Unawareness in Dynamic Psychological Games*

Carsten S. Nielsen† Alexander Sebald‡

November 14, 2011

Abstract

Building on Battigalli and Dufwenberg (2009)'s framework of dynamic psychological games and the recent progress in the modeling of dynamic unawareness, we provide a general framework that allows for unawareness in the strategic interaction of players motivated by belief-dependent psychological preferences like reciprocity and guilt. We show that unawareness has a pervasive impact on the strategic interaction of psychologically motivated players. Intuitively, unawareness influences players' beliefs concerning, for example, the intentions and expectations of others which in turn impacts their behavior.

Keywords: Unawareness; Extensive-form games; Belief-dependent preferences; Sequential rationality; Sequential equilibrium.

JEL-Classifications: C72, C73, D80

*We would like to thank Geir Asheim, Pierpaolo Battigalli, Martin Dufwenberg, Aviad Heifetz, Georg Kirchsteiger, Burkhard Schipper, Peter Norman Sørensen, Lars Peter Østerdal and participants at the EDGE Jamboree 2010 (Dublin), the DGPE Workshop 2010 (Copenhagen), the Econometric Society's European Winter Meeting 2010 (Rome), the Royal Economic Society's Postgraduate Meeting 2011 (London), and the Seminar in Microeconomics 2011 (Lund) for helpful comments and suggestions. All errors are our own.

†Department of Economics, University of Copenhagen, Øster Farimagsgade 5, Building 26, DK-1353, Copenhagen K, Denmark. Phone: (+45) 3532-3051. Fax: (+45) 3532-3064. E-mail: carsten.nielsen@econ.ku.dk. Web: http://www.econ.ku.dk/phdstudent/nielsen.

‡Department of Economics, University of Copenhagen, Øster Farimagsgade 5, Building 26, DK-1353, Copenhagen K, Denmark. Phone: (+45) 3532-4418. Fax: (+45) 3532-3064. E-mail: alexander.sebald@econ.ku.dk. Web: http://www.econ.ku.dk/sebald.

1 Introduction

Recent lab and field evidence suggests that people not only care about the monetary consequences of their actions, but that their behavior is also driven by belief-dependent preferences [see e.g. Fehr *et al.* (1993), Charness and Dufwenberg (2006), Falk *et al.* (2008), Bellemare *et al.* (2010)]. Two prominent examples of belief-dependent preferences in the hitherto existing literature are reciprocity [see e.g. Rabin (1993), Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006)] and guilt aversion [see e.g. Charness and Dufwenberg (2006), Battigalli and Dufwenberg (2007b)]. Departing from the strictly consequentialist tradition in economics Geanakoplos *et al.* (1989) and Battigalli and Dufwenberg (2009) present general frameworks for analyzing the strategic interaction of people with belief-dependent preferences: 'psychological games'. Roughly speaking, psychological games are games in which players' preferences depend upon players' beliefs about the strategies that are being played, players' beliefs about the beliefs of others about the strategies that are being played, and so on *ad infinitum*.

A widely unspoken assumption that is underlying game-theoretic analyses, and therefore also the analyses of psychological games, is that players are aware of the complete structure of the strategic environment they are in. For example, even though players might not know the strategies played by others and know that they do not know, they nevertheless can identify all of them, i.e. all possible strategies that can potentially be played. In a sense players can never be truly surprised by the strategic choices of others. A rough intuition for this is that in game theoretic models strategies are both the modeler's and the players' description of the plans of action. But, if a player is unaware of (i.e., does not know that he does not know) some action, his strategies should be less complete than the modeler's. Hence, any game-theoretic model that does not explicitly distinguish between the players' description and the modeler's will fail to capture unawareness. Recent efforts have mapped out the framework for, difficulty in, and importance of modeling unawareness in non-psychological games [see e.g. Fagin and Halpern (1988), Dekel *et al.* (1998), Modica and Rustichini (1999), Halpern (2001), Heifetz *et al.* (2006), Halpern and Rêgo (2008), Heifetz *et al.* (2008), Feinberg (2009), Grant and Quiggin (2009), Li (2009) and Heifetz *et al.* (2011)].

