A NOTE ON FAIR THREATS AND PROMISES

UNA NOTA SOBRE AMENAZAS Y PROMESAS JUSTAS

Alejandro Tatsuo Moreno-Okuno

Universidad de Guanajuato

Resumen: Con su Equilibrio de Reciprocidad Secuencial (SRE, por sus siglas en inglés), Dufwenberg y Kirchsteiger (2004) desarrollaron un concepto de solución que incorpora la reciprocidad en juegos secuenciales. Un SRE evalúa la bondad o la falta de bondad de una estrategia solamente en las acciones que prescribe la estrategia en el sendero del equilibrio. Sin embargo, el SRE no toma en cuenta las acciones afuera del sendero del equilibrio, donde están incluidas las amenazas y promesas. Este artículo desarrolla un nuevo concepto de solución, Equilibrio de Amenazas Justas, cuyo principal objetivo es dar predicciones más razonables cuando amenazas y promesas son incluidas.

Abstract: With their Sequential Reciprocity Equilibrium (SRE), Dufwenberg and Kirchsteiger (2004) developed a solution concept that incorporates reciprocity in sequential games. A SRE evaluates the kindness or unkindness of a strategy based purely on the actions it prescribes at the equilibrium path. However, given that it is not the objective of the SRE to evaluate threats and promises, it does not consider the actions outside the equilibrium path, where threats and promises are included. This article develops a new solution concept, Fair Threat Equilibria, which main objective is to give more reasonable predictions when threats and promises are included.

Clasificación JEL/JEL Classification: A13, C70, D63

Palabras clave/keywords: reciprocity; promises; threats

Fecha de recepción: 15 III 2021 *Fecha de aceptación: 17 III 2021*

https://doi.org/10.24201/ee.v37i1.429

 https://doi.org/10.24201/ee.v37i1.429

1. Introduction

The literature on reciprocity in game theory began with Rabin (1993), who, in his seminal paper, introduced his solution concept of Fairness Equilibrium (FE), which he defined for static games. Dufwenberg and Kirchsteiger (2004) extended Rabin's model to sequential games via their Sequential Reciprocity Equilibrium (SRE) solution concept.[1]

The main difference between a static and sequential game is that some strategies (some of which can be interpreted as promises and threats) that are optimal in a static game are no longer optimal (and therefore non-credible) in a sequential game, where a player is able to reconsider her moves as the game advances. As a subgame perfect Nash equilibrium eliminates non-credible threats for standard sequential games, a SRE eliminates non-credible threats for sequential games with the emotion of reciprocity. Defining the kindness of a strategy as a function of other players' strategies in the game, a SRE requires that each player plays their optimal action at each history of the game. However, in equilibrium, a SRE evaluates the kindness of each strategy by evaluating the actions it prescribes solely at the equilibrium path, without taking into consideration the actions that are prescribed off the equilibrium path. As promises and threats include the actions located both at and off the equilibrium path, I argue that, in order to evaluate the role of promises and threats correctly, the strategies have to be evaluated as a whole. Because a SRE evaluates the strategies only through the actions at each history, it accepts some unreasonable solutions when threats and promises are involved. As both subgame perfect Nash equilibrium and the SRE eliminate unreasonable solutions, my solution concept aims to eliminate unreasonable solutions when promises and threats are involved in sequential games with the emotion of reciprocity. However, my model is not a refinement of the SRE, but a different model whose solutions are more suited for the case of threat and promises.

This paper develops a solution concept of reciprocity that provides a more accurate tool for evaluating the fairness of threats and promises, which I have called Fair Threat Equilibrium. My conceptualization of reciprocity evaluates the kindness of a strategy as a whole by defining its kindness not as a function of the strategies of other players, but rather as a function of the maximum payoff the

[1] There are other models of reciprocity, with examples found in Falk and Fischbacher (2006), Levine (1998), Fehr and Schmidt (1999), and Bolton and Ockenfels (2000).

opposing players can receive (i.e., if the opposing players play the strategy that maximizes their "material" payoffs). This approach has the advantage of taking better into consideration the role of promises and threats.

Section 1.1 of the present study briefly sets out the argument that the emotion of reciprocity can make threats and promises credible. Moreover, that a SRE does not adequately take threats and promises into account. Section 2 summarizes the results of a survey conducted on the fairness of the threats implicit in the Ultimatum Game. The survey's results suggest that individuals evaluate the strategies of one player independently of the actions of other players and consider the promises and threats implicit in those strategies.

Section 3 develops my model and defines my concept of fair threat equilibrium (FTE), so named because it incorporates the idea that individuals not only want to reciprocate the kind or unkind actions of other players, but also want to reciprocate kind or unkind promises and threats. Section 4 of the present paper compares my model with the SRE. Section 5 applies my model to the Prisoner's Dilemma, the Battle of the Sexes, the Dictator Game, and the Mini-ultimatum Game, showing not only how a SRE allows some equilibria that, I argue, are unreasonable, and that the FTE eliminates these solutions. Section 6 details my conclusions.

1.1 *Does the emotion of reciprocity make promises and threats credible?*

Threats and promises play a fundamental role in game theory, given that the predictions of many sequential games depend on their credibility. As Klein and O'Flaherty (1993) and Schelling (1960) state, in order for a threat or a promise to be credible, there needs to be a commitment, one they argue can be a psychological commitment to keep one's word. However, the emotion of reciprocity can also be a powerful commitment. In this section, I show that the emotion of reciprocity can lead to players making credible promises and threats that are normally considered non-credible.

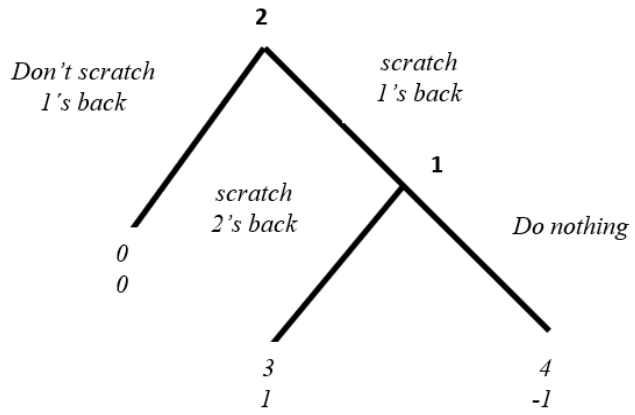
1.1.1 *Promise*

The game set out in figure 1[2] includes a *promise* from Player 1 to scratch Player 2's back if Player 2 scratches Player 1's back. Below, I

[2] This is an example of a pure promise taken from Klein and O'Flaherty (1993).

develop a concept of reciprocity that evaluates the kindness of Player 1's promise. If Player 2 scratches Player 1's back, then he is considered to have been kind to Player 1, and Player 1 may want to carry out her promise in order to be kind in return to Player 2.[3] As Klein and O'Flaherty (1993) and Schelling (1960) describe, a promise is successful if it is carried out. When promises are unsuccessful, the actions that form part of them are not in the equilibrium path and, thus, a SRE does not evaluate their kindness. Therefore, I argue that a SRE incorrectly assess unsuccessful promises.

Figure 1
If you scratch my back, I will scratch your back



Source: Based on Fig. 1 by Klein and O'Flaherty (1993).

