* $NG = (N, (S_i)_{i \in N}, U, I, \kappa)$ *be a normative game. A* history *(for NG) is a tuple* $H = ((\omega_{i,j})_{i,j \in N}, (c_i)_{i \in N})$ *such that, for all* $i, j \in N$*:*

– $\omega_{i,j} : \mathbb{N} \longrightarrow \Delta(S_j)$ *is a function assigning to every time* $t \in \mathbb{N}$ *a probability distribution on* $S_j$*,*
– $c_i : \mathbb{N} \longrightarrow S_i$ *is a choice function specifying the choice of agent* $i$ *at each time point* $t \in \mathbb{N}$*.*

For every $t \in \mathbb{N}$ and $s_j \in S_j$, $\omega_{i,j}(t)(s_j)$ denotes agent $i$'s subjective probability at time $t$ about the fact that agent $j$ will choose action $s_j \in S_j$. For notational convenience, we write $\omega^t_{i,j}(s_j)$ instead of $\omega_{i,j}(t)(s_j)$. For all $i, j \in N$, $t \in \mathbb{N}$ and $s_{-i} \in S_{-i}$ we moreover define:

$$\omega^t_i(s_{-i}) = \prod_{j \in N \setminus \{i\}} \omega^t_{i,j}(s_j)$$

$\omega^t_i(s_{-i})$ denotes agent $i$'s subjective probability at time $t$ about the fact that the other agents will choose the joint action $s_{-i}$.

The following definition introduces the concept of agent $i$'s expected utility at time $t$. Notice that the concept of utility used in the definition is the one of guilt-dependent utility of Definition 5. Indeed, we assume a rational agent is an agent who maximizes her expected guilt-dependent utility reflecting both the agent's desires and the agent's moral considerations determined by her sensitivity to the norm.

**Definition 9 (Expected utility at time $t$).** *Let* $NG = (N, (S_i)_{i \in N}, U, I, \kappa)$ *be a normative game, let* $H = ((\omega_{i,j})_{i,j \in N}, (c_i)_{i \in N})$ *be a history for* $NG$ *and let* $t \in \mathbb{N}$*. Then, the expected utility of action* $s_i \in S_i$ *for the agent* $i$ *at time* $t$*, denoted by* $EU^t_i(s_i)$*, is defined as follows:*

$$EU^t_i(s_i) = \sum_{s'_{-i} \in S_{-i}} \omega^t_i(s'_{-i}) \times U^*_i(s_i, s'_{-i})$$

As the following definition highlights, an agent is rational at a given time point $t$, if and only if her choice at time $t$ maximizes expected utility.

**Definition 10 (Rationality at time $t$).** *Let* $NG = (N, (S_i)_{i \in N}, U, I, \kappa)$ *be a normative game, let* $H = ((\omega_{i,j})_{i,j \in N}, (c_i)_{i \in N})$ *be a history for* $NG$ *and let* $t \in \mathbb{N}$*. Then, agent* $i$ *is rational at time* $t$ *if and only if* $EU^t_i(c_i(t)) \geq EU^t_i(s_i)$ *for all* $s_i \in S_i$*.*

We assume that agents learn via fictitious play [7], a learning algorithm introduced in the area of game theory and widely used in the area of multi-agent systems (see, *e.g.*, [22]). The idea of fictitious play is that each agent best responds to the empirical frequency of play of her opponents. The assumption underlying fictitious play is that each agent believes that her opponents are playing stationary strategies that do not depend from external factors such as the other agents' last moves.

**Definition 11 (Learning via fictitious play).** *Let* $NG = (N, (S_i)_{i \in N}, U, I, \kappa)$ *be a normative game and let* $H = ((\omega_{i,j})_{i,j \in N}, (c_i)_{i \in N})$ *be a history for* $NG$*. Then,*

*agent $i$ learns according to fictitious play (FP) along $H$, if and only if for all $j \in N \setminus \{i\}$, for all $s_j \in S_j$ and for all $t > 0$ we have:*

$$\omega_{i,j}^t(s_j) = \frac{obs_{i,j}^t(s_j)}{\sum_{s'_j \in S_j} obs_{i,j}^t(s'_j)}$$

*where $obs_{i,j}^0(s_j) = 0$ and for all $t > 0$:*

$$obs_{i,j}^t(s_j) = \begin{cases} obs_{i,j}^{t-1}(s_j) + 1 & \text{if } c_j(t-1) = s_j \\ obs_{i,j}^{t-1}(s_j) & \text{if } c_j(t-1) \neq s_j \end{cases}$$

Note that $obs_{i,j}^t(s_j)$ in the previous definition denotes the number of agent $i$'s past observations at time $t$ of agent $j$'s strategy $s_j$.

