# Analysis of the Sadism-Egoism-Altruism Model and Comparison of it to Related Models*

Discussion Paper

Michael Heinrich Baumann†

October $27^{th}$, 2025

*Abstract*—We analyze the Sadism-Egoism-Altruism (SEA) Model from Baumann and Baumann (2025) [3] analytically. We prove that every finite game does not only have an outcome that is plausible in the SEA model when allowing for randomized strategies, but also the existence of such an outcome in pure strategies. Further, we show that all fairness equilibria according to Rabin (1993) [24] are plausible in the SEA model. Although typically many or even all pure-strategy outcomes can be SEA plausible in a game, this model gives deep insights into the structure of the game. Since along with the fact that outcomes are plausible there come ranges for parameters modeling sadism or altruism making the respective outcome plausible, the possible behavior of the agents can be understood. Comparisons to the Fehr-Schmidt und Bolton-Ockenfels models are done. Via the well-known Prisoner's dilemma the SEA model is illustrated for mixed strategies, too. This work opens doors for manifold future research.



*This work builds upon [3] (and also on [2]) and, hence, the author of the work at hand is grateful to his co-author of [2] and [3], Michaela Baumann.

†Department of Mathematics, University of Bayreuth, Germany,
michael.baumann@uni-bayreuth.de

*Keywords*—Game Theory; Fairness; Altruism; Sadism; Egoism; Prisoner's Dilemma; Anti-Social Punishment; Public Goods Game; Reciprocity; One-Shot Game

*JEL codes*—C72, D9
*UDC*—519.83
*MSC2020*—91A05, 91A10, 91A40

*This paper is dedicated to Amalrich.*



# 1 Motivation

The so-called Nash equilibrium [22] is one of the standard tools to solve games. However, in many experiments and real-world situations, behavior that does not fit to the predictions from Nash's idea is observed (see, e.g., [1]). There is a vast body of literature on that. One specifically interesting topic therein is so-called cooperation, which means esp. cooperation in games that resemble the prisoner's dilemma. See, very prominently, [10] (and cf. [9]). These so-called anomalies describe situations where—in the language of Rabin [24]—agents behave kind towards each other, although behaving mean would individually increase their (material) payoff. Many ideas have been developed to explain such an "abnormal" behavior. For example, explanations use distributions of the payoffs among the agents or psychological fairness concepts, see, e.g., [2, 6, 13, 14, 16, 24].

Table 1: Prisoner's Dilemma [27] with scaling parameter $\chi > 0$ [24]. The higher $\chi$, the more important is the material payoff compared to the so-called fairness payoff for the agents

| $u_1(\cdot)\|u_2(\cdot)$ | $a_2^{(1)}$ | $a_2^{(2)}$ |
|---|---|---|
| $a_1^{(1)}$ | $3\chi\|3\chi$ | $0\|5\chi$ |
| $a_1^{(2)}$ | $5\chi\|0$ | $\chi\|\chi$ |

In [3], another interesting anomaly presented in the literature is discussed, namely "antisocial punishment." The idea behind "punishment" (or, more specific, "costly punishment") is the following: When having a look at the prisoner's dilemma depicted in Table 1 it is obvious that $(a_1^{(2)}, a_2^{(2)})$ is the only Nash equilibrium, i.e., both agents do not cooperate, although cooperation, i.e. $(a_1^{(1)}, a_2^{(1)})$, is Pareto superior to non-cooperation. It is argued that often games in the real world are not pure prisoner's dilemmas (or the very related public goods games), but that there is a second round with another game, namely costly punishment, i.e., all agents can pay for some punishment of opponents (whereas the costs are positive but small compared to the punishment). In [3], this is formalized, cf. Table 2. Then, cooperation and non-punishment can be one Nash equilibrium—though punishment is an empty threat, see [27].[1]

Table 2: Costly Punishment [3] with scaling parameter $\varpi > 0$ [24]

| $u_1(\cdot)\vert u_2(\cdot)$ | $a_2^{(3)}$ | $a_2^{(4)}$ |
|---|---|---|
| $a_1^{(3)}$ | $0\vert 0$ | $-10\varpi\vert -\varpi$ |
| $a_1^{(4)}$ | $-\varpi\vert -10\varpi$ | $-11\varpi\vert -11\varpi$ |

The interesting anomaly 'anti-social punishment' is now that in experiments it can be observed that some agents cooperate in the first round, but punish the other—one may say: for fun—in the second round [23]. In [3], a model called Sadism-Egoism-Altruism (SEA) model is constructed which is motivated by this anomaly and inspired by [39]. The general idea behind the SEA model is that the given payoffs (agents usually try to maximize), which are then called 'material payoffs,' are transformed according to a rule which includes also the payoffs of the respective opponent. This way, it can be modeled that an agent wants something good or bad for her or his opponent and does not necessarily only think on her or his own payoff.

The model in the form of [3] does not fit to the setting of [23] perfectly, since in [23] the public goods game with costly punishment is somehow a game with incomplete information, due to the fact that agents do not know

[1] For literature on that and on whether and when this idea works in practice, see [7, 23, 25, 26, 37, 38].

the psychological preferences of the other agents, while in [3] the setting is modeled with complete information, i.e., agents know whether the opponent "has some fun" when punishing others—i.e. does anti-social punishment. We also note that in [23] this is a four-agents game while it is a two-agents game in [3]. In [3] it is computed that, for example, for specific values of the parameters modeling sadism respectively altruism, one-sided anti-social punishment is an equilibrium in the SEA model, but social punishment is not for these specific parameters, which are such that one agent is 'half egoist, half altruist' and the other agent is 'half egoist, half sadist.'

In [3], the two-agents prisoner's dilemma with costly punishment is transformed via the von Neumann-Morgenstern transformation into a one-shot game with the following notation like in Table 3.[2]

- action in the Prisoner's dilemma;(
  - action in the costly punishment if $(a_1^{(1)}, a_2^{(1)})$ was played in the Prisoner's dilemma;
  - action in the costly punishment if $(a_1^{(2)}, a_2^{(1)})$ was played in the Prisoner's dilemma;
  - action in the costly punishment if $(a_1^{(1)}, a_2^{(2)})$ was played in the Prisoner's dilemma;
  - action in the costly punishment if $(a_1^{(2)}, a_2^{(2)})$) was played in the Prisoner's dilemma

  )

The research questions of the work at hand are: *Does in all finite games a SEA Nash equilibrium [3] in pure strategies exist? How are the SEA Nash equilibria related to other concepts like Pareto, mutual max, mutual min? Is the SEA model related to other fairness concepts, such as to the one of Rabin [24] (and to others)?*

This section (Section 1) motivates the SEA model and its analysis by means of anti-social punishment, see [23]. In Section 2, Rabin's fairness model and the SEA model are presented, see [2, 3, 24]. The main part comes in Section 3 where the SEA model is analyzed analytically, its connection to Rabin's fairness model is proven, and, finally, connections to the models of

[2]We note that analyzing this game by means of the Selten transformation, cf. [21], would also be interesting.

Table 3: Prisoner's Dilemma with Costly Punishment [3] with scaling $\chi, \varpi > 0$, see [24]; values from Tables 1 and 2; truncated

| $u_1(\cdot)\|u_2(\cdot)$ | $a_2^{(1)};(a_2^{(3)};a_2^{(3)};a_2^{(3)};a_2^{(3)})$ | $\dots$ | $a_2^{(2)};(a_2^{(4)};a_2^{(4)};a_2^{(4)};a_2^{(4)})$ |
|---|---|---|---|
| $a_1^{(1)};(a_1^{(3)};a_1^{(3)};a_1^{(3)};a_1^{(3)})$ | $3\chi\|3\chi$ | $\dots$ | $-10\varpi\|5\chi-\varpi$ |
| $a_1^{(1)};(a_1^{(3)};a_1^{(3)};a_1^{(3)};a_1^{(4)})$ | $3\chi\|3\chi$ | $\dots$ | $-10\varpi\|5\chi-\varpi$ |
| $a_1^{(1)};(a_1^{(3)};a_1^{(3)};a_1^{(4)};a_1^{(3)})$ | $3\chi\|3\chi$ | $\dots$ | $-11\varpi\|5\chi-11\varpi$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $a_1^{(2)};(a_1^{(4)};a_1^{(4)};a_1^{(4)};a_1^{(4)})$ | $5\chi-\varpi\|-10\varpi$ | $\dots$ | $\chi-11\varpi\|\chi-11\varpi$ |

Fehr-Schmidt and Bolton-Ockenfels are presented, see [3, 6, 13, 24]. That the solution structure of games becomes rather complicated when allowing mixed strategies in the SEA model is illustrated by means of the Prisoner's dilemma in Section 5. Section 6 concludes and presents various directions for future research.

