Games With Possibly Naive Present-Biased Players[*]

Marco A. Haan[†] Dominic Hauck[‡]

April 29, 2021

Abstract

We propose a solution concept for games that are played among players with present-biased preferences that are possibly naive about their own, or about their opponent's future time inconsistency. Our perception-perfect outcome essentially requires each player to take an action consistent with the subgame perfect equilibrium, given her perceptions concerning future types, and under the assumption that other present and future players have the same perceptions. Applications include a common pool problem and Rubinstein bargaining. When players are naive about their own time inconsistency and sophisticated about their opponent's, the common pool problem is exacerbated, and Rubinstein bargaining breaks down completely.

Keywords: Present-biased preferences, naivety, common pool, bargaining

JEL Codes C72, C78, D03, D91

[*] Earlier versions of this paper circulated under the title *Games with Possibly Naive Hyperbolic Discounters.*

[†] Corresponding author. Faculty of Economics and Business, University of Groningen, PO Box 800, 9700 AV Groningen, the Netherlands. `m.a.haan@rug.nl`

[‡] Faculty of Economics and Business, University of Groningen, the Netherlands.

1 Introduction

Time-inconsistent present-biased preferences are among the most prominent and persistent behavioral biases in economics. For example, most people would prefer to do an unpleasant task on May 1 rather than on May 15 when faced with that choice on April 1. But on May 1, almost everyone will be inclined to postpone it to May 15. This type of time inconsistency (often also referred to as hyperbolic discounting) has been put forward as an explanation of, for example, why economic agents would choose to use commitment devices to restrict their future selves.[1] O'Donoghue and Rabin (1999) provide a model for behavior with such present-biased preferences. In their model, an individual decision-maker can be time-consistent, or she can have present-biased preferences. Importantly, she can either be sophisticated concerning her time inconsistency, or she can be naive. A sophisticated individual knows that she will have present-based preferences in the future, and hence will have an incentive today to restrict the choices of that future self. If she is naive, then she believes that although her current self has present-biased preferences, her future self will behave in a time-consistent manner.

Many situations that are of interest to economists, however, concern the interaction between economic agents. Consider for example the case in which two individuals, A and B, bargain over the distribution of some future payoff. As in the simple one-player model, player A's behavior will depend on whether she is present-biased and, if so, whether she is naive or sophisticated about that. However, her behavior will also depend on whether she perceives player B to be time-consistent or not, and whether she believes player B is naive or sophisticated. It may even depend on her perceptions concerning player B's perceptions about player A. Where the one-player model implies a game played between a current and future self, a two-player model effectively implies a game played between both A and B's current and future selves.

In this paper, we study such games. We introduce a solution concept for games played between possibly present-biased players. As a starting point, we take O'Donoghue and Rabin (1999). They consider a one-player game played by a current self against her future self. In their model, players have to decide whether to do a task now, or to do it later. The authors introduce the concept of a perception-perfect strategy, which essentially is a course of action that maximizes the current player's utility given her perception about the type of her future self, and the behavior she rationally expects from such a type. Possible types then refer to whether the future self will or will not have present-biased preferences. We

[1] For a survey, see e.g. Frederick et al., 2002.

first extend the analysis to one-player games with a richer strategy space, both in the two-period case as well as in a set-up with more periods. We introduce a perception-perfect outcome,[2] which is an extension of O'Donoghue and Rabin (1999)'s perception perfect strategy, that can also be applied to a multi-player set-up. We then analyze games with two players. We apply our solution concept to a common pool problem (the overproduction that results when competitors seek to exploit an exhaustible resource), and to a model of Rubinstein bargaining (where two players take turns in either accepting their counterpart's offer or making a counteroffer).

Players' perceptions concerning types are going to play a crucial role. Moreover, behavior will depend not only on player A's perception about player B, but also on player A's perception about player B's perception about player A, etcetera. To deal with this complication we impose, first, that players assume that future incarnations of themselves have the same perceptions as their current self has (*intraplayer perception naivety*). Second, we impose that players assume that *other* players have the same perceptions as they themselves have (*interplayer perception naivety*).

Our concept of the perception perfect outcome then entails the following. Consider player A. She has certain perceptions about her own future type, and about the future type of the other player. Given those perceptions, and under the assumption that all other present and future players have the same perceptions, we can derive the subgame perfect equilibrium that player A perceives to be played. We call this the equilibrium as perceived by player A. Similarly, we can derive the equilibrium as perceived by player B. The perception perfect outcome in period $t = 1$ then consists of an action taken by player A that is consistent with an equilibrium as perceived by A, and an action taken by player B that is consistent with an equilibrium as perceived by B. In all later periods the same is true, but given the actions that were played in the past.

From our two main applications, the common pool problem and Rubinstein bargaining, we derive the following insights. First, suppose that players are naive about their own future selves, but are sophisticated about the future self of others. This is consistent with psychological evidence, as e.g. Kahneman (2011) argues. Fedyk (2017) also gives some (quasi)-experimental evidence for this hypothesis. Most strikingly, in a classroom survey she finds that students expect themselves to finish their interim work 22 days before the deadline, while they expect their fellow students to do so 9 days before the deadline. In

[2] In an earlier version of this paper, we referred to our solution concept as the perception-perfect equilibrium. However we now feel that perception-perfect outcome is more appropriate as our solution concept is not an equilibrium in the traditional sense.

fact, the average student hands in her work 7 days before the deadline. Assuming that players are indeed naive about their own future selves, but are sophisticated about the future self of others, we find that the common pool problem becomes much worse than in a standard world with rational actors. This can be seen as follows. Suppose player A perceives B to have present-biased preferences in the future. That implies that B will then claim a large share of the common pool. Given that that is the case, A has an incentive to preempt B and claim a large share today. But the same holds for B. As a result, both players claim a large share of the pool today, completely exhausting it. We show that this effect is even stronger than in a case where both players know their future selves to also be present-biased.

In the case of Rubinstein bargaining, we show that the assumption that players are naive about their own future selves, but are sophisticated about the future self of others, implies a breakdown in bargaining. Suppose that it is player A's turn to make an offer. She will base that offer on the assumption that B will present-biased preferences in the future. Yet B perceives herself to be time-consistent in the future, and hence turns down A's offer. This process will continue indefinitely.

We are neither the first to develop approaches to solve games with possibly naive present-biased players, nor are we the only scholars to solve Rubinstein bargaining with such players. In an unpublished working paper, Sarafidis (2006) proposes "naive backward induction" with possibly naive present-biased players. His naive players are similar to ours, but his sophisticated players know everything, including the perceptions of the naive players. Akin (2007) shows that the bargaining process breaks down if naive bargainers meet, but in his definition of naivety players are sophisticated about the time inconsistency of their opponents. Schweighofer-Kodritsch (2018) studies Rubinstein bargaining allowing for any time preference. However, he does assume both bargainers are sophisticated about their own time preference, and that of their counterpart. Consistent with our results in Section 9, he finds no delay for any form of present bias; a future bias for at least one of the bargainers is necessary for equilibrium delay. From our analysis, we have that naivety and present bias can also cause delay, or even a bargaining breakdown. Lu (2016) studies Rubinstein bargaining between two present-biased but sophisticated players that may have a different degree of present-biasedness.

Other related literature includes Akin (2009), in which a naive player plays against a sophisticated player but learns about her naivety in the course of play, Chade et al. (2008) who analyze repeated games between sophisticated present-biased players, and Gans and Landry (2019) who focus on how initially naive present-biased players may update their beliefs concerning time inconsistency in a dynamic game. In Weinschenk (2021) present-biased players play a dynamic game in which they can collectively win a prize and the probability of doing so is increasing in total effort exerted. In that context, present-biased preferences increase the incentive to exert effort to try to secure the prize quickly, hence helping to overcome the incentive to free ride. Naive players do better than sophisticated ones. Weinschenk (2021) implicitly assumes that players are equally naive (or sophisticated) concerning other's present-bias as they are concerning their own. Turan (2019) studies a common-pool problem where one player perceives the other to have time-biased preferences with some probability, while the other player can manipulate those preferences through its actions. Compared to earlier work, our framework is more general. In particular, we are the first to allow players to be naive about themselves, but sophisticated about others.

The remainder of this paper is structured as follows. Section 2 looks at the case of one player. We first look at the case of a three-period model in which the player has to make two sequential decisions, and generalize the solution concept introduced by O'Donoghue and Rabin (1999). Section 4 further generalizes to a model with more than three periods, and Section 5 gives examples in the context of intertemporal consumption decisions. We then extend the analysis to a two-player game, and introduce the concept of a perception-perfect outcome. We do so for the three-period case in section 6, and apply our solution concept to a common pool problem in section 7. Section 8 looks at a multi-period model, and section 9 applies our analysis to Rubinstein bargaining. Section 10 concludes.

2 The one-player case: three periods

In this section, we consider the simplest set-up. Suppose that one player has to make decisions at times $t = 1$ and $t = 2$. Those decisions determine what will happen in the final period 3. Yet, the player may have intertemporal preferences that are present-biased. Moreover, she may not be aware that her future self (i.e. the one that makes the decision at $t = 2$) may also be present-biased. The problem of the current self then is what action to take at $t = 1$, taking into account her perceptions concerning the preferences of the future self.

Throughout this paper we consider the following preferences. Let u_t be a player's *instantaneous utility* or *felicity* in period t. In a model with T periods, we let $U_t\left(u_t, u_{t+1}, \ldots, u_T; \beta^i\right)$

represent a player's intertemporal preferences, where β^i is a parameter. We assume

$$U_t\left(u_t, u_{t+1}, \ldots, u_T; \beta^i\right) \equiv u_t + \beta^i \sum_{\tau=t+1}^{T} \delta^\tau u_\tau \tag{1}$$

with $0 < \beta^i, \delta \leq 1$. Note that with $\beta^i = 1$, equation (1) collapses into the standard exponential discounting function with discount factor δ. With $\beta^i < 1$, we have the canonical model of hyperbolic discounting introduced by Phelps and Pollak (1968). In that case, the player has present-biased preferences, where β^i represents the bias for the present. In other words, she is time inconsistent.

In this context, player A consider a one-player game with 3 periods, $t = 1, 2, 3$, in which player A makes two sequential decisions at $t = 1$ and $t = 2$. In $t = 1$, she chooses action $a_1 \in \mathcal{A}_1$, with $\mathcal{A}_1$ the set of feasible actions that the current self has. In $t = 2$, she chooses action $a_2 \in \mathcal{A}_2(a_1)$, with $\mathcal{A}_2(a_1)$ the set of feasible actions available at $t = 2$, that may depend on a_1. Her felicity in period 1 will depend on her action in period 1; that in periods 2 and 3 will depend on all actions. Thus $u_1^A = u_1^A(a_1)$, while $u_2^A = u_2^A(a_1, a_2)$ and $u_3^A = u_3^A(a_1, a_2)$.

The present-bias of the current self (that at $t = 1$) is denoted β^A. Following O'Donoghue and Rabin (1999), we allow for two possibilities: she either has present-biased preferences, so $\beta^A = \beta$, where $\beta < 1$ is some exogenously given fixed value, or she is time-consistent and has $\beta^A = 1$. For ease of discussion, we denote the true present-bias of the *future* self (i.e. that at $t = 2$) as γ^A, where we also assume that $\gamma^A \in \{\beta, 1\}$. Using (1) A's lifetime utility at both dates is thus given by

$$U_1^A\left(a_1, a_2; \beta^A\right) = u_1^A(a_1) + \beta^A \delta u_2^A(a_1, a_2) + \beta^A \delta^2 u_3^A(a_1, a_2) \tag{2}$$

$$U_2^A\left(a_1, a_2; \gamma^A\right) = u_2^A(a_1, a_2) + \gamma^A \delta u_3^A(a_1, a_2) \tag{3}$$

where we have now written utilities as functions of actions.