Two notions of convergence for fictitious play are given in the literature, one for pure strategies and one for mixed strategies. Let $H = ((\omega_{i,j})_{i,j \in N}, (c_i)_{i \in N})$ be a history. Then, $H$ converges in the pure strategy sense if and only if there exists a pure strategy $s \in S$ and $\bar{t} \in \mathbb{N}$ such that for all $i \in N$:

$$c_i(t) = s_i \text{ for all } t \geq \bar{t}$$

On the contrary, $H$ converges in the mixed strategy sense if and only if there exists a mixed strategy $\sigma \in \Sigma$ such that for all $i \in N$ and for all $s_i \in S_i$:

$$\lim_{\bar{t} \to \infty} \frac{|\{t \leq \bar{t} : c_i(t) = s_i\}|}{\bar{t} + 1} = \sigma_i(s_i)$$

Clearly, convergence in the pure strategy sense is a special case of convergence in the mixed strategy sense.

It has been proved [18] that for every *non-degenerate* $2 \times 2$ game (*i.e.*, two-player game where each player has two strategies available) and for every history $H$ for this game, if all agents are rational and learn according to fictitious play along $H$, then $H$ converges in the mixed strategy sense. The fact that the game is *non-degenerate* just means that, for every strategy of the second player there are no different strategies of the first player which guarantee the same payoff to the first player, and for every strategy of the first player there are no different strategies of the second player which guarantee the same payoff to the second player.[2] A generalization of this result to $2 \times n$ games has been given by [5].

## 4 Mathematical Analysis in the PD with Fairness Norm à la Rawls

In this section, we provide convergence results for fictitious play in the case of iterated Prisoner's Dilemma in which players are more or less sensitive to the fairness norm à la Rawls.

[2] Miyazawa [17] assumed a particular tie-breaking rule to prove convergence of fictitious play in $2 \times 2$ games.

| | C | D |
|---|---|---|
| C | $R, R$ | $-\kappa_1 P, T - \kappa_2 R$ |
| D | $T - \kappa_1 R, -\kappa_2 P$ | $P, P$ |

**Fig. 2.** Prisoner's Dilemma with transformed utilities according to fairness norm à la Rawls.

The first thing we can observe is that for any possible combination of norm sensitivity values for the two players, the behaviors of both players will converge to mixed strategies. In particular:

**Theorem 1.** *Let $NG = (N, (S_i)_{i \in N}, U, I, \kappa)$ be a normative game with fairness norm à la Rawls such that $(N, (S_i)_{i \in N}, U)$ is the Prisoner's Dilemma and let $H = ((\omega_{i,j})_{i,j \in N}, (c_i)_{i \in N})$ be a history for $NG$. Moreover, assume that every agent in $N$ learns according to fictitious play along $H$ and is rational for all $t \geq 0$. Then, $H$ converges in the mixed strategy sense.*

*Proof.* For all possible values of $\kappa_1$ and $\kappa_2$, the transformed PD in which the utility function $U_i$ is replaced by $U_i^*$ for all $i \in \{1, 2\}$ is non-degenerate. The transformed PD is represented in Figure 2. Hence, the theorem follows from the fact that, as observed in the previous section, fictitious play is guaranteed to converge in the class of non-degenerate $2 \times 2$ games. □

Our second result is the following theorem about convergence in the pure strategy sense. The theorem highlights that if at the beginning of the learning process every player has a uniform probability distribution over the strategies of the other player and the value of norm sensitivity is lower than the following threshold for cooperativeness

$$\theta_{tc} = \frac{P + T - R}{R - P}$$

for both players, then the two players will always play mutual defection. On the contrary, if at the beginning of the learning process every player has a uniform probability distribution over the strategies of the other player and the value of norm sensitivity is higher than the threshold $\theta_{tc}$ for both players, then the two players will always play mutual cooperation.

**Theorem 2.** *Let $NG = (N, (S_i)_{i \in N}, U, I, \kappa)$ be a normative game with fairness norm à la Rawls such that $(N, (S_i)_{i \in N}, U)$ is the Prisoner's Dilemma and let $H = ((\omega_{i,j})_{i,j \in N}, (c_i)_{i \in N})$ be a history for $NG$. Moreover, assume that every agent in $N$ learns according to fictitious play along $H$ and is rational for all $t \geq 0$, and that $\omega_{i,j}^0(s_j) = 0.5$ for all $i, j \in \{1, 2\}$ and for all $s_j \in \{C, D\}$. Then:*

- *if $\kappa_1 < \theta_{tc}$ and $\kappa_2 < \theta_{tc}$ then $c_1(t) = c_2(t) = D$ for all $t \geq 0$,*
- *if $\kappa_1 > \theta_{tc}$ and $\kappa_2 > \theta_{tc}$ then $c_1(t) = c_2(t) = C$ for all $t \geq 0$.*

*Proof.* Assume that every agent in $N$ learns according to fictitious play along $H$ and is rational for all $t \geq 0$, and that $\omega^0_{i,j}(s_j) = 0.5$ for all $i, j \in \{1, 2\}$ and for all $s_j \in \{C, D\}$. We are going to prove that, for all $i, j \in \{1, 2\}$, if $\kappa_i > \frac{P+T-R}{R-P}$ then $EU^0_i(C) > EU^0_i(D)$ and that if $\kappa_i < \frac{P+T-R}{R-P}$ then $EU^0_i(D) > EU^0_i(C)$.