# 2 Models

The model of [3], the so-called Sadism-Altruism-Egoism model can explain why exactly one agent punishes the other in a prisoner's dilemma with costly punishment, although there is cooperation in the first round, but it cannot explain that both agents do anti-social punishment—which would be possible in a game with incomplete information. In [3], it is shown that one-sided anti-social punishment is neither Nash [22] nor fair in the sense of Rabin [24] (see also [2]). Whether distribution-dependent fairness concepts may explain anti-social punishment is not analyzed in [3], but may be done in future work. For implementations of Rabin's fairness model and of the SEA model see [2] and [3], where Python with SymPy is used [20, 29].

In the work at hand, we consider—if not stated otherwise—pure strategies (actions) and equilibria in those pure strategies. Though it should not be a big deal to enlarge the concepts to mixed (i.e. randomized) strategies, cf. [24].

## 2.1 Rabin's Fairness Equilibria

One possibility to explain cooperation is by adding some fairness payoff based upon reciprocity to the (material) payoff. This fairness concept is based on the following ideas by Rabin [24], namely: Agents increase their fairness payoff when they try to be kind to someone who is believed to be kind; agents decrease their fairness payoff when they try to be mean to someone who is believed to be mean; when material payoffs are larger, fairness is less important to the agents; this concept uses strategies (resp. actions), first- and second-order beliefs that have to match in an equilibrium. See [2] for an in-depth description of this concept or [24] for Rabin's paper.

Formally, Rabin [24] defines a fairness equilibrium as $(s_1, s_2) \in S_1 \times S_2$ with

$$s_i \in argmax_{s_i' \in S_i} \tilde{U}_i(s_i', b_{-i}, c_i)$$

and $c_i = b_i = s_i$ for $i = 1, 2$, where $s_i \in S_i$ is the mixed strategy $i$ uses, $b_{-i} \in S_{-i}$ is the strategy $i$ believes that $-i$ uses, and $c_i \in S_i$ is the strategy $i$ believes that $-i$ believes that $i$ uses. In [24],

$$\tilde{U}_i(s_i, b_{-i}, c_i) = u_i(s_i, b_{-i}) + \tilde{f}_{-i}(b_{-i}, c_i)(1 + f_i(s_i, b_{-i}))$$

is called the expected utility, which consists of the so-called kindness functions:

$$f_i(s_i, b_{-i}) = \begin{cases} \frac{u_{-i}(b_{-i}, s_i) - u^e_{-i}(b_{-i})}{u^h_{-i}(b_{-i}) - u^{min}_{-i}(b_{-i})} & \text{if } u^h_{-i}(b_{-i}) - u^{min}_{-i}(b_{-i}) \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\tilde{f}_{-i}(b_{-i}, c_i) = \begin{cases} \frac{u_i(c_i, b_{-i}) - u^e_i(c_i)}{u^h_i(c_i) - u^{min}_i(c_i)} & \text{if } u^h_i(c_i) - u^{min}_i(c_i) \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Here, the following functions are used.

- First-order believed, expected material payoff of $-i$ when $-i$ is believed to use $b_{-i}$ when $i$ uses $s_i$: $u_{-i}(b_{-i}, s_i)$
- Equitable payoff: $u^e_{-i}(b_{-i}) = \frac{u^h_{-i}(b_{-i}) - u^l_{-i}(b_{-i})}{2}$, see [24] p. 1286
- Set of all possible outcomes when $-i$ really uses the believed strategy $b_{-i}$: $\mathbf{u}(b_{-i}) = \{(u_i(s_i, b_{-i}), u_{-i}(b_{-i}, s_i)) \mid s_i \in S_i\}$

- Highest payoff $-i$ can receive within $\mathbf{u}(b_{-i})$: $u_{-i}^{h}(b_{-i})$
- Lowest payoff $-i$ can receive within $\mathbf{u}(b_{-i})$: $u_{-i}^{min}(b_{-i})$
- Lowest payoff $-i$ can receive inside the Pareto optimal subset of $\mathbf{u}(b_{-i})$: $u_{-i}^{l}(b_{-i})$, see [2] Comment 8

Please note that $s, b, c$ can be randomized strategies/beliefs making the $u$s expected payoffs. For $\tilde{f}_i, \tilde{f}_{-i}$ the terms are defined analogously with first-order beliefs instead of strategies and second-order beliefs instead of first-order beliefs. And again see [2] for details.

## 2.2 The SEA Model

In [3], another model, the so-called Sadism-Egoism-Altruism model, abbreviated: SEA model, is developed, motivated by the anomaly "anti-social punishment" and inspired by a model presented in [39]. In a first step, altruism is incorporated in the game model. Instead of Rawlsian functions, which are used in [39] to model (partial) altruism, in [3] a linear model is utilized, which according to [5] dates back to the year 1881, see [11].

$$U_i'(\lambda_i, a_i, a_{-i}) = (1 - \lambda_i)u_i(a_i, a_{-i}) + \lambda_i u_{-i}(a_{-i}, a_i)$$

for $i = 1, 2$ with $(\lambda_1, \lambda_2) \in [0, 1]^2$ and $a_{-i}, a_i$ being the actions (pure strategies). Thus, the higher the values of the $\lambda$s are, the more altruistic the agents behave; i.e., the more they want to maximize the other's outcome.

In a second step, this model is mirrored in order to account for (partial) sadism, leading to respecified (psychological) payoffs

$$U_i(\lambda_i, a_i, a_{-i}) = (1 - |\lambda_i|)u_i(a_i, a_{-i}) + \lambda_i u_{-i}(a_{-i}, a_i)$$

for $i = 1, 2$ with $(\lambda_1, \lambda_2) \in [-1, 1]^2$ and again $a_{-i}, a_i$ being the actions (pure strategies). Hence, agents may, depending on the values of the $\lambda$s, aim to minimize or to maximize the other's outcome—additionally to the target of maximizing the agent's own outcome. In the extreme cases of $\lambda_i = -1$, agent i is a pure sadist, $\lambda_i = 1$, he or she is a pure altruist, and $\lambda_i = 0$, she or he is a pure egoist.

Similar models have been utilized in the literature, e.g., in [19], where models for experiments are investigated.[3] The SEA model has the advantages that it comes with only two additional parameters, that it is symmetric for sadism and altruism, and that it uses on both sides (i.e. for sadism and for altruism) convex mixtures of the agents and his or her opponent's (material) payoffs (with minus or plus). Additionally, as we will see in Proposition 5 and the remarks thereafter, the pole cases $\lambda = -1, 1$ as well as the midpoint $\lambda = 0$ are well-known game-theoretic concepts. All in all, this makes results easily interpretable.

Whether, how, and to which degree this model can explain anti-social punishment is explained in [3] and summarized in the introduction of the work at hand, Section 1. In the course of this paper, theoretical findings concerning the SEA model are presented, which include also comparisons to Rabin's fairness model, Fehr-Schmidt, and Bolton-Ockenfels. While the focus of [3] was on modeling and coding (in Python), the work at hand focuses on theory and proofs.

# 3 Theoretical Findings

Now, we present theoretical analyses of the SEA model. In [3], an outcome is defined as SEA plausible, i.e., as plausible under the SEA model, if and only if there exists $(\lambda_1, \lambda_2) \in [-1, 1]^2$ s.t. the outcome is a Nash equilibrium under the transformed payoffs. One basic result is already stated in [3], which immediately follows by setting $\lambda_1 = \lambda_2 = 0$ in the model.