Following Strotz (1956) and Pollak (1968), we allow A either to be sophisticated (knowing her future preferences exactly), or to be naive (believing her future biases to be be identical to her current ones). Crucially, we do *not* allow players to use probability distributions over their future present-biasedness, believing for example that they will be present-biased with a 50% probability. That would complicate the analysis even further.[3]

[3] However, it would be no problem for our analysis if players would be partially sophisticated, in the sense that they are sophisticated about their future present-bias, but underestimate the extent to which

First, suppose that $\beta^A = 1$. In that case, she must believe that $\gamma^A = 1$ as well. It makes no sense for the current self to believe that she will have present-biased preferences in the future if that is not the case today. Second, suppose that $\beta^A = \beta$. In that case, the current self is present-biased. By construction, a player that has present-biased preferences will not only have those today, but also at any point in the future. Yet, she may not be aware of that. Naive present-biased players know that they have a present-bias today, but do not realize that they also have such a bias in the future. Such a naive player will assume that $\gamma^A = 1$. Sophisticated present-biased players know that they will also have a present-bias in the future, and will assume that $\gamma^A = \beta$.

We denote by $\mu^A(\gamma)$ the player's belief that she has $\gamma^A = \gamma$ in the future. Thus, a naive player has $\mu^A(1) = 1$, a sophisticated player $\mu^A(\beta) = 1$. In what follows, we use "perception" rather than "belief" to clearly differentiate from most of the literature where beliefs are rationally formed using Bayes' rule. That is clearly not the case here. As noted, a player that has no present-bias today will also not have such a bias in the future. Thus $\beta^A = 1$ must imply $\mu^A(1) = 1$.

We now introduce a formal solution concept for this game. Note that the model we have is a generalization of O'Donoghue and Rabin (1999).[4] They define a *perception-perfect strategy* as one in which in all periods a player chooses the optimal action given her current preferences and her perceptions of future behavior. Define $\mu^{\mathbf{A}}$ as the vector of perceptions: $\mu^{\mathbf{A}} \equiv \left(\mu^A(\beta), \mu^A(1)\right)$. In our set-up, we then have the following:

Definition 1 *In the three-period one-player game, a perception-perfect strategy at $t = 1$ for a present-biased player, given her perceptions $\mu^{\mathbf{A}}$, is a strategy profile (a_1^*, a_2^*) such that*

$$a_2^*(a_1; \mu^{\mathbf{A}}) \equiv \arg \max_{a_2 \in \mathcal{A}_2(a_1)} \sum_{\gamma \in \{\beta, 1\}} \mu^A(\gamma)\, U_2^A(a_1, a_2; \gamma), \forall a_1 \in \mathcal{A}_1; \tag{4}$$

$$a_1^*(\beta; \mu^{\mathbf{A}}) = \arg \max_{a_1 \in \mathcal{A}_1} U_1^A\left(a_1, a_2^*\left(a_1; \mu^{\mathbf{A}}\right); \beta\right) \tag{5}$$

Trivially, a perception-perfect strategy for a time-consistent player is a strategy profile

they will be present-biased, i.e. they perceive to have a future β that is larger that their true β, but smaller than 1.

[4] In that paper, a possibly present-biased player has to perform an action once, and has to choose some date in the future when to perform that action. Yet, she has the possibility to renege on her plan in the future. Hence, if today she plans to do it tomorrow, when tomorrow comes she may decide to postpone the action for another day. A sophisticated player will foresee this future tendency; a naive player will not.

(a_1^*, a_2^*) *such that*

$$\begin{aligned} a_2^*\left(a_1;(0,1)\right) &= \arg\max_{a_2\in\mathcal{A}_2(a_1)} U_2^A\left(a_1,a_2;1\right) \\ a_1^*\left(1;(0,1)\right) &= \arg\max_{a_1\in\mathcal{A}_1} U_1^A\left(a_1,a_2^*\left(a_1;(0,1)\right);1\right) \end{aligned}$$

The perception-perfect strategy for the present-biased player can be understood as follows. First, given a_1, the current self assumes that the future self is going to take the action that maximizes the future self's utility. In the current self's perception, with probability $\mu^A(\beta)$, the future self's utility is given by $U_2^A\left(a_1,a_2;\beta\right)$, while with probability $\mu^A\left(1\right)$, it is given by $U_2^A\left(a_1,a_2;1\right)$. The maximizer is thus given by equation (4) and denoted $a_2^*(a_1;\mu^{\mathbf{A}})$. In period 1, given her perceptions, the current self's lifetime utility if she takes action a_1 today is given by $U_1^A\left(a_1,a_2^*\left(a_1;\mu^{\mathbf{A}}\right);\beta\right)$. The current self thus chooses a_1 to maximize this expression, hence (5). The perception-perfect strategy for the present-biased player follows directly from backward induction.

Definition 2 *In the three-period one-player game, a perception-perfect outcome is a strategy profile (a_1^*, a_2^*) such that a_1^* is part of a perception-perfect strategy at period 1, while a_2^* maximizes the future self's utility at $t=2$, given the action a_1^* that was taken in period 1.*

Note that there is a crucial difference between a perception-perfect strategy and a perception-perfect outcome; a perception-perfect strategy is a strategy profile that a player *perceives* to be played, while a perception-perfect outcome is the strategy profile that *will* be played. There may be a difference between the two if the player is present-biased and naive. This distinction will become even more important in the T-period case.

3 Example: intertemporal consumption, 3 periods

Consider a player that lives for 3 periods, and starts out with wealth 1 in period 1. Felicity in each period is given by $u_t^A\left(a_t\right)=\sqrt{a_t}$, with a_t consumption in period t. For simplicity, the discount factor δ equals 1. The standard model, with time-consistent preferences, would have the player maximizing

$$U_1^A\left(a_1,a_2\right)=\sqrt{a_1}+\sqrt{a_2}+\sqrt{1-a_1-a_2}$$

which would obviously result in $a_1^*=a_2^*=1/3$. This simple decision problem satisfies our set-up. Two decisions are made; the consumption decision a_1 and the consumption decision

a_2, with $\mathcal{A}_1 = [0,1]$ and $\mathcal{A}_2(a_1) = [0, 1-a_1]$. In period 3, she consumes whatever is left of her initial wealth. Obviously, both the perception-perfect strategy and the perception-perfect outcome of a time-consistent player would be to have $a_1^* = a_2^* = 1/3$ as well.

We now solve for the perception-perfect strategy of the present-biased player. Using (4), at $t = 2$, and given first-period consumption a_1 and future present-biasedness γ, the player will choose a_2 as to maximize

$$U_2^A\left(a_1, a_2; \gamma^A\right) = \sqrt{a_2} + \gamma^A \sqrt{1 - a_1 - a_2}.$$

This yields

$$a_2^*\left(a_1; \mu^{\mathbf{A}}\right) = \frac{1 - a_1}{1 + \left[\beta \mu^A(\beta) + \mu^A(1)\right]^2} = \frac{1 - a_1}{1 + \tilde{\beta}^2},$$

where, for ease of exposition, we write

$$\tilde{\beta} \equiv \beta \mu^A(\beta) + \mu^A(1). \tag{6}$$

Perceived consumption in the last period is then given by

$$a_3^*\left(a_1; \mu^{\mathbf{A}}\right) = \frac{\tilde{\beta}^2 (1 - a_1)}{1 + \tilde{\beta}^2}.$$

Plugging this back into the lifetime utility of the current self yields

$$\begin{aligned} U_1^A\left(a_1, a_2^*\left(a_1; \mu^{\mathbf{A}}\right); \beta\right) &= \sqrt{a_1} + \beta \sqrt{\frac{1 - a_1}{1 + \tilde{\beta}^2}} + \beta \sqrt{\frac{\tilde{\beta}^2 (1 - a_1)}{1 + \tilde{\beta}^2}} \\ &= \sqrt{a_1} + \beta \frac{1 + \tilde{\beta}}{\sqrt{1 + \tilde{\beta}^2}} \sqrt{1 - a_1} \end{aligned}$$

The current self thus sets

$$a_1^*\left(\beta; \mu^{\mathbf{A}}\right) = \frac{1 + \tilde{\beta}^2}{\beta^2 \left(1 + \tilde{\beta}\right)^2 + 1 + \tilde{\beta}^2}.$$

A sophisticated present-biased player has $\mu^A(\beta) = 1$ and $\mu^A(1) = 0$, so $\tilde{\beta} = \beta$. She would thus choose

$$a_1^*(\beta; (1, 0)) = \frac{1 + \beta^2}{\beta^2 (1 + \beta)^2 + 1 + \beta^2}.$$

and plan to have

$$a_2^*(a_1;(1,0)) = \frac{1-a_1^*}{1+\beta^2} = \frac{\beta^2(1+\beta)^2}{(1+\beta^2)(2\beta^2+2\beta^3+\beta^4+1)}.$$

As the future self indeed has $\gamma^A = \beta$, the profile $(a_1^*(\beta;(1,0)), a_2^*(a_1;(1,0)))$ is both the perception-perfect strategy in period 1, and the perception-perfect outcome of the game.

It is easy to see[5] that $a_1^*(1;(1,0)) < a_1^*(\beta;(1,0))$; a time-consistent player consumes less in the first period than a sophisticated present-biased player. As a present-biased player effectively has a higher short-run discount rate, she will choose to consume more today.

Now consider a naive present-biased player. She has $\mu^A(\beta) = 0$ and $\mu^A(1) = 1$, so $\tilde{\beta} = \beta$. Hence

$$a_1^*(\beta;(0,1)) = \frac{1}{1+2\beta^2}$$

and she plans to have

$$a_2^*(a_1;(0,1)) = \frac{1-a_1^*}{2} = \frac{\beta^2}{1+2\beta^2}.$$

In period 2, however, she will find herself with $\gamma^A = \beta$ rather than $\gamma^A = 1$ as she expected. Hence, true second-period consumption will be

$$a_2^*(a_1,\beta) = \frac{1-a_1}{1+\beta^2} = \frac{1}{1+2\beta^2}.$$

Thus, in this case, a perception-perfect strategy in period 1 is to choose $(a_1^*, a_2^*) = \left(\frac{1}{1+2\beta^2}, \frac{\beta^2}{1+2\beta^2}\right)$, while the perception-perfect outcome will turn out to be $(a_1^*, a_2^*) = \left(\frac{1}{1+2\beta^2}, \frac{1}{1+2\beta^2}\right)$. It is interesting to note that $a_1^*(\beta;(0,1)) < a_1^*(\beta;(1,0))$. Hence, a naive player will choose a lower first-period consumption than a sophisticated one. This result is in line with O'Donoghue and Rabin (1999), who in a simpler framework find a "sophistication effect": when the reward of an action is immediate, naive players suffer less from the time inconsistency problem than sophisticated players.

Obviously, in our application, the rewards from consumption are also immediate. Here, the sophistication effect can be explained as follows. Different from naive players, sophisticated players are pessimistic about their future selves; they know that future selves will be present-biased and squander most of their wealth quickly. As a consequence, sophisticated

[5] We can write the inverse of $a_1^*(\beta;(1,0))$ as $1+\beta^2\left(1+\frac{2\beta}{1+\beta^2}\right)$, which is increasing in β on $[0,1]$, hence $a_1^*(\beta;(1,0))$ is decreasing in β, which implies the result.

players restrict the tendency of the future self to over-consume by increasing immediate consumption, which restricts the availability of the resource in the future. In other words, rather than allowing future selves to squander the wealth, current selves prefer to do this themselves. Hence, in our example, first period consumption is higher if there is a sophistication effect. Of course, if current selves can commit to a future consumption path, this result does not necessarily hold. In the presence of a commitment device, sophisticated players benefit from their knowledge because it enables them to restrict future consumption by committing to a certain consumption path.

4 The one-player case: more periods

We now generalize the two-period decision problem we described in Section 2, to one with $T + 1 > 3$ periods, so T is the number of decisions to be made. This complicates the problem. Consider the simplest case, with $T = 3$. Then the decision made by our player at $t = 1$ will be influenced by her perceptions concerning her type at $t = 2$. We will denote these perceptions as $\mu_{12}^{\mathbf{A}}$, where the first subscript reflects the time period in which perceptions are formed, and the second superscript reflects the time period that these perceptions apply to. But the decision made at $t = 1$ will also be influenced by her perceptions concerning her type at $t = 3$, denoted $\mu_{13}^{\mathbf{A}}$. Complicating matters further, the optimal decision at $t = 1$ will be influenced by her perception of the action that the future self will make at $t = 2$, which will in turn be determined by the perceptions that the future self at $t = 2$ will have, or rather, the perceptions that the current self at $t = 1$ will perceive that future self to have. Denote these perceptions as $\mu_1^{\mathbf{A}}\left(\mu_{23}^{\mathbf{A}}\right)$; these are the perceptions that at $t = 1$, player A perceives her future self at $t = 2$ to have concerning her type at $t = 3$.