First of all, let us compute the values of $EU^0_i(D)$ and $EU^0_i(C)$:

$$\begin{aligned} EU^0_i(D) &= 0.5 \times P + 0.5 \times (T - \kappa_i \times (R - S)) \\ &= 0.5 \times P + 0.5 \times (T - \kappa_i \times R) \\ &= 0.5 \times (P + T - \kappa_i \times R) \end{aligned}$$

$$\begin{aligned} EU^0_i(C) &= 0.5 \times R + 0.5 \times (S - \kappa_i \times (P - S)) \\ &= 0.5 \times R + 0.5 \times (-\kappa_i \times P) \\ &= 0.5 \times (R - \kappa_i \times P) \end{aligned}$$

It follows that $EU^0_i(D) > EU^0_i(C)$ if and only if $P + T - \kappa_i \times R > R - \kappa_i \times P$. The latter is equivalent to $\kappa_i < \frac{P+T-R}{R-P}$. Therefore, we have $EU^0_i(D) > EU^0_i(C)$ if and only if $\kappa_i < \frac{P+T-R}{R-P}$. By analogous argument, we can prove that $EU^0_i(C) > EU^0_i(D)$ if and only if $\kappa_i > \frac{P+T-R}{R-P}$.

It is routine task to verify that, for all possible values of $\kappa_1$ and $\kappa_2$ in the original normative game $NG$, the strategy profile $(D, D)$ is a *strict Nash equilibrium* in the transformed PD depicted in Figure 2 in which the utility function $U_i$ is replaced by $U^*_i$ for all $i \in \{1, 2\}$. Hence, by Proposition 2.1 in [8] and the fact that every agent is rational for all $t \geq 0$, it follows that if $\kappa_1 < \frac{P+T-R}{R-P}$ and $\kappa_2 < \frac{P+T-R}{R-P}$ then $c_1(t) = c_2(t) = D$ for all $t \geq 0$.

It is also a routine to verify that, if $\kappa_i > \frac{T-R}{R-S}$ for all $i \in \{1, 2\}$, then the strategy profile $(C, C)$ is a *strict Nash equilibrium* in the transformed PD depicted in Figure 2. Hence, by Proposition 2.1 in [8], the fact that every agent is rational for all $t \geq 0$ and the fact that $\frac{P+T-R}{R-P} > \frac{T-R}{R-S}$, it follows that if $\kappa_1 > \frac{P+T-R}{R-P}$ and $\kappa_2 > \frac{P+T-R}{R-P}$ then $c_1(t) = c_2(t) = C$ for all $t \geq 0$. □

## 5 Computational Results in the PD with Fairness Norm à la Rawls

Theorem 2 shows that if both $\kappa$-values are smaller than the threshold for cooperativeness $\theta_{tc}$, both players converge to mutual defection, whereas if both $\kappa$-values are greater than this threshold, both players converge to mutual cooperation. Note that this does not cover the whole space of tuples of $\kappa$-values, c.f. how do agents operate, if one value is smaller and the other value is greater that $\theta_{tc}$? In these terms we are faced with the more general question: for which combination of $\kappa$-values do agents converge to mutual cooperation or to mutual defection under fictitious play?

To examine the convergence behavior of players under fictitious play for different $\kappa$-values, we conducted multiple computations of repeated interactions, for different game parameters and a large subset of the $\kappa^2$-space. We recorded the results and we managed to deduce the conditions determining the convergence behavior that pertain perfectly with the data. These conditions are as follows.

For all normative games with fairness norm à la Rawls $NG = (N, (S_i)_{i\in N}, U, I, \kappa)$ and history $H = ((\omega_{i,j})_{i,j\in N}, (c_i)_{i\in N})$ for $NG$ that we computed such that $(N, (S_i)_{i\in N}, U)$ is the Prisoner's Dilemma, every agent in $N$ learns according to fictitious play along $H$, is rational for all $t \geq 0$, and $\omega^0_{i,j}(s_j) = 0.5$ for all $i, j \in \{1, 2\}$ and for all $s_j \in \{C, D\}$, the following three conditions were satisfied:

1. if $(\kappa_1 - lim_{mx}) \times (\kappa_2 - lim_{mx}) < curv_{mx}$ then $\exists t' \in \mathbb{N} : c_1(t) = c_2(t) = D$ for all $t \geq t'$,
2. if $(\kappa_1 - lim_{mx}) \times (\kappa_2 - lim_{mx}) > curv_{mx}$ then $\exists t' \in \mathbb{N} : c_1(t) = c_2(t) = C$ for all $t \geq t'$,
3. if $(\kappa_1 - lim_{mx}) \times (\kappa_2 - lim_{mx}) = curv_{mx}$ then both players converge to a mixed strategy,

whereby:

$$curv_{mx} = \left(\frac{PT}{(R+P)(R-P)}\right)^2$$

$$lim_{mx} = \frac{P^2 + R(T-R)}{(R+P)(R-P)}$$

Note that the equation $(\kappa_1 - lim_{mx}) \times (\kappa_2 - lim_{mx}) = curv_{mx}$ defines a separating curve between the convergence to mutual cooperation and mutual defection: for at least one of both $\kappa$-values being less than given, the first condition holds and fictitious play converges to mutual defection, whereas for at least one of both $\kappa$-values being greater than given, the second condition holds and fictitious play converges to mutual cooperation. For each pair of $\kappa$-values that fulfills the equation, the third condition holds and fictitious play converges to a mixed strategy for each player. This curve can be defined as a function for the convergence to a mixed strategy $f_{mx}$ over $\kappa_1$-values[3]:

$$f_{mx}(\kappa_1) = \frac{curv_{mx}}{\kappa_1 - lim_{mx}} + lim_{mx}$$

The function $f_{mx}$ is depicted in Figure 3. A necessary condition of function $f_{mx}$ to be correct is that it has an intersection point for $\kappa_1 = \kappa_2 = \theta_{tc}$, as proved in Theorem 3. An implication of function $f_{mx}$ to be correct is the fact that the value $lim_{mx}$ is the asymptote of the function $f_{mx}$, as proved in Theorem 4, and therefore determines a lower bound for $\kappa$-values that enable the convergence to mutual cooperation. Finally, note that the value $curv_{mx}$ determines the curvature of the function. Since $lim_{mx}$ and $curv_{mx}$ both depend on the parameters of

[3] Note that the function forms an *anallagmatic* curve, c.f. it inverts into itself.