**Proposition 1.** *Each Nash equilibrium (of the original game) is SEA plausible.*

Thus, every finite game has a SEA plausible outcome in mixed strategies (cf. [22]), but does it also have one in pure strategies? For that, we

[3] At the very beginning of Chapter 2 of [19], a payoff transformation formula is given, which reads in our notation for the two agents case as follows:

$$\bar{U}_i(\nu, \mu_i, \mu_{-i}, a_i, a_{-i}) = u_i(a_i, a_{-i}) + \frac{\mu_i + \nu\mu_{-i}}{1 + \nu} \cdot u_{-i}(a_{-i}, a_i),$$

with $\nu \in [0, 1]$ and $\mu_i, \mu_{-i} \in (-1, 1)$. There, $\mu_i$ ($\mu_{-i}$) reflects how altruistic (positive values) or spiteful (negative values) agent $i$ ($-i$) is (cf. the references in [19]). The value $\nu$ is for incorporating "fairness" in a reciprocal sense—see [19]. If $\nu = 0$, the model is similar to the SEA model with $\lambda_i \in (-0.5, 0.5)$.

Table 4: Assurance Game

| $u_1(\cdot); u_2(\cdot)$ | $a_2^{(1)}$ | $a_2^{(2)}$ |
|---|---|---|
| $a_1^{(1)}$ | $2; 2$ | $0; 0$ |
| $a_1^{(2)}$ | $0; 0$ | $1; 1$ |

analyze the relationship between Pareto optima and the SEA model, where Pareto is always a property of the material game i.e. of the game without any transformation of the utilities. When having a look at the assurance game in Table 4[4] it is quite obvious that not every SEA plausible outcome is globally Pareto optimal (the payoffs $(1, 1)$ correspond to a Nash equilibrium, which is according to Proposition 1 SEA plausible, however, they are Pareto inferior to the payoffs $(2, 2)$). The next result tells us something about the converse. Especially, $(1, 1)$ corresponds to an equilibrium of the SEA model with $(\lambda_1, \lambda_2) = (0.5, 0.5)$.

**Proposition 2.** *At least one of the globally Pareto optimal, pure-strategy outcomes is SEA plausible.*

*Proof.* Let $(a_1^{(*)}, a_2^{(*)}) \in argmax_{(a_1,a_2)\in A_1\times A_2}(u_1(a_1, a_2) + u_2(a_2, a_1))$. At first, we are going to show that this outcome is globally Pareto optimal. If this was not the case, there was $(a_1^{(**)}, a_2^{(**)})$ s.t. w.l.o.g. $u_1(a_1^{(**)}, a_2^{(**)}) > u_1(a_1^{(*)}, a_2^{(*)})$ and $u_2(a_2^{(**)}, a_1^{(**)}) \geq u_2(a_2^{(*)}, a_1^{(*)})$. However, this contradicts that $(a_1^{(*)}, a_2^{(*)})$ is in the $argmax$. Now, let $\lambda_1 = \lambda_2 = 0.5$. We claim that $(a_1^{(*)}, a_2^{(*)})$ is SEA plausible for that choice of $\lambda_1, \lambda_2$. If $(a_1^{(*)}, a_2^{(*)})$ was not SEA plausible for $\lambda_1 = \lambda_2 = 0.5$ there would be $a_1^{(**)} \in A_1$ s.t. $U_1(0.5, a_1^{(**)}, a_2^{(*)}) > U_1(0.5, a_1^{(*)}, a_2^{(*)})$ or there would be $a_2^{(**)} \in A_2$ s.t. $U_2(0.5, a_2^{(**)}, a_1^{(*)}) > U_2(0.5, a_2^{(*)}, a_1^{(*)})$. Since $U_1(0.5, a_1, a_2) = U_2(0.5, a_2, a_1) = 0.5 \cdot (u_1(a_1, a_2) + u_2(a_2, a_1))$, $(a_1^{(**)}, a_2^{(*)})$ resp. $(a_1^{(*)}, a_2^{(**)})$ would result in a sum higher than the $max$, which is not possible. □

An outcome as in the proof may be called *sum optimum.* Note that the same proof holds true in a mixed-strategy setting. That means, when replacing $(a_1, a_2) \in A_1 \times A_2$ by $(s_1, s_2) \in S_1 \times S_2$.

[4] `https://en.wikipedia.org/wiki/Coordination_game` (2024-04-05)

Table 5: An Exemplary Game

| $u_1(\cdot); u_2(\cdot)$ | $a_2^{(1)}$ |
|---|---|
| $a_1^{(1)}$ | $9; 2$ |
| $a_1^{(2)}$ | $5; 5$ |
| $a_1^{(3)}$ | $2; 9$ |

Although Proposition 2 may raise hope that all Pareto optimal outcomes in the material game are SEA plausible, we have to disappoint the reader, at least for pure-strategy outcomes. The example in Table 5 shows that there may be outcomes that are Pareto in pure strategies, but not SEA plausible. It is obvious that all outcomes are Pareto in pure strategies, however, there is no $\lambda_1$ such that agent 1 would choose $a_1^{(2)}$.

We remark that the outcome in Table 5 which results in the payoff vector $(5; 5)$ is not Pareto in mixed strategies since $(0.5a_1^{(1)}) + 0.5a_1^{(3)}, a_2^{(1)})$ is Pareto superior to it. When we allow for randomized (i.e. mixed) strategies, the set of admissible outcomes is convex (albeit not strictly convex in a finite setting), leading to a right skewed Pareto frontier (since we maximize).[5] Via the slope of a (in detail: of each) tangent that lies right-above the admissible outcomes, one can compute $\lambda_1$ and $\lambda_2 := 1-\lambda_1$ s.t. each outcome on the Pareto frontier (regarding mixed strategies) is SEA Nash: Let an expected outcome on the frontier be given with tangent slope $m \in [-\infty, 0]$ (which does not have to be unique). With $\lambda_1 = \frac{1}{1-m} \in [0, 1]$ (with $\frac{1}{\infty} := 0$) the maximization leads to the given outcome. Note that due to the choice $\lambda_2 = 1 - \lambda_1$ the two maximization terms for being SEA Nash (i.e. Nash in the transformed game) are equal. Note that these choices of $\lambda_1, \lambda_2$ are special cases in the definition of SEA Nash; and that for the same combination of $\lambda$s various outcomes can be SEA Nash, cf. [3].

It is known that there are finite games without any Nash equilibrium in pure strategies (e.g., rock-scissors-paper, see [27]). From the proof of Proposition 2 it follows directly the following Proposition.

[5] One could imagine a concave function as the Pareto frontier, however, in the graph there can be a vertical (i.e., parallel to the $u_2$-axis) piece such that the relation is not a function at all.

**Proposition 3.** *Every finite two-agent game has at least one SEA plausible outcome in pure strategies.*

Proposition 3 follows also from the remarks after Proposition 2, namely, every outcome and esp. every pure-strategy outcome on the Pareto frontier in mixed strategies is SEA Nash; esp. the cases where only $u_i$ is maximized for $i = 1$ or $i = 2$ and where the sum is maximized—but note that the Pareto frontier can be a singleton. According to [24], a mutual-max resp. mutual-min outcome is an outcome where agents mutually maximize or minimize the opponent's outcome. We note that not every finite game has a mutual-max outcome or a mutual-min outcome in pure strategies, counterexamples are "matching pennies" and "rock scissors paper" (cf. [24]). However, if there is a mutual-max or a mutual-min outcome in pure strategies, it is SEA plausible, as the next proposition shows.

**Proposition 4.** *Every mutual-max outcome and every mutual-min outcome is SEA plausible.*

*Proof.* Let $(a_1, a_2)$ be a mutual-max outcome. By setting $\lambda_1 = \lambda_2 = 1$ the definition of (SEA) Nash equilibrium is exactly the definition of mutual-max outcome. For mutual-min outcomes the same is true for $\lambda_1 = \lambda_2 = -1$ and noting that minimizing a function is the same as maximizing minus the function. □

In [3], the parameters $\lambda_1, \lambda_2$ were defined to be in $[-1, 1]$. One may ask, why no values with $|\lambda_i| > 1$ are allowed in the SEA model, although this would be possible in theory. The proposition above gives a good argument for that: $\lambda_i$ are gradually shifting from $-1$, which is the mutual-min (see [24]), via 0, which is the Nash equilibrium (see [22]), to 1, which is the mutual-max (see again [24]). Thus, the SEA model is a generalization or the hull over these three important concepts with Nash as its midpoint.

