Some Thoughts on Rabin Fairness (Discussion Paper)

Michael Heinrich Baumann* and Michaela Baumann†,‡

May 9^{th}, 2025

Abstract—In this work, we reconsider Rabin fairness, published in 1993.[1] On the one hand, we explain the concept, ask and potentially discuss some (mathematical) questions, and clarify via proofs some points that were unclear. On the other hand, we show whether and how fairness equilibria can be calculated via Python/SymPy. We do not focus on economical or psychological discussions, but on mathematical ones.

Keywords—Game Theory; Fairness; Prisoner's Dilemma; Reciprocity; One-Shot Game

JEL codes—C72, D9
UDC—519.83
MSC2020—91A05, 91A10, 91A40, 91-04

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

†NÜRNBERGER Versicherung, Nuremberg, Germany,
michaela.baumann@nuernberger.de

‡Opinions expressed here are her own and not necessarily those of her employer.

[1] [19]: Rabin, Matthew: Incorporating Fairness into Game Theory and Economics. *The American Economic Review,* **83**(5): 1281-1302, December 1993.

1 Introduction

It is nothing new that outcomes in experiments as well as in the real world do not necessarily fit to predictions of classical game theory, which is based on material payoffs only. For example, in the prisoner's dilemma in Table 1,[2] where both prisoners each have the options of cooperating or defecting, the only Nash equilibrium is that both defect, both in pure and in mixed strategies.[3] Further, this is an equilibrium in strictly dominated strategies. However, in experiments[4] and real world observations it turns out that cooperation in this game is possible.[5] There are several ways how such a cooperation can be explained, however, this is often based on finitely or infinitely repeated prisoner's dilemmas. Rabin [19] presents a theory how this outcome resp. these outcomes can be explained via a concept of reciprocity and fairness.

Table 1: Prisoner's Dilemma: Material Payoffs (higher values are favorable for the agents; i.e., the numbers could be interpreted as, e.g., "5 years minus jail term." The values are taken from [21]).

$u_1(\cdot)\vert u_2(\cdot)$	$a_2^{(1)}$: cooperate	$a_2^{(2)}$: defect
$a_1^{(1)}$: cooperate	$3\vert 3$	$0\vert 5$
$a_1^{(2)}$: defect	$5\vert 0$	$1\vert 1$

Being cited more than 7,800 times according to Google Scholar (`www.scholar.google.de`; 2024-10-15) indicates the huge impact of Rabin's work "Incorporating Fairness into game Theory and Economics" and its many versions (among them the American Economic Review paper [19] and a book chapter) on the science and economics community. Rabin fairness [19] was and still is important. A lot of research has been done on showing whether and to what extent Rabin fairness can explain outcomes in experiments and

[2] One could "defect" also call "confess" and "cooperate" "stay silent."

[3] See [17] for Nash equilibrium and Section 2.2 for our notations.

[4] Confer [3], esp. Footnote 2, which refers to [4, 5] of the work at hand.

[5] Confer esp. the example "Commentary" in [5] on pp. 195f and the corresponding discussion of this example conducted in [19], which we discuss in Section 2.1 of the work at hand.

in the real world. Many ideas have been published to enhance resp. extent the concept of Rabin fairness, see, e.g., [6, 18]. There, [6] allows for sequential games. Fairness approaches—including beliefs or (relative) outcome distributions or both[6]—are highly discussed, see [1, 2, 8, 9, 10]. In contrast to [19], Charness and Rabin [2] explain that also without positive feelings people may be willing to sacrifice themselves in order to help others (see Footnote 1 in [2]). In [1], it is criticized that Rabin's fairness model 'applies to two-person, normal-form games of complete information' [1] p. 167, and that it is not clear how to generalize that. That distributions of outcomes (among the agents) are (also) important and not captured by [19] is discussed in [8] and that *intentions* are explicitly modeled (which needs psychological game theory instead of standard game theory) is criticized in [10], which leads to a similar critique like in [1]. Further notable is the work that introduces "[. . .] Personal Equilibria," see [11, 14].

In contrast, the work at hand targets neither economical improvements of Rabin fairness nor critiques that it may not fit to experiments, but the very basic concept itself. That means, we are going "back to the roots" in order to fully understand this fairness concept based on beliefs. We are esp. interested in the mathematics of these concepts. We ask questions where things are unclear and provide tools to calculate and use this concept. Some of the *comments* we mention are far-reaching to the concept of fairness itself, some are just small remarks or questions. And sometimes we comment when implications of Rabin Fairness seem in our humble opinion to be obviously questionable compared to real-world experiences.

In the following, we present, summarize, and explain important parts of the theory of Rabin [19]. Further, we point out some thoughts that are worth future discussion. We illustrate open questions as well as show and proof new results, e.g., propositions showing equivalences between pure- and mixed-strategy definitions. Additionally, we present how Rabin's so-called fairness equilibria can be calculated via set theory. We use these findings to write a Python/SymPy code [16, 23] for calculating fairness equilibria. Although these calculations work in pure strategies only, we will see in Section 3 that this is not any problem since—as we will show as one of the main points of this work—for fairness equilibria in pure strategies it does not matter whether we are optimizing over pure or mixed strategies.

[6] See, e.g., Footnote 1 of [6] and the paragraph it refers to.

2 Rabin's Fairness Equilibria

Rabin explains why cooperation is possible using a concept of beliefs and fairness. Section 2 summarizes and outlines the work of Rabin [19].

2.1 Motivation and Assumptions

The theory of Rabin is constructed in such a way that the following three so-called stylized facts are fulfilled [19]: i) Agents accept smaller material payoffs in order to help agents who behave kind. ii) Agents accept smaller material payoffs in order to hurt agents who do not behave kind. iii) When material costs for helping or hurting are (relatively) small, agents are more willing to do so. Rabin builds his concept of fairness equilibria in [19], see Sections 2.3 and 2.4, upon the work on psychological Nash equilibria by Geanakoplos, Pearce, and Stacchetti [12].

Rabin [19] cites an example from Dawes and Thaler [5] where farmers sell vegetables on a table near the road without any seller who controls whether the buyers really pay. The money box in that example is constructed and mounted in such a way that the money cannot be stolen easily. The essential part of the story when Rabin discusses this example is—up to our humble opinion—that the buyers behave kind (i.e., they pay), the farmers behave kind (they deliver the vegetables), both believe that the respective other is kind, and both believe that the respective other believes that he or she behaves kind. This example is supposed to motivate Rabin's concept of fairness equilibria, however, it is neither formalized nor calculated in [5] or [19]. After having presented the techniques to automatically calculate (some rational) fairness equilibria in arbitrary games, we show in Section 8.5 that the modeling resp. formalization of Rabin's introductory example such that it fits to the fairness equilibrium concept is not a straightforward one.

Rabin explains that it is important to note that the money box is fixed on the table since the material payoff of stealing the money would outweigh the fairness cost for hurting someone who is kind. The relatively small payoff of stealing vegetables does not outweigh fairness. However, we mention that there is no unique interpretation. For that, we note that Rabin cites from [5] that the farmers know that if the box would not be fixed, it would be stolen by *someone*. Following the chain of reasoning of Rabin [19], everyone would steal the money if the material payoff is high enough.[7] One interpretation is that buyers would not steal the money box as long as there is not enough money in the box, but buy some vegetables. However, when a buyer comes and there is enough money in the box, he or she would steal it. This implicitly assumes that everyone "has a price," i.e., that everyone would steal the money if there is enough money in the box (where "enough" depends on his or her own valuation $\chi > 0$).[8] We also formalize, model, and analyze this stealing vs. not stealing game in detail in Section 8.5.

However, there may be another interpretation, namely that there is a fixed share of unfair people in the world who steal some vegetables (but not all vegetables, since they have no usage for them and selling them inhibits the risk of getting caught) and who would also steal the money if possible. If only a small part of the people would steal (only a small part of the) vegetables, this would not ruin the farmers. However, if only one unfair buyer would steal all the money, the farmers may be ruined. This idea is important for future work: there may be fixed shares of people that are fair resp. unfair. By the way: clearly, we neglected in this example and discussion that stealing is a crime and people should be afraid of getting caught.

Comment 1. *In the farmers example Rabin [19] cites from [5], it is mentioned that stealing vegetables has such a small material payoff that it does not outweigh the negative fairness payoff originating from stealing. But if it would be possible to steal the money, the material payoff would outweigh the negative fairness payoff. Hence, someone would steal the money. However, following the theory of Rabin [19], not only someone—cf. [5]—should steal the money, but all buyers would eventually do so.*

2.2 Some Basic Wording

To enhance comprehensibility of some of the following discussions, we recapitulate some basic wording and notations. Rabin denotes with a resp. a_i both the pure strategies, i.e. action, and the mixed strategies of an agent (i),

[7] For clear, in a two-agent game with one farmer and one buyer, one might ask whether the scaling parameter that determines when the material payoff of stealing is high enough to outweigh the fairness payoff should depend on the specific buyer, i.e. on his or her personality. However, see Comment 13, this does neither change the structure of the game nor the structure of the equilibria.

[8] For agents who would never steal, one would need $\chi = 0 \not> 0$. However, than two different parameters would be necessary, see Comment 13.

with b, b_i the belief of the opponent of what agent i chooses, and with c, c_i resp. the belief of the agent of what the opponent believes of what the agent chooses. For the sake of readability, esp. in the proofs, we differ sometimes from Rabin's notations [19] (e.g., he uses j instead of $-i$ and π instead of u).

Comment 2. *Note that Rabin explains in Footnote 8 of [19] that he focuses on pure strategies in his work. The examples "battle of the sexes," "prisoner's dilemma," or "chicken" fit to that focus on pure strategies since there, only pure strategies, i.e. actions are considered. However, we note that in Rabin's [19] definition of the so-called fairness equilibria, it is optimized over mixed strategies. In his calculations for "battle of the sexes" ([19], p. 1288) he maximizes over pure strategies only. In Proposition 2, we will prove that this does not matter. Please note that the examples "monopoly pricing" and "labor economics" ([19] Section IV.) do not fit exactly to Rabin's definitions, which he himself notes in his work, cf. Comment 3.*

Comment 3. *Rabin [19] says on p. 1286 that his model is applicable to all two-person, finite-strategy games. And he defines the (mixed) strategy sets as derived from* finite *pure-strategy sets. Rabin notes that the examples in Section IV. of [19], namely monopoly pricing and labor economics, use infinite strategy sets or an infinite set (see esp. Footnote 18 of [19] and the sentence it refers to). A thorough analysis of whether and—if so—when a bounded infinite strategy set (e.g., an interval) can be interpreted as a mixed strategy (e.g., with the two poles of the interval as pure strategies) would be interesting, though.*

The pure vs. mixed strategy topic will be discussed in Section 3 in great detail. Next, we explain our notation:

With a_i we denote the action, i.e. the pure strategy, agent i chooses from a finite set of actions $A_i = \{a_i^{(1)}, \ldots, a_i^{(n_i)}\}$. With $-i$ we name the respective other agent. An agent's material payoff depends on both the own and the other's actions $u_i(a_i, a_{-i}), u_{-i}(a_{-i}, a_i)$. The variable s_i denotes the (possibly mixed) strategy of agent i, that is, as usual, s_i can be represented via a vector $p_i \in [0,1]^{n_i}$ s.t. $\sum_{j=1}^{n_i} p_i^{(j)} = 1$ of probabilities denoting the chances that agent i uses $a_i^{(j)}$. The agents' expected payoff is then $u_i(s_i, s_{-i}) = \sum_{j_i=1}^{n_i} \sum_{j_{-i}=1}^{n_{-i}} p_i^{(j_i)} p_{-i}^{(j_{-i})} u_i(a_i^{(j_i)}, a_{-i}^{(j_{-i})})$. The set of strategies of agent i is S_i. We denote with b_i agent $-i$'s belief about what agent i chooses, and with c_i agent i's (second-order) belief about the belief of agent $-i$ about what agent

i chooses. Please note that for beliefs b and second-order beliefs c also our notation does not distinguish between pure and mixed strategies—which is for our purpose not necessary since the optimizations (see further below) are conducted over the actions resp. strategies and not over any beliefs.

We call an outcome, i.e. a pair of strategies, (globally) Pareto optimal if no other outcome is Pareto superior to the optimal one, i.e., if there is no outcome where one agent gets more while the other does not get less.[9] We call an outcome Pareto optimal *in* a specific set if there is no outcome in this set which is Pareto superior to the optimal one. An outcome is a Nash equilibrium if no agent can get a higher payoff by changing his or her strategy when the other sticks to the strategy from the respective Nash equilibrium [17].

Mutual-max outcomes and mutual-min outcomes are defined by Rabin ([19], Def. 4 and 5) as follows:[10] A mutual-max outcome is an outcome (s_i, s_{-i}) s.t. $s_i \in argmax_{s\in S_i} u_{-i}(s_{-i}, s)$, $i = 1, 2$. A mutual-min outcome is an outcome (s_i, s_{-i}) s.t. $s_i \in argmin_{s\in S_i} u_{-i}(s_{-i}, s)$, $i = 1, 2$.

Comment 4. *In [19], the arguments of the payoff functions are switched in its Definitions 4 and 5. However, this does not fit to the convention in [19] that the first argument corresponds to that agent whose payoff is evaluated.*

2.3 Kindness and Beliefs

The concept of fairness equilibria Rabin uses in [19] is based upon [12]. Rabin shows that this setting cannot be projected to classical game theory ([19], last paragraph on p. 1285 ending on p. 1286). He explains this via the example "battle of the sexes," where two outcomes (with the same scaling parameter $\chi < 1$) in the same column are fairness equilibria with strict inequalities in their respective definition.[11]

Comment 5. *We mention that this example is fully valid, since it is impossible that two outcomes in the same column are both strictly preferable for the row player (in the classical Nash sense). Rabin [19] shows that it is in general impossible to map payoff values to new payoff values (i.e., to alter*

[9] See: Pareto, Vilfredo: Manuale d'economia Politica. *Societa Editrice Libraria,* Milano, 1906. *(in Italian)* Respectively: Pareto, Vilfredo: Manuel d'économie politique. *Giard et Brière,* Paris, 1909. *(in French).* See [7].

[10] As always, we write down the cited definitions, theorems, etc. in our notation.