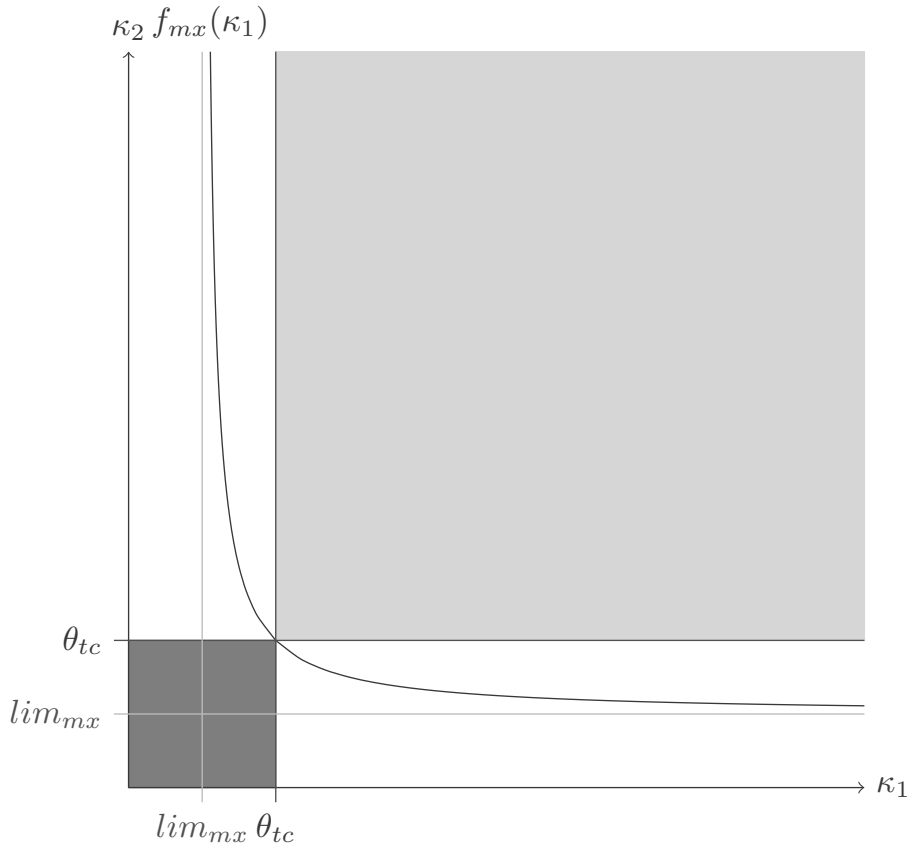



**Fig. 3.** The dark gray/light gray area shows in accordance with Theorem 2 that if both $\kappa$-values are smaller than the threshold for cooperativeness $\theta_{tc} = \frac{P+T-R}{R-P}$, both players behave according to mutual defection (dark gray area), whereas if both $\kappa$-values are greater than this threshold, both players behave according to mutual cooperation (light gray area). The curve represents the function for non-convergence $f_{mx}$ and defines for which combination of $\kappa$-values players converge to mutual cooperation (right of/above the curve), converge to mutual defection (left of/below the curve) or converge to a combination of mixed strategies (points of the curve). Note that i) $lim_{mx}$ is the asymptote of the function $f_{mx}$, thus it defines a lower bound for mutual cooperation, and ii) $\kappa_1 = \kappa_2 = \theta_{tc}$ is an intersection point of the curve.

the PD game, the asymptote and curvature of a function $f_{mx}$ can strongly differ among different games. Figure 4 shows the different curves of function $f_{mx}$ for different game parameters.

**Theorem 3.** *$\kappa_1 = \kappa_2 = \theta_{tc}$ is an intersection point of function $f_{mx}$.*

*Proof.* We are going to show that $f_{mx}(\theta_{tc}) = \theta_{tc}$:

$$\begin{aligned} f_{mx}(\theta_{tc}) &= \frac{curv_{mx}}{\theta_{tc} - lim_{mx}} + lim_{mx} \\ &= \frac{\left(\frac{PT}{(R+P)(R-P)}\right)^2}{\frac{P+T-R}{R-P} - \frac{P^2+R(T-R)}{(R+P)(R-P)}} + lim_{mx} \\ &= \frac{\left(\frac{PT}{(R+P)(R-P)}\right)^2}{\frac{(R+P)(P+T-R)}{(R+P)(R-P)} - \frac{P^2+R(T-R)}{(R+P)(R-P)}} + lim_{mx} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\left(\frac{PT}{(R+P)(R-P)}\right)^2}{\frac{(R+P)(P+T-R)-(P^2+R(T-R))}{(R+P)(R-P)}} + lim_{mx} \\
&= \left(\frac{PT}{(R+P)(R-P)}\right)^2 \times \frac{(R+P)(R-P)}{(R+P)(P+T-R)-(P^2+R(T-R))} + lim_{mx} \\
&= \left(\frac{PT}{(R+P)(R-P)}\right)^2 \times \frac{(R+P)(R-P)}{PT} + lim_{mx} \\
&= \frac{PT}{(R+P)(R-P)} + lim_{mx} \\
&= \frac{PT}{(R+P)(R-P)} + \frac{P^2+R(T-R)}{(R+P)(R-P)} \\
&= \frac{PT+P^2+R(T-R)}{(R+P)(R-P)} \\
&= \frac{PT+P^2+RT-R^2}{(R+P)(R-P)} \\
&= \frac{(P+T-R)(R+P)}{(R+P)(R-P)} \\
&= \frac{P+T-R}{R-P} = \theta_{tc}
\end{aligned}
$$

□

**Theorem 4.** *$lim_{mx}$ is the asymptote of function $f_{mx}$.*

*Proof. We are going to show that $\lim_{\kappa \to +\infty} f_{mx}(\kappa) = lim_{mx}$:*

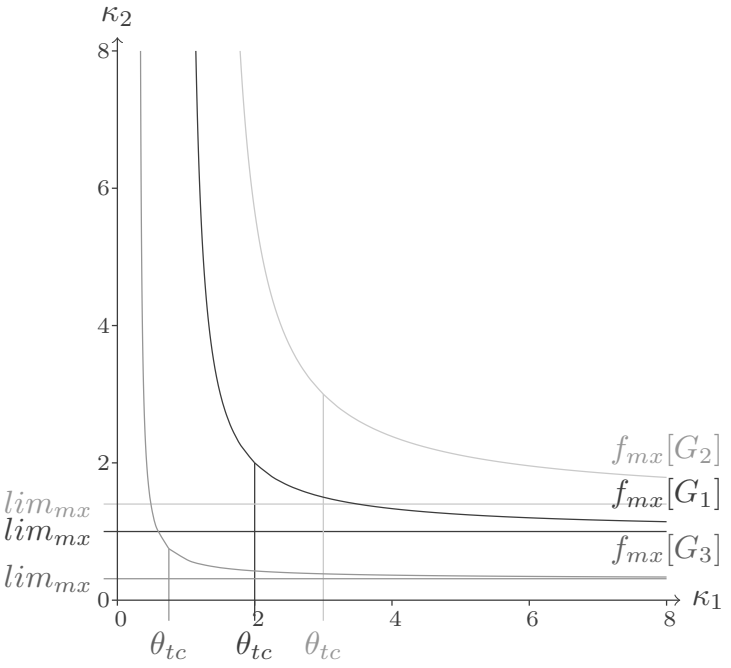
$$
\begin{aligned}
\lim_{\kappa \to +\infty} f_{mx}(\kappa) &= \lim_{\kappa \to +\infty} \left( \frac{curv_{mx}}{\kappa - lim_{mx}} + lim_{mx} \right) \\
&= \lim_{\kappa \to +\infty} \left( \frac{curv_{mx}}{\kappa - lim_{mx}} \right) + lim_{mx} \\
&= lim_{mx}
\end{aligned}
$$