Note that the very same proof for the proposition above also holds for mixed strategies. When transforming a finite game with $\lambda_1 = \lambda_2 = 1$ $(-1)$, the Nash theorem [22] tells us that the transformed game has at least one Nash equilibrium in randomized (i.e. mixed) strategies, which is, in fact, a mutual-max (mutual-min) outcome of the original game. Thus, such outcomes always exist in finite games in randomized strategies.

Here we mention that it is no coincidence that for the example that not every pure-strategy Pareto optimal outcome is SEA plausible there were

Table 6: The General Game ($d, e, f, g, h, i, j, k \in \mathbb{R}$)

| $u_1(\cdot); u_2(\cdot)$ | $a_2^{(1)}$ | $a_2^{(2)}$ |
|---|---|---|
| $a_1^{(1)}$ | $d; e$ | $f; g$ |
| $a_1^{(2)}$ | $h; i$ | $j; k$ |

three strategies for one of the agents needed. The next theorem tells us something about SEA plausibility in small games. However, even if in a game all outcomes are SEA plausible, that does not mean necessarily that every outcome is SEA Nash for all combinations of $\lambda$. When having a look at the sets **E** from the definition of SEA Nash equilibrium in [3]—these are the subsets of $[-1, 1]^2 \ni \lambda_1, \lambda_2$ for which an outcome of the transformed game is Nash—we can still learn a lot about the nature of the game and the outcomes. Especially, not every outcome has to be SEA plausible under so-called social types (as an example; see [3]).

**Proposition 5.** *In two-agent games with one or two strategies per agent, every outcome is SEA plausible.*

*Proof.* If an agent has only one strategy, this strategy is a best response to everything. Thus, we have to prove the proposition for $2 \times 2$-games. For that, we consider the game in Table 6 and we show that there exists for every outcome $(a_1, a_2)$ a tuple $(\lambda_1, \lambda_2) \in [-1, 1]^2$ s.t. $a_1$ is a best response to $a_2$ and vice versa. Since all payoffs are variable and we may interchange payoffs, strategies, and even agents, it is in fact enough to show that the first strategy of agent 1 is a best response to the first strategy of agent 2 for some $\lambda_1 \in [-1, 1]$. Hence, we compare $U_1(\lambda_1, a_1^{(1)}, a_2^{(1)}) = (1 - |\lambda_1|)d + \lambda_1 e$ with $U_1(\lambda_1, a_1^{(2)}, a_2^{(1)}) = (1 - |\lambda_1|)h + \lambda_1 i$. Now, if $e \geq i$, $U_1(1, a_1^{(1)}, a_2^{(1)}) = e \geq i = U_1(1, a_1^{(1)}, a_2^{(1)})$, and if $e \leq i$, $U_1(-1, a_1^{(1)}, a_2^{(1)}) = -e \geq -i = U_1(-1, a_1^{(1)}, a_2^{(1)})$. □

Here, we highlight that for the result above negative values of the $\lambda$s may be needed. For example, in a game where agent 2 has only one option and agent 1 has two options that result in the payoff vectors $(1, 1), (0, 0)$ with non-negative values of $\lambda_1$, the second option would never be an equilibrium in the SEA model.

## 3.1 Fairness and the Function $\Lambda$

Next, we compare fairness as defined by Rabin [24] and SEA Nash equilibria, cf. [3]. We define the function $\Lambda : (-1,1) \to (-\infty,\infty), x \mapsto \Lambda(x) = \frac{x}{1-|x|}$. This function is well-defined on $(-1,1)$ and a compostion of $\mathcal{C}^0$ functions and, thus, a $\mathcal{C}^0$ function, too. On $(-1,0)$ it holds $\Lambda(x) = \frac{x}{1+x} < 0$, which is a compostion of $\mathcal{C}^\infty$ functions and, thus, a $\mathcal{C}^\infty$ function, too. There, $\Lambda'(x) = \frac{1}{(1+x)^2} > 0$ and $\lim_{x\to -1} \Lambda(x) = -\infty$ hold true. On $(0,1)$ it is $\Lambda(x) = \frac{x}{1-x} > 0$. This is a compostion of $\mathcal{C}^\infty$ functions and, hence, also a $\mathcal{C}^\infty$ function. There, $\Lambda'(x) = \frac{1}{(1-x)^2} > 0$ and $\lim_{x\to 1} \Lambda(x) = +\infty$ hold. Taking this and $\Lambda(0) = 0$ together, we conclude that $\Lambda$ is strictly monotonously increasing on $(-1,1)$ and, thus, invertible. We note that $\Lambda$ is an odd function: $\Lambda(-x) = -\Lambda(x)$. The inverse is:

$$\Lambda^{-1}(y) = \frac{y}{1+|y|}, \quad \text{for } y \in \mathbb{R}$$

Note: $\Lambda(\Lambda^{-1}(y)) = \frac{\Lambda^{-1}(y)}{1-|\Lambda^{-1}(y)|} = \frac{\frac{y}{1+|y|}}{1-\left|\frac{y}{1+|y|}\right|} = \frac{\frac{y}{1+|y|}}{\frac{1+|y|-|y|}{1+|y|}} = y$ and $\Lambda^{-1}(\Lambda(x)) = \frac{\Lambda(x)}{1+|\Lambda(x)|} = \frac{\frac{x}{1-|x|}}{1+\left|\frac{x}{1-|x|}\right|} = \frac{\frac{x}{1-|x|}}{\frac{1-|x|+|x|}{1-|x|}} = x$ for all $x \in (-1,1)$ and $y \in \mathbb{R}$.

Additionally, although not needed here, one calculates:

$$\lim_{x\to 0-0} \Lambda'(x) = 1 = \lim_{x\to 0+0} \Lambda'(x)$$

Hence, $\Lambda$ is a $\mathcal{C}^1$ function with $\Lambda'(0) = 1$. Taking the formulae from above together, it holds $\Lambda'(x) = \frac{1}{(1-|x|)^2}$ for all $x \in (-1,1)$. Further, we calculate for $-1 < x < 0$ $\Lambda''(x) = \frac{-2}{(1+x)^3} < 0$ and for $0 < x < 1$ $\Lambda''(x) = \frac{2}{(1-x)^3} > 0$. Hence, $\lim_{x\to 0-0} \Lambda''(x) = -2 \neq 2 = \lim_{x\to 0+0} \Lambda'(x)$ and, thus, $\Lambda \notin \mathcal{C}^2$. On $(-1,1) \setminus \{0\}$ we can write $\Lambda''(x) = \frac{2sgn(x)}{(1-|x|)^3}$. On $(-1,0)$, $\Lambda$ is right-curved and on $(0,1)$ it is left-curved.

**Proposition 6.** *Every fairness equilibrium, i.e. every pure-strategy outcome for which a $\chi > 0$ exists s.t. the outcome is a fairness equilibrium, is SEA plausible.*

*Proof.* On the one hand, an outcome $(a_1, a_2)$ is a fairness equilibrium for some $\chi > 0$—consider again [24]—if

$$\chi u_1(a_1,a_2)+\tilde{f}_2(a_2,a_1)\left(1+\frac{u_2(a_2,a_1)-u_2^e(a_2)}{u_2^h(a_2)-u_2^{min}(a_2)}\mathbf{1}_{u_2^h(a_2)-u_2^{min}(a_2)\neq 0}\right)$$
$$\geq \chi u_1(a_1',a_2)+\tilde{f}_2(a_2,a_1)\left(1+\frac{u_2(a_2,a_1')-u_2^e(a_2)}{u_2^h(a_2)-u_2^{min}(a_2)}\mathbf{1}_{u_2^h(a_2)-u_2^{min}(a_2)\neq 0}\right)$$
$$\forall a_1' \in A_1$$

and

$$\chi u_2(a_2,a_1)+\tilde{f}_1(a_1,a_2)\left(1+\frac{u_1(a_1,a_2)-u_1^e(a_1)}{u_1^h(a_1)-u_1^{min}(a_1)}\mathbf{1}_{u_1^h(a_1)-u_1^{min}(a_1)\neq 0}\right)$$
$$\geq \chi u_2(a_2',a_1)+\tilde{f}_1(a_2,a_1)\left(1+\frac{u_1(a_1,a_2')-u_1^e(a_1)}{u_1^h(a_1)-u_1^{min}(a_1)}\mathbf{1}_{u_1^h(a_1)-u_1^{min}(a_1)\neq 0}\right)$$
$$\forall a_2' \in A_2.$$