[11] See Tables 26 and 27 and note the strict inequality "$\chi < 1$."

the payoff values; in a game of the same type and size) s.t. these new values would allow to reformulate fairness equilibria as (classical) Nash equilibria.

We add that it is in general also not easily possible to extent the game in such a way that all actions and beliefs from the original game are taken as actions of the new game with the same payoffs—in such a way that the fairness equilibria of the original game would equal the Nash equilibria of the new game. This is not easily possible since in the new game one could change beliefs, but in the definition of fairness equilibria (as we will see further below) only the actions are in the set over which it is maximized; see Table 2 and Section 2.4. However, in [13] it is shown that it is possible to change the size and *the utility structure of the game using two (lexicographically ordered) utility functions u and v s.t. psychological equilibria can be rewritten as classical Nash equilibria.*[12] *There may be another way to extent games such that fairness equilibria can be reformulated as Nash equilibria, which allows for a more convenient interpretation.*[13]

Table 2: Expansion of a 2 × 2 game with first- and second-order beliefs. Pure strategies only. With $\star$ the matching-beliefs-and-actions outcomes are marked. With the arrows the outcomes that have to be compared to the stars they are pointing at are denoted. This illustrates that fairness equilibria are not Nash equilibria in a simply transformed game—see Section 2.4.

$U_1(\cdot)\|U_2(\cdot)$		$s_2 =$	$a_2^{(1)}$	$a_2^{(1)}$	$a_2^{(1)}$	$a_2^{(1)}$	$a_2^{(2)}$	$a_2^{(2)}$	$a_2^{(2)}$	$a_2^{(2)}$
		$b_1 =$	$a_1^{(1)}$	$a_1^{(1)}$	$a_1^{(2)}$	$a_1^{(2)}$	$a_1^{(1)}$	$a_1^{(1)}$	$a_1^{(2)}$	$a_1^{(2)}$
$s_1 =$	$b_2 =$	$c_1 = \backslash c_2 =$	$a_2^{(1)}$	$a_2^{(2)}$	$a_2^{(1)}$	$a_2^{(2)}$	$a_2^{(1)}$	$a_2^{(2)}$	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	$a_2^{(1)}$	$a_1^{(1)}$	$\star$				←			
$a_1^{(1)}$	$a_2^{(1)}$	$a_1^{(2)}$			↓					
$a_1^{(1)}$	$a_2^{(2)}$	$a_1^{(1)}$		→				$\star$		
$a_1^{(1)}$	$a_2^{(2)}$	$a_1^{(2)}$								↓
$a_1^{(2)}$	$a_2^{(1)}$	$a_1^{(1)}$	↑							
$a_1^{(2)}$	$a_2^{(1)}$	$a_1^{(2)}$			$\star$				←	
$a_1^{(2)}$	$a_2^{(2)}$	$a_1^{(1)}$						↑		
$a_1^{(2)}$	$a_2^{(2)}$	$a_1^{(2)}$				→				$\star$

Next, in [19], via so-called kindness functions f_i and $\tilde{f}_{-i}$, the material

[12] An interesting topic for future research is to evaluate the connections between [13] and open questions of the work at hand.

[13] This is part of the ongoing work.

payoffs u_i are transformed to so-called expected utilities U_i. There, f_i is agent i's belief of the kindness of him- or herself towards $-i$, which depends on what i believes what $-i$ does. Further, $\tilde{f}_{-i}$ is agent i's (second-order) belief about how kind agent $-i$ is towards him or her, which depends on what i believes what $-i$ does and on what i believes that $-i$ believes what i does.

Comment 6. *We note that the term* expected utilities *could be misleading, since this is not a (stochastic) expected value of some utility that depends solely on the material payoff.*

The kindness functions are:

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

and

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

Note that s, b, c can be mixed strategies/beliefs.

Comment 7. *In [19], in the formula for $\tilde{f}_{-i}(b_{-i}, c_i)$ it is $u^e_i(c_{-i})$ instead of $u^e_i(c_i)$, which has—in our opinion—to be a typo, since Rabin writes that f and $\tilde{f}$ are formally equivalent and, furthermore, $\tilde{f}_{-i}(b_{-i}, c_i)$ does not depend on c_{-i}.*

The kindness functions incorporate the following parts:

- $u_{-i}(b_{-i}, s_i)$ is the material payoff of agent $-i$ when playing b_{-i} when agent i plays s_i
- $u^e_{-i}(b_{-i}) = \frac{u^h_{-i}(b_{-i}) - u^l_{-i}(b_{-i})}{2}$ is the so-called equitable payoff (for an interpretation, see [19] p. 1286)
- $u^h_{-i}(b_{-i})$ is the highest payoff agent $-i$ can receive in $\mathbf{u}(b_{-i})$
- $\mathbf{u}(b_{-i}) = \{(u_i(s_i, b_{-i}), u_{-i}(b_{-i}, s_i)) \mid s_i \in S_i\}$ is the set of possible outcomes when agent $-i$ plays b_{-i}
- $u^{min}_{-i}(b_{-i})$ is the lowest payoff agent $-i$ can receive in $\mathbf{u}(b_{-i})$

- $u^l_{-i}(b_{-i})$ is the lowest payoff agent $-i$ can receive within the Pareto optimal subset of $\mathbf{u}(b_{-i})$

Comment 8. *We do understand Rabin [19] in such a way that the subset mentioned in the definition of $u^l_{-i}(b_{-i})$ does not have to be globally Pareto optimal, but* in $\mathbf{u}(b_{-i})$ *no outcome shall be Pareto superior to $u^l_{-i}(b_{-i})$.*

Please note: the set of globally Pareto optimal outcomes that lie in $\mathbf{u}(b_{-i})$ can be empty.

The definitions for the other agent ($u^e_i(c_i)$ etc.) are analogous. Function f_i is a measure for the (believed) kindness of agent i towards agent $-i$ and $\tilde{f}_{-i}$ describes what agent i believes how kind he or she is treated by agent $-i$. When agent i calculates $\tilde{f}_{-i}$, i.e., his or her belief how kind $-i$ is to him or her, he or she does not use s_i since agent i is sophisticated enough to know that agent $-i$ does not have to know s_i, i.e., what agent i is actually doing, but agent i keeps in mind that he or she has to believe what agent $-i$ believes what agent i does. It holds $u^h_{-i}(b_{-i}) \geq u^e_{-i}(b_{-i}) \geq u^l_{-i}(b_{-i}) \geq u^{min}_{-i}(b_{-i})$.

As explained in [19] p. 1287, if the Pareto frontier in $\mathbf{u}(b_{-i})$ is a singleton (cf. Comment 8 of the work at hand), then $u^h_{-i} = u^l_{-i} = u^e_{-i}$—which becomes clear when drawing a set $\mathbf{u}(b_{-i})$ with a Pareto frontier which is a singleton—and, thus, $f_i(s_i, b_{-i}) \leq 0$. In general, the highest possible value f_i can have is 0.5. If $u_{-i}(b_{-i}, s_i) = u^h_{-i}(b_{-i})$, we get $f_i(s_i, b_{-i}) = 0.5 \cdot \frac{u^h_{-i}(b_{-i})-u^l_{-i}(b_{-i})}{u^h_{-i}(b_{-i})-u^{min}_{-i}(b_{-i})} \in [0, 0.5]$ because $u^l_{-i}(b_{-i}) \geq u^{min}_{-i}(b_{-i})$. This equals 0.5 if and only if $u^l_{-i}(b_{-i}) = u^{min}_{-i}(b_{-i}) \wedge u^h_{-i}(b_{-i}) \neq u^l_{-i}(b_{-i})$.

Comment 9. *In the calculations above, we see an interesting point for discussions, why Rabin distinguishes l and min and why the kindness of an agent, when he or she gives h to his or her opponent, depends on whether l and min are equal or not.*

If $u^h_{-i}(b_{-i}) \neq u^l_{-i}(b_{-i})$ and $u_{-i}(b_{-i}, s_i) = u^l_{-i}(b_{-i})$, it holds $f_i(s_i, b_{-i}) = -0.5 \cdot \frac{u^h_{-i}(b_{-i})-u^l_{-i}(b_{-i})}{u^h_{-i}(b_{-i})-u^{min}_{-i}(b_{-i})} \in [-0.5, 0)$. It equals -0.5 if and only if $u^l_{-i}(b_{-i}) = u^{min}_{-i}(b_{-i})$.

Comment 10. *It is interesting, too, that the kindness of an agent giving l takes into account if l equals min, see Comment 9.*

To reach the smallest possible value for f_i, namely minus one, $u_{-i}(b_{-i}, s_i) = u^{min}_{-i}(b_{-i}) < u^l_{-i}(b_{-i}) = u^h_{-i}(b_{-i})$ has to hold: $f_i(s_i, b_{-i}) = -\frac{u^e_{-i}(b_{-i})-u^{min}_{-i}(b_{-i})}{u^h_{-i}(b_{-i})-u^{min}_{-i}(b_{-i})}$.

Comment 11. *Again, it is worth a discussion, why the kindness of an agent when giving min depends on whether h equals l or not—see Comment 9.*

We note that Rabin [19] does not only use the kindness functions explained above but—since these functions may be discontinuous, see Sections 5.4 and 5.5—explicates much more general kindness functions (see Footnote 11 and Apendix A of [19]). However, since Rabin focuses on these "standard" kindness functions, we also stick to them.[14]

If one gives less than u^{ℓ}_{-i} to his or her opponent, the kindness value is negative, but not necessarily smaller or equal $-\frac{1}{2}$. Kindness values below $-\frac{1}{2}$ are "unreasonably" unkind, but values between $-\frac{1}{2}$ and 0 do not have to be "reasonable."

2.4 Expected Utility and Fairness Equilibria

Rabin defines the expected utility in [19] as

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

For details and interpretations see [19], especially Footnote 10, and cf. [12] ("psychological Nash equilibrium"). Note that $u_i(s_i, b_{-i})$ is—in contrast to $u_i(s_i, s_{-i})$—not what agent i receives, but what agent i believes what he or she receives. (We will see that in the definition of fairness equilibrium, beliefs have to match.)

Using this framework, Rabin defines a fairness equilibrium as $(s_1, s_2) \in S_1 \times S_2$ with

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

and $c_i = b_i = s_i$ for $i = 1, 2$ (i.e., optimization under matching beliefs and strategies/actions). We highlight that in the function to be optimized, there is c_i, which equals the equilibrium strategy s_i, but which does not necessarily equal the optimization argument s'_i. Since $\tilde{f}_{-i}$ is agent i's belief about how kind agent $-i$ is, when checking whether s'_i is in *argmax*, agent i does not change his or her belief about how kind agent $-i$ will be to him or her. Just the material payoff of agent i and the kindness of him or her to agent $-i$ are affected by the maximization.

[14] A detailed analysis of other or general kindness functions (Appendix A of [19]) and the implications for all that work which is done in the work at hand is postponed to future work. Confer Comment 19.

Comment 12. *A discussion of the implications if one would demand $c_i = s'_i$ (in the sense that if she or he changes her or his strategy/action, she or he also changes her or his second order belief) would be interesting.*

In the solution concept of [19], if an agent believes that the other is doing something and believes that the other is doing this to help him or her, the agent also wants to help the other. If the agent believes that the other is doing the very same action (or strategy) but if the agent also believes that the other is doing this to hurt him or her, then the agent also wants to hurt the other, at least when the material costs for helping or hurting are not too large. Please note: If there are several fairness equilibria, it is not clear why agents should choose a nice one, i.e. (cooperation, cooperation) exemplarily in the prisoner's dilemma. If agents know each other, this question might be linked to sympathy. But in a one-shot game with strangers, it is not clear which fairness equilibrium will likely be played—just as it is the case when dealing with Nash equilibria.

2.5 Results from Rabin [19]

Here, we shortly summarize some propositions and some other findings of Rabin [19], whereby the propositions' proofs can be found in Appendix B of [19] (cf. also [12]).

- All Nash equilibria that are either mutual-max or mutual-min outcomes are fair for all $\chi > 0$ ([19], Proposition 1).

- In all fairness equilibria, either both kindness functions f_1, f_2 are positive or both are non-positive ([19], Proposition 2).

- Mutual-max outcomes with $f_1, f_2 > 0$ and mutual-min outcomes with $f_1, f_2 < 0$ are fair for all χ small enough ([19], Proposition 3). Compare Comment 14.

- All strict Nash equilibria are fair for all large enough χ ([19], Proposition 5 Part 1). Again, see Comment 14.

- All outcomes that are not Nash cannot be fair for large enough χ ([19], Proposition 5 Part 2).

- Proposition 6 of [19] states that ‚In every game, there exists a weakly negative fairness equilibrium.‘ What is meant by that is discussed in Comment 20. Furthermore, there can be problems with continuity, see Footnote 24, Appendix A, and the corresponding proof in Appendix B, all in [19].
- The fairness concept is also applicable to (some) examples with infinite pure strategy sets (Section IV of [19]). Confer Comment 3.
- The concept of ‚trust‘ is challenging for fairness, see p. 1296f of [19].