□

## 6 Tournaments and Experimental Results

Let's assume we have a mixed population in terms of sensitivity to fairness norm $\kappa$. There might be individuals with high $\kappa$-values, with low $\kappa$-values or with no sensitivity to that norm at all. In such a setup it is reasonable to ask how beneficial fairness norm sensitivity might be. Is a low, middle or high sensitivity rather detrimental or profitable - especially in comparison with the outcome of the other individuals of the population?

To get a general idea of how beneficial a particular degree of sensitivity to the fairness norm might be, we tested the performance of agents with different $\kappa$-values in a tournament. Such a tournament was inspired by Axelrod's tournament of the repeated Prisoner's Dilemma [4]. In Axelrod's tournament a number of agents play the repeated Prisoner's Dilemma - pairwise each agent against every other agent - for a particular number of repetitions. Each agent



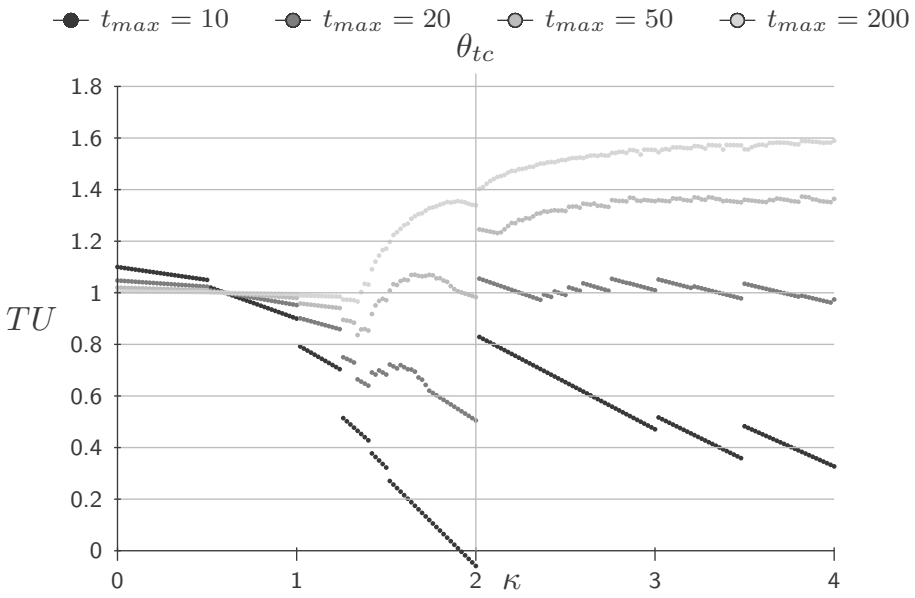

| Game | $G_1$ | $G_2$ | $G_3$ |
|---|---|---|---|
| T | 3 | 4 | 7 |
| R | 2 | 3 | 6 |
| P | 1 | 2 | 2 |
| $\theta_{tc}$ | 2 | 3 | .75 |
| $lim_{mx}$ | 1 | 1.4 | .125 |
| $curv_{mx}$ | 1 | 2.56 | $\approx .2$ |

**Fig. 4.** Exemplary Prisoner's Dilemma games $G_1$ ($T = 3, R = 2, P = 1$), $G_2$ ($T = 4, R = 3, P = 2$) and $G_3$ ($T = 7, R = 6, P = 2$) and their corresponding values $\theta_{tc}$, $lim_{mx}$ and $curv_{mx}$ (right table). The graph shows the corresponding curves of the function $f_{mx}$ for each game. Note that the value $curv_{mx}$ behaves anti-proportional to the curvature of the function.

updates her behavior according to a rule defined by its creator. The score of each encounter is recorded, and the agent with the highest average utility over all encounters wins the tournament.