Please note that—with a small abuse of the notation—we assume that if a nominator is zero, the indicator function is evaluated first, causing the fraction to vanish. On the other hand, an outcome $(a_1, a_2)$ is SEA plausible if for some $\lambda_1, \lambda_2 \in [-1, 1]$

$$\begin{aligned}(1-|\lambda_1|)u_1(a_1,a_2)+\lambda_1 u_2(a_2,a_1) &\geq (1-|\lambda_1|)u_1(a_1',a_2)+\lambda_1 u_2(a_2,a_1') \\ &\forall a_1' \in A_1 \\ \Leftrightarrow \quad (1-|\lambda_1|)(u_1(a_1,a_2)-u_1(a_1',a_2)) &\geq \lambda_1(u_2(a_2,a_1')-u_2(a_2,a_1)) \\ &\forall a_1' \in A_1\end{aligned}$$

and

$$\begin{aligned}(1-|\lambda_2|)u_2(a_2,a_1)+\lambda_2 u_1(a_1,a_2) &\geq (1-|\lambda_2|)u_2(a_2',a_1)+\lambda_2 u_1(a_1,a_2') \\ &\forall a_2' \in A_2 \\ \Leftrightarrow \quad (1-|\lambda_2|)(u_2(a_2,a_1)-u_2(a_2',a_1)) &\geq \lambda_2(u_1(a_1,a_2')-u_1(a_1,a_2)) \\ &\forall a_2' \in A_2\end{aligned}$$

We rewrite the first of the two fairness inequalities.

$$
\begin{aligned}
&\chi(u_1(a_1,a_2)-u_1(a_1',a_2))\\
&\geq \tilde{f}_2(a_2,a_1)\Bigg(1+\frac{u_2(a_2,a_1')-u_2^e(a_2)}{u_2^h(a_2)-u_2^{min}(a_2)}\mathbf{1}_{u_2^h(a_2)-u_2^{min}(a_2)\neq 0}\\
&-1-\frac{u_2(a_2,a_1)-u_2^e(a_2)}{u_2^h(a_2)-u_2^{min}(a_2)}\mathbf{1}_{u_2^h(a_2)-u_2^{min}(a_2)\neq 0}\Bigg)\\
&=(u_2(a_2,a_1')-u_2(a_2,a_1))\cdot\frac{\tilde{f}_2(a_2,a_1)}{u_2^h(a_2)-u_2^{min}(a_2)}\mathbf{1}_{u_2^h(a_2)-u_2^{min}(a_2)\neq 0}\ \forall a_1'\in A_1
\end{aligned}
$$

And analogously the second one.

$$
\begin{aligned}
&\chi(u_2(a_2,a_1)-u_2(a_2',a_1))\\
&\geq (u_1(a_1,a_2')-u_1(a_1,a_2))\cdot\frac{\tilde{f}_1(a_1,a_2)}{u_1^h(a_1)-u_1^{min}(a_1)}\mathbf{1}_{u_1^h(a_1)-u_1^{min}(a_1)\neq 0}\ \forall a_2'\in A_2
\end{aligned}
$$

We set

$$\tilde{F}_2(a_2,a_1,\chi)=\frac{\tilde{f}_2(a_2,a_1)}{u_2^h(a_2)-u_2^{min}(a_2)}\mathbf{1}_{u_2^h(a_2)-u_2^{min}(a_2)\neq 0}\cdot\chi^{-1}\in\mathbb{R}$$

and

$$\tilde{F}_1(a_1,a_2,\chi)=\frac{\tilde{f}_1(a_1,a_2)}{u_1^h(a_1)-u_1^{min}(a_1)}\mathbf{1}_{u_1^h(a_1)-u_1^{min}(a_1)\neq 0}\cdot\chi^{-1}\in\mathbb{R}.$$

If $(a_1,a_2)$ is for a specific $\chi$ a fairness equilibrium, it is a SEA Nash equilibrium for $(\lambda_1,\lambda_2)=(\Lambda^{-1}(\tilde{F}_2(a_2,a_1,\chi)),\ \Lambda^{-1}(\tilde{F}_1(a_1,a_2,\chi)))\in(-1,1)^2\subset[-1,1]^2$.

□

The converse is obviously—when having a look at various examples—not true. However, from the proof of Proposition 6, we learn that if $(a_1,a_2)$ is a SEA Nash equilibrium for $\mathbf{E}\subset(-1,1)^2$, it is fair if and only if one can find $(\lambda_1,\lambda_2)\in\mathbf{E}$ and $\chi>0$ s.t.

$$\Lambda(\lambda_1)=\tilde{F}_2(a_2,a_1,\chi)\text{ and }\Lambda(\lambda_2)=\tilde{F}_1(a_1,a_2,\chi).$$

Maybe, in the future more structure can be found for analyses of the question which SEA plausible outcomes are fair, e.g., for symmetric games. Also the relationship between being SEA plausible under social (i.e. $\lambda_1, \lambda_2 > 0$) or anti-social (i.e. $\lambda_1, \lambda_2 < 0$) types and positive (i.e. $f_1, f_2 > 0$) and negative (i.e. $f_1, f_2 < 0$) fairness equilibria (see [3] and [24]) shall be investigated. We highlight again that in general not every outcome is plausible under the SEA model and, even if so, not for all values of $\lambda_1, \lambda_2$.

# 4 Connections to other Fairness Models

At this point, connections and similarities to other fairness models will be shown. These models and their differences to the SEA model are especially interesting when interpreting results.

## 4.1 The Fehr-Schmidt Model

We discuss an interesting connection to another model, which also deals with fairness and which is structurally not too different to our SEA model: The Fehr-Schmidt model [13]. This model also uses a respecification of the payoffs and does not deal with believed kindness as defined by Rabin [24]. To distinguish the utility function in the Fehr-Schmidt model from that in our SEA model, we use $\hat{U}_i$ for agent $i$'s respecified utility in the Fehr-Schmidt model. The idea behind the model of Fehr and Schmidt is that agents do not only account for their personal payoff, but also for how fair the distribution of the total payoff among the agents is. In the two agent case, this is done via:

$$\begin{aligned}\hat{U}_i(a_i, a_{-i}) = u_i(a_i, a_{-i}) \\ - \alpha_i(u_{-i}(a_{-i}, a_i) - u_i(a_i, a_{-i}))^+ \\ - \beta_i(u_i(a_i, a_{-i}) - u_{-i}(a_{-i}, a_i))^+,\end{aligned}$$

which is Equation (2) from [13] just in our notation. We use the definition $(\cdot)^+ := \max\{\cdot, 0\}$. The $\alpha$-term captures the idea that agents are averse against when the opponent gets more than the agent him- or herself. The $\beta$-term is an aversion against an unfair outcome in the sense that the agent gets more than the opponent. It is $\alpha_i \geq \beta_i$ assumed, which means that unfairness against oneself is not considered less important than unfairness against the

opponent. And finally, $0 \leq \beta_i < 1$, which leads to $\alpha_i \geq 0$. Values $\beta_i \geq 1$ are argued to be "implausible" [13].[6]

Values $\beta_i < 0$ are considered to be realistic, but "have virtually no impact on equilibrium behavior" [13]. However, the case of $\beta_i < 0$ or—in general—$a_i, b_i \in \mathbb{R}$ will be important for the understanding of the difference between the Fehr-Schmidt model and the SEA model. A negative $\beta$ would mean that someone wants to have more than her or his opponent, a negative $\alpha$ that he or she wants to have less than his or her opponent. The question is whether such preferences are similar to altruistic or sadistic preferences as expressed in the SEA model?