To simplify matters, we make the following assumptions[6]

Assumption 1 ***Perception consistency***. *Perceptions concerning the type of a future self are identical for all future selves: $\mu_{ij}^{\mathbf{A}} = \mu_{ik}^{\mathbf{A}}$ for all $i < T$, $j, k \in \{i + 1, \ldots, T\}$.*

Assumption 2 ***Intraplayer perception naivety.*** *Perceptions of a future self are identical to perceptions of the current self: $\mu_i^{\mathbf{A}}(\mu_{jk}^{\mathbf{A}}) = \mu_{ik}^{\mathbf{A}}$ for all $T \geq k > j > i$.*

[6] Note that these assumptions also implicitly made by O'Donoghue and Rabin (1999). They assume that a naive player not only beliefs that she will be time-consistent in the next period, but also in any future period. Effectively, this is our perception consistency. Also, they implicitly rule out complications that may be caused by, say, a sophisticated player that maintains the possibility that he may be naive in the future. This is explicitly ruled out by our intraplayer perception naivety.

Note that there is a subtle difference between these two assumptions. Perception consistency implies that a player rules out that her type will change at some point in the future; if she perceives herself to be time-consistent at some point in the future, then she should perceive herself to be time-consistent at any point in the future. This seems a natural assumption to make; it is hard to justify a case in which, say, a player is naive concerning her future self in even periods but sophisticated concerning herself in odd periods.[7] Intraplayer perception naivety implies that a player rules out that her future self will change her opinion about selves that are in the further future. Thus, we rule out that a player perceives today that her future self in two weeks is sophisticated, but maintains the possibility that one week from now she perceives that same future self to be naive.

Note that this also implies that we assume that a naive player will never learn to be more sophisticated through e.g. some kind of Bayesian updating. This greatly simplifies the analysis and seems consistent with casual observation. Still, it is feasible to enrich our framework to allow for such learning, but we leave that for future research.

At time t, define history $\mathbf{H}_t \equiv (a_1, \ldots, a_{t-1})$. Similar to (2) and (3), lifetime utility at time $t \leq T$ can then be written

$$U_1^A\left(\mathbf{a}; \beta^A\right) = u_1(a_1) + \beta^A \sum_{k=2}^{T} \delta^{k-1} u_k^A(\mathbf{H}_k, a_k) + \beta^A \delta^{T+1} u_{T+1}^A(\mathbf{H}_{T+1}),$$

$$U_t^A\left(\mathbf{a}; \gamma^A\right) = u_t\left(\mathbf{H}_t, a_t\right) + \gamma^A \sum_{k=t+1}^{T} \delta^{k-t} u_k^A(\mathbf{H}_k, a_k) + \gamma^A \delta^{T+1} u_{T+1}^A(\mathbf{H}_{T+1}) \ \forall 1 < t \leq T,$$

with $\mathbf{a}$ the vector of all decisions: $\mathbf{a} \equiv (a_1, a_2, \ldots, a_T)$, and where we allow felicity in period $T+1$ to also play a role, just as we did in the case that $T = 2$. Given the assumptions above, $\mu^{\mathbf{A}}$ now reflects the perceptions at any time t concerning the type of the future self at any time $k > t$. More precisely $\mu^{\mathbf{A}}(\gamma) = \Pr\left(\gamma^A = \gamma | \beta^A = \beta\right)$ with γ^A the present-biasedness at any future period.[8]

Definition 3 *In the $T+1$-period one-player game, a perception-perfect strategy at time τ for a present-biased player, given her perceptions $\mu^{\mathbf{A}}$ and history $\mathbf{H}_t$ is a strategy profile*

[7] It is conceivable though that a player is sophisticated concerning the near future (say, up to some $t \leq t^*$), but naive concerning the more distant future ($t > t^*$). It is straightforward to extend the analysis to allow for such a possibility. That, however, is beyond the scope of this paper.

[8] Hence, we do not need a subscript t on either γ or γ^A.

$(a_\tau^*, a_{\tau+1}^*, \ldots a_T^*)$ *such that*

$$
\begin{aligned}
a_T^*(\mathbf{H}_T; \mu^{\mathbf{A}}) &= \arg \max_{a_T \in \mathcal{A}_T(\mathbf{H}_T)} \sum_{\gamma \in \{\beta, 1\}} \mu^A(\gamma)\, U_T^A(\mathbf{H}_T, a_T; \gamma); \qquad (7) \\
a_t^*(\mathbf{H}_t; \mu^{\mathbf{A}}) &= \arg \max_{a_t \in \mathcal{A}_t(\mathbf{H}_t)} \sum_{\gamma \in \{\beta, 1\}} \mu^A(\gamma)\, U_t^A\Big(\mathbf{H}_t, a_t, a_{t+1}^*\left(\mathbf{H}_{t+1}; \mu^{\mathbf{A}}\right), \\
&\qquad \ldots, a_T^*(\mathbf{H}_T; \mu^{\mathbf{A}}); \gamma\Big)\ \forall \tau < t < T; \\
a_\tau^*(\beta; \mu^{\mathbf{A}}) &= \arg \max_{a_\tau \in \mathcal{A}_\tau(\mathbf{H}_\tau)} U_\tau^A\left(\mathbf{H}_\tau, a_\tau, a_{\tau+1}^*\left(\mathbf{H}_{\tau+1}; \mu^A\right), \ldots, a_T^*(\mathbf{H}_T; \mu^{\mathbf{A}}); \beta\right).
\end{aligned}
$$

Trivially, a perception-perfect strategy for a time-consistent player is a strategy profile $(a_\tau^*, a_{\tau+1}^*, \ldots a_T^*)$ *such that*

$$
\begin{aligned}
a_T^*\left(\mathbf{H}_T; (0,1)\right) &= \arg \max_{a_T \in \mathcal{A}_T(\mathbf{H}_T)} U_T^A\left(\mathbf{H}_T; 1\right) \\
a_t^*\left(\mathbf{H}_t; (0,1)\right) &= \arg \max_{a_t \in \mathcal{A}_t(\mathbf{H}_t)} U_t^A\left(\mathbf{H}_t, a_{t+1}^*\left(\mathbf{H}_{t+1}; (0,1)\right), \ldots, a_T^*(\mathbf{H}_T; (0,1)); 1\right) \\
&\forall \tau \le t < T.
\end{aligned}
$$

The perception-perfect strategy for the present-biased player can be understood much along the same lines as that for the case $T = 2$. We solve with backward induction. First, given $\mathbf{H}_T$, the current self assumes that the future self is going to take the action that maximizes the future self's utility. In the current self's perception, with probability $\mu^A(\beta)$, the future self's utility is given by $U_T^A\left(\mathbf{H}_t, a_t; \beta\right)$, while with probability $\mu^A(1)$, it is given by $U_T^A\left(\mathbf{H}_t, a_t; 1\right)$. The maximizer is thus given by (7) and denoted $a_T^*(\mathbf{H}_T; \mu^{\mathbf{A}})$. In period $T-1$, with probability $\mu^A(\beta)$, the future self's utility is given by $U_{T-1}^A\left(\mathbf{H}_{T-1}, a_{T-1}, a_T^*\left(\mathbf{H}_T; \mu^{\mathbf{A}}\right); \beta\right)$, with probability $\mu^A(1)$, it is given by $U_{T-1}^A\big(\mathbf{H}_{T-1}, a_{T-1}, a_T^*\left(\mathbf{H}_T; \mu^{\mathbf{A}}\right);$ In both cases, $\mathbf{H}_T = (\mathbf{H}_{T-1}, a_{T-1})$. For ease of exposition, this dependence of future history on current action is not explicitly taken into account in our notation above. Again, the current self assumes the future self at $t = T-1$ to take the action that maximizes her utility. This process unravels until period 1, where the current self chooses the a_1 that maximizes her lifetime utility given her perceptions about future selves and given her true β^A in period 1.

Definition 4 *In the $T+1$-period one-player game, a perception-perfect outcome is a strategy profile $(a_1^*, a_2^*, \ldots, a_T^*)$ such that a_τ^* is part of a perception-perfect strategy at time τ for all $\tau = 1, \ldots, T$.*

Note again the crucial difference between a perception-perfect strategy and a perception-perfect outcome; a perception-perfect strategy is a strategy profile that a player *perceives* to be played, while a perception-perfect outcome is the strategy profile that *will* be played.

It is relatively straightforward to extend the analysis to a case with infinitely many periods. Solving such a model would be similar to solving an infinite-horizon maximization problem in the case of time-consistent preferences, but under the assumption that all future selves have the type the current self perceives them to have.

5 Example: intertemporal consumption $T+1$ periods

To give a flavor of the analysis, we consider the same consumption example as above, but now with $T+1$ periods;

$$U_1^A(\mathbf{a}) = \sqrt{a_1} + \sqrt{a_2} + \ldots \sqrt{a_T} + \sqrt{1 - \sum_{t=1}^{T} a_t}.$$

In this case, a time-consistent player would set $a_1^* = \ldots = a_T^* = \frac{1}{T+1}$

We now solve for the perception-perfect strategy of the present-biased player. Define total consumption in the past at time τ as $h_\tau = \sum_{t=1}^{\tau-1} a_t$. At $t = T$, and given first-period consumption a_1 and future present-biasedness γ, the player will choose a_2 as to maximize

$$U_T^A\left(\mathbf{H}_T, a_T; \gamma^A\right) = \sqrt{a_T} + \gamma^A \sqrt{1 - h_T - a_T}.$$

This yields

$$a_T^*\left(\mathbf{H}_t; \mu^{\mathbf{A}}\right) = \frac{1 - h_T}{1 + \left[\beta\mu^A(\beta) + \mu^A(1)\right]^2} = \frac{1 - h_T}{1 + \tilde{\beta}^2},$$

where again $\tilde{\beta}$ is given by (6). Now move back to $T-1$.

$$\begin{aligned} U_{T-1}^A\left(\mathbf{H}_{T-1}, a_{T-1}, a_T^*\left(\mathbf{H}_t; \mu^{\mathbf{A}}\right); \gamma^A\right) = \sqrt{a_{T-1}} + \gamma^A \sqrt{\frac{1 - h_{T-1} - a_{T-1}}{1 + \tilde{\beta}^2}} \\ + \gamma^A \sqrt{1 - h_{T-1} - a_{T-1} - \frac{1 - h_{T-1} - a_{T-1}}{1 + \tilde{\beta}^2}} \end{aligned}$$

Take advantage of perception consistency to note that the future self at $t = T - 2$ is thus expected to maximize

$$U_{T-1}^{A}=\sqrt{a_{T-1}}+\tilde{\beta}\sqrt{\frac{1-h_{T-1}-a_{T-1}}{1+\tilde{\beta}^{2}}}+\tilde{\beta}\sqrt{1-h_{T-1}-a_{T-1}-\frac{1-h_{T-1}-a_{T-1}}{1+\tilde{\beta}^{2}}}$$

This yields

$$a_{T-1}^{*}=\frac{1+\tilde{\beta}^{2}}{\tilde{\beta}^{2}\left(1+\tilde{\beta}\right)^{2}+1+\tilde{\beta}^{2}}\left(1-h_{t-1}\right).$$

Solving the model further is conceptually straightforward but analytically tedious.

6 Two-player case: three periods

We now come to the main aim of this paper: to extend the analysis above to a case with multiple players. Needless to say, this will greatly complicate the analysis. The current decisions of a player will now not only depend on her perceptions concerning her own future type, but also on her perceptions concerning the other player's future type, and possibly even about her perceptions of the other player's perceptions, plus how those perceptions will affect her own and the other player's future actions.