It is, however, not only in non-psychological games that unawareness is important. In line with recent experimental evidence suggesting that people are more prone to selfish choices if they believe that others will remain unaware of them [see e.g. Dana *et al.* (2006), Dana *et al.* (2007), Broberg *et al.* (2007), Tadelis (2008), Andreoni and Bernheim (2009), Lazear *et al.* (2009)], we show in our analysis here that unawareness also has a profound impact on the strategic interaction of players in psychological games. To see this consider the following intuitive example: Imagine two friends, *Ann* and *Bob*. Assume it is *Bob*'s birthday, he is planning a party and would be very happy, if *Ann* could come. Unfortunately *Bob*'s birthday coincides with the date of *Ann*'s final exam at the university. She can either decide to take the exam the morning after *Bob*'s party or two weeks later at a second date. *Ann* is certain that *Bob* would feel let down, if she were to cancel his party without having a very good excuse. Quite intuitively, although *Ann* would really like to get over her exam as soon as possible, she might anticipate a feeling of guilt towards *Bob* if she canceled his party to take the exam the following morning. *Ann* knows that *Bob* would feel let down as he expects her to come. As a consequence, *Ann* might choose the second date in order to avoid feeling guilty. In contrast, consider now the following variant of the same example: *Ann* knows that *Bob* is unaware of the second date. In this situation *Ann* might choose to take the exam on the first date and not feel guilty. Since *Bob* is unaware of the second date and the final exam is a good excuse, he does not expect *Ann* to come. *Ann* knows this and, hence, does not feel guilty as *Bob* is not let down. In fact, if she were certain that *Bob* would never become aware of the second date, she probably had a strong emotional incentive to leave him unaware in order not to raise his expectations. That is, she had a strong incentive not to make him aware of the fact that she actually has the time to come to his party, but just wants to get over her exam. Interestingly, if *Ann* were only interested in her own payoff in this strategic situation with unawareness, she would not care whether *Bob* is or will become aware of the second date. She would simply not attend his party irrespective of his awareness. Only her belief-dependent feeling of guilt towards *Bob* creates the strong emotional incentive not to make him aware.

Bob's unawareness concerning *Ann*'s ability to come to his party and, connectedly, *Ann*'s incentive not to tell him about the second date intuitively highlight the focus of our analysis here. We analyze the influence and importance of unawareness concerning feasible paths of play for the strategic interaction of players in psychological games. Building on Battigalli and Dufwenberg (2009)'s framework of dynamic psychological games and the recent progress in the modeling of unawareness [i.e. Heifetz *et al.* (2006), Heifetz *et al.* (2008) and Heifetz *et al.* (2011)], we define a framework that allows for unawareness in the strategic interaction of players motivated by belief-dependent preferences. We characterize a version of the sequential equilibrium as a solution concept for our framework and provide different examples highlighting the role of unawareness in the strategic interaction of players motivated by reciprocity à la Dufwenberg and Kirchsteiger (2004) and guilt aversion à la Battigalli and Dufwenberg (2007b).

More precisely, in the formulation of our framework we first concentrate on extensive forms with complete information, observable actions and no chance moves. To allow for unawareness we divide standard extensive forms into subtrees consisting of paths of play and define extensive forms with subjective views. These extensive forms are in essence a special case of Heifetz *et al.* (2011)'s generalized extensive forms, and therefore embeddable in their generalized setting. Extensive forms with subjective views are complete strategic environments describing for each history in any subtree the histories players perceive to be at. Of course, such a complete dynamic awareness structure is typically not commonly known among players, and therefore should be interpreted from the modeler's point of view. To describe what a player considers to be the strategic environment in any particular history, we define T-partial extensive forms (henceforth; T-partial forms). That is, we divide extensive forms with subjective views into T-partial forms each defining the frame of mind of a player in a particular history.