1.1.2 *Threat*

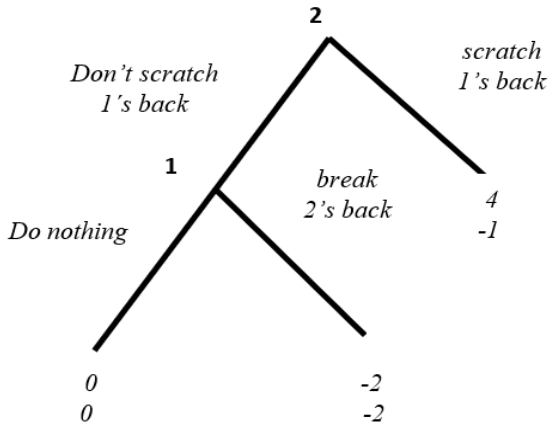
As the emotion of reciprocity can also make threats credible, I argue that the *threat* made by Player 1 to break Player 2's back in figure 2[4] should be evaluated as unkind. If Player 2 doesn't scratch Player 1's back, then he is unkind to Player 1, who may want to carry out

[3] In the present paper, I will refer to Player 1 as a woman and to Player 2 as a man, while any other player will be referred to as a woman. This will allow me to maintain gender neutrality and, at the same time, to be clearer in my explanations.

[4] This is an example of a pure threat from Klein and O'Flaherty (1993).

her threat in return. While a SRE only evaluates the threat of Player 1 if it is on the equilibrium path, namely if it is carried out, as both Klein and O'Flaherty (1993) and Schelling (1960) note, a threat is successful if it is not carried out. Because a threat that is not carried out is off the equilibrium path, a SRE does not assess the kindness of threats that are successful. I argue that the threat made by Player 1 should be evaluated as unkind, even if it is not carried out.

Figure 2
Scratch my back or else I will break your back



Source: Based on Fig. 2 by Klein and O'Flaherty (1993).

2. Survey

In a small survey conducted at the University of Guanajuato, a group of students were asked a series of questions related to the Ultimatum Game, which is played with two players. At the beginning of the game, the first player (the Proposer) chooses how to divide a given amount of money between herself and a second player (the Responder). The Responder chooses to accept their share or reject it, while, if he accepts, both players are paid the share chosen by the Proposer. If the Responder rejects the offer, both receive zero.

The participants were asked to indicate the share that they consider to be fair without being informed as to the Responder's specific strategy. Of the 63 students surveyed, 49 answered that a 50-50 share of the money was fair, while five students opted for a 60-40 share and

four opted for a 99-1 share in favor of the Proposer. Of the remaining students, one answered that an 80-20 share in favor of the Proposer was fair, one stated that he did not believe a fair share was possible, and one stated that any share offering a positive amount to the Responder was fair. Only two stated that a fair offer would correspond to the highest amount the Responder would accept, provided that this was less than 50%. Given that the respondents were not informed as to the Responder's strategy, these answers are consistent with the notion that the fairness of a strategy is independent of the strategies used by the other players of a game.

In terms of the Responder's participation in the game, the students were asked as to the fairness of strategies of the form accept any offer that is higher than or equal to a constant k, and reject any offer that is lower than k and what they felt the fair value of k would be. Of the 63 students, 46 gave a k value of 50%, with the remaining students giving k values lower than 50%, aside from one who opted for the highest possible value of k that would result in the share offered by the Proposer being accepted. This finding suggests that most individuals can evaluate the strategies used by the Responder independently of the strategies used by the Proposer, given that the respondents were not aware of any of the strategies used by the Proposer. Moreover, it also shows that they are able to evaluate the fairness of threats, as the strategies used can be interpreted as threats to reject offers lower than k.

Respondents were then asked a pair of questions regarding the credibility of two threats. The first question details a threat made by the Responder to reject any offer lower than 80% of the total amount, while the second describes a threat made by the Responder to reject any offer lower than 20%. While only 17 of the 63 students found the threat of rejecting any offer lower than 80% of the total amount to be credible, 43 found the threat credible when the threshold was 20%. This suggests that the credibility of a threat is related to its fairness, as most participants considered the threat to reject offers lower than 50% of the total amount as fair and the threat to reject offers higher or equal to 50% of the money as unfair.

3. The model

I define the kindness of a strategy as a function of the maximum payoff that it offers to the opposing players. I maintain that, in order to evaluate the kindness of a strategy, we have to take into account

not only the payoff we believe an opposing player would receive, but also the opportunity to choose a high payoff.

The present study analyzes the case of finite games of perfect information (see Fudenberg and Tirole, 1991, Chapter 3). A_i is the set of (possible mixed) strategies for player i, $a_i \in A_i$ is a strategy for individual i, and $b_{ij} \in B_{ij}$ are the beliefs of individual i regarding the strategy of individual j. The space of actions is the same as the space of beliefs: $A_i = B_{ji}$. $\pi_i : A \to R$ are individual i's material payoffs given that $A = \prod_{i \in N} A_i$, and N is the set of players in the game. $a_i(h)$ is the action that the strategy a_i prescribes at history h for player i, while $a_i(-h)$ is the set of the actions that strategy a_i prescribes at every history with the exception of history h. We have that $a_i = (a_i(h), a_i(-h))$. $a_i|h$ is the same as strategy a_i, but involves playing history h with probability one.

The present study first defines an equitable payoff, in order to use it as a reference point for evaluating the kindness of a strategy. I propose that strategies that provide opposing players with a potentially higher payoff than the equitable payoff, be evaluated as kind by playing the strategy that maximizes one's own "material" payoffs. Moreover, I propose that strategies providing opposing players with a potentially lower payoff than the equitable payoff be evaluated as unkind. The present study solely identified the equitable payoffs from the set of efficient strategies, wherein an inefficient payoff cannot be equitable, given that the player is offering a lower payoff to both the opposing player and herself.

I define a player's strategy as efficient when no other strategy always provides every player a higher or equal payoff with strict inequality for at least one player. I use the definition of Dufwenberg and Kirchsteiger (2004) of an efficient strategy.

The set of efficient strategies, E_i, is given by the following conditions:

$$E_i = \{a_i \in A_i \mid \text{there exists no } a_i' \in A_i \text{ such that for all } h \in H \text{ and } i, j \in \{1,2\} \text{ we have that } \pi_j(a_i'(h), a_{-i}(h)) \geq \pi_j(a_i(h), a_{-i}(h)), \text{ with strict inequality for at least one } (h, a_{-i}(h), j)\}$$

Rabin (1993) and Dufwenberg and Kirchsteiger (2004) define the equitable payoff of a player as the average of the highest and the lowest material payoffs she can receive, given her own actions. For example, the SRE would evaluate an offer of 100% of the total amount as kind, unkind, or neutral, depending on the Responder's strategy, which is

not consistent with the results from the survey above.[5] I argue that an offer of 100% of the total amount in the Ultimatum Game should always be evaluated as kind, regardless of the Responder's strategy. In the survey discussed above, most of the participants chose the 50-50 split as the fair offer, suggesting that people think that the fair offer is 50% of the money, independently of their own actions. In order for the equitable payoff to be independent of the Responder's strategy, the definition of an equitable payoff used in the present study is based on the maximum payoff a player can potentially receive by playing the strategy that maximizes her own payoff. In order to simplify the analysis presented here, the model used in the present study was developed for two players.