For simplicity, we start with the case of three periods and two players, denoted A and B. For ease of exposition, in what follows we will refer to player A as being female, and to player B as being male. Again, player i's present-bias is denoted $\beta^{i}\in\{1,\beta\}$. The true present-bias of the future self of player i (i.e. player i's type) is $\gamma^{i}\in\{1,\beta\}$. There are 3 periods, $t=1,2,3$. In the first two periods both A and B make a simultaneous decision. In $t=1$, player A chooses action $a_1\in\mathcal{A}_1$, while B chooses action $b_1\in\mathcal{B}_1$. At $t=2$, players learn the actions taken at $t=1$, and player A chooses action $a_2\in\mathcal{A}_2(a_1,b_1)$, while B chooses $b_2\in\mathcal{B}_2(a_1,b_1)$. We now have

$$\begin{aligned}U_1^i\left(a_1,b_1,a_2,b_2;\beta^i\right)&=u_1^i(a_1,b_1)+\beta^i\delta u_2^i\left(a_1,b_1,a_2,b_2\right)+\beta^i\delta^2u_3^i\left(a_1,b_1,a_2,b_2\right)\\U_2^i\left(a_1,b_1,a_2,b_2;\gamma^i\right)&=u_2^i\left(a_1,b_1,a_2,b_2\right)+\gamma^i\delta u_3^i\left(a_1,b_1,a_2,b_2\right),i\in\{A,B\}.\end{aligned}$$

In period 1, what A expects to happen in period 2 depends on her perceptions concerning her own future type, as well as those concerning B's future type. For simplicity, we assume players can observe each other's current type, so both A and B observe β^A and β^B. This simplifies the exposition, but it is conceptually straightforward to relax this assumption and allow players to have perceptions concerning their competitor's current type.

A straightforward extension of the one-player case is as follows. In the perception of player A we have $\mu^{AA}(\gamma) = \Pr^{A}\left(\gamma^{A} = \gamma|\beta^{A} = \beta\right)$, where the first superscript denotes perceptions held by player A, and the second denotes perceptions concerning player A. The superscript on Pr denotes that this is the probability as perceived by player A. Similarly, we have $\mu^{AB}(\gamma) = \Pr^{A}\left(\gamma^{B} = \gamma|\beta^{B} = \beta\right)$. Naturally, $\mu^{BA}(\gamma) = \Pr^{B}\left(\gamma^{A} = \gamma|\beta^{A} = \beta\right)$ and $\mu^{BB}(\gamma) = \Pr^{B}\left(\gamma^{B} = \gamma|\beta^{B} = \beta\right)$.

In principle we now have to be concerned about what A perceives B to perceive about A, for example, i.e. we need to be concerned about $\mu^{\mathbf{AB}}\left(\mu^{\mathbf{BA}}\right)$. We also assume naivety in this respect, in the sense that what A perceives B to perceive about A is the same what A perceives about herself, thus $\mu^{\mathbf{AB}}\left(\mu^{\mathbf{BA}}\right) = \mu^{\mathbf{AA}}$. More generally, we assume

Assumption 3 ***Current interplayer perception naivety.*** *Perceptions that the other player has are identical to one's own perceptions:* $\mu^{\mathbf{ij}}(\mu^{\mathbf{jk}}) = \mu^{\mathbf{ik}}$ *for all* $i, j, k \in \{A, B\}$.

Note that this is a natural extension of the intraplayer perception naivety we assumed in the one-player case. That assumption implies that a player rules out that her future self will change her opinion about selves that are even further in the future. This assumption implies that, say, player A rules out that player B has perceptions about the future self of player A that are different from what player A herself has. In other words, player A is so convinced about the type of her future self that she cannot perceive that the other player has different perceptions.

Again, we solve the game using backward induction. For ease of exposition, we restrict attention to the case where both players are present-biased. Consider player A. When deciding upon her first-period action, she again has to form some perception as to what will happen in period 2, given the actions taken in period 1. In the one-player case, she could simply derive the action her future self would be taking in period 2, given her perceptions about her future self. Now the analysis becomes more involved, as she also has to take the type and possible actions of player B into account. Suppose that the actions taken in period 1 are (a_1, b_1). Given these actions, we now look for a Nash equilibrium for the subgame at $t = 2$ as perceived by player A. As an example suppose player A perceives both players to be time-consistent in the future, so $\mu^{AA}(1) = \mu^{AB}(1) = 1$. She will then expect a Nash equilibrium (a_2^A, b_2^A) to be played which is such that a_2^A maximizes her future self's utility given b_2^A and given her perception that her future self is time-consistent, and such that b_2^A maximizes the future self's utility of player B, given A's perception that B's future self is time-consistent. Thus

$$\begin{aligned} a_2^A &= \arg\max_{a_2} U_2^A \left(a_1, b_1, a_2, b_2^A; 1\right) \\ b_2^A &= \arg\max_{b_2} U_2^A \left(a_1, b_1, a_2^A, b_2; 1\right) \end{aligned}$$

where superscripts denote that we study the perceptions of player A. More generally,

Definition 5 *Consider the three-period two-player game played by present-biased players. In period 2, given (a_1, b_1) an equilibrium as perceived by player $i \in \{A, B\}$ is an outcome $\left(a_2^i(a_1, b_1; \mu^{\mathbf{iA}}), b_2^i(a_1, b_1; \mu^{\mathbf{iB}})\right)$ that forms a Nash equilibrium of the second-period game, given the perceptions of player i. Hence*

$$\begin{aligned} a_2^i &= \arg\max_{a_2 \in \mathcal{A}_2(a_1,b_1)} \sum_{\gamma \in \{\beta,1\}} \mu^{iA}(\gamma)\, U_2^A \left(a_1, b_1, a_2, b_2^A; \gamma\right) \\ b_2^i &= \arg\max_{b_2 \in \mathcal{B}_2(a_1,b_1)} \sum_{\gamma \in \{\beta,1\}} \mu^{iB}(\gamma)\, U_2^B \left(a_1, b_1, a_2^A, b_2; \gamma\right) \end{aligned}$$

Moving back to period 1, given that player A has a perception of the play that will ensue in period 2 for any (a_1, b_1) in period 1, it is straightforward to write down the conditions for a subgame perfect Nash equilibrium as perceived by player A. We will refer to this simply as an equilibrium as perceived by player A.

Definition 6 *In period 1, an equilibrium as perceived by player i is an outcome*

$$\left(a_1^i\left(\beta; \mu^{\mathbf{iA}}, \mu^{\mathbf{iB}}\right), b_1^i\left(\beta; \mu^{\mathbf{iA}}, \mu^{\mathbf{iB}}\right)\right)$$

that is part of a subgame perfect Nash equilibrium of the entire game, given the perceptions of player i. Thus,

$$\begin{aligned} a_1^i &= \arg\max_{a_1 \in \mathcal{A}_1} U_1^A \left(a_1, b_1^i, a_2^i(a_1, b_1^i; \mu^{\mathbf{iA}}), b_2^i(a_1, b_1; \mu^{\mathbf{iB}}); \beta\right) \\ b_1^i &= \arg\max_{b_1 \in \mathcal{B}_1} U_1^B \left(a_1^i, b_1, a_2^i(a_1, b_1; \mu^{\mathbf{iA}}), b_2^i(a_1, b_1; \mu^{\mathbf{iB}}); \beta\right). \end{aligned} \tag{8}$$

Using these definitions, and considering play in period 1, we thus expect player A to take an action that she perceives to be part of a subgame perfect equilibrium for the entire game, while we expect player B to take an action that he perceives to be part of a subgame perfect equilibrium for the entire game.

Definition 7 *A perception-perfect outcome of the game is an outcome* $(a_1^*, b_1^*, a_2^*, b_2^*)$ *such that* a_1^* *is part of an equilibrium as perceived by player* A; b_1^* *is part of an equilibrium as perceived by player* B; a_2^* *is an equilibrium as perceived by player* A *given* (a_1^*, b_1^*)*; and* b_2^* *is an equilibrium as perceived by player* B *given* (a_1^*, b_1^*).

Needless to say, the actions a_1^* and b_1^* do not have to be consistent with each other, in the sense that they do not have to be part of the same equilibrium. Also, we assume that players do not learn anything about the perceptions or type of the other player upon observing first-period actions. Of course, we do allow a player to adapt her strategy in the second period upon observing the other player's action in period 1. In other words, we assume that, say, player A takes the action that she feels is part of an equilibrium of the second period based on the actions that she actually observed to be played in period 1, rather than the actions that she expected to be played in period 1.

It is straightforward to extend the analysis to a case where, for example, one player is known to be time-consistent,[9] or one where players cannot observe the other player's current type.[10]

7 Application to the common pool problem

We consider a common pool problem similar to the example that we gave for the one-player model. Consider two players, A and B, that live for 3 periods, and start out with joint wealth 1. Felicity in each period is given by $u_t^i(c) = c^\rho$, $i \in \{A, B\}$,and $\rho < 1$, so consumers are risk averse. For simplicity, the discount factor δ equals 1. In each of the

[9] Suppose that player B is known to be time-consistent. In that case, his future self will necessarily also be time-consistent, so $\mu^{AB}(1) = \mu^{BB}(1) = 1$. Moreover, the conditions (8) then modify to

$$a_1^i = \arg\max_{a_1 \in \mathcal{A}_1} U_1^A\left(a_1, a_2^i(a_1, b_1^i; \mu^{\mathbf{iA}}), b_1^i, b_2^i(a_1, b_1; \mu^{\mathbf{iB}}); \beta\right)$$
$$b_1^i = \arg\max_{b_1 \in \mathcal{B}_1} U_1^B\left(a_1^i, a_2^i(a_1, b_1; \mu^{\mathbf{iA}}), b_1, b_2^i(a_1, b_1; \mu^{\mathbf{iB}}); 1\right)$$

[10] The conditions (8) then modify to

$$a_1^i = \arg\max_{a_1 \in \mathcal{A}_1} \sum_{\gamma \in \{\beta, 1\}} \mu^{iA}(\gamma) U_1^A\left(a_1, a_2^i(a_1, b_1^i; \mu^{\mathbf{iA}}), b_1^i, b_2^i(a_1, b_1; \mu^{\mathbf{iB}}); \beta\right)$$
$$b_1^i = \arg\max_{b_1 \in \mathcal{B}_1} \sum_{\gamma \in \{\beta, 1\}} \mu^{iA}(\gamma) U_1^B\left(a_1^i, a_2^i(a_1, b_1; \mu^{\mathbf{iA}}), b_1, b_2^i(a_1, b_1; \mu^{\mathbf{iB}}); 1\right)$$

2 periods each player takes some amount for immediate consumption out of the common pool. Whatever is left in the last period will be equally shared among the two.[11]

Time-consistent players First consider the case in which both players are time-consistent. Their respective lifetime utility functions in period 1 then equal

$$\begin{aligned} U_1^A\left(a_1, b_1, a_2, b_2\right) &= a_1^\rho + a_2^\rho + \left(\frac{1 - a_1 - a_2 - b_1 - b_2}{2}\right)^\rho \\ U_1^B\left(a_1, b_1, a_2, b_2\right) &= b_1^\rho + b_2^\rho + \left(\frac{1 - a_1 - a_2 - b_1 - b_2}{2}\right)^\rho \end{aligned} \tag{9}$$

In period 1 player A correctly perceives the equilibrium in period 2 to satisfy

$$a_2^A = \arg\max a_2^\rho + \left(\frac{W_2 - a_2 - b_2}{2}\right)^\rho \tag{10}$$

$$b_2^A = \arg\max b_2^\rho + \left(\frac{W_2 - a_2 - b_2}{2}\right)^\rho . \tag{11}$$

with $W_2 \equiv 1 - a_1 - b_1$ the amount of wealth left at the start of period 2. Taking the first-order condition

$$\rho a_2^{\rho-1} - \frac{1}{2}\rho\left(\frac{1 - a_1 - a_2 - b_1 - b_2}{2}\right)^{\rho-1} = 0,$$

yields the reaction function

$$a_2 = \Gamma_{tc}\left(W_2 - b_2\right) \quad \text{with} \quad \Gamma_{tc} \equiv \tfrac{2^{\frac{1}{1-\rho}}}{2+2^{\frac{1}{1-\rho}}}.$$

Imposing symmetry, this yields the Nash equilibrium

$$a_2^* = b_2^* = \theta_{tc} W_2 \quad \text{with} \quad \theta_{tc} \equiv \tfrac{\Gamma_{tc}}{1+\Gamma_{tc}}. \tag{12}$$

Now move back to period 1. Plugging (12) back into (9),

$$U_1^A\left(a_1, b_1, a_2, b_2\right) = a_1^\rho + \left(\theta_{tc} W_2\right)^\rho + \left(\frac{1}{2}\left(1 - 2\theta_{tc}\right) W_2\right)^\rho$$

[11] For simplicity, we assume parameters are such that the common pool is not depleted after period 2.

maximizing with respect to a_1 :

$$\rho a_1^{\rho-1} - \rho\theta_{tc}\left(\theta_{tc}W_2\right)^{\rho-1} - \frac{1}{2}\left(1-2\theta_{tc}\right)\rho\left(\frac{1}{2}\left(1-2\theta_{tc}\right)W_2\right)^{\rho-1} = 0$$

hence

$$a_1 = \Omega_{tc}\left(1-a_1-b_1\right) \quad \text{with} \quad \Omega_{tc} \equiv \left[\theta_{tc}^{\rho} + \left(\tfrac{1}{2}\left(1-2\theta_{tc}\right)\right)^{\rho}\right]^{\frac{1}{\rho-1}}.$$

This implies that first period consumption choices equal

$$a_1^{tc} = b_1^{tc} = \frac{\Omega_{tc}}{1+2\Omega_{tc}}, \tag{13}$$

where the superscripts tc denote equilibrium values for the time-consistent case.