Having defined our class of extensive forms with unawareness, we formally characterize belief-dependent preferences in our framework. In synthesis, we define players' strategies in each T-partial form and show that there exist beliefs about others' pure strategies (first-order beliefs), beliefs about the beliefs of others (second-order beliefs), and so on. That is, we show that infinite hierarchies of conditional beliefs exist in every T-partial form and use them for the general specification of the belief-dependent preferences and, hence, the characterization of our class of dynamic psychological games with unawareness. As mentioned above, specific types of belief-dependent preferences that can be embedded in our framework are among others reciprocity and guilt aversion. Importantly, in each history players think the strategic interaction is as described by the T-partial form they perceive to be in. Therefore, belief-dependent preferences defined in our framework are influenced by the awareness of players, the awareness players attribute to other players, the awareness players belief others attribute to them, and so on.

Given this characterization, we propose a sequential equilibrium solution concept and prove its existence. In our equilibrium analysis we assume that a profile of first-order beliefs (i.e., conjectures) in a T-partial form is derived from a behavioral strategy profile in the same T-partial form. This implies, that in equilibrium any two players confined to the same T-partial form will independently hold the same first-order beliefs about any third player. An assessment in a T-partial form, a behavioral strategy profile and a profile of infinite hierarchies of beliefs, is consistent if the profile of first-order beliefs is derived from the behavioral strategy profile and each higher-order belief assigns probability one to lower-order beliefs. Intuitively, players aware of the same must in equilibrium hold common, correct beliefs about each others infinite belief hierarchies. A consistent assessment in a T-partial form and sequential rationality (based on belief-dependent preferences) induce a sequential equilibrium in the same T-partial form. As players are unaware of any situation in which other players are aware of more than themselves, they believe that the strategic situation they perceive to be in is the most expressive. This implies that there exists an equilibrium strategy in which players confined to a T-partial form fix the equilibrium strategies of other players, whom they believe are confined to less descriptive partial forms, and then choose an equilibrium strategy based on this belief.

We focus on equilibrium reasoning in our analysis here to be able to clearly highlight the implications of our framework with unawareness in the context of e.g. Dufwenberg and Kirchsteiger (2004)'s model of sequential reciprocity and Battigalli and Dufwenberg (2007b)'s model of guilt aversion. Clearly, assuming equilibrium play is very demanding in a dynamic setting with unawareness in which every increase of awareness by definition is a shock or surprise. Given the importance of this issue we develop and discuss a non-equilibrium solution concept which embodies forward induction reasoning in the context of our dynamic framework with unawareness and belief-dependent preferences in a companion paper: Nielsen and Sebald (2011).

After having defined our framework and sequential equilibrium, we describe two examples to highlight the role of unawareness in the interaction of agents motivated by reciprocity and guilt aversion. First, we consider the sequential prisoners dilemma also analyzed by Dufwenberg and Kirchsteiger (2004) featuring a reciprocal second mover, *Bob*, who is unaware that the first mover, *Ann*, can defect. It is shown that *Bob* defects even when *Ann* cooperates as long as he is unaware of the fact that *Ann* could have defected. Intuitively, the way he perceives *Ann*'s kindness does not only depend on what she does, but also on what *Bob* thinks she could have done given his awareness of the strategic situation. Interestingly, *Ann* anticipates this and defects as well, if she cannot cooperate and simultaneously make *Bob* aware of the fact that she could have defected. As a second example we investigate the trust game with guilt aversion also analyzed in Battigalli and Dufwenberg (2009). We assume that the second mover, *Bob*, is aware of everything, but the first mover, *Ann*, is unaware that *Bob* can actually 'share part of the pie'. That is, *Ann* thinks that *Bob*'s only option is to 'grab the pie' if she 'trusts'. *Bob* who thinks that *Ann* does not expect him to share, since she is unaware of this option, does not feel any guilt towards *Ann* from grabbing the entire pie, if *Ann* trusts him. Of cause, as *Ann* thinks that *Bob* will grab everything, she chooses not to trust *Bob*–an unique equilibrium. *Ann*'s behavior implies that *Bob* has an incentive to make *Ann* simultaneously aware of the possibility that they can share the pie and the fact that he would feel too guilty to just grab everything, if she trusts him. Also in this example it is the level of unawareness in combination with *Bob*'s belief-dependent preference that shapes the strategic interaction. In synthesis, both examples highlight that unawareness in the interaction of players with belief-dependent preferences leads to very intuitive behavioral predictions distinct from predictions using non-psychological preferences or no unawareness. Furthermore, it becomes evident that managing others' awareness levels is an important and integral part of strategic interactions of players motivated by belief-dependent preferences.