3 Fairness Equilibria in Pure Strategies

As expressed in Comment 2, the usage of pure and mixed strategies in [19] is worth further discussion. In the definitions of fairness equilibria, mutual max outcomes, etc., Rabin [19] uses S, i.e. mixed strategies—and so do we. However, in [19], in most of the examples, the calculations and discussions are done for pure strategies only. In the following, we give an argument, why this is meaningful in these examples. Proposition 1 provides a technical result:

Proposition 1. *When agent i plays s_i, which is playing $a_i^{(1)}, \dots, a_i^{(n_i)}$ each with probability $p_i^{(1)}, \dots, p_i^{(n_i)}$, it holds:*

$$U_i(s_i, b_{-i}, c_i) = \sum_j p_i^{(j)} U_i(a_i^{(j)}, b_{-i}, c_i)$$

when b_{-i}, c_i are fixed.[15]

Proof.

$$\begin{aligned} U_i(s_i, b_{-i}, c_i) &= u_i(s_i, b_{-i}) + \tilde{f}_{-i}(b_{-i}, c_i)(1 + f_i(s_i, b_{-i})) \\ &= u_i(s_i, b_{-i}) + \tilde{f}_{-i}(b_{-i}, c_i) \\ &\quad \left(1 + \frac{u_{-i}(b_{-i}, s_i) - u^e_{-i}(b_{-i})}{u^h_{-i}(b_{-i}) - u^{min}_{-i}(b_{-i})} \mathbb{I}_{u^h_{-i}(b_{-i}) - u^{min}_{-i}(b_{-i}) \neq 0}\right) \\ &= u_i(s_i, b_{-i}) + \tilde{f}_{-i}(b_{-i}, c_i) \end{aligned}$$

[15] Please note that we use the indicator function in that way that this is evaluated first s.t. divisions by zero cannot happen.

$$
\begin{aligned}
&+\frac{u_{-i}(b_{-i}, s_i)}{u^h_{-i}(b_{-i}) - u^{min}_{-i}(b_{-i})}\tilde{f}_{-i}(b_{-i}, c_i)\mathbb{I}_{u^h_{-i}(b_{-i})-u^{min}_{-i}(b_{-i})\neq 0}\\
&-\frac{u^e_{-i}(b_{-i})}{u^h_{-i}(b_{-i}) - u^{min}_{-i}(b_{-i})}\tilde{f}_{-i}(b_{-i}, c_i)\mathbb{I}_{u^h_{-i}(b_{-i})-u^{min}_{-i}(b_{-i})\neq 0}\\
=&\sum_j p_i^{(j)}(u_i(a_i^{(j)}, b_{-i}) + \tilde{f}_{-i}(b_{-i}, c_i))\\
&+\sum_j p_i^{(j)} \cdot \frac{u_{-i}(b_{-i}, a_i^{(j)})}{u^h_{-i}(b_{-i}) - u^{min}_{-i}(b_{-i})}\tilde{f}_{-i}(b_{-i}, c_i)\mathbb{I}_{u^h_{-i}(b_{-i})-u^{min}_{-i}(b_{-i})\neq 0}\\
&-\sum_j p_i^{(j)} \cdot \frac{u^e_{-i}(b_{-i})}{u^h_{-i}(b_{-i}) - u^{min}_{-i}(b_{-i})}\tilde{f}_{-i}(b_{-i}, c_i)\mathbb{I}_{u^h_{-i}(b_{-i})-u^{min}_{-i}(b_{-i})\neq 0}\\
=&\dots\\
=&\sum_j p_i^{(j)}(u_i(a_i^{(j)}, b_{-i}) + \tilde{f}_{-i}(b_{-i}, c_i)(1 + f_i(a_i^{(j)}, b_{-i})))\\
=&\sum_j p_i^{(j)} U_i(a_i^{(j)}, b_{-i}, c_i)
\end{aligned}
$$

□

Proposition 2. *It holds with fixed* b_{-i}, c_i*:*

$$A_i \cap argmax_{s'_i \in S_i} U_i(s'_i, b_{-i}, c_i) = argmax_{a'_i \in A_i} U_i(a'_i, b_{-i}, c_i)$$

Proof. "$\subset$": Let $a_i \in A_i$ s.t. $U_i(a_i, b_{-i}, c_i) = max_{s'_i \in S_i} U_i(s'_i, b_{-i}, c_i)$. Since $A_i \subset S_i$ (in the sense that one can represent every element of A_i by a vector $p = (0, \dots, 0, 1, 0, \dots, 0)$, it holds: $U_i(a_i, b_{-i}, c_i) \geq max_{a'_i \in A_i} U_i(a'_i, b_{-i}, c_i)$, i.e., $U_i(a_i, b_{-i}, c_i) \geq U_i(a'_i, b_{-i}, c_i)$ $\forall a'_i \in A_i$, which shows the first inclusion.

"$\supset$": Now, let $a_i \in A_i$ s.t. $U_i(a_i, b_{-i}, c_i) = max_{a'_i \in A_i} U_i(a'_i, b_{-i}, c_i)$. Assume that it exists an $s_i \in S_i \setminus A_i$ with $U_i(a_i, b_{-i}, c_i) \leq U_i(s_i, b_{-i}, c_i)$. We compute under usage of Propostion 1:

$$
\begin{aligned}
U_i(s_i, b_{-i}, c_i) &= \sum_j p_i^{(j)} U_i(a_i^{(j)}, b_{-i}, c_i)\\
&\leq \sum_j p_i^{(j)} U_i(a_i, b_{-i}, c_i)
\end{aligned}
$$

$$
\begin{aligned}
&= U_i(a_i, b_{-i}, c_i)\\
&\leq U_i(s_i, b_{-i}, c_i),
\end{aligned}
$$

which shows by the Sandwich lemma that it does not exist an $s_i \in S_i \setminus A_i$ with $U_i(a_i, b_{-i}, c_i) < U_i(s_i, b_{-i}, c_i)$. □

We emphasize that this proof builds essentially on the affine-linear structure of the expected utility arising from the use of the standard kindness functions of [19].[16] When using other kindness functions (see Appendix A of [19]), we do not know whether Propositions 1, 2, and also 4 hold true.

But please note that this *does not* mean that $argmax_{s'_i \in S_i} U_i(s'_i, b_{-i}, c_i) = argmax_{a'_i \in A_i} U_i(a'_i, b_{-i}, c_i)$ has to hold, but

$$argmax_{s'_i \in S_i} U_i(s'_i, b_{-i}, c_i) \supset argmax_{a'_i \in A_i} U_i(a'_i, b_{-i}, c_i).$$

From this, it directly follows:

$$max_{s'_i \in S_i} U_i(s'_i, b_{-i}, c_i) = max_{a'_i \in A_i} U_i(a'_i, b_{-i}, c_i)$$

This latter equality could also be derived from a generalization of the finding that an arithmetic average becomes larger/smaller when adding/deleting a data point above the average or deleting/adding a data point below this average in an analogous way to Footnote 17 of [19] (where this is done for the case of Nash equilibria).

Proposition 3. *If* $1 < m \in \mathbb{N}$*,* $p_1, \ldots, p_m > 0$*,* $\sum_k p_k = 1$*,* $x_1, \ldots, x_m \in \mathbb{R}$*,* $\bar{x} := \sum_{k=1}^{m} p_k x_k$*, and* $x_j < \bar{x}$ *for some* $j \in \{1, \ldots, m\}$*, it holds:*

$$\frac{1}{1-p_j} \sum_{k \in \{1,\ldots,m\} \setminus \{j\}} p_k x_k > \bar{x}$$

and

$$\sum_{k \in \{1,\ldots,m\} \setminus \{j\}} \frac{1}{1-p_j} p_k = 1$$

[16] Confer Footnote 14 and Comment 19 of the work at hand and Definition A3 of Appendix A of [19].

Proof. We start with the probability transformation.

$$\begin{aligned} \frac{1}{1-p_j} &= \frac{1}{\sum_{\ell \in \{1,\ldots,m\}} p_\ell - p_j} \\ &= \frac{\sum_{\ell \in \{1,\ldots,m\}} p_\ell}{\sum_{\ell \in \{1,\ldots,m\} \setminus \{j\}} p_\ell} \\ &= \frac{\sum_{\ell \in \{1,\ldots,m\} \setminus \{j\}} p_\ell + p_j}{\sum_{\ell \in \{1,\ldots,m\} \setminus \{j\}} p_\ell} \\ &= \frac{p_j}{\sum_{\ell \in \{1,\ldots,m\} \setminus \{j\}} p_\ell} + 1 \end{aligned}$$

That means, via $\frac{1}{1-p_j}$ the probability p_j is proportionally distributed to the remaining probabilities. Now, the second line (i.e. the first equation) holds, since:

$$\begin{aligned} &\sum_{k \in \{1,\ldots,m\} \setminus \{j\}} \left(\frac{p_j}{\sum_{\ell \in \{1,\ldots,m\} \setminus \{j\}} p_\ell} + 1 \right) p_k \\ &= \frac{p_j}{\sum_{\ell \in \{1,\ldots,m\} \setminus \{j\}} p_\ell} \sum_{k \in \{1,\ldots,m\} \setminus \{j\}} p_k + \sum_{k \in \{1,\ldots,m\} \setminus \{j\}} p_k \\ &= \sum_{k \in \{1,\ldots,m\}} p_k = 1 \end{aligned}$$

Next, we start with $\bar{x} > x_j$. Here we note that, because of $m > 1$, $\sum p = 1$, $p_k > 0$ it follows that $1 - p_j \in (0,1)$. Furthermore, we need $p_j > 0$ for the inequality.

$$\begin{aligned} \bar{x} &= \sum_{k \in \{1,\ldots,m\}} p_k x_k \\ &= p_j x_j + \sum_{k \in \{1,\ldots,m\} \setminus \{j\}} p_k x_k \\ &< p_j \bar{x} + \sum_{k \in \{1,\ldots,m\} \setminus \{j\}} p_k x_k \\ \bar{x}(1 - p_j) &< \sum_{k \in \{1,\ldots,m\} \setminus \{j\}} p_k x_k \end{aligned}$$

□

We note that the proposition and its proof do not hold if one replaced the probabilities by general weights w_k that do not sum up to one.[17]

From Proposition 2 it directly follows:

Proposition 4. *For fairness equilibria in pure strategies, i.e. actions, it does not matter whether to optimize over S or over A.*

It is widely known that Nash equilibria cannot only be computed via the definition—i.e., that no agent can improve his or her payoff if the others do not change their strategy—via checking for all outcomes if some can improve his or her payoff, but also by calculating best responses and searching for outcomes where these best-response functions intersect, which corresponds to a definition with argmax functions. Keeping this and the definition of mutual max and mutual min in mind, Proposition 4 also applies in an analogous way to Nash equilibria, mutual max outcomes, and mutual min outcomes, since (with fixed strategies of the resp. opponent) it holds $u_i(s_i, s_{-i}) = \sum_j p_i^j u_i(a_i^{(j)}, s_{-i})$ and $u_{-i}(s_{-i}, s_i) = \sum_j p_i^j u_{-i}(s_{-i}, a_i^{(j)})$. Analogous versions of Proposition 2 can be proven similarly,[18] where for the mutual min the inequality signs are the other way around.

From the discussion after the proof of Propostion 2—which refers to Footnote 17 of [19], where this is explained for Nash—it follows that strict fairness equilibria, i.e. fairness equilibria with strict inequalities in the *argmax* of the definition, can only appear in pure strategies. That strict equilibria cannot be in proper mixed strategies does not only hold for Nash and Rabin fairness, but—as a consequence of the above calculations—also for mutual min and mutual max.

4 Fairness Equilibria in the Prisoner's Dilemma

The prisoner's dilemma in [19] is slightly different to ours in Table 1 (from [21]) in so far Rabin uses a scaled version of Table 3. The structure—of

[17] For example: $w = (9, 5, 9)$, $x = (4, 1, 4)$. Then: $x_1 = 4 < \bar{x} = 36 + 5 + 36 = 77 > \left(\frac{w_1}{w_2+w_3} + 1\right)(w_2x_2 + w_3x_3) = \left(\frac{9}{14} + 1\right)(5 + 36) \approx 67.4$ (This example was found using MS Excel)

[18] ... which may be written down in the near future.

Table 3: Prisoner's Dilemma: material payoffs (with values from [19] with scaling parameter equal to one)

$u_1(\cdot)\vert u_2(\cdot)$	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	$4\vert 4$	$0\vert 6$
$a_1^{(2)}$	$6\vert 0$	$1\vert 1$

course—is the same. This scaling factor serves for weighting between material payoffs and emotions.

For the scaled prisoner's dilemma in [19] it is shown that the Nash equilibrium (defect, defect) is always, i.e. for all $\chi > 0$, a fairness equilibrium. However there is also the fairness equilibrium (cooperate, cooperate) for a small enough scaling factor. Next, we calculate the fairness equilibria for our prisoner's dilemma in Table 1 in a scaled version, where we scale all material payoffs with the same scaling factor $\chi > 0$, see Table 4.

Table 4: Scaled Prisoner's Dilemma: material payoffs (higher values are favorable for the agents. The values are taken from [21] scaled by $\chi > 0$, cf. [19])

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

We calculate the fairness equilibria in our scaled prisoner's dilemma, cf. [19, 21], by use of Proposition 4, that is, we optimize over A. We start with the question whether $(a_1^{(2)}, a_2^{(2)})$ is a fairness equilibrium. For that, we have to check under the conditions $c_1 = b_1 = a_1$ and $c_2 = b_2 = a_2$ whether

$$U_1(a_1^{(2)}, a_2^{(2)}, a_1^{(2)}) \overset{?}{\geq} U_1(a_1^{(1)}, a_2^{(2)}, a_1^{(2)})$$

(since our game is symmetric and our investigated strategy pair is on the diagonal, we can omit the condition for agent 2). We calculate:

$$u_1(a_1^{(2)}, a_2^{(2)}) = \chi,$$

as well as

$$u_2(a_2^{(2)}, a_1^{(2)}) = \chi,$$

$$u_2^h(a_2^{(2)}) = 5\chi,$$

$$u_2^l(a_2^{(2)}) = u_2^{min}(a_2^{(2)}) = \chi,$$

$$u_2^e(a_2^{(2)}) = 3\chi,$$

$$f_1(a_1^{(2)}, a_2^{(2)}) = \frac{1-3}{5-1} = -0.5,$$

and

$$u_1^h(a_1^{(2)}) = 5\chi,$$

$$u_1^l(a_1^{(2)}) = u_1^{min}(a_1^{(2)}) = \chi,$$

$$u_1^e(a_1^{(2)}) = 3\chi,$$

$$\tilde{f}_2(a_2^{(2)}, a_1^{(2)}) = \frac{1-3}{5-1} = -0.5,$$

and, thus,

$$U_1(a_1^{(2)}, a_2^{(2)}, a_1^{(2)}) = \chi - 0.5(1-0.5) = \chi - 0.25.$$

Next, we compute

$$u_1(a_1^{(1)}, a_2^{(2)}) = 0,$$

as well as

$$u_2(a_2^{(2)}, a_1^{(1)}) = 5\chi,$$

$$f_1(a_1^{(1)}, a_2^{(2)}) = \frac{5-3}{5-1} = 0.5,$$

and, thus,

$$U_1(a_1^{(1)}, a_2^{(2)}, a_1^{(2)}) = (0 - 0.5(1+0.5)) = -0.75 < U_1(a_1^{(2)}, a_2^{(2)}, a_1^{(2)}).$$