In order to see the structural similarity of the Fehr-Schmidt model and the SEA model, we rewrite the Fehr-Schmidt model:

$$\hat{U}_i(a_i, a_{-i}) = \begin{cases} (1+\alpha_i)u_i(a_i, a_{-i}) - \alpha_i u_{-i}(a_{-i}, a_i), & u_{-i}(a_{-i}, a_i) \geq u_i(a_i, a_{-i}), \\ (1-\beta_i)u_i(a_i, a_{-i}) + \beta_i u_{-i}(a_{-i}, a_i), & u_{-i}(a_{-i}, a_i) \leq u_i(a_i, a_{-i}). \end{cases}$$

For a two-agents game in the Fehr-Schmidt model, the material game is enlarged by four parameters, namely $\alpha_1, \beta_1, \alpha_2, \beta_2$. We may call an outcome Fehr-Schmidt plausible, if and only if their exist $\alpha_1, \beta_1, \alpha_2, \beta_2 \in \mathbb{R}$ such that the outcome is a Nash equilibrium in the respecified game using Fehr and Schmidts' formulae for $\hat{U}_i$ (although, in [13] they assume $\beta < 1$ and argue that $\beta \geq 0$ is not a hard assumption from which it follows $\alpha \geq 0$).

Since the SEA model

$$U_i(a_i, a_{-i}) = (1 - |\lambda_i|)u_i(a_i, a_{-i}) + \lambda_i u_{-i}(a_{-i}, a_i)$$

is described by two parameters $\lambda_1, \lambda_2$, we cannot hope for all Fehr-Schmidt plausible outcomes to be SEA plausible. However, also the reverse is not true, as we will see. But for social types, it is true.

**Proposition 7.** *Every outcome (in a two-agents game) that is SEA plausible under social types (i.e. $\lambda_i \geq 0$) is Fehr-Schmidt plausible.*

*Proof.* Setting $-\alpha_i := \lambda_i =: \beta_i$ makes $\hat{U}_i$ and $U_i$ the same since $\lambda_i = |\lambda_i|$ for $i = 1, 2$. □

[6] In the recent work [12], also negative values and spiteful resp. competitive preferences are analyzed.

Table 7: Exemple for Fehr-Schmidt

| $u_1(\cdot); u_2(\cdot)$ | $a_2^{(1)}$ |
|---|---|
| $a_1^{(1)}$ | $0; 2$ |
| $a_1^{(2)}$ | $2; 4$ |

In this case, $\alpha_i \leq 0$ and $\beta_i \geq 0$ hold, which does not fit to the original Fehr-Schmidt model since there, agents are more concerned about unfairness when they get less than the opponent than about unfairness when they get more. However, here, we allow for real $\alpha$ and $\beta$ in any order, that means, an agent in a SEA plausible outcome under social types get some fairness payoff from the $\alpha$ term when getting less than the opponent ($\alpha_i \leq 0$) and does not get one from the $\beta$ term if she or he is getting more than the opponent ($\beta_i \geq 0$). That $|\alpha_i| = \beta_i$ accounts for the fact that the SEA model comes with two parameters while the Fehr-Schmidt model uses four parameters, i.e., the SEA model does not distinguish who gets more.

However, if at least one $\lambda_i$, let us say w.l.o.g. $\lambda_1$, is negative $\alpha_1$ would have to be positive and $\beta_1$ negative to match the return from $u_{-i}$. Both would increase the revenue from $u_i$, but since $-|\lambda_i| < 0$, this does not fit to the SEA model. In order to see that this is not only a difference in the representing values but also in behavior, let us have a look at the following example, see Table 7. To make the analysis more convenient, we use an example where in all cases $u_1 < u_2$ holds, such that we do not need $\alpha_1$. Additionally, agent 2 has only one strategy, thus we need no $\alpha_2, \beta_2$, too.

Since the distance between agent 1's and agent 2's payoffs is always two, according to the Fehr-Schmidt model, agent 1 would always (for all values of $\alpha_1, \beta_1, \alpha_2, \beta_2 \in \mathbb{R}$) prefer option $a_1^{(2)}$ over $a_1^{(1)}$. For the SEA model we calculate:

$$(1 - |\lambda_1|) \cdot 0 + \lambda_1 \cdot 2 \overset{?}{>} (1 - |\lambda_1|) \cdot 2 + \lambda_1 \cdot 4$$
$$\lambda_1 < -\frac{1}{2}$$

Thus, in the SEA model, agent 1 prefers option $a_1^{(2)}$ over $a_1^{(1)}$ if and only if $\lambda_1 \in [-\frac{1}{2}, 1]$ and strictly if and only if $\lambda_1 \in (-\frac{1}{2}, 1]$. Hence, if agent 1 is 'sadistic enough,' her or she would choose option $a_1^{(1)}$. Since the Fehr-Schmidt

Table 8: Exemple 1 for Bolton-Ockenfels

| $u_1(\cdot); u_2(\cdot)$ | $a_2^{(1)}$ |
|---|---|
| $a_1^{(1)}$ | $2; 0$ |
| $a_1^{(2)}$ | $2; 3$ |

model cannot mimic this behavior, but can distinguish the cases where one agents gets more or less than the other, the models are not similar in general in behavior.[7]

## 4.2 The ERC Model from Bolton and Ockenfels

Lastly, we mention that there is another famous fairness model, the ERC model by Bolton and Ockenfels [6]. There, utilities are respecified via functions that depend (for agent $i$ in the two-agents case) on $u_i(a_i, a_{-i})$, $c := u_i(a_i, a_{-i}) + u_{-i}(a_{-i}, a_i)$, $\sigma_i := \mathbb{I}_{c\neq 0} u_i(a_i, a_{-i}) c^{-1} + \mathbb{I}_{c=0} 2^{-1}$, and $n = 2$. Several assumptions on this respecified utility have to be fulfilled. For our paper, Equation (2) of [6] is interesting:

$$\check{U}(a_i, a_{-i}) = d_i u_i(a_i, a_{-i}) - e_i(\sigma_i - 2^{-1})^2 2^{-1}$$

with $d_i \geq 0, e_i > 0$—where we altered the notation. We highlight that Bolton and Ockenfels [6] use a model where parameters named $r, s$, which depend i.a. on $d_i, e_i$, are private information, i.e. a model of incomplete information. As mentioned in the work at hand, also the SEA model shall be extended to a model of incomplete information in future work. However, for now, to compare the Bolton-Ockenfels model and the SEA model, we define an outcome as Bolton-Ockenfels plausible, if it is a Nash equilibrium of the Bolton-Ockenfels respecified model where $d_i, e_i$ are common knowledge.

When we have a close look on Tables 8 and 9, we observe—due to the facts that agent 2 has no decision option, agent 1 gets always two utility

[7]The author is grateful to Fabian Herweg, who suggested at an internal "Graduate Seminar in Economics" (University of Bayreuth, Germany; July 4th-5th, 2024) to investigate the relationship between negative $\beta$s in the Fehr-Schmidt model and negative $\lambda$s in the SEA model.

Table 9: Exemple 2 for Bolton-Ockenfels

| $u_1(\cdot); u_2(\cdot)$ | $a_2^{(1)}$ |
|---|---|
| $a_1^{(1)}$ | $2; 1$ |
| $a_1^{(2)}$ | $2; 4$ |

units, and there is no outcome where $c = 0$—we do not have to analyze agent 2's payoffs and we can rewrite

$$\begin{aligned}\check{U}_1(a_i, a_{-i}) &= -0.5e_i \left( \frac{2}{2 + u_2(a_{-i}, a_i)} - \frac{1}{2} \right)^2 + const \\ &= -0.125e_i \left( \frac{2 - u_2(a_{-i}, a_i)}{2 + u_2(a_{-i}, a_i)} \right)^2 + const\end{aligned}$$

That means, even if we allow for $d_i, e_i \in \mathbb{R}$, which is not like in [6], the preferences expressed through $U_i$ and $\check{U}_i$ do not correspond: while an altruistic agent would in both examples (Tables 8 and 9) prefer the second option and a sadistic one the first one, an agent with a positive $e_1$ would in Table 8 prefer the second option and an agent with a negative $e_i$ would have preferences the other way around; though, in Table 9, both such agents (i.e. with positive or negative $e_i$) would be indifferent.

# 5 Example: Prisoner's Dilemma

As explained in the introduction Section 1, we do model neither altruism nor sadism in a reciprocal manner, but as intrinsic motivations. This is done by means of respecifying the material payoffs $u_i(s)$ $(i = 1, 2)$ into psychological payoffs with parameters $\lambda_i \in [-1, 1]$.