The equitable payoff that player j believes is fair for player i to receive is given by:

$$\pi_{ji}^{e} = \frac{1}{2}\left[\max_{a_j \varepsilon A_j} \max_{a_j \varepsilon A_j} \{\pi_i (a_i, a_j)\} + \min_{a_j \varepsilon E_j} \max_{a_j \varepsilon A_j} \{\pi_i (a_i, a_j)\}\right]$$

I define the equitable payoff as the average of the highest and the lowest possible payoffs that player j can potentially give to player i (if player i chooses to maximize her own payoff). I use the equitable payoff of a player as a reference point to define what is the kindness toward that player. If a player receives the equitable payoff, then she is treated fairly. If a player receives a payoff higher than the equitable payoff, then she is treated kindly, and if she receives a payoff lower than the equitable payoff, then she is treated unkindly. I define the kindness of a player as the maximum payoff her strategy could potentially offer their counterpart, minus the equitable payoff.

Definition 1: the kindness of player i towards player j when playing strategy a_i is given by:

$$f_{ij} (a_i) = \max_{a_j \varepsilon A_j} \pi_j (a_i, a_j) - \pi_{ij}^{e}$$

Definition 2: player i's belief as to the level of kindness shown by player j toward her when she believes player i is playing strategy b_{ij} is given by the following:

[5] The SRE would evaluate an offer of 100% of the total amount as unkind when the Responders strategy is to reject an offer of 100% of the total amount and accept any offer below 100%. In this case, an offer of 100% would give the Responder a payoff of 0, which compares negatively to a 99% payoff were it to be offered by the Proposer.

$$\tilde{f}_{iji}\left(b_{ij}\right)=\max_{a_i \varepsilon A_i}\pi_i\left(a_i,b_{ij}\right)-\pi_{ji}^{e}$$

As explained above, an advantage of the model proposed in the present study is that it evaluates threats and promises, and not only independent actions. In the case of the Ultimatum Game, a threat to only accept offers of more than 50% of the total amount would be evaluated as unkind, as the maximum amount the Proposer can receive is lower than 50% if the Responder carries out her threat, while the most unfair threat of all is to reject any offer below 100% and only accept a 100% payoff. A threat to reject any offer lower than 50% and accept 50% or more would be evaluated as fair, as it gives the Proposer the opportunity of receiving 50% of the total amount; furthermore, threats that include accepting offers below 50% of the total amount would be evaluated by this model as kind.

In the Ultimatum Game, for example, an offer above 50% of the total amount would be defined as kind, given that the maximum amount the recipient can receive by accepting the offer is over 50%. Furthermore, offers lower than 50% would be evaluated as unkind, given that the maximum amount a recipient can receive is lower than 50%, which is consistent with the results of the survey conducted on the Ultimatum Game, the results of which are discussed above.

The definitions used in the present study have the advantage of correctly taking into account threats and promises, even if they are outside the equilibrium path. They have the further advantage of being simpler, as they do not evaluate the kindness of a strategy as a function of the Responder's strategy (that is why my definitions do not include second order beliefs.) One complication of my model is that it evaluates the kindness of a strategy not by the actions at separate histories, but by the strategy as a whole, thus complicating the analysis, as the actions must be ascertained at every possible history. As detailed in the subsequent section, the SRE tries to avoid this complication by evaluating a strategy by means of the actions it prescribes at one history, without considering the actions at other histories. This results in the SRE evaluating the same strategy differently, depending on the history it is evaluating, which, I argue, leads to the SRE to predict unreasonable solutions.

Once kindness and beliefs regarding kindness have been defined, the discussion then moves on to a definition of an individual's utility function.

Definition 3: The utility of individual i is given by:

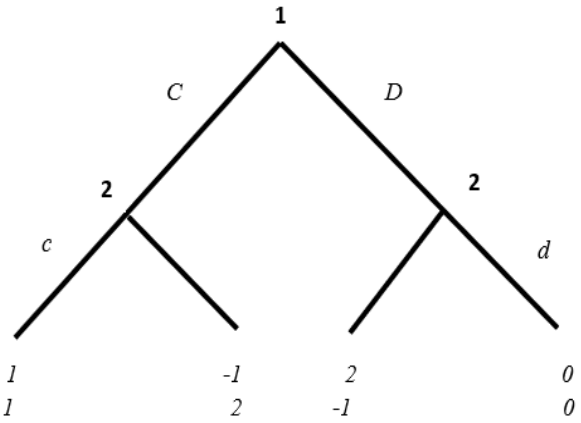
$$U_i\left(a_i, b_{ij}\right) = \pi_i\left(a_i, b_{ij}\right) - \lambda_i\Big(f_{ij}\left(a_i\right) - \tilde{f}_{iji}\left(b_{ij}\right)\Big)^2$$

where λ_i is a measure of how much importance individual i gives to emotions of reciprocity, which the present study will refer to as the emotional payoff, as it includes the utility from fairness. The emotional payoff enters the utility function as a subtraction of the square of the difference between the fairness of the players, an assumption made firstly for the sake of simplicity and secondly to represent the idea that individuals seek to reciprocate kindness and unkindness to the same degree. According to the SRE, individuals want to repay any offense with the most severe punishment possible.

One drawback of the utility function identified in the present study is that the existence of the FTE cannot be proven, as it cannot be guaranteed that the utility function is quasiconcave for every game, while an equilibrium has been found for every example examined.

The present study assumes that individuals are sophisticated, in the sense that they are fully aware of their own preferences and behavior and know how they would have behaved at other histories. For example, in the Prisoner's Dilemma presented in figure 3, at history C, Player 2 knows that by history D, she will be angry with Player 1.

Figure 3
Sequential Prisoners' Dilemma



Source: Based on Fig. 2 by Dufwenberg and Kirchsteiger (2004).

Definition 4: A Fair Threat Equilibrium of an extensive game with perfect information is a strategy profile a^*, such that for every player $i \in N$ and every non-terminal history $h \in H?Z$ for which $P(h) = i$, we have:

1) $a_i{}^*(h) \in argmax_{a_i(h) \in A_i(h)} U_i((a_i(h), a_i^*(-h)), b_{ij}|h)$

2) $b_{ij} = a_i{}^*$ for $j \neq i$

4. Comparison with the Sequential Reciprocity Equilibrium

This section presents a rewrite of Dufwenberg and Kirchsteiger's (2004) SRE and then compares it with the model used in the present study. In order to make it easier the comparison with my model, I will rewrite the SRE for two players (as I did for the FTE).[6]

In the sequential Prisoner's Dilemma presented in figure 3, Dufwenberg and Kirchsteiger (2004) argue that the emotion of reciprocity cannot render plausible unconditional cooperation by Player 2, as Player 2's strategy includes the promise to cooperate even if Player 1 defects. They develop their solution concept of a SRE order to eliminate this incredible promise, with a SRE requiring each player to optimize at every history. In the aforementioned sequential Prisoner's Dilemma, the SRE eliminates the incredible promise of unconditional cooperation by requiring that when Player 1 defects, Player 2 optimizes by also defecting.

One of the disadvantages of a SRE is that it evaluates the kindness of a strategy based purely on the player's actions at each history.[7]

For example, the SRE evaluates Player 2's *cd* strategy as kind if Player 1 plays C, or unkind if Player 1 plays D. The kindness of the strategy employed by Player 2 depends on the strategy employed by Player 1. In order to correctly take threats and promises into account, I argue that we should evaluate the kindness of a strategy as a whole, independently of other players' strategies.