So far we have restricted our analysis to strategic environments with observable actions and no chance moves. As will become clear, doing so restricts the type of unawareness scenarios that can be described. In particular, it implies that the framework presented up to now limits itself to strategic situations in which there is certainty of the unawareness of other players. To also allow for situations in which players are uncertain of the subjective views of others, situations in which players are uncertain regarding the uncertainties of others about yet other players' frames of mind, and so forth we go on and extend our setting to include imperfect information, chance moves and asymmetric information. To demonstrate the potential role chance plays in our framework, we sketch a simple principal-agent example in which a principal and an agent enter a business venture which either has a big or small potential. We assume the principal is only interested in his own profit, but the agent is motivated by reciprocity. Interestingly, in this strategic situation, if the agent is unaware of the fact that the business venture has a big potential, the principal has a material incentive to leave the agent unaware. Intuitively, by leaving the agent unaware the principal can appear kind by offering a wage payment to the agent to incentivize him to work hard that would be perceived as unkind were the agent aware of the true potential of the venture. Again, unawareness influences the 'menu' as perceived by the agent and, hence, his perception concerning the kindness of the principal.

The organization of the paper is as follows: In section 2 we introduce our framework. Following this, in section 3 we define psychological games with unawareness. Section 4 contains the definition of our equilibrium concept: sequential equilibrium. In section 5 we add uncertainty to our framework. Section 6 contain our conclusion.

2 Extensive forms with unawareness

In this section we present a class of extensive forms with unawareness. These extensive forms heavily rely on Heifetz *et al.* (2011)'s class of generalized extensive-forms, and should be seen as a special case of their setting that simplifies the properties needed to capture unawareness. As a starting point we describe a finite extensive form (2.1), define subtrees thereof (1) and explain how these can be used to capture players' subjective views regarding the feasible paths of play (2.3). Subsequently, we define T-partial extensive forms which combine the players' subjective frames of mind regarding the feasible paths of play to extensive forms with unawareness (2.4).

2.1 Extensive forms

A finite extensive form with complete information, observable actions and no chance moves is a tuple $\langle N, H\rangle$ where $N = \{1, \ldots, n\}$ is the set of players, and H is the finite set of histories. A history of length l is a sequence $h = (a^1, \ldots, a^l)$ where each $a^t = (a_1^t, \ldots, a_n^t)$ represents the profile of actions chosen at stage t $(1 \leq t \leq l)$. The history $\tilde{h} = (\tilde{a}^1, \ldots, \tilde{a}^k)$ precedes $h = (a^1, \ldots, a^l)$, written $\tilde{h} < h$, if $\tilde{h}$ is a prefix of h (i.e., $k < l$ and $(\tilde{a}^1, \ldots, \tilde{a}^k) = (a^1, \ldots, a^k)$). The empty history (or root) of length 0, denoted h^0, is an element of H. Let Z denote the set of terminal histories in H.

The set of feasible actions for player $i \in N$ at history h is $A_i(h) = \{a_i' : (h, a_i', a_{-i}) \in H\}$. Players move simultaneously at each stage. This is without loss of generality, since the set of feasible actions of players may depend on actions chosen in previous stages and may be singleton. We say player i is active at h whenever $A_i(h)$ contains more than one element, otherwise he is passive. At a terminal history z, $A_i(z) = \varnothing$ for every player i.

In what follows the extensive form $\langle N, H\rangle$ represents the objectively feasible paths of play.