Sophisticated present-biased players In this case, at $t = 1$, the current self of player A perceives an equilibrium in period 2 to satisfy

$$\begin{aligned} a_2^A &= \arg\max a_2^{\rho} + \beta\left(\frac{1}{2}\left(W_2-a_2-b_2\right)\right)^{\rho} \\ b_2^A &= \arg\max b_2^{\rho} + \beta\left(\frac{1}{2}\left(W_2-a_2-b_2\right)\right)^{\rho}. \end{aligned}$$

Taking the first-order condition of her own problem:

$$\rho a_2^{\rho-1} - \frac{1}{2}\beta\rho\left(\frac{1}{2}\left(W_2-a_2-b_2\right)\right)^{\rho-1} = 0,$$

which implies the reaction function

$$a_2 = \Gamma_s\cdot\left(W_2-b_2\right) \quad \text{with} \quad \Gamma_s \equiv \frac{\left(\frac{1}{2}\beta\right)^{\frac{1}{\rho-1}}}{2+\left(\frac{1}{2}\beta\right)^{\frac{1}{\rho-1}}}.$$

Hence, along the same lines as above, this yields

$$a_2^* = b_2^* = \theta_s W_2 \quad \text{with} \quad \theta_s \equiv \tfrac{\Gamma_s}{1+\Gamma_s}.$$

Moving back to period 1, note the following. After period 1, W_2 is left. Player A perceives both players to consume $\theta_s W_2$ in period 2, hence in period 3 there is $\left(1-2\theta_s\right)W_2$ left, which is equally shared among both players. Hence, using (9), the equilibrium in period 1 as anticipated by player 1 satisfies

$$\begin{aligned} a_1^A &= \arg\max_{a_1} a_1^\rho + \beta\left(\theta_s W_2\right)^\rho + \beta\left(\frac{1}{2}\left(1-2\theta_s\right)W_2\right)^\rho . \\ b_1^A &= \arg\max_{b_1} b_1^\rho + \beta\left(\theta_s W_2\right)^\rho + \beta\left(\frac{1}{2}\left(1-2\theta_s\right)W_2\right)^\rho . \end{aligned} \tag{14}$$

The first-order condition for player A equals

$$\rho a_1^{\rho-1} - \beta\rho\theta_s\left(\theta_s W_2\right)^{\rho-1} - \frac{1}{2}\beta\rho\left(1-2\theta_s\right)\left(\frac{1}{2}\left(1-2\theta_s\right)W_2\right)^{\rho-1} = 0$$

or

$$a_1 = \Omega_s\left(1-a_1-b_1\right) \quad \text{with} \quad \Omega_s \equiv \beta^{\frac{1}{\rho-1}}\left[\theta_s^\rho + \left(\tfrac{1}{2}\left(1-2\theta_s\right)\right)^\rho\right]^{\frac{1}{\rho-1}} .$$

Imposing symmetry:

$$a_1^s = b_1^s = \frac{\Omega_s}{1+2\Omega_s}, \tag{15}$$

where superscript s denotes equilibrium values in the sophisticated case. As B faces the same problem and has the same perceptions, she will have the same perceived equilibrium in periods 1 and 2 as A. Moreover, as players' perceptions turn out to be correct, what they perceive to be played in period 2 is also what is actually played in period 2.

Comparing (13) and (15), we can show the following;

Theorem 1 *In the common pool problem, time-consistent players will claim a smaller share in the first period than players that are present-biased but sophisticated:* $a_1^{tc} < a_1^s$.

Proof. In Appendix. ■

Note that, qualitatively, this result is similar to what we found in Section 3; players that are present-biased but sophisticated consume more than those that are time-consistent.

Naive present-biased players Now consider the case in which both players are naive concerning all future selves. At $t = 1$, player A perceives the equilibrium in period 2 to satisfy (10) and (11). so $a_2^A = b_2^A = \theta_{tc}W_2$. Moving back to period 1, the equilibrium perceived by player A should thus satisfy

$$\begin{aligned} a_1^A &= \arg\max_{a_1} a_1^\rho + \beta\left(\theta_{tc} W_2\right)^\rho + \beta\left(\frac{1}{2}\left(1-2\theta_{tc}\right)W_2\right)^\rho \\ b_1^A &= \arg\max_{b_1} b_1^\rho + \beta\left(\theta_{tc} W_2\right)^\rho + \beta\left(\frac{1}{2}\left(1-2\theta_{tc}\right)W_2\right)^\rho . \end{aligned}$$

This problem is essentially the same as in (14) – but with θ_{tc} rather than θ_s. Maximizing thus yields

$$a_1 = \Omega_n \left(1 - a_1 - b_1\right) \quad \text{with} \quad \Omega_n = \beta^{\frac{1}{\rho-1}} \left[\theta_{tc}^{\rho} + \left(\tfrac{1}{2}\left(1 - 2\theta_{tc}\right)\right)^{\rho}\right]^{\frac{1}{\rho-1}}.$$

Imposing symmetry:

$$a_1^n = b_1^n = \frac{\Omega_n}{1 + 2\Omega_n}. \tag{16}$$

As player B faces the same problem and the same perceptions, she will have the same perceived equilibrium in periods 1 and 2 as player A. However, players' perceptions turn out to be incorrect: the equilibrium in the second period has them both consuming $\theta_s W_2$ (as we saw in the previous analysis) rather than $\theta_{tc} W_2$. Hence, actual consumption in period 2 will turn out to be

$$a_2^n = b_2^n = \theta_s \left(1 - \frac{2\Omega_n}{1 + 2\Omega_n}\right).$$

We now have

Theorem 2 *In the common pool problem, naive present-biased players will claim a smaller first-period share than sophisticated present-biased players, but a larger share than time-consistent players: $a_1^{tc} < a_1^n < a_1^s$.*

Proof. In Appendix. ■

Again, qualitatively this is the same result that we found in Section 3; there is a sophistication effect in that sophisticated players that are sophisticated suffer more from present-biasedness than those that are naive. Again, sophisticated players know that their future selves will be present-biased and squander resources. However, they now also know that their future competitors will do the same. That gives them an additional reason to consume more today. In addition, there is a strategic effect in that knowing that their competitor will also be inclined to consumer more today, will induce them to increase their current consumption even further.

Naive about yourself, sophisticated about the other The most interesting case is the one in which both players perceive themselves to be time-consistent in the future, but perceive their competitor to be present-biased in the future. In other words, each player is naive concerning her own future self, but sophisticated concerning the future self of the other player. As noted in the introduction, Kahneman (2011) argues that this is the typical situation. In period 1, Player A will then perceive a second-period equilibrium

$$\begin{aligned} a_2^A &= \arg\max a_2^\rho + \left(\frac{W_2 - a_2 - b_2}{2}\right)^\rho \\ b_2^A &= \arg\max b_2^\rho + \beta\left(\frac{W_2 - a_2 - b_2}{2}\right)^\rho . \end{aligned}$$

From the analysis above, this yields reaction functions (as perceived by player A) of

$$\begin{aligned} a_2^A &= \Gamma_{tc}\left(W_2 - b_2^A\right) \\ b_2^A &= \Gamma_s\left(W_2 - a_2^A\right), \end{aligned}$$

hence

$$\begin{aligned} a_2^A &= \theta_{ns} W_2 \quad \text{with} \quad \theta_{ns} = \tfrac{\Gamma_{tc}(1-\Gamma_s)}{1-\Gamma_{tc}\Gamma_s}, \\ b_2^A &= \theta_{sn} W_2 \quad \text{with} \quad \theta_{sn} = \tfrac{\Gamma_s(1-\Gamma_{tc})}{1-\Gamma_{tc}\Gamma_s}. \end{aligned}$$

Below, we show that $\theta_{ns} < \theta_{sn}$. Thus, player A perceives to consume much less in period 2 than player B does. Note that the reaction functions are strategic substitutes, in the sense of Bulow et al. (1985). Player A perceives player B to be very aggressive in period 2, due to A's perception that B will be very impatient and thus claim a high amount of then current consumption. As A (again in her perception) will be much more patient, she will claim a very low share of the available wealth then as future consumption is still important for her.

In period 1, player A perceives the following game to be played:

$$\begin{aligned} a_1^A &= \arg\max_{a_1} a_1^\rho + \beta\left(\theta_{ns}W_2\right)^\rho + \beta\left(\frac{1}{2}\left(1 - \theta_{ns} - \theta_{sn}\right)W_2\right)^\rho \\ b_1^A &= \arg\max_{b_1} b_1^\rho + \beta\left(\theta_{sn}W_2\right)^\rho + \beta\left(\frac{1}{2}\left(1 - \theta_{ns} - \theta_{sn}\right)W_2\right)^\rho \end{aligned}$$

Taking first order conditions:

$$\begin{aligned} \rho a_1^{\rho-1} - \beta\theta_{ns}\rho\left(\theta_{ns}W_2\right)^{\rho-1} &- \frac{1}{2}\beta\rho\left(1 - \theta_{ns} - \theta_{sn}\right)\left(\frac{1}{2}\left(1 - \theta_{ns} - \theta_{sn}\right)W_2\right)^{\rho-1} = 0 \\ \rho b_1^{\rho-1} - \beta\theta_{sn}\rho\left(\theta_{sn}W_2\right)^{\rho-1} &- \frac{1}{2}\beta\rho\left(1 - \theta_{ns} - \theta_{sn}\right)\left(\frac{1}{2}\left(1 - \theta_{ns} - \theta_{sn}\right)W_2\right)^{\rho-1} = 0 \end{aligned}$$

so

$$
\begin{aligned}
a_1^{\rho-1} &= \beta\left(\theta_{ns}^{\rho}+\left(\frac{1}{2}\left(1-\theta_{ns}-\theta_{sn}\right)\right)^{\rho}\right)\left(1-a_1-b_1\right)^{\rho-1}\\
b_1^{\rho-1} &= \beta\left(\theta_{sn}^{\rho}+\left(\frac{1}{2}\left(1-\theta_{ns}-\theta_{sn}\right)\right)^{\rho}\right)\left(1-a_1-b_1\right)^{\rho-1}
\end{aligned}
$$

This implies

$$
\begin{aligned}
a_1 = \Omega_{ns}\left(1-a_1-b_1\right) \quad &\text{with} \quad \Omega_{ns}=\beta^{\frac{1}{\rho-1}}\left[\theta_{ns}^{\rho}+\left(\tfrac{1}{2}\left(1-\theta_{ns}-\theta_{sn}\right)\right)^{\rho}\right]^{\frac{1}{\rho-1}},\\
b_1 = \Omega_{sn}\left(1-a_1-b_1\right) \quad &\text{with} \quad \Omega_{sn}=\beta^{\frac{1}{\rho-1}}\left[\theta_{sn}^{\rho}+\left(\tfrac{1}{2}\left(1-\theta_{ns}-\theta_{sn}\right)\right)^{\rho}\right]^{\frac{1}{\rho-1}}.
\end{aligned}
$$

This implies

$$
\begin{aligned}
a_1^{ns} &= \frac{\Omega_{ns}}{1+\Omega_{ns}+\Omega_{sn}}\\
b_1^{ns} &= \frac{\Omega_{sn}}{1+\Omega_{ns}+\Omega_{sn}}
\end{aligned}
$$

Hence, in period 1, A expects these shares to be played. But B expects the opposite shares. Both players will thus consume $\Omega_{ns}/\left(1+\Omega_{ns}+\Omega_{sn}\right)$, so after period 1, $W_2 = 1-2\Omega_{ns}/\left(1+\Omega_{ns}+\Omega_{sn}\right)$ will be left. In period 2, both players will consume a share θ_{ns} of that wealth, although they both expect their opponent to consume θ_{sn}. We now have

Theorem 3 *In the common pool problem, present-biased players that are naive about themselves but sophisticated about others, will claim a larger first-period share than any other type of player we considered;* $a_1^{ns} > a_1^{s} > a_1^{n} > a_1^{tc}$.