The SRE uses the same definition for the set of efficient strategies as that defined above in Equation 1 for the FTE. Dufwenberg and

[6] I will also change some of their notation to make it easier to compare to my model.

[7] Falk and Fischbacher's (2005) solution concept of Reciprocity Equilibrium also evaluates each strategy based only on the actions it prescribes at each history.

 https://doi.org/10.24201/ee.v37i1.429

Kirchsteiger (2004) define an equitable payoff as the average of the highest and lowest of the payoffs in the set of efficient strategies.

What Player i believes is the equitable payoff for individual j, given that player i's beliefs about player j's action, is given by:

$$\pi_j^{e_i}(b_{ij}) = \frac{1}{2}\left[\max\left\{\pi_j(a_i, b_{ij}) \mid a_i \in A_i\right\} + \min\left\{\pi_j(a_i, b_{ij}) \mid a_i \in E_i\right\}\right]$$

The equitable payoff for Player j is a function of her own strategy (or what other players believe her strategy to be: b_{ij}), according to which definition, the equitable payoff for a player is the average of the highest and lowest of the efficient payoffs, given the same player's strategy. For example, the equitable payoff in the Dictator Game is for the recipient to receive 50% of the total amount. However, in the case of the Ultimatum Game, for the SRE, an equitable payoff is not necessarily a 50% share, given that the payoffs for the Responder depend not only on the Proposer's offer, but also on the Responder's decision on whether to accept or reject an offer, and any offer would pay zero to the Responder if she were to reject the offer. I contend that an offer of 50% should be regarded as the equitable payoff in the Ultimatum Game, independently of the strategy used by the Responder, wherein their responsibility for rejecting an offer lies with him and not the Proposer.

The SRE measures the kindness of a strategy based on the equitable payoff. If a strategy offers other players a higher payoff than the equitable payoff, it is classified as a kind strategy. In contrast, if it offers other players a lower payoff than the equitable payoff, it is classified as an unkind strategy.

Definition 5: the kindness of player i towards player j at history h is given by:

$$K_{ij}\left(a_i(h), b_{ij}(h)\right) = \pi_j\left(a_i(h), b_{ij}(h)\right) - \pi_j^{e_i}\left(b_{ij}(h)\right)$$

The above definition implies that the degree of kindness Player i believes she is extending to Player j depends on Player i's beliefs about the actions of Player j. The SRE evaluates how kind a player is toward another player based on that player's strategy.

Dufwenberg and Kirchsteiger's definition of the degree to which individual i believes other players are kind to her is set out below.

Definition 6: player i's beliefs about how kind player j is with her at history h is:

$$\tilde{K}_{iji}\left(b_{ij}\left(h\right),c_{iji}\left(h\right)\right)=\pi_{i}\left(b_{ij}\left(h\right),c_{iji}\left(h\right)\right)-\pi_{j}^{e_{j}}\left(c_{iji}\left(h\right)\right)$$

The above definition implies that the degree to which Player i believes Player j is kind to her depends on the beliefs of Player i about the actions of Player j and on c_{iji}, the beliefs of Player i about the beliefs of Player j about the actions of Player i, namely second order beliefs (the space of second order beliefs (C_{iji}) is the same as the space of beliefs and the space of actions: $C_{iji}=B_{ij}=A_{i}$).

Once kindness and belief have been defined with regard to kindness, an individual's utility function can then be defined.

Definition 7: The utility of individual i at history h is given by:

$$\begin{aligned}&U_{i}\left(a_{i}\left(h\right),b_{ij}\left(h\right),c_{iji}\left(h\right)\right)=\pi_{i}\left(a_{i}\left(h\right),b_{ij}\left(h\right)\right)\\&+\lambda_{i}\cdot K_{ij}\left(a_{i}\left(h\right),b_{ij}\left(h\right)\right)\cdot\tilde{K}_{iji}\left(b_{ij}\left(h\right),c_{iji}\left(h\right)\right)\end{aligned}$$

where λ_{i} is a measure of how much importance individual i gives to the emotion of reciprocity.

In the equilibrium, the SRE requires that all players optimize at every history and that their beliefs and second order beliefs are correct.

Definition 8: The strategy profile a^{*} is a Sequential Reciprocity Equilibrium if all $i\in N$ and for every $h\in H$ it holds that:

1) ${a_{i}}^{*}\left(h\right)\in argmax_{a_{i}\in A_{i}(h,a^{*})}U_{i}\left(a_{i}\left(h\right),b_{ij}\left(h\right),c_{iji}\left(h\right)\right)$

2) $b_{i}j={a_{j}}^{*}$ for $j\neq i$,

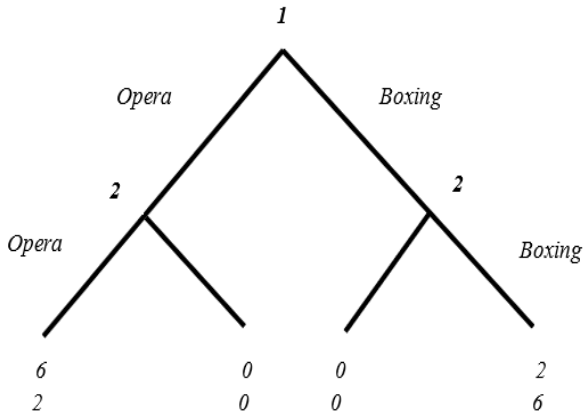
3) $c_{iji}={a_{i}}^{*}$ for $j\neq i$.

The main difference between a SRE and an FTE is that a SRE evaluates the kindness of a strategy by the separate actions it prescribes at each history, while an FTE evaluates the strategy as a whole. The problem with using a SRE as an approach is that it may evaluate the same strategy differently, depending on the history at which it is evaluated. For example, in the Battle of the Sexes game presented in figure 4, the SRE determines the strategy $(Opera,Opera)$ to

be kind at history *Opera*, but as unkind at history *Box*. Although the SRE requires that players optimize at every history, the fact that the strategies are evaluated as independent parts enables the SRE to predict some solutions that I argue are unreasonable.

A further difference between a SRE and an FTE is that the former evaluates the kindness of each strategy as a function of the strategies of other players, while the FTE evaluates each strategy independently of the strategies of other players. If a player receives a low material payoff in the SRE, the blame falls on the opposing players, even if it is the player receiving the low payoff that is minimizing her own payoff. I argue that this characteristic of the SRE also leads it to predict some unreasonable solutions.

Figure 4
Battle of the Sexes



Source: As reported by Muller and Sadanand (2003).

5. Examples

5.1 *Prisoners' Dilemma*

This section presents an analysis of the FTE applied for the Prisoner's Dilemma presented in figure 3. Player 2 has two emotional states, one at history C (cooperate) and one at history D (defect), which will be referred here as Emotional State C and Emotional State D,

respectively. In Emotional State C, Player 2 wants to be kind in return for Player 1's kindness, while, in Emotional State D, Player 2 wants to be unkind in return for Player 1's unkindness.

Observation 1: in every FTE in the sequential Prisoners' Dilemma shown in figure 3, if Player 1 plays D, Player 2 will play d. All proofs are in the Appendix.