2.2 Subtrees

To allow for the possibility that players are unaware, we define subjective views regarding the feasible paths of play. Subjective views are described by subtrees of H. More formally, consider a family of subtrees $\boldsymbol{T}$ of H where each subtree $T \in \boldsymbol{T}$ is defined as follows:

Definition 1. A set of histories $T \in \boldsymbol{T}$ is a subtree if for some nonempty subset of terminal histories $E \subseteq Z$:

$$T = \{h \in H : h \leq z \text{ for some } z \in E\}.$$

As can be seen from Definition 1, all $T \in \boldsymbol{T}$ start at the root and end at one or more terminal histories in Z. Each subtree $T \in \boldsymbol{T}$ thus represents a set of feasible paths of play. The family of subtrees $\boldsymbol{T}$ forms a join semi-lattice under the inclusion partial order relation $\preceq$ (that is, it is ordered by the inclusion of paths of play).

We assume the maximal subtree $T = H$ which represents the objectively feasible paths of play is an element of $\boldsymbol{T}$. Over and above this, we do not consider all possible subtrees of H, but assume that the family $\boldsymbol{T}$ selectively contains all subtrees that are relevant to the strategic situation. Each subtree $T \in \boldsymbol{T}$ represents some subjective view regarding the feasible path of play. The subjective view may be a frame of mind of some player i, or a frame of mind assigned to i by some other player j, whose own frame of mind might be represented by a subtree different from player i's, and so on. It might even represent a player's own view on his own frame of mind at an earlier stage of the game, after his awareness regarding the feasible paths of play has evolved.

As all $T \in \boldsymbol{T}$ start at the root and end at one or more terminal histories in Z, players' subjective views regarding the feasible path of plays are well-defined. A player may, however, subjectively think that the strategic interaction starts with all players being passive effectively moving the start of the interaction, as he perceives it, to a history at which one or more players are active.

To highlight that a player's subjective view of a history is framed by the subtree it is in, we write h_T whenever a history h is in subtree T. This implies that although histories may have copies (histories for which the sequence of actions coincide) in other subtrees, they are still treated as distinct elements. Finally, notice that each terminal history z_T in each subtree, by definition, has a copy in the objective set of terminal histories Z.

2.3 Extensive forms with subjective views

We define an extensive form with subjective views using the extensive form $\langle N, H \rangle$ and the family of subtrees $\boldsymbol{T}$. This structure is typically not commonly known among players, and therefore has to be interpreted from the modeler's point of view.

Denote by $\mathcal{T} = \bigcup_{T \in \boldsymbol{T}} T$ the union of subtrees. An extensive form with different subjective views is a tuple $\langle N, \mathcal{T}, \phi \rangle$ where $\phi = (\phi_1, \ldots, \phi_n)$ is a sequence of possibility functions associating each player at each history with a (perhaps different) history subjectively considered as true/possible. The interpretation is that when the history is h_T, player i knows only that the history is the outcome of $\phi_i(h_T)$.

We define the possibility function by:

Definition 2. For each player $i \in N$ in $\langle N, \mathcal{T}, \phi \rangle$ there exists a possibility function

$$\phi_i : \mathcal{T} \to \mathcal{T}.$$

We denote by Φ_i the set of histories known by player i at all histories in $\langle N, \mathcal{T}, \phi \rangle$.[1] That is, Φ_i is the complete set of all mappings of the possibility function of player i at all histories in all subtrees in $\boldsymbol{T}$.[2] Let Φ_i^Z be the set of mappings from terminal histories in $\mathcal{T}$.

We follow Heifetz *et al.* (2011) by assuming that the set of functions Φ_i adheres to properties that regulate what history players can perceive to be in. When actions are observable they become common knowledge once they have been taken. This implies that at a history

[1] I.e., $\Phi_i = \{\phi_i(h) : \text{ for all } h \in T, \text{ for all } T \in \boldsymbol{T}\}$

[2] Note that since updated beliefs at terminal histories matter in dynamic psychological games (Battigalli and Dufwenberg, 2009), also the players' perceptions at these histories need to be defined.

each player knows the sequence of actions that have led to it, knows that the other players know, and so on. Observable actions thus places two natural properties on the possibility functions that the modeler shall take into account. First, the history that a player subjectively perceives to be the true history must be described by the sequence of actions observed. That is, $\phi_i(h_T)$ must map into a copy of h_T (**Property 1**) ($P1$[3] *Generalized reflexivity*). Second, observable actions imply that $\phi_i(h_T)$ maps into a single history (**Property 2**) ($P2$ *Introspection*). I.e., since a player knows the sequence of actions and therefore the history, he does not need to make any inferences about its legitimacy.