Proof. In Appendix. ■

In this set-up, each player perceives her future self to be time-consistent, but her future competitor to be present-biased. Hence, each player perceives that in the game to be played in the second period, her competitor will claim most of the available resources. Knowing that she will receive little in the future will give both players an incentive to already make a large claim today. The unfounded fear to get an unequal share in the future, gives both players an incentive to already make a large claim today. As we will show in the numerical example below, this seriously exacerbates the common pool problem.

A Numerical Example For $\beta = 1/2$ and $\rho = 1/3$, Table 1 gives consumption per player in periods 1, 2, and 3, and total utility from the perspective of period 1 for all scenarios we considered.[12] From the Table, time consistent players indeed take much less from the common pool in the first period than players in any other scenario. The difference between sophisticated and naive is relatively small, but first-period consumption is much higher in the case one is sophisticated about the other player, but naive about oneself. As noted, players end up better off when they are naive rather than when they are sophisticated.

Table 1: Numerical example common pool problem; $\beta = 1/2$, $\rho = 1/3$.

	a_1	a_2	a_3	U
time consistent	0.2981	0.1492	0.0528	
sophisticated	0.4110	0.0791	0.0099	1.0654
naive	0.4033	0.0859	0.0107	1.0698
soph other	0.4780	0.0196	0.0024	0.9840

Perception perfect outcomes with time consistent players, sophicated present-biased players, naive present-biased players, and present-biased players that are naive about their own present-biasedness but sophisticated about that of the other player (soph other). Columns give the consumption per player in the first (a_1), second (a_2) and third period (a_3) as well as total discounted lifetime utility in period 1.

8 Two-player case: more periods

We now extend the two-player two-period model that we analyzed above, to a setting with $T + 1$ periods, $T > 2$. This is conceptually straightforward, but notationally tedious. We solve with backward induction. In period T, what the current player A expects to be played is a game between herself and player B, with both players having the type that she currently expects them to have. Moving back to period 1, and given her perceptions concerning the players in period $T - 1$, she can then derive her perceived equilibrium play in that period. Continuing in this manner yields a perceived equilibrium in period 1, and hence a course of action for player A in period 1, with a similar analysis for player B.

To analyze this problem, we again need to make simplifying assumptions concerning the perceptions of players. Not only do we need that A has to believe that she has the

[12] We do not report total utility in the time consistent case, since players have a different utility function in that scenario compared to the other scenarios.

same perceptions as B concerning future types, we also need that higher-order perceptions are perceived to be equal. In other words we also need that the perceptions that A has in period l concerning the perceptions of B in period m concerning the perceptions of A in period n, equal the perceptions that A thinks she herself has in period l concerning herself in period n. Thus

Assumption 4 ***Future interplayer perception naivety.*** *Perceptions that the other player has concerning future perceptions, are assumed identical to one's own:* $\mu_{lm}^{\mathbf{ij}}(\mu_{mn}^{\mathbf{jk}}) = \mu_{lm}^{\mathbf{ii}}(\mu_{mn}^{\mathbf{jk}})$ *for all* $i, j, k \in \{A, B\}$.

Without this assumption, we would have to allow for the possibility that, at any time in the future player A maintains the possibility that player B has different perceptions concerning future types than she herself has. This possibility would force player A to also form higher order beliefs concerning perceptions – a possibility that would highly complicate the analysis. Together with the previous assumptions, future interplayer perception naivety implies that all perceptions are always constant – and are always assumed to be constant.

History at time t is now defined as $\mathbf{H}_t \equiv (a_1, b_1; \ldots; a_{t-1}, b_{t-1})$. Lifetime utility at time $t \leq T$ for player i can be written

$$U_1^i\left(\mathbf{a}, \mathbf{b}; \beta^i\right) = u_1^i(a_1, b_1) + \beta^i \sum_{k=t+1}^{T} \delta^{k-t} u_k^i(\mathbf{H}_k, a_k, b_k) + \beta^i \delta^{T+1} u_{T+1}^i(\mathbf{H}_{T+1})$$

$$U_t^i\left(\mathbf{a}, \mathbf{b}; \gamma^i\right) = u_t^i(\mathbf{H}_t, a_t, b_t) + \gamma^i \sum_{k=t+1}^{T} \delta^{k-t} u_k^i(\mathbf{H}_k, a_k, b_k) + \gamma^A \delta^{T+1} u_{T+1}^A(\mathbf{H}_{T+1})$$

$\forall 1 < t \leq T$ for $i \in \{A, B\}$, $\mathbf{a} = (a_1, \ldots, a_T)$ and $\mathbf{b} = (b_1, \ldots, b_T)$.

The analysis for $T = 2$ naturally extends to one with more periods. Consider period T. In an equilibrium as perceived by player i, actions taken in the last period will be mutual best responses given the perceptions player i has about the future type of both players, and given the history of play up to period T. Again, we can write player i's perceptions about player j's future type as $\mu^{\mathbf{ij}}$, the only difference with the analysis in Section 6 being that this now refers to perceptions about types in any future period, rather than just the next. Given perceived play in period T, player i can then move back to period $T-1$ and derive a perceived equilibrium for that period. This process unravels until period 1, allowing us to write down the conditions for a subgame perfect Nash equilibrium as perceived by player i. We will refer to this simply as an equilibrium as perceived by player i.

Definition 8 *In the $T+1$-period, 2-player game with present-biased players, an equilibrium at time τ as perceived by player i, given her perceptions $\mu^{\mathbf{i}}$ and history $\mathbf{H}_t$ is a sequence $(a^i_\tau, b^i_\tau, a^i_{\tau+1}, b^i_{\tau+1}, \ldots, a^i_T, b^i_T)$ such that*

1. *For period T*

$$a^i_T = \arg \max_{a_T \in \mathcal{A}_T(\mathbf{H}_T)} \sum_{\gamma \in \{\beta,1\}} \mu^{iA}(\gamma)\, U^A_T\left(\mathbf{H}_T, a_T, b^A_T; \gamma\right)$$

$$b^i_T = \arg \max_{b_T \in \mathcal{B}_T(\mathbf{H}_T)} \sum_{\gamma \in \{\beta,1\}} \mu^{iB}(\gamma)\, U^B_T\left(\mathbf{H}_T, a^A_T, b_T; \gamma\right)$$

2. *For periods t with $\tau < t < T$*

$$\begin{aligned} a^i_t\left(\mathbf{H}_{\tau+1}; \mu^{\mathbf{A}}\right) = \arg \max_{a_t \in \mathcal{A}_t(\mathbf{H}_t)} \sum_{\gamma \in \{\beta,1\}} \mu^{iA}(\gamma)\, U^A_t \Big(&\mathbf{H}_t, a_t, b^i_t, a^i_{t+1}\left(\mathbf{H}_{t+1}; \mu^{\mathbf{A}}\right), \\ &b^i_{t+1}\left(\mathbf{H}_{t+1}; \mu^{\mathbf{A}}\right), \ldots, a^i_T(\mathbf{H}_T; \mu^{A}), b^i_T\left(\mathbf{H}_T; \mu^{\mathbf{A}}\right); \gamma\Big) \\ b^i_t\left(\mathbf{H}_{\tau+1}; \mu^{\mathbf{A}}\right) = \arg \max_{b_t \in \mathcal{B}_t(\mathbf{H}_t)} \sum_{\gamma \in \{\beta,1\}} \mu^{iB}(\gamma)\, U^B_t \Big(&\mathbf{H}_t, a^i_t, b_t, a^i_{t+1}\left(\mathbf{H}_{t+1}; \mu^{\mathbf{A}}\right), \\ &b^i_{t+1}\left(\mathbf{H}_{t+1}; \mu^{\mathbf{A}}\right), \ldots, a^i_T(\mathbf{H}_T; \mu^{\mathbf{A}}), b^i_T\left(\mathbf{H}_T; \mu^{\mathbf{A}}\right); \gamma\Big) \end{aligned}$$

3. *For $t = \tau$*

$$\begin{aligned} a^i_\tau &= \arg \max_{a_\tau \in \mathcal{A}_\tau(\mathbf{H}_\tau)} U^A_\tau \left(\mathbf{H}_\tau, a_\tau, b^i_\tau, a^i_{\tau+1}\left(\mathbf{H}_{\tau+1}; \mu^{\mathbf{A}}\right), b^i_{\tau+1}\left(\mathbf{H}_{\tau+1}; \mu^{\mathbf{A}}\right), \right. \\ &\quad \left. \ldots, a^i_T(\mathbf{H}_T; \mu^{\mathbf{A}}), b^i_T\left(\mathbf{H}_T; \mu^{\mathbf{A}}\right); \beta\right) \\ b^i_\tau &= \arg \max_{b_\tau \in \mathcal{B}_\tau(\mathbf{H}_\tau)} U^B_\tau \left(\mathbf{H}_\tau, a^i_\tau, b_\tau, a^i_{\tau+1}\left(\mathbf{H}_{\tau+1}; \mu^{\mathbf{A}}\right), b^i_{\tau+1}\left(\mathbf{H}_{\tau+1}; \mu^{\mathbf{A}}\right), \right. \\ &\quad \left. \ldots, a^i_T(\mathbf{H}_T; \mu^{\mathbf{A}}), b^i_T\left(\mathbf{H}_T; \mu^{\mathbf{A}}\right); \beta\right) \end{aligned}$$

Using these definitions, and considering play in period 1, we thus expect player A to take an action that she perceives to be part of a subgame perfect equilibrium for the entire game, while we expect player B to take an action that he perceives to be part of a subgame perfect equilibrium for the entire game.

Definition 9 *A perception-perfect outcome of the game is an outcome $(a^*_1, b^*_1, a^*_2, b^*_2, \ldots, a^*_T, b^*_T)$ such that $\forall \tau \in \{1, \ldots, T\}$ a^*_τ is part of an equilibrium at time τ as perceived by player A; b^*_τ is part of an equilibrium at time τ as perceived by player B.*

Note again that players do not learn anything about the perceptions or type of the other player upon observing her actions. Of course, we do allow a player to adapt her strategy upon observing the other player's action in a previous period. In other words, we assume that, say, player A takes the action that she feels is part of an equilibrium of the second period based on the actions that she actually observed to be played in period 1, rather than the actions that she expected to be played in period 1. Also, it is again straightforward to extend the analysis above to a case where, for example, one player is known to be time-consistent, or to a case where players cannot observe the other player's current type.

9 Application to Sequential Bargaining

In this section, we apply our framework to a dynamic bargaining game as proposed by Stahl (1972) and Rubinstein (1982).[13] In a Rubinstein bargaining game, two players, A and B, bargain over the division of a pie of size 1. There are $T+1$ periods. In odd-numbered periods ($t = 1, 3, 5, \ldots$) A proposes a sharing rule $(x_t, 1-x_t)$ that B can accept or reject. The first argument of the sharing rule always represents the share that A obtains, while the second is the share that B obtains. If B accepts an offer, the game ends and the proposed division is implemented. If B rejects, he makes a counteroffer in the next period that A can accept or reject. In this standard specification of the game, both players have the usual time-consistent preferences. Suppose that A uses discount factor δ_A, while B uses δ_B. Hence if $(x, 1-x)$ is accepted at time t, the payoffs to the players are $(\delta_A^t x, \delta_B^t (1-x))$.