Note that the FTE determines strategy D, which is employed by Player 1, as unkind, given that the highest possible material payoff Player 2 can receive is zero, which is lower than the equitable payoff. Player 2 maximizes both his material and emotional payoffs by playing d at history D, in which Player 2 is unkind in response to the unkindness of Player 1 and also maximizes his own material payoffs.

Observation 2: in every FTE in the sequential Prisoners' Dilemma shown in figure 3, if Player 1 plays C, Player 2 plays c if $\lambda_2 > 1/3$.

The FTE determines strategy C, used by Player 1, as kind, given that it offers Player 2 the possibility of obtaining the highest possible material payoffs by playing d. If Player 2 cares enough about fairness, then he would be willing to sacrifice his own material payoffs and play c, in order to repay Player 1's kindness, where, if not, he would play d in order to maximize his material payoffs.

While the SRE makes the same predictions as my model for the Prisoner's Dilemma, the following examples show that the SRE predicts some unreasonable solutions that the FTE does not allow.

5.2 *Battle of the Sexes*

The SRE enables the strategy profile $(Box, (Opera, Opera))$ to be an equilibrium for the Battle of the Sexes shown in figure 4, which I maintain is not a reasonable prediction.

Proposition 1: In the Battle of the Sexes shown in figure 4, if $\lambda_1 \geq 3/2$ and $\lambda_2 \geq 3$, the strategy profile $(Box, (Opera, Opera))$ is a SRE.

I argue that the foregoing solution is unreasonable, in that, while Player 2's strategy $(Opera, Opera)$ includes a promise to play $Opera$ if Player 1 plays $Opera$, the SRE does not take this promise into account, as it is outside the equilibrium path. Because this strategy

includes the promise that provides Player 1 the possibility of obtaining the highest possible material payoffs available in the game, if Player 1 plays *Opera*, I maintain that this should be evaluated as a kind strategy. Moreover, I argue that the Strategy Box used by Player 1 should be evaluated as kind, as it gives Player 2 the opportunity to obtain the highest material payoffs.

The SRE determines the strategies employed by both players to be unkind to each other, as both are offering each other a payoff of zero when they could have offered each other a positive payoff. As both players are choosing to minimize their own material payoffs, I argue that it is unreasonable that the SRE determines both players' strategies to be unkind. However, the SRE evaluates the kindness of each strategy as a function of other players' strategies, wherein, if a player sabotages her own material payoffs, the SRE would not blame her but would blame the opposite player, as that player played the strategy that enabled this sabotage.

The FTE eliminates the possibility of the aforementioned solution and predicts that Player 2 would play *Box* when Player 1 plays *Box*.

Observation 3: in every FTE, if Player 1 plays *Box*, Player 2 plays *Box*.

It should be noted that, in the FTE, the strategy *Box*, used by Player 1, would be determined to be kind, as it gives Player 2 the possibility of maximizing his own material payoffs by playing *Box* himself. The FTE predicts that Player 2 would play *Box* at history *Box*, in order not to be unkind to Player 1, while, at the same time, maximizing his own material payoffs.

Muller and Sadanand (2003) conducted an experiment with the sequential Battle of the Sexes shown in figure 4,[8] using the direct method, where Player 2 chooses the action he will take after observing the action taken by Player 1.[9] Muller and Sadanand found that 115 of the 120 participants acting as Player 1 chose to play *Opera*, while, of the 115 participants acting as Player 2 that observed *Opera*, 103

[8] The strategies in their experiment are named A and B for Player 1, and a and b for Player 2.

[9] An alternative method is the strategy method, in which Player 2 selects all the actions he would choose at every history, after which, Player 1 selects her action. The advantage of the strategy method over the direct method is that it reveals Player 2's complete strategy, while the direct method reveals only one action for Player 2.

chose to play *Opera* and 12 chose to play *Box*. Only five of those acting as Player 1 chose to play *Box*, in all of which cases, their opposing players chose to play *Box*, as predicted in Observation 1.

Although there are few instances in which Player 1 plays *Box*, the actions that correspond to the SRE, namely $(Box, (Opera, Opera))$, are not observed in this experiment.

5.3 *The Dictator Game*

In the Dictator Game, a player, the Dictator, chooses to divide an amount of money between herself and another player, the Recipient, after which the game ends, and both players are paid the share decided upon by the Dictator. The SRE predicts that the Dictator would keep all the money for herself, while, given that the Recipient does not have any choice to make, his kindness toward the Dictator is zero. Therefore, the emotional payoff of the utility function is zero for the Dictator, and according to the SRE, she should maximize her own material payoff and keep all the money for herself.

I will represent the Dictator's offer as s. Then the material payoff of the Dictator is $\pi_D = 100 - s$, and the material payoff of the Recipient (R) is $\pi_R = s$. The FTE stipulates that the equitable payoff for the Recipient is 50% of the money available. I will assume that the Dictator chooses to divide 100 units of money. Then, $\pi^e_{DR} = 50$ with every offer higher than 50 considered kind and every offer lower than 50 considered unfair. The kindness of the Dictator towards the Recipient is $f_{DR} = s - 50$, and given that the Recipient does not take any action, the beliefs of the Dictator of the kindness of the Recipient is $\tilde{f}_{DRD} = 0$.

The utility of the Dictator is $U_D = \pi_D - \lambda_D\left(f_{DR} - \tilde{f}_{DRD}\right)^2$. Solving the First Order Conditions reveals that the Dictator maximizes her utility when $s = 50 - \frac{1}{2\lambda_D}$. The more important the emotion of reciprocity is for the Dictator (λ_D), the less unkind she wants to be to the Recipient.

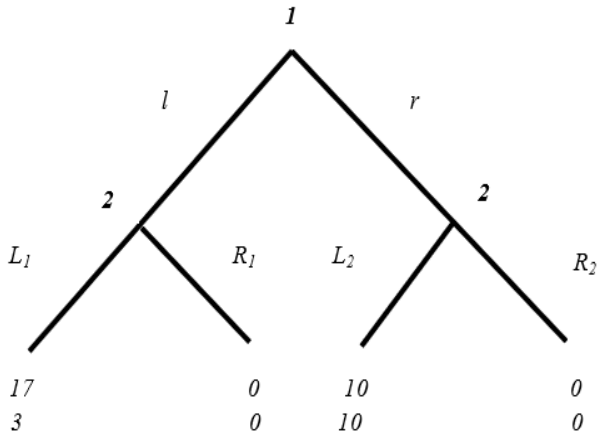
The results obtained using my model correspond to the experimental findings reported in the literature more accurately than those obtained by means of the SRE, which predicts that the Dictator will not give anything to the Recipient. According to Engel (2011), Dictators give 28% of the money available on average, with 36% of Dictators giving nothing to the Recipient and 64% giving something. Of those who give something, 16% opt for an equal split with the Recipient. In a game where the amount of money to be shared is 100, a Dictator

 https://doi.org/10.24201/ee.v37i1.429

giving the average according to Engel (28% of the total amount), with a utility function calculated via the FTE, would have a λ_D of 1/44. Representing a Dictator who decides not to share any of the money requires $1/100 > \lambda_D$, while representing a Dictator who opts to split the money requires λ_D to be arbitrarily large, which I recognize is unrealistic.[10]

5.4 *Mini-ultimatum Game*

Figure 5
Mini Ultimatum Game



Game "Equal". Source: Güth et al. (2001).