In our tournament we also define a number of $n$ agents $0, 1, 2, ..., n-1$, where each agent plays against each other agent for a number of repetitions $t_{max}$. In distinction from Axelrod's tournament, all agents i) play the Prisoner's Dilemma as a normative game with fairness norm à la Rawls, and ii) have the same update rule: fictitious play. Although the agents have the same update rule, they differ in another crucial aspect: their sensitivity to the fairness norm. To keep things simple, we predefine that their sensitivity is i) bounded above by a value $\kappa_{max} \in \mathbb{R}_{>0}$, and ii) equally distributes among the $n$ agents, just by ascribing sensitivity to the fairness norm $\kappa_i = \frac{i \times \kappa_{max}}{n-1}$ to agent $i$.[4] A tournament works as follows: for each pair of agents $i, j$ we conducted a normative game with fairness norm à la Rawls based on the Prisoner's Dilemma for a number of $t_{max}$ repetitions, whereby agents $i$ and $j$ learn according to fictitious play along their common history. For each agent $i$ her average utility $TU_i$ - called *tournament*

[4] Note that to ascribe a value of fairness norm sensitivity $\kappa = \frac{i \times \kappa_{max}}{n-1}$ to agent $i$ ensures that agent 0 has a sensitivity to the fairness norm of 0, agent $n-1$ has one of $\kappa_{max}$, and all other agents' sensitivity to the fairness norm are equally distributed between these boundaries.




| $t_{max}$ | $\kappa^*$ |
|---|---|
| 10 | 0.0 |
| 20 | 2.02 |
| 50 | 3.82 |
| 200 | 4.0 |

**Fig. 5.** The resulting tournament utilities of four different tournaments with 200 agents each, pairwise playing a normative game with fairness norm à la Rawls based on a Prisoners Dilemma game with $T = 3$, $R = 2$, $P = 1$ and $S = 0$. The right table shows for each $t_{max}$ parameter the appropriate optimal sensitivity to the fairness norm $\kappa^*$ of the tournament's winner.

*utility* - is computed, which is the average utility value an agent scored over all interactions.

For a given set of agents $A$ that participate in such a tournament, the winner is the agent $i \in A$ who obtains the maximal tournament utility $TU_i$. We refer to the winner's $\kappa_i$ value as the optimal fairness norm sensitivity $\kappa^*$, with respect to her tournament utility:

$$\kappa^* = \kappa_i \text{ with } i = \arg\max_{j \in A} TU_j$$

We computed 4 tournaments, each with 200 agents playing a normative game $NG$ with fairness norm à la Rawls based on a Prisoner's Dilemma with $T = 3$, $R = 2$, $P = 1$ and $S = 0$. For such a game $\theta_{tc}$ is 2, and to ensure an equal portion of cooperative and non-cooperative agents, we set $\kappa_{max} = 2 \times \theta_{tc} = 4$. The tournaments differed in the parameter for $t_{max}$, here we chose the values 10, 20, 50 and 200. Figure 5 shows the performance of each agent in the appropriate tournament and the appended table shows the $\kappa^*$ value of each tournament's winner.

The results of the tournaments indicate that the optimal sensitivity to the fairness norm is by any means dependent of $t_{max}$ and $\theta_{tc}$. To verify this indication we computed a great number of further tournaments with different $t_{max}$ and $\kappa_{max}$ values. The results support de facto - without any exception - the following two observations:

1. For two tournaments that differ solely in the parameters $t_{max}$ and $t'_{max}$, whereby the tournaments' optimal values of sensitivity to the fairness norm are $\kappa^*$ and $\kappa'^*$, respectively, the following fact holds:

$$\text{if } t_{max} > t'_{max} \text{ then } \kappa^* \geq \kappa'^*$$

2. For every tournament it holds that:

$$\kappa^* = 0 \text{ or } \kappa^* > \theta_{tc}$$

The first observation unveils one condition for which a high sensitivity to the fairness norm à la Rawls is beneficial. It tells us that $\kappa^*$ is monotonically increasing in dependence of $t_{max}$, i.o.w. the value of the optimal sensitivity to the fairness norm increases with the number of repetitions of a repeated game in such a tournament. This result is in line with former insights, since i) we showed in Section 4 that a high value of fairness norm sensitivity supports cooperative behavior, and ii) we know from studies of repeated Prisoner's Dilemma that cooperative behavior is especially beneficial in combination with reputation [19], a virtue that needs repetition to be established.

The second observation says that it is optimal either to have no sensitivity to the fairness norm at all, or to have a sensitivity to the fairness norm that ensures *preliminary cooperativeness*[5]. Which of both cases holds depends inter alia on the number of repetitions $t_{max}$. By all means, it is never the case that $0 < \kappa^* \leq \theta_{tc}$. This stresses the fact that a great fairness norm sensitivity is only beneficial if it not only enables a line of mutual cooperation, but it also implies preliminary willingness to start it.