Due to the symmetry, (defect,defect) is a fairness equilibrium for all $\chi > 0$. Next, we analyze (defect,cooperate), that is, we check whether and if so, for which $\chi > 0$, $(a_1^{(2)}, a_2^{(1)})$ is a fairness equilibrium. Caused by the symmetry, (cooperate,defect) does not have to be analyzed separately. We check under the conditions $c_1 = b_1 = a_1$ and $c_2 = b_2 = a_2$ whether

$$U_1(a_1^{(2)}, a_2^{(1)}, a_1^{(2)}) \overset{?}{\geq} U_1(a_1^{(1)}, a_2^{(1)}, a_1^{(2)})$$

$\wedge$

$$U_2(a_2^{(1)}, a_1^{(2)}, a_2^{(1)}) \overset{?}{\geq} U_2(a_2^{(2)}, a_1^{(2)}, a_2^{(1)}).$$

It is

$$u_1(a_1^{(2)}, a_2^{(1)}) = 5\chi,$$

as well as

$$u_2(a_2^{(1)}, a_1^{(2)}) = 0,$$
$$u_2^h(a_2^{(1)}) = 3\chi,$$
$$u_2^l(a_2^{(1)}) = u_2^{min}(a_2^{(1)}) = 0,$$
$$u_2^e(a_2^{(1)}) = 1.5\chi,$$
$$f_1(a_1^{(2)}, a_2^{(1)}) = \frac{0 - 1.5}{3 - 0} = -0.5,$$

and

$$\tilde{f}_2(a_2^{(1)}, a_1^{(2)}) = 0.5,$$

and, thus,

$$U_1(a_1^{(2)}, a_2^{(1)}, a_1^{(2)}) = 5\chi + 0.5(1 - 0.5) = 5\chi + 0.25.$$

(Please consult Footnote 10 in [19] for a comment why this value is higher than 5χ.)

$$u_1(a_1^{(1)}, a_2^{(1)}) = 3\chi,$$

as well as

$$u_2(a_2^{(1)}, a_1^{(1)}) = 3\chi,$$
$$f_1(a_1^{(1)}, a_2^{(1)}) = \frac{3 - 1.5}{3 - 0} = 0.5,$$

and, thus,

$$U_1(a_1^{(1)}, a_2^{(1)}, a_1^{(2)}) = 3\chi + 0.5(1 + 0.5) = 3\chi + 0.75.$$

Since,

$$3\chi + 0.75 \leq 5\chi + 0.25 \Leftrightarrow \chi \geq 0.25$$

For all $\chi \geq 0.25$, the argmax condition for agent 1 is fulfilled. In this case, agent 1 is defecting, i.e., he or she is mean to agent 2. When agent 1 was not defecting, the fairness payoff would be higher ($0.75 > 0.25$), however, if the material payoff is large enough, it is profitable for agent 1 to defect nonetheless. For agent 2 it turns out that the argmax condition is not fulfilled:

$$u_1^h(a_1^{(1)}) = 3\chi,$$
$$u_1^l(a_1^{(1)}) = u_1^{min}(a_1^{(1)}) = 0,$$
$$u_1^e(a_1^{(1)}) = 1.5\chi,$$
$$\tilde{f}_2(a_2^{(2)}, a_1^{(1)}) = \frac{0 - 1.5}{3 - 0} = -0.5,$$

and, thus,

$$U_1(a_1^{(1)}, a_2^{(2)}, a_1^{(1)}) = (0 - 0.5(1 + 0.5)) = -0.75$$

$$U_2(a_2^{(2)}, a_1^{(2)}, a_2^{(1)}) = \chi - 0.5(1 - 0.5) = \chi - 0.25 > U_1(a_1^{(1)}, a_2^{(2)}, a_1^{(1)}).$$

Thus, the last possibility for a fairness equilibrium is (cooperate,cooperate), hence, we check—again under the conditions $c_1 = b_1 = a_1$ and $c_2 = b_2 = a_2$—whether

$$U_1(a_1^{(1)}, a_2^{(1)}, a_1^{(1)}) \overset{?}{\geq} U_1(a_1^{(2)}, a_2^{(1)}, a_1^{(1)})$$

(again: since our game is symmetric and our investigated strategy pair is on the diagonal, we omit the condition for agent 2). It holds:

$$U_1(a_1^{(1)}, a_2^{(1)}, a_1^{(1)}) = 3\chi + 0.5(1 + 0.5) = 3\chi + 0.75$$

and

$$U_1(a_1^{(2)}, a_2^{(1)}, a_1^{(1)}) = 5\chi + 0.5(1 - 0.5) = 5\chi + 0.25$$

which leads to the statement that for $\chi \leq 0.25$, staying mutually silent is a fairness equilibrium (just as in [19]), but keep Comment 14 in mind. Thus, if being kind is relatively important compared to material payoffs, cooperation is a fairness equilibrium.[19]

Comment 13. *Rabin [19] scales every material payoff in a game with the same parameter $\chi > 0$. Introducing two parameters for the two agents would complicate the analysis of the game, but maybe allow for better interpretations. Nonetheless, in [19] it is somehow implicitly assumed that both value the scaled material payoff in the same way.*

[19] For explanations concerning the calculations the interested reader may also consult [15, 20].

It turns out that (defect,defect) is always a fairness equilibrium and (cooperate,cooperate) is one if and only if $\chi \leq 0.25$, i.e., if the material costs for being kind are relatively small. There are no other fairness equilibria in pure strategies.

Comment 14. *We mention that Rabin [19] writes that (cooperate, cooperate) is a fairness equilibrium if the scaling parameter is less than* 0.25*. In our calculations (see the example in Section 4 and the computations in Sections 7 and 8 via Python using SymPy) it is* $\chi \leq 0.25$*. Thus, when Rabin says that if the scaling parameter is less than 0.25, the strategy pair is a fairness equilibrium, this is correct in all cases. However, maybe there is more to it: maybe the argmax in his definition of fairness is meant to be unique resp. strict? (...if this is meaningful and possible at all.) The same* $<$ *vs.* $\leq$ *problem occurs in the battle of the sexes with the value one for (opera, boxing) right under Definition 3 [19], in Propositions 3 and 5 [19] and possibly at other locations, too. Thus, a discussion of this* $<$ *vs.* $\leq$ *topic would be very fruitful.*

Contrary to this comment that the definition of argmax is meant to be strict is that Rabin [19] accents in the last paragraph on p. 1285 ending on p. 1286, where the "battle of the sexes" is used as an example, that there are two fairness equilibria with each strict *inequalities.*

Comment 15. *Note that in the explaining sentence right before the '*$X < 1$*' in the battle of the sexes [19]—up to our opinion—the* U *has to be replaced by the word opera.*

Table 5: Prisoner's Dilemma (with values from [19] or [21]): Nash Equilibria

Nash	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	no	no
$a_1^{(2)}$	no	yes

5 Examples, Existence, and Continuity

In this section we start with two examples where no pure-strategy fairness equilibrium exists, leading to the question of existence. Further, we discuss

Table 6: Prisoner's Dilemma (with values from [19] or [21]): Pure-Strategy Fairness Equilibria (i.e. all values of $\chi > 0$ s.t. the resp. outcome is a fairness equilibrium)

Rabin	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	$(0, 0.25]$	$\emptyset$
$a_1^{(2)}$	$\emptyset$	$(0, \infty)$

whether such a non-existence is related to the mixed-vs.-pure strategy topic and/or to discontinuities of kindness functions.

5.1 Rock-Scissors-Paper

A classical game in game theory (and also played often in real life, e.g., by pupils) is Rock-Scissors-Paper, which can be formalized like in Table 7 as done by Sieg [21]. It is well known (or, if not, easy to see) that there is no Nash equilibrium in pure strategies, see Table 8. It is not that easy to conclude that there is also no fairness equilibrium in pure strategies, cf. Table 9.

Proposition 5. *There is no fairness equilibrium in pure strategies in Rock-Scissors-Paper.*

Proof. In this proof, we do not consider mixed strategies. Note that here are two different outcomes: a tie, e.g., $(rock, rock)$ or a loss-win situation, e.g., $(scissors, rock)$. Irrespective of both the acting agent and the analyzed strategy, it holds: $u^h = \chi$, $u^l = u^{min} = -\chi$, $u^e = 0$. In the tie case all $\tilde{f} = 0$ and (scaled with $\chi > 0$), e.g., $U_1(paper, rock, rock) = \chi > 0 = U_1(rock, rock, rock)$, i.e., a tie is never fair.

We consider now $(scissors, rock)$. With $a_1 = b_1 = c_1 = scissors$ and $a_2 = b_2 = c_2 = rock$ we get $\tilde{f}_2(rock, scissors) = -0.5$, $f_1(scissors, rock) = 0.5$, $f_1(paper, rock) = -0.5$. Thus, $U_1(scissors, rock, scissors) = -\chi - 0.75 < \chi - 0.25 = U_1(paper, rock, scissors)$ $\forall \chi > 0$. Also a win-loss outcome is never fair. □

5.2 The Question of Existence

Here, an interesting fact shall be noted, namely, in [19] Proposition 6 states that—under specific assumptions on the continuity of the kindness functions—every game has at least one fairness equilibrium (in detail: a so-called weakly negative one, i.e. one where $f_1, f_2 \leq 0$ holds). We highlight (again) that Rabin [19] *does use* mixed strategies in the definition of fairness equilibrium. However, since in Footnote 8 in [19] he states that he emphasizes pure strategies and in most of the examples (battle of the sexes, chicken, prisoner's dilemma, ...) and the corresponding explanations only pure strategies are considered, the reader might guess that Proposition 6 of [19] talks about pure strategies, too. However, this is not true: as can be seen in Proposition 5 or Table 9, Rock-Scissors-Paper does not have a fairness equilibrium in pure strategies.

There are two possible answers to this problem: First, there is simply no fairness equilibrium in pure strategies but (at least) one in mixed ones, or second, there is neither one in pure nor in mixed strategies.

Concerning the second point, we may refer the reader to Proposition 6 and especially to Footnote 24 and the corresponding Appendix A and the proof of Proposition 6 in Appendix B of [19]. There, it is explained that the proof of the existence of a (weakly negative) fairness equilibrium is based on the existence theorem in [12] that assumes continuous kindness functions (see Footnote 24 in [19]). In Appendix A of [19], it is explained that the existence of fairness equilibria may not hold if the kindness function is not continuous. The continuity topic will be discussed later on (in Section 5.3) using a simpler game. Before that, we discuss the first point in more detail.

Proposition 6. *There is a fairness equilibrium in Rock-Scissors-Paper, namely the mixed outcome* $((1/3, 1/3, 1/3), (1/3, 1/3, 1/3))$

Proof. When one agent plays $(1/3, 1/3, 1/3)$, irrespective of what the other does, the expected material payoff is 0, thus, $u^h = u^l = u^{min} = u^e = \tilde{f} = 0$ for all agents. Hence, the expected utility is the expected material payoff and $((1/3, 1/3, 1/3), (1/3, 1/3, 1/3))$ is a fairness equilibrium. (See Comment 14.) □

Comment 16. *Please note that formally [19] uses mixed strategies since the argmax in the definition of fairness equilibria are taken over S (which is the convex hull of A when representing A as a set of vectors with the canonical*

Table 7: Rock-Scissors-Paper: material payoffs (with values from [21])

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

Table 8: Rock-Scissors-Paper: Nash equilibria (with values from [21])

Nash	$a_2^{(1)}$	$a_2^{(2)}$	$a_2^{(3)}$
$a_1^{(1)}$	no	no	no
$a_1^{(2)}$	no	no	no
$a_1^{(3)}$	no	no	no

unit vectors). Although the reader might guess that Footnote 8 in [19] suggests that in [19] only pure strategies are considered, this is not true.

That [19] does not use pure strategies only becomes clear when looking at the seeming contradiction that Rock-Scissors-Paper does not hove a fairness equilibrium in pure strategies but in mixed ones or when noting that both the work [12] (e.g., p. 64) and the proof of Proposition 6 in Appendix B of [19] use mixed strategies. Thus, the existence proposition (Proposition 6) (and its proof) in [19] show the existence of a fairness equilibrium in mixed strategies and not necessarily in pure ones; which is analogous to Nash equilibria [17].

Comment 17. *That [19] Proposition 6 makes a statement about mixed strategies becomes also obvious when thinking about what "continuous functions" in a room of discrete strategies shall be? We do not take into account degenerate things like the trivial metric to define continuity on discrete spaces.*

5.3 Matching-Pennies

Absolutely the same reasoning as for Rock-Scissors-Paper holds for the game Matching-Pennies, see Tables 10, 11, 12.

Table 9: Rock-Scissors-Paper: fairness equilibria (with values from [21])

Rabin	$a_2^{(1)}$	$a_2^{(2)}$	$a_2^{(3)}$
$a_1^{(1)}$	$\emptyset$	$\emptyset$	$\emptyset$
$a_1^{(2)}$	$\emptyset$	$\emptyset$	$\emptyset$
$a_1^{(3)}$	$\emptyset$	$\emptyset$	$\emptyset$

Proposition 7. *In Matching-Pennies, no outcome in pure strategies is fair.*

Proof. In this proof, we do not consider mixed strategies. There is basically only one type of outcome. Let us have a look at $(head, tail)$. For all agents and all strategies it holds $u^h = \chi, u^l = u^{min} = -\chi, u^e = 0$. Thus, $\tilde{f}_2(tail, head) = -0.5$, $f_1(head, tail) = 0.5$, $f_1(tail, tail) = -0.5$, and $U_1(head, tail, head) = -\chi - 0.75 < \chi - 0.25 = U_1(tail, tail, head)$. Hence, this is not fair. □

Proposition 8. *In Matching-Pennies,* $((1/2, 1/2)|(1/2, 1/2))$ *is fair. In that equilibrium all kindness functions are zero.*

Proof. If one agent plays $(1/2, 1/2)$, it does not matter what the other does, the material output will be zero, thus $\tilde{f}_2 = 0$. The material output equals the fairness output and all of them are zero for all strategies of the first agent. Thus all strategies are in the *argmax* and, hence, also $(1/2, 1/2)$, cf. Comment 14. □

5.4 Continuity

Matching-Pennies, although it was not the first example we found for the question of existence, has the advantage (other than Rock-Scissors-Paper) that continuity can be analyzed more easily. In [19], Appendix A, the possible discontinuity of $f_1(a_1, b_2)$ in b_2 at points where $u_2^h(b_2) = u_2^{min}(b_2)$ or of $u_2^e(b_2)$ also in b_2 could cause problems, which could lead to the non-existence of a fairness equilibrium, is discussed (cf. Comment 19).