Figure 5 shows the Mini-ultimatum Game. I found one SRE in this game: $(r, (L_1, R_2))$ to be unreasonable, wherein Player 2 is accepting the low offer (L_1) and rejecting the high offer (R_2). If Player 2 rejects the high offer, thus receiving a material payoff of zero, and accepts the low offer, thus receiving a positive amount, the SRE determines the high offer to be unkind and the low offer to be kind.

[10] In order to represent that some people want to be nice to others, even when they are not nice to them in return, their utility function could be modified by adding a constant to the emotional component. For example, if the utility function of the Dictator were $U_D = \pi_D - \lambda_D\left(f_{DR} - \tilde{f}_{DRD} - \alpha\right)^2$, where α is a positive constant, some individuals with a high λ_D and high α would want to split the money with the recipient.

Therefore, given his own actions, the low offer gives Player 2 the highest possible payoff, while the high offer gives Player 2 the lowest possible payoff. If fairness is sufficiently important to him, Player 2 would want to be unkind to Player 1 (by rejecting the high offer) when Player 1 is unkind to him and would, moreover, want to be kind to Player 1 (by accepting the low offer) when she is kind to him.

Proposition 2: In the Mini-ultimatum game shown in figure 5 the strategy profile $(r, (L_1, R_2))$ is a SRE when $\lambda_2 \geq \frac{2}{3}$ and $\lambda_1 \geq \frac{17}{30}$.

Player 2's strategy (L_1, R_2) includes the promise to accept the low offer, which I argue should be determined to be kind, as it offers Player 1 the possibility of receiving the highest possible payoffs in the game. However, because this promise is outside the equilibrium path, the SRE does not take it into consideration and only evaluates the strategy by the action in the equilibrium path, which is to reject the high offer. The FTE does not allow this equilibrium, as it determines the low offer to be unkind and the high offer to be kind and evaluates the strategy used by Player 1 independently of the strategy used by Player 2.

Observation 4: in every FTE in the Mini-ultimatum Game shown in figure 5, if Player 1 plays r, Player 2 plays L_2.

The FTE determines the high offer (r) to be kind. Given that Player 1 is being kind by playing r, Player 2 wants to be kind in return by accepting the offer, which also maximizes his material payoffs. However, as he may already have shown kindness if his strategy included the promise of accepting the low offer, Player 1's behavior at history ℓ must be ascertained. Ascertaining all the possible actions Player 2 may take at history ℓ, including the promise of accepting the low offer, reveals that Player 2 always wants to accept the high offer (L_2).

Observation 5: in every FTE in the Mini-ultimatum Game shown in figure 5, if Player 1 plays ℓ, Player 2 plays L_1.

According to the FTE, if Player 1 makes the low offer (ℓ), she is being unkind and, in order to maximize his emotional payoffs, Player 2 wants to repay Player 1's unkindness. However, in this game, Player 2 is left with the sole possibility of being much unkinder than Player 1 could possibly be, which also reduces Player 2's emotional payoff. Therefore, Player 2 would prefer to accept the low offer.

Güth *et al.* (2001) conducted an experiment with three variations of the Mini-ultimatum game shown in figure 5, using both the direct and strategy methods, in which they did not observe the SRE $(r, (L_1, R_2))$ being played. Twenty-six participants used the strategy method to play the game shown in figure 5, while none of the 13 participants that played as recipients used the strategy (L_1,R_2), nor did any of the 62 participants who played as recipients using all variations in the strategy method.

Of all of the choices made by the 267 recipients across all variations of the Mini-ultimatum Game and all the elicitation methods used in the experiment, only five chose the action R_2 at history r, while 262 choices taken at history r were L_2, as predicted by Observation 5 above. This suggests that the SRE $(r, (L_1, R_2))$ is not a good prediction of how individuals play the Mini-ultimatum Game.

6. Conclusions

Reciprocity is a complex concept. Dufwenberg and Kirchsteiger's (2004) extend Rabin's (1993) concept of fairness equilibrium to sequential games. However, the SRE evaluates the kindness or unkindness of a strategy based purely on the actions it prescribes at the equilibrium path. However, the actions that a strategy prescribes outside the equilibrium path include threats and promises, which the SRE does not take into account.

The aim of this article was to develop a concept of sequential reciprocity that provides more reasonable predictions when threats and promises are involved. My concept of FTE evaluates the kindness, not of actions, but of whole strategies, which I argue is fundamental in order to correctly account for threats and promises.

The present article has shown how a SRE predicts some unreasonable equilibria for the Battle of the Sexes, the Dictator Game, and the Mini-ultimatum Game. Moreover, it has shown how my FTE solution concept does not allow such unreasonable equilibria. It also shows that the experimental results reported in the literature support my solution concept.

My solution concept may be used in future research to analytically ascertain whether explicit promises and threats provide a commitment. It may be possible for individuals to use this commitment to force themselves to behave in line with their desired behavior.

My solution concept may also be useful for analyzing repeated games, in that, as the collusion in repeated games is sustained by threats, the FTE is a good solution concept for these types of games.

Alejandro Tatsuo Moreno-Okuno: atatsuo@ugto.mx

References

Bolton G.E. and A. Ockenfels. 2000. ERC: A theory of equity, reciprocity, and competition, *American Economic Review*, 90(1): 166-193.

Dufwenberg, M. and G. Kirchsteiger. 2004. A theory of sequential reciprocity, *Games and Economic Behavior*, 47(2): 268-298.

Engel, C. 2011. Dictator games: A meta study, *Experimental Economics*, 14: 583-610.

Falk, A. and U. Fischbacher. 2006. A theory of reciprocity, *Games and Economic Behavior*, 54(2): 293-315.

Fehr E. and K. Schmidt. 1999. A theory of fairness, competition, and cooperation, *The Quarterly Journal of Economics*, 114(3): 817-868.

Fudenberg, D. and J. Tirole. 1991. *Game Theory*, Cambridge MA, MIT Press.

Güth, W., S. Huck, and W. Müller. 2001. The relevance of equal splits in ultimatum games, *Games and Economic Behavior*, 37(1): 161-169.

Klein, D.B. and B. O'Flaherty. 1993. A game-theoretic rendering of promises and threats, *Journal of Economic Behavior and Organization*, 21(3): 295-314.

Levine, D.K. 1998. Modeling altruism and spitefulness in experiments, *Review of Economic Dynamics*, 1(3): 593-622.

Muller, R.A. and A. Sadanand. 2003. Order of play, forward induction, and presentation effects in two-person games, *Experimental Economics*, 6(1): 5-25.

Rabin, M. 1993. Incorporating fairness into game theory and economics, *American Economic Review*, 83(5): 1281-1302.

Schelling, T.C. 1960. *The Strategy of Conflict*, Cambridge MA, Harvard University Press.

Appendix

All the propositions set out below relate to the SRE and are proven using the definitions of the SRE outlined in section 4 above, while all the observations set out below relate to the FTE and are proven using the definitions of the FTE outlined in section 5 above.

Prisoners' Dilemma

Proof of Observation 1

Note that, for the FTE, the equitable payoffs in the Prisoner's Dilemma for both players is 1, as the highest possible payoff that a player can

offer the other player is 2 and the lowest is 0. The kindness of Player 1 when playing C is 1, while the kindness of Player 1 when playing D is -1. The kindness of Player 2 when playing cc is 1, while this is -1 when playing dd, 0 when playing cd, and 1 when playing dc.