## 7 Conclusion and Perspectives

Our study presents a game-theoretic model of guilt in relation to sensitivity to the norm of fairness à la Rawls. We i) employed this model on the Prisoner's Dilemma, and ii) worked out the convergence behavior under fictitious play for any combination of the fairness norm sensitivity of both players. We found out that a particular threshold for cooperation $\theta_{tc}$ plays a crucial role: it defines for which combinations both agents cooperate of defect from the beginning, and for which combinations they might learn to cooperate or to defect. In a final experimental setup, we analyzed the performance of multiple agents with different values of sensitivity to the fairness norm involved in a tournament of repeated games. We revealed that i) a great sensitivity to the fairness norm is the more beneficial, the higher the number of repetitions of the repeated game is, and ii) the threshold for cooperation $\theta_{tc}$ defines a lower bound for a great sensitivity to the fairness norm to be beneficial at all.

[5] As we have shown in Theorem 2, if agents have a sensitivity to fairness $\kappa > \theta_{tc}$, their first move is to cooperate. This behavioral characteristic can be seen as preliminary cooperativeness.

A further observation - that was not elaborated here - was the fact that a great sensitivity to a fairness norm is the more beneficial in a population, the more other agents have a great sensitivity to that norm. This fact let us presume that fairness norm sensitivity i) is a reasonable value for explaining multiple cooperation in multi-player games like the public goods game, and ii) is a good candidate to be analyzed under stability aspects of population dynamics, e.g. as an *evolutionary stable strategy* [15], a standard concept in evolutionary game theory. Such analyses are currently in progress and will be part of subsequent studies. This line of future work will allow us to relate our analysis with existing naturalistic theories of fairness according to which sensitivity to fairness norm might be the product of evolution (see, *e.g.*, [6]).

## References


1. Andrighetto, G., Villatoro, D., Conte, R.: Norm internalization in artificial societies. AI Communications **23**, 325–339 (2010)
2. Andrighetto, G., Villatoro, D., Conte, R.: The role of norm internalizers in mixed populations. In: Conte, R., Andrighetto, G., Campenni, M. (eds.) Minding Norms: Mechanisms and Dynamics of Social Order in Agent Societies, pp. 153–174. Oxford University Press, Oxford (2013)
3. Aronfreed, J.M.: Conduct and Conscience: The Socialization of Internalized Control Over Behavior. Academic Press, New York (1968)
4. Axelrod, R.: The Evolution of Cooperation. Basic books (1984)
5. Berger, U.: Fictitious play in 2xn games. Journal of Economic Theory **120**, 139–154 (2005)
6. Binmore, K.: Natural justice. Oxford University Press, New York (2005)
7. Brown, G.W.: Iterative solution of games by fictitious play. In: Koopmans, T.C. (ed.) Activity Analysis of Production and Allocation, pp. 374–376. John Wiley, New York (1951)
8. Fudenberg, D., Levine, D.K.: The Theory of Learning in Games. MIT Press, Cambridge (1998)
9. Gaudou, B., Lorini, E., Mayor, E.: Moral guilt: an agent-based model analysis. In: Kamiński, B., Koloch, G. (eds.) Advances in Social Simulation. AISC, vol. 229, pp. 95–106. Springer, Heidelberg (2014)
10. Gintis, H.: The hitchhiker's guide to altruism: Gene-culture co- evolution, and the internalization of norms. Journal of Theoretical Biology **220**(4), 407–418 (2003)
11. Gintis, H.: The genetic side of gene-culture coevolution: internalization of norms and prosocial emotions. Journal of Economic Behavior and Organization **53**, 57–67 (2004)
12. Haidt, J.: The moral emotions. In: Davidson, R.J., Scherer, K.R., Goldsmith, H. H. (eds.) Handbook of Affective Sciences, pp. 852–870. Oxford University Press (2003)
13. Harsanyi, J.: Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. Journal of Political Economy **63**, 309–321 (1955)
14. Loomes, G., Sugden, R.: Regret theory: An alternative theory of rational choice under uncertainty. Economic Journal **92**(4), 805–824 (1982)
15. Smith, J.M., Price, G.R.: The logic of animal conflict. Nature **246**, 15–18 (1973)
16. Mill, J.S.: Utilitarianism. Parker, Son & Bourn, West Strand (1863)

17. Miyazawa, K.: On the convergence of the learning process in a 2x2 non-zero-sum two-person game, vol. 3. Princeton University Econometric Research Program (1961)
18. Monderer, D., Shapley, L.S.: Fictitious play property for games with identical interests. Journal of Economic Theory **68**, 258–265 (1996)
19. Nowak, M., Sigmund, K.: Evolution of indirect reciprocity. Nature **437**, 1291–1298 (2005)
20. Rawls, J.: A Theory of Justice. Harvard University Press, Cambridge (1971)
21. Searle, J.: Rationality in Action. MIT Press, Cambridge (2001)
22. Vidal, J.M.: Learning in multiagent systems: an introduction from a game-theoretic perspective. In: Alonso, E., Kudenko, D., Kazakov, D. (eds.) Adaptive Agents and Multi-agent Systems, pp. 202–215. Springer-Verlag, Berlin (2003)