Before analyzing the continuity issues, we recall that we know that in Matching-Pennies there exists one fairness equilibrium in mixed strategies. Additionally, we recall that both in [12] and in the proof of Proposition 6 in [19] mixed strategies are used.

Table 10: Matching-Pennies: material payoffs (with values from [21], "Matching-Euros", and scaled by $\chi > 0$, see [19])

$u_1(\cdot)\|u_2(\cdot)$	head	tail
head	$\chi\| -\chi$	$-\chi\|\chi$
tail	$-\chi\|\chi$	$\chi\| -\chi$

Table 11: Matching-Pennies: Nash equilibria (with values from [21], "Matching-Euros"). (Only pure strategies are considered here)

Nash	head	tail
head	no	no
tail	no	no

So, let—because of the symmetry—$\frac{1}{2} \leq \sigma_2 \in [0,1]$ and Agent 2 play "head" with probability σ_2 and "tail" with probability $1-\sigma_2$, resp. let Agent 1 believe that Agent 2 does this. Then it holds: $u_2^h(b_2) = \chi(2\sigma_2 - 1)$, $u_2^l(b_2) = u_2^{min}(b_2) = -\chi(2\sigma_2 - 1)$, $u_2^e(b_2) = 0 \ \forall \sigma_2$. Thus, $u_2^e(b_2)$ is continuous in b_2. Now, let us consider f_1. For that, Agent 1 plays head with probability $\sigma_1 \in [0,1]$ and tail with probability $1 - \sigma_1$.

$$f_1(s_1, b_2) = \begin{cases} \frac{-2\sigma_1+1}{2}, & \sigma_2 > \frac{1}{2}, \\ 0, & \sigma_2 = \frac{1}{2}. \end{cases}$$

Proposition 9. *Thus, in Matching-Pennies, f_1 is for all $\sigma_1 \neq 1/2$ discontinuous in b_2; but $u^e(b_2)$ is continuous in b_2 in Matching-Pennies.*

Taken together, f_1 is discontinuous, but there exists a fairness equilibrium in mixed strategies. If mixed strategies would not be considered, we would not know how to apply continuity.[20]

Proposition 10. *In Matching-Pennies $((1/2, 1/2), (1/2, 1/2))$ is indeed the only fairness equilibrium.*

[20] However, the question whether a game with continuous f and u^e functions (in mixed strategies—where the concept of continuity is applicable) does always have a pure-strategy fairness equilibrium remains.

Table 12: Matching-Pennies: fairness equilibria (with values from [21], "Matching-Euros"), i.e. all values for $\chi > 0$ for which this outcome is a fairness equilibrium. (Only pure strategies are considered here)

Rabin	head	tail
head	$\emptyset$	$\emptyset$
tail	$\emptyset$	$\emptyset$

Proof. Due to the symmetry, there are only two cases we have to check. First, consider $((1/2, 1/2), (\sigma_2, 1-\sigma_2))$ with $\sigma_2 \in [0,1] \setminus \{0.5\}$. Thus, since $c_1 = (1/2, 1/2)$, it follows $u_1^h(c_1) = u_1^l(c_1) = u_1^{min}(c_1) = u_1^e(c_1) = 0 = \tilde{f}_2((\sigma_2, 1-\sigma_2), c_1)\ \forall \sigma_2$. Thus, the expected material payoffs are the expected utilities. That means, playing head is better for Agent 1 if the probability for Agent 2 to play head is larger than 1/2 and the analogous strategy is true for tail.

Now, consider $((\sigma_1, 1-\sigma_1), (\sigma_2, 1-\sigma_2))$. For the reason of symmetry we only look at $\sigma_1, \sigma_2 > 1/2$. The other cases are analogous. In this case, Agent 1 is mean to Agent 2, but Agent 2 is nice to Agent 1. Hence, Agent 2 could increase his or her fairness payoff and at the same time his or her material payoff (in expectation) by reducing σ_2. As a consequence, $((1/2, 1/2), (1/2, 1/2))$ is the only fairness equilibrium. □

Comment 18. *We have not proven that* $((1/3, 1/3, 1/3)|(1/3, 1/3, 1/3))$ *is the only fairness equilibrium in Rock-Scissors-Paper.*

5.5 Discontinuity of the Equitable Payoff

At first, since the Matching-Pennies example does not lead to any discontinuity of $u_2^e(b_2)$ we give another example in Table 13.

Proposition 11. *The equitable payoff* $u_2^e(b_2)$ *in the example in Table 13 is discontinuous.*

Proof. In Table 13, $\mathbf{u}((\sigma_2, 1-\sigma_2)) = conv(\{(4\chi\sigma_2, 2\chi), (\chi + 2\chi\sigma_2, 0)\})$. If $\sigma_2 < 1/2$, no point in $\mathbf{u}((\sigma_2, 1-\sigma_2))$ is Pareto inferior to another point in $\mathbf{u}((\sigma_2, 1-\sigma_2))$. But if $\sigma_2 \geq 1/2$, the only Pareto superior point left is

Table 13: Discontinuity Example: material payoffs (scaled by $\chi > 0$, see [19]). This example was found with MS Excel.

$u_1(\cdot)\vert u_2(\cdot)$	mean	nice
nice but risky	$0\vert 2\chi$	$4\chi\vert 2\chi$
mean but safe	$\chi\vert 0$	$3\chi\vert 0$

$(4\chi\sigma_2, 2\chi)$. Thus, $u_2^l((\sigma_2, 1-\sigma_2)) = 0$ if $\sigma_2 < 1/2$, and $\ldots = 2\chi$ otherwise. This shows the discontinuity of u^l and in turn u^e (when mixed strategies are considered). (Confer also Comment 8.) □

However, u^h and u^{min} are continuous, since, if Agent i plays the mixed strategy $s_i = (\sigma_i^{(1)}, \sigma_i^{(2)}, \ldots, \sigma_i^{(n_i)})$, it holds:

$$u_2(b_2, s_1) = \sum_{j_2=1,\ldots,n_2} \sum_{j_1=1,\ldots,n_1} \sigma_2^{(j_2)} \sigma_1^{(j_1)} u_2(b_2^{(j_2)}, a_1^{(j_1)})$$

This is a composition of continuous functions and, thus, a continuous function in all $\sigma_1^{(1)}, \ldots, \sigma_2^{(n_2)}$. The set of all possible outcomes in mixed strategies is the convex hull of the outcomes in pure strategies. Taking *max* or *min* leads to continuous functions again.[21]

Comment 19. *Rabin [19] states in Footnote 11 (in the first part of this footnote) that up to the time he wrote that paper, all games he analyzed had at least one fairness equilibrium although the "standard kindness" is not continuous. For that reason of discontinuity, he used more general (continuous) kindness functions in Appendix A of [19]. We are neither aware of an example where really no fairness equilibrium in mixed strategies exists due to the discontinuity of the standard kindness function nor of a proof that irrespective of such a discontinuity of the standard kindness functions always a fairness equilibrium exists.*

[21] Please note that taking *max* or *min*, if the respective set is not convex, does in general not lead to continuous functions.

6 Further Comments and Results on Rabin Fairness

Before going on to computational issues, we discuss a basic question on the definition of fairness equilibrium. Further, we prove a minor result on mutual min resp. mutual max outcomes, which are also important in [19].

6.1 Two Types of Fairness Equilibria

Proposition 6 in [19] states (which becomes clear when checking its proof) that there is for all $\chi > 0$ an outcome that is a weakly negative fairness equilibrium (later on, we will call that Type II equilibrium). In the case of the Prisoner's Dilemma (see [19]) for χ large enough the only fairness equilibrium is the (strictly) negative one in pure strategies (even when allowing for mixed strategies), since both players increase their material payoff and their fairness payoff when both are mean by being as mean as possible. This negative fairness equilibrium is for all $\chi > 0$ a fairness equilibrium (Type I) and it is strictly negative for all $\chi > 0$. From our Matching-Pennies example we learn that there are games where for every $\chi > 0$ neither a strictly negative nor a strictly positive fairness equilibrium exists.

Comment 20. *At this point, we discuss an important question, namely:* what is a fairness equilibrium? *Is it meant to be in the sense that an outcome (in pure or in mixed strategies) is a fairness equilibrium if this definition is fulfilled for all $\chi > 0$—we call that Type I—or if there exists a $\chi > 0$ s.t. the definition of [19] is fulfilled (Type II)? Proposition 6 and its proof in [19] show—as far as we can see—that there is always a (weakly negative) Type II fairness equilibrium.*

Here, the chicken game is interesting. As can be seen in Tables 22 and 23, in pure strategies, there is no Type I fairness equilibrium. Whether there is one in mixed strategies is not yet answered. However, using the extension of our Python code from the very end of Section 7 (with $N = 8$) suggests that the chicken game does not have any Type I equilibrium. Is it possible to prove this for all mixed strategies? Or, if not, is it possible to show that every game—given continuity—has a Type I fairness equilibrium?

There are games where all Type I equilibria are strictly negative (i.e. $f_1, f_2 < 0$), e.g., the prisoner's dilemma. There are games where all Type I *and Type II equilibria have zero kindness, namely Matching-Pennies, Rock-Scissors-Paper, trivial games.*

Unclear is the following topic: Are there games where all Type II equilibria are strictly negative? Or: is there always a Type II equilibrium that has zero kindness?

In [19], it is shown that there are games without a (strictly) positive fairness equilibrium. This is interesting since in games where both agents are kind, the fairness payoff should increase the agents' payoffs, but this seems not always to be enough to be kind compared to an increase in material payoffs.

6.2 Mixed-Strategy Mutual-Max and Mutual-Min

Before explaining how fairness equilibria can be calculated, we give some insights on mutual-max and mutual-min outcomes, since Proposition 1 in [19] states that all Nash equilibria that are either mutual-max or mutual-min outcomes are fair for all $\chi > 0$.

Proposition 12. *Not every game has a mutual-max or a mutual-min outcome in pure strategies, however, every game has a mutual-max as well as a mutual-min outcome in mixed strategies.*

Proof. For part one of the proposition we consider Matching-Pennies: none of the pure outcomes is mutual-min or mutual-max since one of the agents could make the material payoff of the other higher by deviating and the other one could make the material payoff of the resp. other one smaller by deviating.

That every game has a mutual-max outcome follows by Nash [17] when switching the material payoffs $(u_1, u_2) \mapsto (u_2, u_1)$ in every cell of the table representing a game. For mutual-min we have to multiply additionally by minus one: $(u_1, u_2) \mapsto (-u_2, -u_1)$ □

However, note that these mutual-max or mutual-min outcomes do not have to be Nash equilibria.

6.3 Computation of Fairness Equilibria

Since in all kindness functions the scaling parameter cancels out, the expected utilities are affine-linear functions in χ. Two affine-linear functions in $\mathbb{R}^+$ can either have no intersection point (when they are parallel or when they would intersect in $\mathbb{R}_0^-$) or they intersect exactly once or they are equal.

Proposition 13. *The sets for χ to make a pure-strategy outcome a fairness equilibrium is always a closed subset of* $(0, \infty)$.

We note that, e.g., $(0, 1]$ is of course not a closed subset of $\mathbb{R}$, but it is a closed subset of $\mathbb{R}^+$ since every accumulation point of $(0, 1]$ that is an element of $\mathbb{R}^+$ lies in $(0, 1]$.

Proof. Let $\chi((a_1, a_2)) : A_1 \times A_2 \to \mathcal{P}((0, \infty))$ be the function that assigns to a given outcome (a_1, a_2) all values for the scaling variable χ s.t. this outcome is fair. Then it holds:

$$\begin{aligned}\chi((a_1, a_2)) =&(0, \infty) \setminus \Big(\bigcup_{a_1' \in A_1} \{\chi > 0 \mid U_1(a_1', a_2, a_1) > U_1(a_1, a_2, a_1)\} \\ &\cup \bigcup_{a_2' \in A_2} \{\chi > 0 \mid U_2(a_2', a_1, a_2) > U_2(a_2, a_1, a_2)\} \Big)\end{aligned}$$

Since U_1, U_2 are affine-linear functions in χ, the sets $\{\chi > 0 \mid \dots\}$ are of the form $(0, \infty)$, $\emptyset$, $(0, \beta)$, or (α, ∞). Every union of such sets results in a set of the form $(0, \infty)$, $\emptyset$, $(0, \beta)$, (α, ∞), or $(0, \beta) \cup (\alpha, \infty)$. Again, unions of such sets have the same structure. Then, $(0, \infty) \setminus \dots$ has the structure $\emptyset$, $(0, \infty)$, $[\beta, \infty)$, $(0, \alpha]$, or $[\beta, \alpha]$. Here, it is assumed that $0 < \beta \leq \alpha$ and $[\beta, \beta] := \{\beta\}$. □

As stated above, a fairness equilibrium is strict if the *argmax* functions in its definition lead to a unique strategy, each. Those strict fairness equilibria can only be in pure strategies and they do not have to exist, e.g., in Rock-Scissors-Paper. In the following we repeat Proposition 13 for strict fairness equilibria.

Proposition 14. *The sets for χ to make an outcome a strict fairness equilibrium is always an open subset of* $(0, \infty)$.

Proof. Let $\chi_s((a_1, a_2)) : A_1 \times A_2 \to \mathcal{P}((0, \infty))$ be the function that assigns to a given outcome (a_1, a_2) all values for the scaling variable χ s.t. this outcome is strictly fair. Then it holds:

$$\begin{aligned}\chi_s((a_1,a_2)) =& (0,\infty) \setminus \Big(\bigcup_{a_1 \neq a'_1 \in A_1} \{\chi > 0 \mid U_1(a'_1, a_2, a_1) \geq U_1(a_1, a_2, a_1)\} \\ & \cup \bigcup_{a_2 \neq a'_2 \in A_2} \{\chi > 0 \mid U_2(a'_2, a_1, a_2) \geq U_2(a_2, a_1, a_2)\} \Big)\end{aligned}$$

Since U_1, U_2 are affine-linear functions in χ, the sets $\{\chi > 0 \mid \dots\}$ are of the form $(0,\infty)$, $\emptyset$, $(0,\beta]$, or $[\alpha,\infty)$. Every union of such sets results in a set of the form $(0,\infty)$, $\emptyset$, $(0,\beta]$, $[\alpha,\infty)$, or $(0,\beta] \cup [\alpha,\infty)$. Again, unions of such sets have the same structure. Then, $(0,\infty) \setminus \dots$ has the structure $\emptyset$, $(0,\infty)$, (β,∞), $(0,\alpha)$, or (β,α). Here, it is assumed that $0 < \beta \leq \alpha$. □

From the proofs of Propositions 13 and 14 the next proposition follows.