Below, I show that, at history D, the material and emotional payoffs for Player 2 are maximized by playing d, no matter what Player 2 would have done at history C. Firstly, I show that Player 2 would play d at history D if he believed that he would have played c at history C. I then show that Player 2 would also play d when he would have played d at history C.

∘ Player 2 at history D:

At history D, Player 1 is being unkind to Player 2 and Player 2 is maximizing his emotional payoffs by being unkind in return. If Player 2 believes he would have played c at history C, Player 2 can choose between strategies cc and cd. The utility of Player 2 playing cc at history D is $U_2(D, cc) = -1 - \lambda_2(-1-1)^2 = -1 - 4\lambda_2$ and $U_2(D, cd) = 0 - \lambda_2(-1-0)^2 = 0 - \lambda_2$ when playing cd. The utility for Player 2 is always higher when playing cd.

If Player 2 believes he would have played d at history C, he can choose between strategies dc and dd. The utility for Player 2 of playing dc at history D is $U_2(D, dc) = -1 - \lambda_2(-1-1)^2 = -1 - 4\lambda_2$ and is $U_2(D, dd) = 0 - \lambda_2(-1-(-1))^2 = 0$ when playing dd. The utility for Player 2 is always higher when playing dd.

The material and emotional payoffs are higher when he plays d at history D, as he wants to repay unkindness with unkindness. This proves Observation 1.

Proof of Observation 2

As set out above, Player 2 will always play d at history D; therefore, we only have to ascertain which of cd and dd is the optimal strategy.

∘ Player 2 at history C:

At history C, Player 1 is kind to Player 2, and Player 2 maximizes his emotional payoff by being kind in return. However, Player 2 will be kind to Player 1 only if it is not too costly. If Player 2 plays d at history D, he can only choose between playing strategies cd and dd. The utility for Player 2 at history C if he plays cd is $U_2(C, cd) =$

$1 - \lambda_2(1-0)^2$ and is $U_2\,(C, dd) = 2 - \lambda_2(1-(-1))^2 = 2 - \lambda_2(2)^2$ if he plays *dd*. The utility for Player 2 is higher when he plays *cd* only when $\lambda_2 > 1/3$. When $\lambda_2 < 1/3$ the utility for Player 2 is higher when he plays *dd* and presents indifference to either strategy when $\lambda_2 = 1/3$. This proves Observation 2.

Battle of the Sexes

Proof of Proposition 1

To show that the strategy profile $(Box, (Opera, Opera))$ is a SRE in the Battle of the Sexes when $\lambda_1 > 3/2$ and $\lambda_2 > 3$, we have to show that the best strategy for Player 1 is to play *Box* and that, at both histories (*Box* and *Opera*), playing *Opera* is the best strategy for him.

∘ Player 1:

Analysis of Player 1's strategy reveals that, given that Player 2 is playing $(Opera, Opera)$, Player 1 can give Player 2 a payoff of 2, by playing *Opera*, or a payoff of 0, by playing *Box*. By playing *Opera*, Player 1 is being fair (neither kind nor unkind), given that, by playing *Opera*, she is maximizing her own payoffs and the payoffs for Player 2. Therefore, the kindness of Player 1 is zero if she plays *Opera*.

By playing *Box*, Player 1 is being unkind to Player 2, given that, by playing *Box*, Player 1 is sacrificing her own material payoffs to reduce the material payoffs for Player 2.

The utility of Player 1 by playing *Box* is:

$$U_1\,(Box, (Opera, Opera)) = 0 + \lambda_2\,(-2)\,(-1) = 2\lambda_1$$

while the utility for Player 1 of playing *Opera* is:

$$U_1\,(Opera, (Opera, Box)) = 6 + \lambda_2\,(-2)\,(1) = 6 - 2\lambda_1$$

From these equations we obtain that Player 1 wants to be unkind, by playing *Opera* if λ_1 is higher or equal to 3/2.

 https://doi.org/10.24201/ee.v37i1.429

∘ Player 2 at history *Opera*:

By playing *Opera*, Player 2 is being fair to Player 1 if Player 1 plays *Opera*. By playing *Box*, Player 2 is being unkind, given that he is sacrificing his own payoffs in order to minimize the payoffs for Player 1.

The utility of Player 2 playing *Opera* at history *Opera* is:

$$U_2\left(Opera,\left(Opera,Opera\right)\right)=2+\lambda_2\left(1\right)\left(0\right)=2$$

while the utility of Player 2 playing *Opera* at history *Box* is:

$$U_2\left(Opera,\left(Box,Opera\right)\right)=0+\lambda_2\left(1\right)\left(-6\right)=-6\lambda_2$$

revealing that the utility of Player 2 by playing *Opera* is higher when Player 1 plays *Opera*.

∘ Player 2 at history *Box*:

By playing *Box*, Player 1 is being unkind to Player 2, giving him a payoff of zero. Because Player 1 is sacrificing her own material payoffs by playing *Box*, Player 2 would interpret Player 1 playing *Box* as a very unkind action, and Player 2 would want to retaliate by being unkind in return, by playing *Opera* if he cares enough about reciprocity.

The utility of Player 2 playing *Opera* at history *Box* is:

$$U_2\left(Box,\left(Opera,Opera\right)\right)=0+\lambda_2\left(-1\right)\left(-2\right)=2\lambda_2$$

while the utility of Player 2 playing *Opera* at history *Box* is:

$$U_2\left(Box,\left(Opera,Box\right)\right)=6+\lambda_2\left(-1\right)\left(0\right)=6$$

The utility of Player 2 playing *Opera* is higher or equal, but only if λ_2 is higher or equal than 3.

Proof of Observation 3

Below, I show that, for the FTE at history *Box*, the material payoffs and emotional payoffs for Player 2 are maximized by playing *Box*, no matter what Player 2 would have done at history C.

Note that the equitable payoff for Player 2 is 4 (as the highest and lowest possible payoffs that Player 1 can give Player 2 are 6 and 2, if she maximizes her payoffs) and 6 for Player 1 (as the only efficient strategy is ($Opera$,Box), which gives a maximum payoff of 6 to Player 2). The kindness of Player 1 when playing $Opera$ is -2, while the kindness of Player 1 when playing Box is 2. For Player 2, the kindness of playing $(Opera, Opera)$ is 0, while this is 0 when playing $(Opera, Box)$, -4 when playing (Box, Box),and -6 when playing $(Box, Opera)$.

∘ Player 2 at history Box:

At history Box, Player 1 is being kind to Player 2, while Player 2 maximizes his emotional payoffs by being kind in return. Given that the optimal action of a player at one history may depend on the actions of the same player at other histories, the optimal action of Player 2 at history Box must be ascertained, as must his two possible actions at history $Opera$.

If Player 2 believes he would have played $Opera$ if Player 1 had played $Opera$, Player 2 can choose between strategies $(Opera, Opera)$ and $(Opera, Box)$. The utility of Player 2 playing $(Opera, Opera)$ at history $Opera$ is $U_2\left(Box, (Opera, Opera)\right) = 0 - \lambda_2(2-0)^2 = 0 - 4\lambda_2$ and $U_2\left(Box, (Opera, Box)\right) = 6 - \lambda_2(2-0)^2 = 6 - 4\lambda_2$ when playing $(Opera, Box)$. The utility of Player 2 is always higher when playing $(Opera, Box)$.