Rubinstein bargaining with time-consistent players To fix ideas, we first consider the well-known solution to the standard model. Consider the case that T is even. We look for a subgame perfect equilibrium. In period T, A will accept any proposal. Player B will thus offer $(x_{T-1}, 1-x_{T-1}) = (0, 1)$. Knowing this, in period $T-1$, player A claims the highest share that would still make player B be willing to accept. Hence, she plays $(x_{T-1}, 1-x_{T-1}) = (1-\delta_B, \delta_B)$. With the same logic, in period $T-2$, player B offers A the lowest share she is still willing to accept, so $(x_{T-2}, 1-x_{T-2}) = (\delta_A (1-\delta_B), 1-\delta_A (1-\delta_B))$, etc. The equilibrium then has player A making an offer in period 1 that is immediately accepted. In the remainder, in a T-period game with T odd where it is common knowl-

[13] As noted in the introduction, we are not the first to consider Rubinstein bargaining with possibly naive present-biased players. Akin (2007) and Sarafidis (2006) derive comparable results by imposing restrictions on possible types and without explicitly modeling systems of perceptions

edge that A and B use discount factors δ_A and δ_B respectively, we denote the equilibrium sharing rule proposed in period t as $(x_t^*(\delta_A, \delta_B), 1 - x_t^*(\delta_A, \delta_B))$.

Rubinstein bargaining with present-biased players Now consider present-biased and possibly naive players. Our solution concept, perception-perfect outcome, requires that in each period each player chooses the action that is part of a subgame-perfect equilibrium, given her perceptions concerning future types of both players. In a sequential move game as we have here, this concept is relatively easy to implement.

Suppose that both players use the discount factor δ, but may differ in the extent to which they are time-consistent. We first derive the equilibrium as perceived by player A. She perceives the future player B to have type $\gamma^{AB} \in \{\beta, 1\}$ and the future player A to have type $\gamma^{AA} \in \{\beta, 1\}$. In other words, she perceives the future player B to use discount factor $\gamma^{AB}\delta$, and the future player A to use discount factor $\gamma^{AA}\delta$. Importantly, she also perceives all other players, present and future, to have those same perceptions. In period T, player A will accept anything. Player B will thus offer $(x_T, 1 - x_T) = (0, 1)$. Knowing this, in period $T-1$, player A claims the highest share she perceives she can get and that would still make player B be willing to accept. Hence, she plays $(x_{T-1}, 1 - x_{T-1}) = (1 - \gamma^{AB}\delta, \gamma^{AB}\delta)$. For period $T-2$, player A perceives player B to have the same perceptions concerning how play will continue in $T-1$. Hence, in period $T-2$, the current player A perceives B to offer A the lowest share she is still willing to accept, so $(x_{T-2}, 1 - x_{T-2}) = \left(\gamma^{AA}\delta\left(1 - \gamma^{AB}\delta_B\right), 1 - \gamma^{AA}\delta\left(1 - \gamma^{AB}\delta\right)\right)$. Hence, player A perceives future selves to act as if player A's true discount factor is $\gamma^{AA}\delta$, while player B's true discount factor is $\gamma^{AB}\delta$. Thus, the equilibrium as perceived by player A is $\left(x_t^*\left(\gamma^{AA}\delta, \gamma^{AB}\delta\right), 1 - x_t^*\left(\gamma^{AA}\delta, \gamma^{AB}\delta\right)\right)$, for all $t \in \{1, \ldots T\}$. Similarly, the equilibrium as perceived by player B is $\left(x_t^*\left(\gamma^{BA}\delta, \gamma^{AB}\delta\right), 1 - x_t^*\left(\gamma^{AA}\delta, \gamma^{AB}\delta\right)\right)$ for all $t \in \{1, \ldots T\}$.[14]

Infinite horizon To derive some qualitative predictions, we look at the case with an infinite horizon. From the literature on Rubinstein bargaining, we know the following. Suppose that players are time-consistent and have discount factors δ_A and δ_B. In a period where it is player A's turn to make an offer, the unique equilibrium then has equilibrium payoff to player A that equal

$$\pi_A(\text{A moves first}) = \frac{1 - \delta_B}{1 - \delta_A\delta_B}.$$

[14] Note that we also need that player A prefers her current offer above what she will get from B in the future, properly discounted. It is easy to show, however, that that is always satisfied.

If it is player B's turn to make an offer, the equilibrium payoff to player A is

$$\pi_A(\text{B moves first}) = \frac{\delta_A \left(1 - \delta_B\right)}{1 - \delta_A \delta_B}.$$

Of course, expressions for π_B are similar. A straightforward proof can be found in Shaked and Sutton (1984) or Fudenberg and Tirole (1991), chapter 4.

Now consider our model with possibly present-biased players. Again consider the equilibrium as perceived by player A. For the finite-horizon case, we saw that that equilibrium is equivalent to one with time-consistent players where $\delta_A = \gamma^{AA}\delta$ and $\delta_B = \gamma^{AB}\delta$. It is straightforward to see that that also applies to the infinite horizon case.[15] Thus, for any future period where A moves first, $i \in \{A, B\}$ perceives the continuation payoffs of player A to be

$$\pi_A^i(\text{A moves first}) = \frac{1 - \gamma^{iA}\delta}{1 - \gamma^{iA}\gamma^{iB}\delta^2}.$$

and those of player B

$$\pi_B^i(\text{A moves first}) = \frac{\gamma^{iA}\left(1 - \gamma^{iB}\right)}{1 - \gamma^{iA}\gamma^{iB}}$$

More generally, for any future period where j moves first, i perceives the continuation payoffs of player k to be

$$\pi_k^i(j \text{ moves first}) = \begin{cases} \frac{1-\gamma^{im}\delta}{1-\gamma^{iA}\gamma^{iB}\delta^2} & j = k, m \neq j \\ \frac{\gamma^{ik}\delta\left(1-\gamma^{ij}\delta\right)}{1-\gamma^{iA}\gamma^{iB}\delta^2} & j \neq k \end{cases}$$

for $i, j, k, m \in \{A, B\}$.

Note however that these expressions apply to any future period. When one player makes an offer to another player in the current time period, she will not base that offer on the perceived future type of that player, but rather on the current type. By assumption, she can observe the true current type β^B of the other player. If player A makes an offer in period 1, she will thus offer B the lowest amount he is willing to accept, given that if B can make a counteroffer in the next period, B's continuation payoff will be

[15] The proof is identical to that in Shaked and Sutton (1984) or Fudenberg and Tirole (1991), but using discount factors $\gamma^{AA}\delta$ and $\gamma^{AB}\delta$ rather that δ_A and δ_B. Hence, we do not repeat it here.

$\left(1-\gamma^{AA}\delta\right)/\left(1-\gamma^{AA}\gamma^{AB}\delta^2\right)$. Thus, A will offer

$$1-x_t\left(\gamma^{AA},\gamma^{AB}\right)=\frac{\beta^B\delta\left(1-\gamma^{AA}\delta\right)}{1-\gamma^{AA}\gamma^{AB}\delta^2}. \tag{17}$$

A similar analysis holds if it is player B's turn to move.

Yet, player A's offer will not always be accepted. If it is not, there will be delay in bargaining. Consider period 1. Player B perceives his continuation payoff in that period to be

$$\pi_B^B(B\text{ moves first})=\frac{1-\gamma^{BA}\delta}{1-\gamma^{BA}\gamma^{BB}\delta^2}.$$

He will thus reject A's offer (17) if he perceives it to give him a lower net present value than holding out and making a counteroffer in the next period, thus if

$$\frac{\beta^B\delta\left(1-\gamma^{AA}\delta\right)}{1-\gamma^{AA}\gamma^{AB}\delta^2}<\frac{\beta^B\delta\left(1-\gamma^{BA}\delta\right)}{1-\gamma^{BA}\gamma^{BB}\delta^2}.$$

We thus have:

Theorem 4 *In the perception-perfect outcome of the Rubinstein bargaining game with possibly naive present-biased players, in period t, player i will make an offer*

$$\frac{\beta^j\delta\left(1-\gamma^{ii}\delta\right)}{1-\gamma^{iA}\gamma^{iB}\delta^2}$$

to player j, $i\in\{A,B\}$, $j\neq i$. Player j will accept if and only if

$$\frac{1-\gamma^{ii}\delta}{1-\gamma^{iA}\gamma^{iB}\delta^2}\geq\frac{1-\gamma^{ji}\delta}{1-\gamma^{jA}\gamma^{jB}\delta^2} \tag{18}$$

Note that this expression does not directly depend on β. Thus, if we have present-biased preferences but no naivety, there will never be a delay in reaching an agreement. More generally, if A and B share the same perceptions (thus $\gamma^{AA}=\gamma^{BA}$ and $\gamma^{BA}=\gamma^{BB}$) the left- and right-hand side of (18) are equal and there is no delay. This is natural: any player offers to the other player what she perceives the other player is just willing to accept. As long as those perceptions are shared we get the same qualitative outcome as in the standard Rubinstein model, in the sense that the first offer will be immediately accepted.

Bargaining breakdown Above, we derived condition (18) for a delay in bargaining to occur.[16] Note that this immediately implies

Corollary 5 *In the Rubinstein bargaining model with present-biased, possibly naive players, negotiations break down in the sense that an agreement is never reached whenever the following conditions hold:*

$$\frac{1-\gamma^{AA}\delta}{1-\gamma^{AA}\gamma^{AB}\delta^2} < \frac{1-\gamma^{BA}\delta}{1-\gamma^{BA}\gamma^{BB}\delta^2} \tag{19}$$

$$\frac{1-\gamma^{BB}\delta}{1-\gamma^{BA}\gamma^{BB}\delta^2} < \frac{1-\gamma^{AB}\delta}{1-\gamma^{AA}\gamma^{AB}\delta^2}. \tag{20}$$

This result allows us to easily derive whether negotiations will break down in various scenarios. Consider for example the case in which both players are sophisticated about the other player, but naive about themselves. Thus, assume $\gamma^{AB} = \gamma^{BA} = \beta$ and $\gamma^{AA} = \gamma^{BB} = 1$. In that case, the denominators of both (19) and (20) are equal, and both conditions simplify to $1 - \delta < 1 - \beta\delta$, which is always satisfied. Hence, bargaining breaks down and the two parties never reach an agreement.

The intuition is as follows. If player A makes an offer to player B, she perceives the future B to be present-biased. Hence, her offer will be relatively low, as she perceives B to be very impatient. Player B however, perceives his future self to be patient. Therefore, he will not accept the current offer of player A, as he perceives to be able to do better. The same is true in the opposite case where player B makes an offer to A. Hence, players keep rejecting each others' offers and an agreement is never reached. Qualitatively, we thus get a similar result to that in the case of the common pool problem discussed earlier. Also there, the game broke down if players correctly anticipated their competitor's present-biasedness but were naive about their own.

Now suppose that each player is sophisticated about her own present-biased, but naive about the other player, so $\gamma^{AB} = \gamma^{BA} = 1$ and $\gamma^{AA} = \gamma^{BB} = \beta$. The conditions then simplify to $1-\beta\delta < 1-\delta$, which is never satisfied. Players immediately reach an agreement, like they do in the standard model. Now, player A perceives a future B to be more patient that B himself perceives his future self to be. Hence, the offer of A is actually better than B was expecting to get, and he will gladly accept.

When players differ in their naivety, the outcome depends on who moves first. Consider

[16] For more reasons why there may be delay in Rubinstein bargaining, see e.g. Yildiz (2004), and the references therein.

a case in which player A is naive about both players, while B is sophisticated about both. Hence $\gamma^{AA} = \gamma^{AB} = 1$ and $\gamma^{BA} = \gamma^{BB} = \beta$. Conditions (19) and (20) then simplify to

$$\frac{1-\delta}{1-\delta^2} < \frac{1-\beta\delta}{1-\beta\delta^2}$$
$$\frac{1-\beta\delta}{1-\beta^2\delta^2} < \frac{1-\delta}{1-\delta^2}$$

It is easy to see that the first condition is always satisfied, while the second never is. We thus get some delay in bargaining: player B rejects the offer of player A, but player A accepts the counteroffer. When B moves first, A accepts immediately, perceiving the offer of B as overwhelmingly generous.