Proposition 15. *It holds for all* $(a_1,a_2) \in A_1 \times A_2$ *that*

$$\chi_s((a_1,a_2)) = interior(\chi((a_1,a_2))).$$[22]

6.4 Mixed-Strategy Fairness

Proposition 16. *The prisoner's dilemma (with the values of [19]), when allowing for mixed strategies, not only* $(defect, defect)$ *for all* $\chi > 0$ *and* $(cooperate, cooperate)$ *for* $\chi < 1/4$ *(or for* $\chi \leq 1/4$*, see again Comment 14), are fairness equilibria, but there are more.*

When calling the probability of Agent i for playing *cooperate* σ_i, then, e.g., $((\sigma_1, 1-\sigma_1), (\sigma_2, 1-\sigma_2)) = ((3/4,1/4),(3/4,1/4))$ is fair for some $\chi > 0$.

Proof. Since the game and the outcome are symmetric, it holds $u^h = 9/2 \cdot \chi$, $u^l = u^{min} = \chi/4$, because when fixing one mixed strategy all outcomes in this fixed set are Pareto optimal for all fixed strategies (other than in Table 13), cf. Comment 8, and $u^e = 19/8 \cdot \chi$. Then, $\tilde{f} = 1/4$ and $f_1((\sigma_1, 1-\sigma_1), b_2) = \sigma_1 - 1/2$, and $U_1((\sigma_1, 1-\sigma_1),(3/4,1/4),(3/4,1/4)) = (-7/4\chi + 1/4)\sigma_1 + (19/4\chi + 1/8)$. Hence, for $\chi = 1/7$ $((3/4,1/4),(3/4,1/4))$ is fair, cf. Comment 14. □

[22] Note that in general for a set A it does not have to hold $interior(\bar{A}) = interior(A)$, but only $interior(\bar{A}) \supset interior(A)$.

Comment 21. *Allowing for mixed strategies increases the set of fairness equilibria, which is worth noting when considering Footnote 8 of [19].*[23]

Up to now, it is unclear how all fairness equilibria in mixed strategies and their respective sets for χ can be calculated if this is possible at all. However, the next proposition gives us a tool for checking whether an outcome is fair (and for which χ). However, please note that this cannot be used to find all fairness equilibria since 1.) one could only search on a grid (due to runtime issues up to some minimal grid size) and 2.) irrational fairness equilibria, i.e., those were $(s_1, s_2) \in \mathbb{R}^2 \setminus \mathbb{Q}^2$, cannot be found when searching numerically.

Proposition 17. *Consider a given outcome $(s_1, s_2) \in S_1 \times S_2$ represented by $(p_1^{(1)}, \ldots, p_1^{(n_1)})$ and $(p_2^{(1)}, \ldots, p_2^{(n_2)})$. Under the matching actions and beliefs assumption, the expected utility of the strategies in the given outcome has only to be compared to expected utilities of pure strategies, i.e. actions, when checking whether the given outcome is in the argmax.*

Proof. If there would be a proper mixed strategy leading to a higher expected utility than the given one, all pure strategies that come with non-zero probability would lead to the same expected utility or an even higher one. Thus, the comparison with actions is sufficient. □

7 Python/SymPy Code

Building upon the proof of Proposition 13, we can implement a code in Python [23] using SymPy [16] that searches for all fairness equilibria in pure strategies.[24] With Proposition 4 it follows that these are the pure-strategy

[23] An in-depth analysis of the general structure of all Type II fairness equilibria of games would be very interesting.

[24] For the inequality solver `solve_poly_inequality` see `https://docs.sympy.org/latest/modules/solvers/inequalities.html` (2024-03-21). For intervals, set operations, and `oo` (∞), see `https://docs.sympy.org/latest/modules/sets.html` (2024-03-21). For the reduce function `functools.reduce (fun,seq)` see `https://www.geeksforgeeks.org/reduce-in-python/` (2024-03-21). For the topics copy, deepcopy, and mutable objects, see `https://stackoverflow.com/questions/8743072/when-adding-to-list-why-does-python-copy-values-instead-of-pointers` (2024-03-25), `https://stackoverflow.com/questions/19210971/python-prevent-copying-object-as-reference` (2024-03-26), and `https://docs.python.org/3/library/copy.html` (2024-03-26). And, finally, for `time`, see `https://www.python-lernen.de/python-modul-time.htm` *(in German;* 2024-03-26).

fairness equilibria even when using mixed strategies in the definition. In detail, we calculate for each pure-strategy outcome the subset (see Proposition 13) of $(0, \infty) \ni \chi$ for which this outcome, if the game is scaled by χ, is a fairness equilibrium.[25]

We start with the header.

```
"""
Computation of Rabin's fairness equilibria,
Nash equilibria, etc.

Michael Hch. Baumann and Michaela Baumann

2024-03-20 -- 2025-01-23
"""

# Possible games:
# "pd_sieg": prisoner's dilemma (Sieg)
# "pd_rabin": prisoner's dilemma (Rabin)
# "rsp_sieg": rock-scissors-paper (Sieg)
# "chicken_rabin": chicken game (Rabin)
# "grabbing_rabin": grabbing game (Rabin)
# "bs_rabin": battle of the sexes (Rabin)
# "pnd_rabin": prisoner's non-dilemma (Rabin)
# "leaving_rabin": leaving a partnership (Rabin)
# "dictator"
# "ultimatum"
# "coordination"
# "assurance"
# "stag-hunt"
# "matching-euros_sieg": # matching euros resp. pennies (Sieg)

import time
import sympy
from sympy import Poly
from sympy import Interval
```

[25] Please note that due to the affine-linear structure of the “standard kindness” functions we could reduce the `reduce` function to “`(...)[0]`” because the list exhibits only one element.

```python
from sympy import oo
import functools
import copy
from pprint import pprint
```

Next, the game has to be chosen that shall be analyzed automatically and all the games, which can be chosen, have to be defined. When after a game's name (Sieg) is written, the values are from [21], when (Rabin) is written, the values are from [19], for the ultimatum game and the dictator game confer https://en.wikipedia.org/wiki/Ultimatum_game (2024-10-30) resp. https://en.wikipedia.org/wiki/Dictator_game (2024-10-30), and for the coordination games (coordination, assurance, stag hunt, and also battle of the sexes cf. https://en.wikipedia.org/wiki/Coordination_game (2024-04-05).

```python
# choose game:
game = "pd_rabin"

def set_game(game):
    # u1: material payoff matrix player 1
    # u2: material payoff matrix player 2
    elif game=="pd_sieg": # prisoner's dilemma (Sieg)
        u1 = [[3,0],[5,1]]
        u2 = copy.deepcopy(u1)
    elif game=="pd_rabin": # prisoner's dilemma (Rabin)
        u1 = [[4,0],[6,1]]
        u2 = copy.deepcopy(u1)
    elif game=="rsp_sieg": # rock-scissors-paper (Sieg)
        u1 = [[0,1,-1],[-1,0,1],[1,-1,0]]
        u2 = copy.deepcopy(u1)
    elif game=="chicken_rabin": # chicken game (Rabin)
        u1 = [[-2,2],[0,1]]
        u2 = copy.deepcopy(u1)
    elif game=="grabbing_rabin": # grabbing game (Rabin)
        u1 = [[1,2],[0,1]]
        u2 = copy.deepcopy(u1)
    elif game=="bs_rabin": # battle of the sexes (Rabin)
        u1 = [[2,0],[0,1]]
        u2 = [[1,0],[0,2]]
    elif game=="pnd_rabin": # prisoner's non-dilemma (Rabin)
        u1 = [[4],[6]]
        u2 = [[4,0]]
    elif game=="leaving_rabin": # leaving a partnership (Rabin)
        u1 = [[6,0],[5,5]]
        u2 = [[6,5],[12,5]]
    elif game=="dictator": # dictator
        u1 = [[10],[9],[8],[7],[6],[5],[4],[3],[2],[1],[0]]
        u2 = [[0,1,2,3,4,5,6,7,8,9,10]]
    elif game=="ultimatum": # ultimatum
        u1 = [[10,0],[9,0],[8,0],[7,0],[6,0],[5,0],[4,0],[3,0],
              [2,0],[1,0],[0,0]]
        u2 = [[0,1,2,3,4,5,6,7,8,9,10],[0,0,0,0,0,0,0,0,0,0,0]]
    elif game=="coordination": # coordination
        u1 = [[1,0],[0,1]]
        u2 = copy.deepcopy(u1)
    elif game=="assurance": # assurance
        u1 = [[2,0],[0,1]]
        u2 = copy.deepcopy(u1)
    elif game=="stag-hunt": # stag hunt
        u1 = [[10,0],[6,4]]
        u2 = copy.deepcopy(u1)
    elif game=="matching-euros_sieg": # matching pennies (Sieg)
        u1 = [[1,-1],[-1,1]]
        u2 = [[-1,1],[1,-1]]
    else:
        print("no predefined game")

    print(u1)
    print(u2)

    if u1 and u2:
        for i in range(len(u1)-1):
            if len(u1[i])!=len(u1[i+1]):
                print("Payoff 1 not specified correctly!")

        for i in range(len(u2)-1):
            if len(u2[i])!=len(u2[i+1]):
                print("Payoff 2 not specified correctly!")

        if len(u1)!=len(u2[0]) or len(u1[0])!=len(u2):
            print("Dimensions do not fit!")

    return u1, u2
```

Everything here is done in pure strategies only. Next, we calculate (strict) Nash equilibria, (strict) mutual-min outcomes, and (strict) mutual-max outcomes. The (strict) Nash equilibria are calculated via best responses (BR, SBR).

```
def nash_equilibria(game):

    t = time.time()
    u1, u2 = set_game(game)

    # strategies of the players
    S1 = range(len(u1))
    S2 = range(len(u1[0]))

    # (strict) Nash equilibria
    # (strictly) best response functions
    # is i in S1 a/the best response if agent 2 plays j in S2?
    BR1 = []
    BR2 = []
    Nash = []
    SBR1 = []
    SBR2 = []
    SNash = []
    for i in S1:
        BR1.append([])
        BR2.append([])
        Nash.append([])
        SBR1.append([])
        SBR2.append([])
        SNash.append([])
        for j in S2:
            BR1[i].append(1)
            BR2[i].append(1)
            Nash[i].append(0)
            SBR1[i].append(1)
            SBR2[i].append(1)
            SNash[i].append(0)
            for k in S1:
                if u1[k][j] > u1[i][j]:
                    BR1[i][j] = 0
                if (k != i and u1[k][j] >= u1[i][j]):
                    SBR1[i][j] = 0
            for k in S2:
                if u2[k][i] > u2[j][i]:
                    BR2[i][j] = 0
                if (k != j and u2[k][i] >= u2[j][i]):
                    SBR2[i][j] = 0
            Nash[i][j]=BR1[i][j]*BR2[i][j]
            SNash[i][j]=SBR1[i][j]*SBR2[i][j]
    # BR1 = [[0, 0, 1], [1, 0, 0], [0, 1, 0]] means that i's
    # first strategy is the best response to -i's third one,
    # i's second one is the best response to -i's first one,
    # and finally i's third one is the best answer to -i's
    # second strategy only pure and no mixed strategies and
    # Nash equilibria are considered. Is j in S2 a best response
    # if agent a plays i in S1? BR2 = [[0, 1, 0], [0, 0, 1],
    # [1, 0, 0]] means that the best -i can do if i does its
    # 1st, is its 2nd, the best -i can do if i plays its 2nd,
    # is its 3rd, ... Note that it might be more easy to compute
    # Nash equilibria in such a way that no one-sided
    # improvement can happen.

    # Additionally, we provide (strict) mutual-min/-max

    sMuMi = copy.deepcopy(u1)
    MuMi = copy.deepcopy(u1)
    sMuMa = copy.deepcopy(u1)
    MuMa = copy.deepcopy(u1)

    for i in S1:
        for j in S2:
            sMuMi[i][j] = 1
            MuMi[i][j] = 1
            sMuMa[i][j] = 1
            MuMa[i][j] = 1
            for k in S1:
                if i != k and u2[j][k] <= u2[j][i]:
                    sMuMi[i][j] = 0
                if u2[j][k] < u2[j][i]:
                    MuMi[i][j] = 0
                if i != k and u2[j][k] >= u2[j][i]:
                    sMuMa[i][j] = 0
                if u2[j][k] > u2[j][i]:
                    MuMa[i][j] = 0
            for l in S2:
                if j != l and u1[i][l] <= u1[i][j]:
                    sMuMi[i][j] = 0
                if u1[i][l] < u1[i][j]:
                    MuMi[i][j] = 0
                if j != l and u1[i][l] >= u1[i][j]:
                    sMuMa[i][j] = 0
                if u1[i][l] > u1[i][j]:
                    MuMa[i][j] = 0

    runtime = time.time()-t
    return Nash, SNash, sMuMi, MuMi, sMuMa, MuMa, MaMa, runtime
```