If Player 2 believes that he would have played Box if Player 1 had played Box, Player 2 can choose between the strategies (Box, Box) and $(Box, Opera)$. The utility of Player 2 playing (Box, Box) at history $Opera$ is $U_2\left(Box, (Box, Box)\right) = 6 - \lambda_2(2-(-4))^2 = 6 - 36\lambda_2$ and $U_2\left(Box, (Box, Opera)\right) = 0 - \lambda_2(2-(-6))^2 = 0 - 64\lambda_2$ when playing $(Box, Opera)$. The utility of Player 2 is always higher when playing (Box, Box).

As we can see, the utility of Player 2 is always higher when playing Box if Player 1 plays Box, no matter what he believes his action at history $Opera$ would be. This proves Observation 3.

Mini-ultimatum Game

Proof of Proposition 2

 https://doi.org/10.24201/ee.v37i1.429

In order to ascertain whether the strategy profile $(r, (L_1, R_2))$ is a SRE in the Mini-ultimatum Game, it must be verified that playing L_1 at history ℓ is the best action and that playing R_1 at history r is the best action for Player 2, and that the best strategy for Player 1 is to play r.

Given that Player 2 is playing (L_1,R_2), Player 1 can give Player 2 a payoff of 3, by playing ℓ, or a payoff of 0, by playing r. At history ℓ, Player 1 is being fair (neither kind nor unkind), given that, by playing ℓ, she is maximizing her payoffs and the payoffs of Player 2. Therefore, the kindness of Player 1 is zero.

∘ Player 2 at history ℓ:

At history ℓ, the utility of Player 2 playing L_1 is:

$$U_2\left(\ell, (L_1, R_2)\right) = 3 + \lambda_2 \left(3 - 1.5\right)\left(17 - 17\right) = 3$$

while the utility of Player 2 playing R_1 at history ℓ is:

$$U_2\left(\ell, (R_1, R_2)\right) = 0 + \lambda_2 \left(3 - 1.5\right)\left(0 - 17\right) = 0$$

As we can see, the utility for Player 2 is higher when playing L_1.

∘ Player 2 at history r:

At history r, Player 1 is being unkind to Player 2. Because Player 1 is sacrificing her own material payoffs by playing r (which gives Player 2 his lowest payoff), Player 2 would interpret r as an unkind action. Given this, Player 2 wants to retaliate by being unkind in return by playing R_2, if λ_2 is large enough.

At history r, the utility for Player 2 of playing R_2 is:

$$U_2\left(r, (L_1, R_2)\right) = 0 + \lambda_2 \left(0 - 1.5\right)\left(0 - 10\right) = 15\lambda_2$$

while the utility for Player 2 of playing L_2 at history ℓ is:

$$U_2\left(r, (L_1, L_2)\right) = 10 + \lambda_2 \left(0 - 1.5\right)\left(10 - 10\right) = 10$$

As we can see, the utility of playing R_2 for Player 2 is higher if $\lambda_2 \geq \frac{2}{3}$.

∘ Player 1:

Given that Player 2 is playing (L_1,R_2), Player 1 can either play ℓ and be fair or play r and be unkind. Note that the material payoff she can give to Player 2 by playing ℓ is 3, while this is zero when playing r. The equitable payoff for Player 2 is 3 (when Player 1 plays ℓ), given that, by doing so, Player 1 is also receiving her highest possible material payoff. If Player 1 plays r, she is being unkind, given that she is giving a payoff of zero, and sacrificing her own material payoffs to do so. The SRE evaluates the strategy (L_1,R_2) to be unkind, given that Player 2 is playing R_2 at history r, minimizing both the material payoff for Player 1 and her material payoffs at the same time.

The utility of Player 1 if playing r is:

$$U_1\left(r,\left(L_1,R_2\right)\right)=0+\lambda_1\left(0-10\right)\left(0-1.5\right)=0+15\lambda_1$$

while the utility for Player 1 of playing ℓ is:

$$U_1\left(\ell,\left(L_1,R_2\right)\right)=17+\lambda_1\left(0-10\right)\left(3-1.5\right)=17-15\lambda_1$$

As we can see, Player 1 plays r if $\lambda_1 \geq \frac{17}{30}$.

Proof of Observation 4

Note that, for the SRE, the equitable payoff for Player 2 is 6.5, as the highest and lowest possible payoffs that Player 1 can give Player 2 are 10 and 3, respectively). The equitable payoff for Player 1 is 17, as the only efficient strategy for Player 2 is (L_1,L_2). The kindness of Player 1 when playing ℓ is -3.5, while the kindness of Player 1 when playing r is 3.5. For Player 2, the kindness of playing (L_1,L_2) is 0, while this is 0 for playing (L_1,R_2), -7 for playing (R_1,L_2), and -17 for playing (R_1,R_2).

Below, I show that, at history r, the material and emotional payoffs for Player 2 are maximized by playing L_2, no matter what Player 2 would have done at history ℓ.

∘ Player 2 at history r:

At history r, Player 1 is being kind to Player 2, while Player 2 maximizes his emotional payoffs by being kind in return. If Player 2 believes he would have played L_1 if Player 1 had played ℓ, he can choose

 https://doi.org/10.24201/ee.v37i1.429

between strategies (L_1,L_2) and (L_1,R_2). The utility for Player 2 of playing (L_1,L_2) at history r is $U_2(r, L_1, L_2) = 10 - \lambda_2(3.5-0)^2 = 10-12.25\lambda_2$ and $U_2(r,(L_1,R_2)) = 0-\lambda_2(3.5-0)^2 = -12.25\lambda_2$ when playing (L_1,R_2). The utility for Player 2 is always higher when playing (L_1,L_2).

If Player 2 believes he would have played R_2 if Player 1 had played ℓ, Player 2 can choose between strategies (R_1,L_2) and (R_1,R_2). The utility for Player 2 of playing (R_1,L_2) at history r is $U_2(r,(R_1,L_2)) = 10 - \lambda_2(3.5-(-7))^2 = 10 - 110.25\lambda_2$ and $U_2(r,(R_1,R_2)) = 0 - \lambda_2(3.5-(-17))^2 = 0 - 420.25\lambda_2$ when playing (R_1,R_2). The utility for Player 2 of playing L_2 at history r is always higher.

As can be seen, Player 2 will always play L_2 if Player 1 plays r, no matter what he believes his own intentions will be at history ℓ. This proves Observation 4.

Proof of Observation 5

At history ℓ, Player 1 is being unkind to Player 2 and Player 2 maximizes her emotional payoffs by being unkind in return. As in Observation 4, Player 2 plays L_2 at history r. Then Player 2 can choose between strategies (L_1,L_2) and (R_1,L_2). The utility for Player 2 of playing (L_1,L_2) at history ℓ is $U_2(r,(L_1,L_2)) = 3 - \lambda_2(-3.5-0)^2 = 3-12.25\lambda_2$ and $U_2(r,(R_1,L_2)) = 0-\lambda_2(-3.5-(-7))^2 = 0-12.25\lambda_2$ when playing (L_1,R_2). The utility for Player 2 of playing (L_1,L_2) is always higher; therefore, Player 2 plays L_1 at history ℓ.