10 Conclusion

In this paper, we proposed a solution concept, perception-perfect outcome, for games played between players with present-biased preferences that are possibly naive about their own future time inconsistency, and/or the time inconsistency of their competitor. A perception-perfect outcome essentially requires each player in each period to play an action that is consistent with subgame perfection, given the perception of that player concerning the time consistency of each player, and under the assumption that all other present and future players have the same perceptions.

We applied our solution concept to the common pool problem and to Rubinstein bargaining. In both cases, we showed that, if we assume that players are sophisticated about their competitor's future present-biasedness but naive about their own, the perfection-perfect equilibria of those games are disastrous. The common pool is exhausted even more quickly than with standard, rational players, and even more quickly than with present-biased but sophisticated players. Bargaining in the Rubinstein model breaks down completely, as each offer is rejected.

Of course, our approach is just the first step in the analysis of such games. There is much room for further analysis. For example, our perception-perfect outcome requires that players are strategically naive, in the sense that they do not take into account the possibility that other players may have different perceptions. Also, they do not learn from past behavior of other players. If offers in a bargaining game are rejected repeatedly, for example, one may expect players to take that into account and choose a somewhat different strategy when making further offers. Also, a highly sophisticated player may take advantage of her knowledge concerning the naivety of the other player to gain a strategic advantage.

Still, our framework is highly flexible and easily allows for extensions and modifications. For example, it is easy to allow for cases in which players are partially naive and realize their future present-biased to some limited extent. Also, it is straightforward to extend our perception-perfect outcome to a case with more than two types, or with more than two players. Our framework may even be applied to other (mis)perceptions and behavioral biases to which players are possibly unaware.

References

Akin, Z., 2007. Time inconsistency and learning in bargaining games. International Journal of Game Theory 36 (2), 275–299.

Akin, Z., 2009. Imperfect information processing in sequential bargaining games with present biased preferences. Journal of Economic Psychology 30 (4), 642–650.

Bulow, J. I., Geneakoplos, J., Klemperer, P. D., 1985. Multimarket oligopoly: Strategic substitutes and strategic complements. Journal of Political Economy 93, 488–511.

Chade, H., Prokopovych, P., Smith, L., 2008. Repeated games with present-biased preferences. Journal of Economic Theory 139 (1), 157–175.

Fedyk, A., 2017. Assymetric naïveté: Beliefs about self-control. mimeo, Harvard University.

Frederick, S., Loewenstein, G., O'Donoghue, T., 2002. Time discounting and time preference: A critical review. Journal of Economic Literature 40 (2), 351–401.

Fudenberg, D., Tirole, J., 1991. Game Theory, 1991. MIT Press.

Gans, J. S., Landry, P., 2019. Self-recognition in teams. International Journal of Game Theory 48, 1169–1201.

Kahneman, D., 2011. Thinking, Fast and Slow. Farrar, Straus and Giroux, New York.

Lu, S., 2016. Self-control and bargaining. Journal of Economic Theory 165, 390–413.

O'Donoghue, T., Rabin, M., 1999. Doing it now or later. American Economic Review, 103–124.

Pollak, R., 1968. Consistent planning. The Review of Economic Studies 35 (2), 201–208.

Pollak, R., Phelps, E., 1968. On second-best national saving and game-equilibrium growth. The Review of Economic Studies 35 (2), 185–199.

Sarafidis, Y., 2006. Games with time inconsistent players. mimeo.

Schweighofer-Kodritsch, S., 1984. Time preferences and bargaining. Econometrica 86 (1), 173–217.

Shaked, A., Sutton, J., 1984. Involuntary unemployment as a perfect equilibrium in a bargaining model. Econometrica: Journal of the Econometric Society 52 (6), 1351–1364.

Strotz, R., 1955. Myopia and inconsistency in dynamic utility maximization. The Review of Economic Studies 23 (3), 165–180.

Turan, A. R., 2019. Intentional time inconsistency. Theory and Decision 86, 41–64.

Weinschenk, P., 2021. On the benefits of time-inconsistent preferences. Journal of Economic Behavior and Organization 182, 185–195.

Yildiz, M., 2004. Waiting to persuade. The Quarterly Journal of Economics 119 (1), 223.

Appendix: Proofs of Section 7

Throughout, we make extensive use of the following straightforward result:

Lemma 1 *The function function $f(x) \equiv \frac{x}{a+bx}$ is strictly increasing in x for $a, b > 0$.*

Proof of Theorem 1. As $\beta, \rho \in (0,1)$, we have that $\left(\frac{1}{2}\beta\right)^{\frac{1}{\rho-1}}$ is decreasing in β, hence (from Lemma 1) Γ_s is decreasing in β. With $\Gamma_s \to \Gamma_{tc}$ as $\beta \to 1$, this implies from Lemma 1 that $\Gamma_s > \Gamma_{tc}$, which in turn implies from Lemma 1 that $\theta_s > \theta_{tc}$. Also note that, with $\rho \in (0,1)$, we have $\left(\frac{1}{2}\right)^{\frac{1}{\rho-1}} > 2$, which implies that $\Gamma_{tc} > 1/2$, hence $\theta_{tc} > 1/3$.

Define the function

$$\omega(\theta) \equiv \theta^{\rho} + \left(\frac{1}{2}(1-2\theta)\right)^{\rho}$$

As $\rho < 1$, we have

$$\frac{\partial\omega(\theta_s)}{\partial\theta_s} = \rho\left[\theta_s^{\rho-1} - \left(\frac{1}{2} - \theta_s\right)^{\rho-1}\right] < 0 \tag{21}$$

as $\theta_s > \theta_{tc} > 1/4$. Note that

$$\frac{\partial\left(\beta\omega\left(\theta_s\right)\right)}{\partial\beta} = \omega\left(\theta_s\right) + \beta\rho\frac{\partial\omega\left(\theta_s\right)}{\partial\theta_s}\frac{\partial\theta_s}{\partial\beta}$$

With $\frac{\partial\theta_s}{\partial\beta} < 0$, the second term is positive. The first term clearly is as well, hence $\frac{\partial(\beta\omega(\theta_s))}{\partial\beta} > 0$. With $\Omega_s = \left[\beta\omega\left(\theta_s\right)\right]^{\frac{1}{\rho-1}}$, this implies $\frac{\partial\Omega_s}{\partial\beta} < 0$, and hence from Lemma 1 that $\frac{\partial a_1^s}{\partial\beta} < 0$. Note that $a_1^{tc} = \lim_{\beta\uparrow 1} a_1^s$. Hence, this implies that $a_1^{tc} < a_1^s$. ■

Proof of Theorem 2. Note that we can write

$$\begin{aligned}\Omega_n &= \beta^{\frac{1}{\rho-1}}\omega\left(\theta_{tc}\right)^{\frac{1}{\rho-1}};\\ \Omega_s &= \beta^{\frac{1}{\rho-1}}\omega\left(\theta_s\right)^{\frac{1}{\rho-1}}.\end{aligned}$$

From (21), ω is decreasing in θ for $\theta > 1/4$. With $1/4 < \theta_{tc} < \theta_s$ and $\rho < 1$, this implies $\Omega_n < \Omega_s$, hence from Lemma 1, $a_1^n < a_1^s$. Also note that $\Omega_n = \beta^{\frac{1}{\rho-1}}\Omega_{tc}$, so $\Omega_n > \Omega_{tc}$, which implies $a_1^n > a_1^{tc}$. ■

Proof of Theorem 3. We first establish a number of lemmas.

Lemma 2 *We have the following:*

1. $\theta_{ns} < \theta_s < \theta_{sn}$.

2. $\theta_{ns} + \theta_{sn} < 2\theta_s$.

Proof. First note that $\Gamma_{tc} < \Gamma_s$ as Γ_s is decreasing in β and $\Gamma_s \to \Gamma_{tc}$ as $\beta \to 1$. Consider

$$\frac{\theta_{ns}}{\theta_s} = \frac{\frac{\Gamma_{tc}(1-\Gamma_s)}{1-\Gamma_{tc}\Gamma_s}}{\frac{\Gamma_s}{1+\Gamma_s}} = \frac{\Gamma_{tc} - \Gamma_{tc}\Gamma_s^2}{\Gamma_s - \Gamma_{tc}\Gamma_s^2}$$

This is smaller than 1 as $\Gamma_s > \Gamma_{tc}$. Next consider

$$\frac{\theta_{sn}}{\theta_s} = \frac{\frac{\Gamma_s(1-\Gamma_{tc})}{1-\Gamma_{tc}\Gamma_s}}{\frac{\Gamma_s}{1+\Gamma_s}} = \frac{\Gamma_s\left(1-\Gamma_{tc}\right)\left(1+\Gamma_s\right)}{\Gamma_s - \Gamma_{tc}\Gamma_s^2}$$

This is larger than 1 if the numerator is larger than the denominator, hence if $\Gamma_s\left(\Gamma_s - \Gamma_{tc}\right) > 0$, which is true as $\Gamma_s > \Gamma_{tc}$. This establishes 1. Next consider

$$\frac{\theta_{ns} + \theta_{sn}}{2\theta_s} = \frac{\frac{\Gamma_{tc}(1-\Gamma_s)+\Gamma_s(1-\Gamma_{tc})}{1-\Gamma_{tc}\Gamma_s}}{2\frac{\Gamma_s}{1+\Gamma_s}} = \frac{\Gamma_{tc} + \Gamma_s - 2\Gamma_s\Gamma_{tc}}{1 - \Gamma_{tc}\Gamma_s}\frac{1+\Gamma_s}{2\Gamma_s}$$

This is smaller than 1 if $(\Gamma_s - \Gamma_{tc})(1 - \Gamma_s) > 0$ which is true. This establishes 2. ■

Lemma 3 $\Omega_{sn} < \Omega_{ns}$.

Proof. The fact that $\theta_{ns} < \theta_{sn}$ and $\rho > 0$ implies that $\theta_{ns}^{\rho} + \left(\frac{1}{2}(1 - \theta_{ns} - \theta_{sn})\right)^{\rho} < \theta_{sn}^{\rho} + \left(\frac{1}{2}(1 - \theta_{ns} - \theta_{sn})\right)^{\rho}$. With $\rho < 1$ this immediately implies the result. ■

Lemma 4 $\Omega_{ns} > \Omega_s$.

Proof. From their definitions, to have $\Omega_{ns} > \Omega_s$, we need

$$\theta_{ns}^{\rho} + \left(\frac{1}{2}(1 - \theta_{ns} - \theta_{sn})\right)^{\rho} < \theta_s^{\rho} + \left(\frac{1}{2}(1 - 2\theta_s)\right)^{\rho}.$$

To prove that this is indeed the case, we proceed as follows. First, define $a \equiv \theta_{ns}$; $b \equiv \frac{1}{2}(1 - \theta_{ns} - \theta_{sn})$; $c \equiv \theta_s$; $d \equiv \frac{1}{2}(1 - 2\theta_s)$. Consider the function $f(x) = x^{\rho}$. Note that f is increasing and concave. We want to show

$$f(a) + f(b) < f(c) + f(d).$$

From Lemma 2.1, $a < c$. Also $b - d = \theta_s - \theta_{ns} > 0$ so $b > d$. Moreover $c + d = \frac{1}{2}$, while $a + b = \frac{1}{2}(1 + \theta_{ns} - \theta_{sn}) < \frac{1}{2}$ so $c + d > a + b$. Define $\Delta \equiv (c + d) - (a + b) > 0$ and consider $B \equiv b + \Delta$. By construction, $a + B = c + d$. However, with $a < c$ and $B > d$, the fact that $f(x)$ is increasing and concave then implies $f(a) + f(B) < f(c) + f(d)$. With $b < B$, this immediately implies that indeed $f(a) + f(b) < f(c) + f(d)$, which establishes the result. ■

The required result now follows;

$$a_1^A = \frac{\Omega_{ns}}{1 + \Omega_{ns} + \Omega_{sn}} > \frac{\Omega_{ns}}{1 + 2\Omega_{ns}} > \frac{\Omega_s}{1 + 2\Omega_s} = a_1^S,$$

where the first inequality follows from $\Omega_{sn} < \Omega_{ns}$ and Lemma 1, while the second follows from $\Omega_{ns} > \Omega_s$ and Lemma 1. ■