And finally, we compute (pure-strategy) fairness equilibria.

```
def fairness_equilibria_with_scaling(game):

    t = time.time()

    u1, u2 = set_game(game)

    # scaling factor
    x = sympy.symbols('x')
    # strategies of the players
    S1 = range(len(u1))
    S2 = range(len(u1[0]))

    # initialize uis for the fis
    u1_h = []
    u1_l = []
    u1_e = []
    u1_min = []
    u2_h = []
    u2_l = []
    u2_e = []
    u2_min = []

    # note that we ignore the x in u1_h, ..., u2_e since
    # the x is cancelling out in the fs

    # computing u1_h for all b1
    for i in S1:
        u1_h.append(u1[i][0])
        for j in range(1,S2[-1]+1):
            if u1[i][j]>u1_h[i]:
                u1_h[i] = u1[i][j]

    # computing u1_min for all b1
    for i in S1:
        u1_min.append(u1[i][0])
        for j in range(1,S2[-1]+1):
            if u1[i][j]<u1_min[i]:
                u1_min[i] = u1[i][j]


    # eliminating non-Pareto outcomes IN THE ROW
    P1 = copy.deepcopy(u1)
    # NOTE that in SymPy there are objects that are mutable,
    # i.e., when using P1 = u1 and altering P1, u1 would be altered, too
    for i in S1:
        for j in S2:
            for k in S2:
                if ((u1[i][k] > u1[i][j] and u2[k][i] >= u2[j][i])
                or (u1[i][k] >= u1[i][j] and u2[k][i] > u2[j][i])):
                    P1[i][j] = oo

    # computing u1_l and u1_e for all b1
    for i in S1:
        u1_l.append(P1[i][0])
        for j in range(1,S2[-1]+1):
            if P1[i][j]<u1_l[i]:
                u1_l[i] = P1[i][j]
        u1_e.append((u1_l[i]+u1_h[i])/2)

    # computing f2 (we don't need f1_tilde since a1, b1, c1
    # are all from S1 and the functions f1 and f1_tilde are
    # formally identical, see Rabin'93)
    f2 = []
    for i in S2:
        f2.append([])
        for j in S1:
            if u1_h[j]-u1_min[j]==0:
                f2[i].append(0)
            else:
                f2[i].append((u1[j][i]-u1_e[j])/(u1_h[j]-u1_min[j]))

    # computing u2_h for all b2
    for i in S2:
        u2_h.append(u2[i][0])
        for j in range(1,S1[-1]+1):
            if u2[i][j]>u2_h[i]:
                u2_h[i] = u2[i][j]

    # computing u2_min for all b2
    for i in S2:
        u2_min.append(u2[i][0])
        for j in range(1,S1[-1]+1):
            if u2[i][j]<u2_min[i]:
                u2_min[i] = u2[i][j]


    # eliminating non-Pareto outcomes IN THE ROW
    P2 = copy.deepcopy(u2)
    for i in S2:
        for j in S1:
            for k in S1:
                if ((u2[i][k] > u2[i][j] and u1[k][i] >= u1[j][i])
                or (u2[i][k] >= u2[i][j] and u1[k][i] > u1[j][i])):
                    P2[i][j] = oo

    # computing u2_l and u2_e for all b2
    for i in S2:
        u2_l.append(P2[i][0])
        for j in range(1,S1[-1]+1):
            if P2[i][j]<u2_l[i]:
                u2_l[i] = P2[i][j]
        u2_e.append((u2_l[i]+u2_h[i])/2)

    # computing f1
    f1 = []
    for i in S1:
        f1.append([])
        for j in S2:
            if u2_h[j]-u2_min[j] == 0:
                f1[i].append(0)
            else:
                f1[i].append((u2[j][i]-u2_e[j])/(u2_h[j]-u2_min[j]))


    for i in S1:
        for j in S2:
            u1[i][j] = u1[i][j]*x
            u2[j][i] = u2[j][i]*x

    # Us
    U1 = []
    for i in S1:
        U1.append([])
        for j in S2:
            U1[i].append([])
            for k in S1:
                U1[i][j].append(u1[i][j]+f2[j][k]*(1+f1[i][j]))
    U2 = []
    for i in S2:
        U2.append([])
        for j in S1:
            U2[i].append([])
            for k in S2:
                U2[i][j].append(u2[i][j]+f1[j][k]*(1+f2[i][j]))

    # checking for which x (i,j) with i in S1 and j in S2 is
    # a fairness equilibrium.
    X = []
    for i in S1:
        X.append([])
        for j in S2:
            X[i].append(sympy.EmptySet)
            for k in S1:
                X[i][j] = functools.reduce(lambda a, b: a.union(b), (
                    sympy.solve_poly_inequality(
                    Poly(U1[k][j][i]-U1[i][j][i],x,domain='RR'), ">")
                    )
                    ).union(X[i][j])
            for k in S2:
                X[i][j] = functools.reduce(lambda a, b: a.union(b), (
                    sympy.solve_poly_inequality(
                    Poly(U2[k][i][j]-U2[j][i][j],x,domain='RR'), ">")
                    )
                    ).union(X[i][j])
            X[i][j] = X[i][j].complement(Interval.open(0,oo))

    runtime = time.time()-t
    return X, runtime
    #end def fairness_equilibria_with_scaling
```

```
352 # Compute Nash equilibria in pure strategies
353 nash, snash, smumi, mumi, smuma, muma, runtime_nash = nash_equilibria(game)
354 pprint(nash)
355 pprint(snash)
356 
357 # Compute fairness equilibria according to Rabin with scaling
358 # of the original game
359 fair, runtime_fair = fairness_equilibria_with_scaling(game)
360 pprint(fair)
361 
362 print("Runtime:")
363 print(runtime_nash+runtime_fair)
```

Note that we can neither prove nor guarantee that our code calculates always the right equilibria etc. since we do not have references in general. Further, there may be numerical issues. The code is conducted with various games, see Section 8. Before going on to these games, we discuss an extension of the program.

The code can (easily) be adapted to search for (some rational) mixed-strategy fairness equilibria—see Proposition 17. For games where both agents have two actions, each, this can be done, e.g., via inserting the following code between lines 87 and 89.

```
364     if True and len(u1)==2 and len(u1[0])==2:
365         N = 2
366         v1 = []
367         v2 = []
368         for i in range(N+1):
369             v1.append([(N-i)*u1[0][0]/N+i*u1[1][0]/N])
370             v2.append([(N-i)*u2[0][0]/N+i*u2[1][0]/N])
371             for j in range(1,N+1):
372                 v1[i].append((N-i)*(N-j)*u1[0][0]/N**2
373                         +i*(N-j)*u1[1][0]/N**2+(N-i)*j*u1[0][1]/N**2
374                         +i*j*u1[1][1]/N**2)
375                 v2[i].append((N-i)*(N-j)*u2[0][0]/N**2
376                         +i*(N-j)*u2[1][0]/N**2+(N-i)*j*u2[0][1]/N**2
377                         +i*j*u2[1][1]/N**2)
378         u1 = v1
```

```
        u2 = v2
        print(u1)
        print(u2)
```

We note that this code is very inefficient since, due to Proposition 17, a comparison to pure strategies would be enough. However, that way, we can reuse all the basic Python code and just add some lines.

With this procedure with $N = 32$, for the prisoner's dilemma with the values of [19], we find that with σ_i being Agent i's probability of playing cooperate: $((\sigma_1, 1-\sigma_1)|(\sigma_2, 1-\sigma_2)) =$

- $((1,0)|(1,0))$ for $\chi \in (0, 1/4]$
- $((31/32, 1/32)|(31/32, 1/32))$ for $\chi = 5/21$
- $((30/32, 2/32)|(30/32, 2/32))$ for $\chi = 7/31$
- $((29/32, 3/32)|(29/32, 3/32))$ for $\chi = 13/61$
- $((28/32, 4/32)|(28/32, 4/32))$ for $\chi = 1/5$
- $((27/32, 5/32)|(27/32, 5/32))$ for $\chi = 11/59$
- $((26/32, 6/32)|(26/32, 6/32))$ for $\chi = 5/29$
- $((25/32, 7/32)|(25/32, 7/32))$ for $\chi = 3/19$
- $((24/32, 8/32)|(24/32, 8/32))$ for $\chi = 1/7$
- $((23/32, 9/32)|(23/32, 9/32))$ for $\chi = 7/55$
- $((22/32, 10/32)|(22/32, 10/32))$ for $\chi = 1/9$
- $((21/32, 11/32)|(21/32, 11/32))$ for $\chi = 5/53$
- $((20/32, 12/32)|(20/32, 12/32))$ for $\chi = 1/13$
- $((19/32, 13/32)|(19/32, 13/32))$ for $\chi = 1/17$
- $((18/32, 14/32)|(18/32, 14/32))$ for $\chi = 1/25$
- $((17/32, 15/32)|(17/32, 15/32))$ for $\chi = 1/49$
- $((0,1)|(0,1))$ for $\chi > 0$

are fairness equilibria for the noted sets of χ. However, we assume that there are much more. Note that the item $((24/32, 8/32)|(24/32, 8/32))$ fits to Propostion 16 and its proof.

Interestingly enough, there is no fairness equilibrium for $\sigma_1 = 1/2, \ldots, 1-31/32$, but for all $\sigma_1 = 1/32, \ldots, 1/2 - 1/32$ there is one for exactly one χ. Maybe, it is possible to describe the structure of (all) mixed-strategy fairness equilibria in a closed way—for this or for all 2×2 games ...

When running this code with $N = 2$ for Matching-Pennies with Sieg's values [21], it turns out that $((1/2, 1/2), (1/2, 1/2))$ is a fairness equilibrium for all $\chi > 0$. This result fits to our calculations in Section 5.3.

8 Further Examples

In this section, we briefly show the material payoff tables (unscaled) for some classical games, namely those you can also find in lines 40-82 in our code. For symmetric games, we state only the material payoff of Agent 1. Further, for all these games, we show the results of our code. There 'N' means Nash equilibrium, 'sN' strict N, 'MuMa' mutual max outcome, 'sMuMa' strict MuMa, 'MuMi' mutual min outcome, 'sMuMi' strict MuMi. Further, a subset of $(0, \infty)$ is given, for which this outcome is a fairness equilibrium. The strict fairness equilibria can easily be found by applying Proposition 15. Values and results can be found in Tables 14, ..., 41. In general, we do not discuss these games or results, only if such a discussion is important. For the material payoff values consider the text written between lines 33 and 34 of our code. All the computations here are done for pure strategies only—by use of our code, see Section 7.

8.1 Examples from above

For completeness, we also show the games that are already shown in Sections 4, 5.1, and 5.3. That way, it can be verified that our results by hand fit to those by code. See Tables 14-21.

8.2 Some Games from Rabin [19]

In this section, we demonstrate the functioning of our code by applying it to the finite-strategy-set games of Rabin [19], i.e., we do not apply it to those

Table 14: Prisoner's dilemma with values of [21]

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

Table 15: Prisoner's dilemma with values of [21]—Results

Results	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	sMuMa, $(0, 1/4]$	$\emptyset$
$a_1^{(2)}$	$\emptyset$	sN, sMuMi, $(0, \infty)$

of [19], Section IV. See Tables 16, 17, 22, ..., 31.

The game "Leaving a Partnership" (Tables 30 and 31) is particularly interesting for the reason of 'trust.' For this topic, see [19], Section IV.

The chicken game (Tables 22 and 23; please check Footnote 16 in [19]) is just like the prisoner's dilemma (Tables 14,..., 17) also very prominent in analyzing political, social, and business (administration, management) topics. In evolutionary game theory it is known as or similar to the hawk-dove game, when social interactions are modeled one reads the name snowdrift game, and in political analysis the so-called brink(s)manship game is related.

Please note that "Battle of the Sexes" (Table 26) is not a symmetric game. Hence, it is no wonder that also the sets for χ in Table 27 are unequal in the non-coordination cases. This game is used in [19] to explain why not only the opponent's action but his or her intention is important for fairness.

8.3 Dictator and Ultimatum Game

The dictator and the ultimatum game are highly interesting (esp. in experiments). Please consider again what we have written between lines 33 and 34 of our Python code. See Tables 32, ..., 35. And also see [19], Section I, and [22].

Please note that the games "Prisoner's Non-Dilemma" (Tables 28 and 29)

Table 16: Prisoner's dilemma with values of [19]

$u_1(\cdot)$	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	4	0
$a_1^{(2)}$	6	1

Table 17: Prisoner's dilemma with values of [19]—Results

Results	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	sMuMa, $(0, 1/4]$	$\emptyset$
$a_1^{(2)}$	$\emptyset$	sN, sMuMi, $(0, \infty)$

and "Dictator" (Tables 32 and 33) are interesting for another reason: only one of the agents has a choice at all. This leads to the fact that the other agent cannot be kind or hostile. Thus, the first agent does not have any fairness reason to be kind or hostile, either.

8.4 Coordination Games

In social analysis, coordination games are important. Those can be found in Tables 26, 27, 36, . . . , 41.

For stag hunt see also `https://en.wikipedia.org/wiki/Stag_hunt` (2024-11-19) and the seminal works of Jean-Jacques Rousseau "Discours sur l'origine et les fondements de l'inégalité parmi les hommes" *(in French;* 1754/55)[26] and David Hume "Book 3, Of Morals" (1740)[27].

8.5 Discussion of Dawes and Thaler's Example

On Page 4 we presented an example, which is given in [5] and which is used in [19] to motivate the concept of fairness equilibria. In the following, we try to

[26] `https://en.wikipedia.org/wiki/Discourse_on_Inequality` (2024-11-19)

[27] `https://davidhume.org/texts/t/3/2/2` (2024-11-19) and `https://davidhume.org/texts/t/3/2/7` (2024-11-19)

Table 18: Rock-Scissors-Paper with values of [21]

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

Table 19: Rock-Scissors-Paper with values of [21]—Results

Results	$a_2^{(1)}$	$a_2^{(2)}$	$a_2^{(3)}$
$a_1^{(1)}$	$\emptyset$	$\emptyset$	$\emptyset$
$a_1^{(2)}$	$\emptyset$	$\emptyset$	$\emptyset$
$a_1^{(3)}$	$\emptyset$	$\emptyset$	$\emptyset$

formalize this example s.t. it serves as a motivation for Rabin's [19] fairness equilibria. However, we also discuss which relatively hard assumptions we have to make in order to actually reach that goal. We leave it up to the reader to judge those assumptions as unrealistic or not. Finally, we also discuss how the extension of this example, where the money may be stolen, might be formalized.

In Table 42, we provide one possibility to formalize this example. We assume that the farmer has production costs of 10, offers the vegetable for sale for 20, and has additional costs of 1 for building the stall. The farmer may alternatively sell the vegetables also for 20 on the market. The car driver wants the vegetables and values them at 20. His or her alternative costs 15. In this example, where the farmer has additional material costs for building the stall, as can be seen in Table 43, $(stall, pay)$ is a fairness equilibrium for a small enough valuation of the material payoff, i.e. small χ, but not Nash.