If $\lambda_i$ is zero, agent $i$ is a (pure) egoist, who is not directly interested or affected by agent $-i$'s payoffs. However, of course, agent $i$ is affected indirectly due to the game by them. Sadism and altruism are supposed to be opposite. We do not allow for values of $|\lambda_i| > 1$ since this would alter the game too much to hold as a reasonable explanation for behavior of the material game, i.e. it would lead to utilities 'outside the box' of the

material game (we do not use the wording 'convex hull' because also values $\lambda_i \in [-1, 0)$ are outside that hull). We stick to common knowledge and to rational agents, although the agents do not necessarily consider own material payoffs. The agents know whether the opponent is altruistic or sadistic and to what degree.

As mentioned in Section 1, the analytical results for the work at hand are stated mainly for pure strategies, i.e. actions. Also the various examples given in [3] are given are analyzed for those actions only. Here, we demonstrate the complexity of the SEA model when allowing for mixed, i.e. randomized strategies. For that, we investigate the prisoner's dilemma from Table 10 again. From Proposition 5 we know that all pure-strategy outcomes are SEA plausible. In [3], it is calculated for which parameters $\lambda_1, \lambda_2$ which pure-strategy outcome is SEA Nash. Next, we show for all parameter combinations $\lambda_1, \lambda_2$ which mixed-strategy outcomes (which clearly include the pure ones) are SEA Nash.

Table 10: Prisoner's Dilemma [27]

| $u_1(\cdot); u_2(\cdot)$ | $a_2^{(1)}$ | $a_2^{(2)}$ |
|---|---|---|
| $a_1^{(1)}$ | $3; 3$ | $0; 5$ |
| $a_1^{(2)}$ | $5; 0$ | $1; 1$ |

If $\lambda_i \geq 0$, $i = 1, 2$, this leads to the psychological payoffs shown in Table 11, when both empathy parameters are negative, the payoffs can be found in Table 12, and when, let's say, the first is positive and the second is negative, the psychological payoffs are given in Table 13—the other combination is analogous, see Table 14.

Table 11: Prisoner's Dilemma: Psychological Payoffs when $\lambda_i \geq 0$, $i = 1, 2$

| $U_1(\lambda_1; \cdot); U_2(\lambda_2; \cdot)$ | $a_2^{(1)}$ | $a_2^{(2)}$ |
|---|---|---|
| $a_1^{(1)}$ | $3; 3$ | $5\lambda_1; -5\lambda_2 + 5$ |
| $a_1^{(2)}$ | $-5\lambda_1 + 5; 5\lambda_2$ | $1; 1$ |

Table 12: Prisoner's Dilemma: Psychological Payoffs when $\lambda_i < 0$, $i = 1, 2$

| $U_1(\lambda_1;\cdot); U_2(\lambda_2;\cdot)$ | $a_2^{(1)}$ | $a_2^{(2)}$ |
|---|---|---|
| $a_1^{(1)}$ | $6\lambda_1 + 3; 6\lambda_2 + 3$ | $5\lambda_1; 5\lambda_2 + 5$ |
| $a_1^{(2)}$ | $5\lambda_1 + 5; 5\lambda_2$ | $2\lambda_1 + 1; 2\lambda_2 + 1$ |

Table 13: Prisoner's Dilemma: Psychological Payoffs when $\lambda_1 \geq 0$, $\lambda_2 < 0$

| $U_1(\lambda_1;\cdot); U_2(\lambda_2;\cdot)$ | $a_2^{(1)}$ | $a_2^{(2)}$ |
|---|---|---|
| $a_1^{(1)}$ | $3; 6\lambda_2 + 3$ | $5\lambda_1; 5\lambda_2 + 5$ |
| $a_1^{(2)}$ | $-5\lambda_1 + 5; 5\lambda_2$ | $1; 2\lambda_2 + 1$ |

Via straight forward, albeit lengthy, computations, one may find all Nash equilibria using standard game theory ([22], cf. [21]). This leads to Table 15.

We use the following notation: $\pi_i$ is the probability for agent $i = 1, 2$ to play $a_i^{(1)}$, hence, $1 - \pi_i$ is the probability of agent $i = 1, 2$ for playing $a_i^{(2)}$. We note for the purpose of interpretation that, e.g., for the outcomes $(0, 1)$, i.e., that one agent stays silent while the other cooperates with the police, one of the agents (with values from [27], cf. [3]) wants to help his or her opponent, while the opponent wants to hurt the agent or at least does not want to help the agent too much. Such a behavior is not fair in the sense of Rabin [2, 24].

Table 14: Prisoner's Dilemma: Psychological Payoffs when $\lambda_1 < 0$, $\lambda_2 \geq 0$

| $U_1(\lambda_1;\cdot); U_2(\lambda_2;\cdot)$ | $a_2^{(1)}$ | $a_2^{(2)}$ |
|---|---|---|
| $a_1^{(1)}$ | $6\lambda_1 + 3; 3$ | $5\lambda_1; -5\lambda_2 + 5$ |
| $a_1^{(2)}$ | $5\lambda_1 + 5; 5\lambda_2$ | $2\lambda_1 + 1; 1$ |

Table 15: The Nash Equilibria of the Empathic Prisoner's Dilemma with $\lambda_i \in [-1, 1], i = 1, 2$

| Nash-Eq. | $\lambda_2 \in (0.4, 1]$ | $\lambda_2 = 0.4$ | $\lambda_2 \in (0.2, 0.4)$ | $\lambda_2 = 0.2$ | $\lambda_2 \in [-1, 0.2)$ |
|---|---|---|---|---|---|
| $\lambda_1 \in (0.4, 1]$ | $\{(1, 1)\}$ | $\{1\} \times [0, 1]$ | $\{(1, 0)\}$ | $\{(1, 0)\}$ | $\{(1, 0)\}$ |
| $\lambda_1 = 0.4$ | $[0, 1] \times \{1\}$ | $([0, 1] \times \{1\})$ $\cup (\{1\} \times [0, 1])$ | $\{(0, 1), (1, 0),$ $(5\lambda_2 - 1, 1)\}$ | $\{(0, 1), (1, 0)\}$ | $\{(1, 0)\}$ |
| $\lambda_1 \in (0.2, 0.4)$ | $\{(0, 1)\}$ | $\{(0, 1), (1, 0),$ $(1, 5\lambda_1 - 1)\}$ | $\{(1, 0), (0, 1),$ $(5\lambda_2 - 1, 5\lambda_1 - 1)\}$ | $\{(0, 1), (1, 0),$ $(0, 5\lambda_1 - 1)\}$ | $\{(1, 0)\}$ |
| $\lambda_1 = 0.2$ | $\{(0, 1)\}$ | $\{(0, 1), (1, 0)\}$ | $\{(0, 1), (1, 0),$ $(5\lambda_2 - 1, 0)\}$ | $([0, 1] \times \{0\})$ $\cup (\{0\} \times [0, 1])$ | $[0, 1] \times \{0\}$ |
| $\lambda_1 \in [-1, 0.2)$ | $\{(0, 1)\}$ | $\{(0, 1)\}$ | $\{(0, 1)\}$ | $\{0\} \times [0, 1]$ | $\{(0, 0)\}$ |

# 6 Conclusion and Future Work

We analyzed the SEA model from [3] and showed that for small games, i.e., for two-agent games with at most two actions per agent, all pure-strategy outcomes are SEA plausible, but not for all parameter combinations, which gives insights into the structure of those games. We showed that all fair outcomes are SEA plausible (see [24]) and we compared the SEA model to the models of Bolton-Ockenfels and Fehr-Schmidt. Additional, mixed-strategy outcomes are analyzed for the example Prisoner's dilemma.

There are various ways for future work concerning the SEA model or related concepts: The most important step will be to implement the SEA model in a framework with incomplete information, i.e., when agents do not know the parameter of the opponent (but can have some belief about it). This way it would be interesting to check whether the two-sided anti-social punishment (cf. [23]) can be mimicked. Note that in principle an intrinsically (partially) altruistic or (partially) sadistic agent does not care about the type of her or his opponent. Whereas for equilibrium behavior this does matter since an outcome can be an equilibrium if one agent is sadistic and the other one is altruistic but may not be one if both agents are, let us say, sadistic. If both agents are sadistic but do not know the type of the respective opponent and both believe the opponent is altruistic, they might both choose the option from the described equilibrium. Challenging, however, may be the Dictator und the Ultimatum game [3, 17, 18] as well as [4, 28] and the references therein.