If the farmer would not have additional costs for building the stall, he or she could not be kind towards the buyer, thus, *stall* is not fair anymore according to [19]. Note that $(10|5)$ is Pareto inferior to $(10|10)$ and, thus, not considered in u^ℓ, but $(10|5)$ is not Pareto inferior to $(9|10)$. In Rabin's fairness concept, one has to sacrifice him-/herself in order to be friendly [19].

Table 20: Matching Euros resp. Pennies with values of [21]

$u_1(\cdot)\vert u_2(\cdot)$	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	$1\vert -1$	$-1\vert 1$
$a_1^{(2)}$	$-1\vert 1$	$1\vert -1$

Table 21: Matching Euros resp. Pennies with values of [21]—Results

Results	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	$\emptyset$	$\emptyset$
$a_1^{(2)}$	$\emptyset$	$\emptyset$

An agent, who can give his resp. her opponent more without refraining from something, this is not considered friendliness (kindness). This is an important insight to (Rabin) fairness. Note that this may not align to someone's personal view of friendliness/kindness. See Tables 44 and 45. Note that we need to add costs for the stall in order to fit the farmers' example into Rabin's fairness framework—however, usually one might think that selling without intermediate dealer might be cheaper for the farmer.

Next, we alter the example in Table 42 in such a way that the money box, which is worth 100, could easily be stolen and neglect the possibility that someone wants to steal vegetables, but no money. The "buyer" does not want the vegetables if he or she wants to steal the money box. Values and results can be found in Tables 46 and 47. Two points are particularly interesting: 1.) Stealing the money box is never fair (in pure strategies). This is because if there is no stall, "wanting the vegetables" instead of "going for the money" would increase the material payoff and the fairness payoff of the "buyer;" if the "buyer" wants to steal, "no stall" increases both payoffs of the farmer. 2.) $(stall, pay)$ is still fair for tiny $\chi > 0$. The scaling parameter just becomes smaller since the money box has a higher value than the vegetables.

All in all, we see that the formalization of the farmers example s.t. it fits to Rabin's [19] fairness concept is quite challenging. Last, we give the code to be inserted in the Python program:

Table 22: Chicken with values of [19]

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

Table 23: Chicken with values of [19]—Results

Results	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	sMuMi, $(0, 1/2]$	sN, $[1/4, \infty)$
$a_1^{(2)}$	sN, $[1/4, \infty)$	sMuMa, $(0, 1/2]$

```
    elif game=="Farmer1":
        u1=[[9,-11],[10,10]]
        u2=[[10,5],[20,5]]
    elif game=="Farmer2":
        u1=[[10,-10],[10,10]]
        u2=[[10,5],[20,5]]
    elif game=="Farmer3":
        u1=[[9,-100],[10,10]]
        u2=[[10,5],[100,0]]
```

9 Conclusion

In this work, we had a close look on Rabin fairness, which uses beliefs to analyze how human agents act and not only technically transformed payoffs. Although Rabin's concept of fairness and reciprocity [19] is more than 30 years old and although it is challenged mainly for experimental, economical, and resp. or psychological reasons, it is still highly interesting, since it uses beliefs and not only actions (because it uses [12]) and since there are many mathematical questions to be answered.

In the work at hand, we asked some of these question in the hope that they will be answered in the future. Some questions we already answered, e.g., on

Table 24: Grabbing with values of [19]

$u_1(\cdot)$	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	1	2
$a_1^{(2)}$	0	1

Table 25: Grabbing with values of [19]—Results

Results	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	sN, sMuMi, $(0, \infty)$	$\emptyset$
$a_1^{(2)}$	$\emptyset$	sMuMa, $(0, 1/2]$

the relationship between pure and mixed strategy fairness equilibria and on the structure of the sets of the scaling parameters for which a pure-strategy outcome is fair. The latter one can be used to automatically compute all pure-strategy and some mixed-strategy fairness equilibria. This is done with Python/SymPy in this work, too.

Acknowledgment

Small parts of this work—mainly the idea of Rabin fairness and its connection to Comment 1—were presented at the international conference "Manifesto and Research frontiers for a Renaissance in Economic" (University of Perugia, Italy; June 20th-21st, 2024) and at the internal "Graduate Seminar in Economics" (University of Bayreuth, Germany; July 4th-5th, 2024) in talks on a related topic. This work was presented at the "15th Day on Computational Game Theory" (MindLabs / Tilburg University, the Netherlands; March 6th-7th, 2025) in the internal "MODUS Seminar Program" (University of Bayreuth, Germany; January 29th, 2025).

The authors of the work at hand thank Melanie Birke, Lars Grüne, Bernhard Herz, and Kilian Pioch for interesting discussions as well as Robert Baier for helping with some technical/Python issues. Further, the authors thank all organizers and participants of the above mentioned conferences and

Table 26: Battle of the Sexes with values of [19]

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

Table 27: Battle of the Sexes with values of [19]—Results

Results	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	sN, sMuMa, $(0, \infty)$	sMuMi, $(0, 1]$
$a_1^{(2)}$	sMuMi, $(0, 1/2]$	sN, sMuMa, $(0, \infty)$

seminars, esp. Marco Cucculelli, Fabian Herweg, Stephan Napel, Armando Pugno, Raimundo Julián Saona Urmeneta.

On May 29th, 2024, we sent a "Letter to M. Rabin" which addressed many of the questions raised in the work at hand. This document is written in LaTeX, programs were written in Python/SymPy and MS Excel.

References

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

[2] Charness, Gary, Matthew Rabin: Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics,* **117**(3): 817-869, August 2002.

[3] 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.

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

Table 28: Prisoner's Non-Dilemma with values of [19]

$u_1(\cdot)\mid u_2(\cdot)$	$a_2^{(1)}$
$a_1^{(1)}$	$4\mid 4$
$a_1^{(2)}$	$6\mid 0$

Table 29: Prisoner's Non-Dilemma with values of [19]—Results

Results	$a_2^{(1)}$
$a_1^{(1)}$	sMuMa, $\emptyset$
$a_1^{(2)}$	sN, sMuMi, $(0, \infty)$

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

[6] Dufwenberg, Martin, Georg Kirchsteiger: A Theory of Sequential Reciprocity. *Games and Economic Behavior,* **47**: 268-298, 2004.

[7] Kirman, Alan: Pareto, Vilfredo (1848-1923). *The New Palgrave Dictionary of Economics,* Palgrave Macmillan, London, 10041-10059, January 2018. `https://doi.org/10.1057/978-1-349-95189-5_1679`

[8] Falk, Armin, Ernst Fehr, Urs Fischbacher: On the Nature of Fair Behavior. *Economic Inquiry,* **41**(1): 20-26, January 2003.

[9] Falk, Armin, Urs Fischbacher: A Theory of Reciprocity. *Games and Economic Behavior,* **54**: 293-315, 2006.

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

[11] Freeman, David J.: Preferred Personal Equilibrium and Simple Choices. *Journal of Economic Behavior & Organization,* **143**: 165-172, 2017.

Table 30: Leaving a Partnership with values of [19]

$u_1(\cdot)\|u_2(\cdot)$	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	6|6	0|12
$a_1^{(2)}$	5|5	5|5

Table 31: Leaving a Partnership with values of [19]—Results

Results	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	sMuMa, $\emptyset$	$\emptyset$
$a_1^{(2)}$	MuMi, $\emptyset$	N, MuMi$(0, \infty)$

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

[13] Kolpin, Van: Equilibrium Refinement in Psychological Games. *Games and Economic Behavior,* **4**(2): 218-231, 1992.

[14] Kőszegi, Botond, Matthew Rabin: A Model of Reference-Dependent Preferences. *The Quarterly Journal of Economics,* **121**(4): 1133-1165, November 2006.

[15] Levin, Jonathan D.: Fairness and Reciprocity. *Working Paper, Stanford University,* June 2006. `https://web.stanford.edu/~jdlevin/Econ\%20286/Fairness.pdf` (2024-03-27)

[16] 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

Table 32: Dictator

$u_1(\cdot)\|u_2(\cdot)$	$a_2^{(1)}$
$a_1^{(1)}$	10|0
$a_1^{(2)}$	9|1
$a_1^{(3)}$	8|2
$a_1^{(4)}$	7|3
$a_1^{(5)}$	6|4
$a_1^{(6)}$	5|5
$a_1^{(7)}$	4|6
$a_1^{(8)}$	3|7
$a_1^{(9)}$	2|8
$a_1^{(10)}$	1|9
$a_1^{(11)}$	0|10

Python. *Python, Computer algebra system, Symbolics,* **3**:e103, January 2017.

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

[18] Nelson, William Robert Jr.: Incorporating Fairness into game Theory and Economics: Comment. *American Economic Review* **91**(4): 1180-1183, 2001.

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

[20] Reciprocity. *Lecture, University of Oslo,* no date. `https://www.uio.no/studier/emner/sv/oekonomi/ECON4260/h17/dokumenter/topic-3-third-lecture-reciprocity.pdf` (2024-03-27)

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

Table 33: Dictator—Results

Results	$a_2^{(1)}$
$a_1^{(1)}$	sN, sMuMi, $(0, \infty)$
$a_1^{(2)}$	$\emptyset$
$a_1^{(3)}$	$\emptyset$
$a_1^{(4)}$	$\emptyset$
$a_1^{(5)}$	$\emptyset$
$a_1^{(6)}$	$\emptyset$
$a_1^{(7)}$	$\emptyset$
$a_1^{(8)}$	$\emptyset$
$a_1^{(9)}$	$\emptyset$
$a_1^{(10)}$	$\emptyset$
$a_1^{(11)}$	sMuMa, $\emptyset$

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

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

Table 34: Ultimatum

$u_1(\cdot)\vert u_2(\cdot)$	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	10|0	0|0
$a_1^{(2)}$	9|1	0|0
$a_1^{(3)}$	8|2	0|0
$a_1^{(4)}$	7|3	0|0
$a_1^{(5)}$	6|4	0|0
$a_1^{(6)}$	5|5	0|0
$a_1^{(7)}$	4|6	0|0
$a_1^{(8)}$	3|7	0|0
$a_1^{(9)}$	2|8	0|0
$a_1^{(10)}$	1|9	0|0
$a_1^{(11)}$	0|10	0|0

Table 35: Ultimatum—Results

Results	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	N, $\emptyset$	N, MuMi, $(0, \infty)$
$a_1^{(2)}$	$\emptyset$	MuMi, $\emptyset$
$a_1^{(3)}$	$\emptyset$	MuMi, $\emptyset$
$a_1^{(4)}$	$\emptyset$	MuMi, $\emptyset$
$a_1^{(5)}$	$\emptyset$	MuMi, $\emptyset$
$a_1^{(6)}$	$\emptyset$	MuMi, $\emptyset$
$a_1^{(7)}$	$\emptyset$	MuMi, $\emptyset$
$a_1^{(8)}$	$\emptyset$	MuMi, $\emptyset$
$a_1^{(9)}$	$\emptyset$	MuMi, $\emptyset$
$a_1^{(10)}$	$\emptyset$	MuMi, $\emptyset$
$a_1^{(11)}$	MuMa, $\emptyset$	MuMa, MuMi, $\emptyset$

Table 36: (Pure) Coordination

$u_1(\cdot)$	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	1	0
$a_1^{(2)}$	0	1

Table 37: (Pure) Coordination—Results

Results	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	sN, sMuMa, $(0, \infty)$	sMuMi, $(0, 1]$
$a_1^{(2)}$	sMuMi, $(0, 1]$	sN, sMuMa, $(0, \infty)$

Table 38: Assurance

$u_1(\cdot)$	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	2	0
$a_1^{(2)}$	0	1

Table 39: Assurance—Results

Results	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	sN, sMuMa, $(0, \infty)$	sMuMi, $(0, 1/2]$
$a_1^{(2)}$	sMuMi, $(0, 1/2]$	sN, sMuMa, $(0, \infty)$

Table 40: Stag Hunt

$u_1(\cdot)$	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	10	0
$a_1^{(2)}$	6	4

Table 41: Stag Hunt—Results

Results	$a_2^{(1)}$	$a_2^{(2)}$
$a_1^{(1)}$	sN, sMuMa, $(0, \infty)$	$\emptyset$
$a_1^{(2)}$	$\emptyset$	sN, sMuMi, $(0, \infty)$

Table 42: Farmer 1

$u_1(\cdot)\|u_2(\cdot)$	$a_2^{(1)}$ "pay"	$a_2^{(2)}$ "steal"
$a_1^{(1)}$ "stall"	$9\|10$	$-11\|20$
$a_1^{(2)}$ "no stall"	$10\|5$	$10\|5$

Table 43: Farmer 1—Results

Results	$a_2^{(1)}$ "pay"	$a_2^{(2)}$ "steal"
$a_1^{(1)}$ "stall"	sMuMa, $(0, 1/20)$	$\emptyset$
$a_1^{(2)}$ "no stall"	N, MuMi, $(0, \infty)$	N, MuMi, $(0, \infty)$

Table 44: Farmer 2

$u_1(\cdot)\|u_2(\cdot)$	$a_2^{(1)}$ "pay"	$a_2^{(2)}$ "steal"
$a_1^{(1)}$ "stall"	$10\|10$	$-10\|20$
$a_1^{(2)}$ "no stall"	$10\|5$	$10\|5$

Table 45: Farmer 2—Results

Results	$a_2^{(1)}$ "pay"	$a_2^{(2)}$ "steal"
$a_1^{(1)}$ "stall"	sMuMa, $\emptyset$	$\emptyset$
$a_1^{(2)}$ "no stall"	N, MuMi, $(0, \infty)$	N, MuMi, $(0, \infty)$

Table 46: Farmer 3

$u_1(\cdot)\|u_2(\cdot)$	$a_2^{(1)}$ "pay"	$a_2^{(2)}$ "steal money box"
$a_1^{(1)}$ "stall"	9|10	−100|100
$a_1^{(2)}$ "no stall"	10|5	10|0

Table 47: Farmer 3—Results

Results	$a_2^{(1)}$ "pay"	$a_2^{(2)}$ "steal money box"
$a_1^{(1)}$ "stall"	sMuMa, $(0, 1/180)$	$\emptyset$
$a_1^{(2)}$ "no stall"	sN, MuMi, $(0, \infty)$	MuMi, $\emptyset$