Another important point is the check for evolutionary stability. Can altruists or sadists or both survive? Related to that is the question what happens if types are not described by one parameter anymore but also via

distributions over his or her descendants. While this topic is likely to be analyzed via simulations at first, maybe also analytical results are possible in the not too near future. Next, also other types of "personality" not only altruistic or sadistic behavior but also other preferences that have influence on the choice of options, on the fitness, or on survival chances can be analyzed that way. Lastly, risk and the possibly limited ability of agents to see through the structure of games and behavior is important—agents might simply not be able to calculate equilibria or this is too expensive for them or they simplify games and play like in the simplified game.[8] Not to mention the question what happens if agents care about the psychological (i.e. the respecified) payoff of the opponent and not (only) about the material one.



# Acknowledgment

The author of the work at hand thanks Michaela Baumann, Melanie Birke, Lars Grüne, and Bernhard Herz.

# References

[1] Andreoni, James, William T. Harbaugh, Lise Vesterlund: Altruism in Experiments. *The New Palgrave Dictionary of Economics,* 1-7, Palgrave Macmillan, London, 2016.

[2] Baumann, Michael H., Michaela Baumann: Some Thoughts on Rabin Fairness. *Universität Bayreuth,* Discussion Paper, May 2025. `https://doi.org/10.15495/EPub_UBT_00008439`

[3] Baumann, Michael H., Michaela Baumann: When Reciprocity is not Enough : Explaining Anti-Social Outcomes via Intrinsic Personality Traits. *Universität Bayreuth,* Discussion Paper, June 2025. `https://doi.org/10.15495/EPub_UBT_00008501`

[4] Camerer, Colin, Richard H. Thaler: Anomalies: Ultimatums, Dictators and Manners. *Journal of Economic Perspectives,* **9**(2): 209-219, Spring 1995.

[8] Confer, e.g., [15, 27, 31, 30, 32, 33, 36, 35, 34].

[5] Bester, Helmut, Werner Güth: Is Altruism Evolutionarily Stable? *Journal of Economic Behavior & Organization,* **34**(2):193-209, 1998.

[6] Bolton, Gary E., Axel Ockenfels: ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review,* **91**(1):166-193, 2000.

[7] Colman, Andrew M., J. Clare Wilson: Antisocial Personality Disorder: An Evolutionary Game Theory Analysis. *Legal and Criminological Psychology,* **2**:23-34, 1997.

[8] Cooper, Russell, Douglas V. DeJong, Robert Forsythe, Thomas W. Ross: Cooperation Without Reputation: Experimental Evidence From Prisoner's Dilemma Games. *Games and Economic Behavior,* **12**(2): 187-218, 1996.

[9] Dawes, Robyn Mason: Social Dilemmas. *Annual Review of Psychology,* **31**(1): 169-193, 1980.

[10] Dawes, Robyn Mason, Richard H. Thaler: Anomalies: Cooperation. *Journal of Economic Perspectives,* **2**(3): 187-197, 1988.

[11] Edgeworth, Francis Ysidro: *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences,* Kegan Paul, London, 1881.

[12] Fehr, Ernst, Gary Charness: Social Preferences: Fundamental Characteristics and Economic Consequences. *Journal of Economic Literature,* **63**(2):440-514, 2025.

[13] Fehr, Ernst, Klaus M. Schmidt: A Theory Of Fairness, Competition And Cooperation. *The Quarterly Journal of Economics,* **114**(3):817-868, August 1999.

[14] Fehr, Ernst, Klaus M. Schmidt: The Economics of Fairness, Reciprocity and Altruism – Experimental Evidence and New Theories. *In: Serge-Christophe Kolm, Jean Mercier Ythier (Eds.), Handbook of the Economics of Giving, Altruism and Reciprocity,* **1**:615-691, 2006.

[15] Föllmer, Hans, Alexander Schied: *Stochastic Finance—An Introduction in Discrete Time,* 3rd Edition, De Gruyter Graduate, Berlin/New York, 2011.

[16] Geanakoplos, John, David Pearce, Ennio Stacchetti: Psychological Games and Sequential Rationality. *Games and Economic Behavior,* **1**: 60-79, March 1989.

[17] Guala, Francesco, Luigi Mittone: Paradigmatic Experiments: The Dictator Game. *The Journal of Socio-Economics,* **39**(5):578-584, October 2010.

[18] Güth, Werner, Rolf Schmittberger, Bernd Schwarze: An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior & Organization,* **3**(4):367–388, 1982.

[19] Levine, David K.: Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics,* **1**(3):593-622, 1998.

[20] Meurer, Aaron, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, Anthony Scopatz: SymPy: Symbolic Computing in Python. *Python, Computer algebra system, Symbolics,* **3**:e103, January 2017.

[21] Myerson, Roger B.: *Game Theory : Analysis of Conflict,* Harvard University Press, Cambridge, Massachusetts/London, England, 1991.

[22] Nash, John Forbes Jr.: Non-Cooperative Games. *Dissertation,* Princeton University, 1950.

[23] Pfattheicher, Stefan, Johannes Keller, Goran Knezevic: Sadism, the Intuitive System, and Antisocial Punishment in the Public Goods Game. *Personality and Social Psychology Bulletin,* **43**(3):337-346, 2017.

[24] Rabin, Matthew: Incorporating Fairness into Game Theory and Economics. *The American Economic Review,* **83**(5): 1281-1302, December 1993.

[25] Rand, David G., Martin A. Nowak: The Evolution of Antisocial Punishment in Optional Public Goods Games. *Nature Communications,* **2**(434):1-7, August 2011.

[26] Roberts, Gilbert: When Punishment Pays. *PLOS ONE,* **8**(3)e57378: 1-8, March 2013.

[27] Sieg, Gernot: *Spieltheorie,* 2nd Edition, R. Oldenburg Verlag, München/Wien, 2005. *(in German)*

[28] Thaler, Richard H.: Anomalies: The Ultimatum Game. *Journal of Economic Perspectives,* **2**(4): 195-207, Fall 1988.

[29] van Rossum, Guido, Fred L. Drake Jr.: *Python Tutorial.* Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995

[30] von Neumann, John: Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. *Ergebnisse eines Mathematischen Kolloquiums,* **8**(1935/36), Leipzig, 1937. *(in German)*

[31] von Neumann, John: Zur Theorie der Gesellschaftsspiele, *Mathematische Annalen,* 100:295-320, December 1928. *(in German)*

[32] von Neumann, John, Oskar Morgenstern: *Theory of Games and Economic Behaviour,* Princeton University Press, 1944.

[33] Wald, Abraham: Contributions to the Theory of Statistical Estimation and Testing Hypotheses. *The Annals of Mathematics,* **10**(4):299-326, 1939.

[34] Wald, Abraham: Generalization of a Theorem of von Neumann Concerning Zero Sum Two Person Games. *The Annals of Mathematics,* **46**(2):281-286, 1945.

[35] Wald, Abraham: *Statistical Decision Functions,* John Wiley, New York, 1950.

[36] Wald, Abraham: Statistical Decision Functions which Minimize the Maximum Risk. *The Annals of Mathematics,* **46**(2):265-280, 1945.

[37] Wu, Jia-Jia, Bo-Yu Zhang, Zhen-Xing Zhou, Qiao-Qiao He, Xiu-Deng Zheng, Ross Cressman, and Yi Tao: Costly Punishment does not Always Increase Cooperation. *Proceedings of the National Academy of Sciences,* **106**(41):17448-17451, October 2009.

[38] Ye, Hang, Fei Tan, Mei Ding, Yongmin Jia, Yefeng Chen: Sympathy and Punishment: Evolution of Cooperation in Public Goods Game. *Journal of Artificial Societies and Social Simulation,* **14**(4)20:1-14, 2011.

[39] Zapata, Asunción , Amparo M. Mármol, Luisa Monroy, M. Ángeles Caraballo: When the Other Matters. The Battle of the Sexes Revisited. In: Patrizia Daniele, Laura Scrimali (Eds.) *New Trends in Emerging Complex Real Life Problems,* 501-509, AIRO Springer Series 1, ODS, Taormina, Italy, September 2018.