Besides these two implied properties, modeling dynamic unawareness requires further properties.[4] First, a player cannot imagine a copy in a subjective view that is not available at the subtree that frames the history:

Property 3 ($P0$ *Confined awareness*)**.** At a history h_T in subtree T, the possibility function can only map into a copy $h_{T'}$ in an embeddable subtree T', $h_{T'} = \phi_i(h_T)$.

Second, a player aware of a subjective view cannot be unaware in an embeddable subtree:

Property 4 ($P3$ *Subtrees preserve awareness*)**.** If at a history h_T in a subtree T the possibility function maps into itself, $h_T = \phi_i(h_T)$, then at a copy $h_{T'}$ in a embeddable subtree T', the possibility function must also map into itself, $h_{T'} = \phi_i(h_{T'})$.

Third, if a player is ignorant of a subjective view in a subtree, then he should be ignorant in any embeddable subtree:

Property 5 ($P4$ *Subtrees preserve ignorance*)**.** Consider subtrees T'', T' and T such that subtree T'' can be embedded in subtree T' which again can be embedded in T. If at a history h_T in subtree T the possibility function maps into a copy $h_{T''}$ in subtree T'', $h_{T''} = \phi_i(h_T)$, and subtree T' also contains a copy $h_{T'}$ of h_T, then the possibility functions at h_T and $h_{T'}$ in subtree T' must both map into $h_{T''}$, $\phi_i(h_{T'}) = \phi_i(h_T)$.

Fourth, a player cannot forget a subjective view he considered possible at some history:

Property 6 ($P5$ *Subtrees preserve knowledge*)**.** Consider subtrees T'', T' and T such that subtree T'' can be embedded in subtree T' which again can be embedded in T. If at a history h_T in subtree T the possibility function maps into a copy $h_{T'}$ in subtree T', $h_{T'} = \phi_i(h_T)$, and subtree T'' also contains a copy $h_{T''}$ of h_T, then the possibility function at $h_{T''}$ must also map into itself, $h_{T''} = \phi_i(h_{T''})$.

[3]The number following each P corresponds to the respective property in Heifetz *et al.* (2006).

[4]A graphical illustration of each property can be found on page 11 and 12

Properties 3-6 are all static properties in the sense that they relate a history in one subtree to a copy in another subtree. However, we also need to take care of the dynamic character of the extensive form with the help of two additional properties.

First, a player cannot expect to forget a subjective view he is currently aware of:

Property 7 (*I3*[5] *No expectation to forget currently conceivable paths*)**.** If at a history h_T in subtree T the possibility function maps into a copy $h_{T'}$ in embeddable subtree T', $h_{T'} = \phi_i(h_T)$, and the copy $h_{T'}$ has a successor $h'_{T'}$, $h'_{T'} < h_{T'}$, then the possibility function at $h'_{T'}$ must map into itself, $h'_{T'} = \phi_i(h'_{T'})$.

Second, a player's awareness can only increase along a path of play:

Property 8 (*DADynamic awareness*)**.** Consider subtrees T'', T' and T such that subtree T'' can be embedded in subtree T' which again can be embedded in T. If at history h_T in subtree T the possibility function maps into copy $h_{T''}$ in subtree T'', $h_{T''} = \phi_i(h_T)$, then at a succeeding history $h'_T < h_T$ in subtree T the possibility function must map into a copy of the succeeding history $h'_{T'}$ in subtree T'.

This concludes the properties of the possibility function needed to describe unawareness in our class of extensive forms. It is worth noticing that in a dynamic setting without asymmetric awareness (implying only one subtree in $\boldsymbol{T}$) we do not need any of the awareness properties, if there are two subtrees in $\boldsymbol{T}$ only property 3 is binding, and if there are three or more subtrees in $\boldsymbol{T}$ all the properties 3-8 are binding. In the special case in which there is only one stage (one-shot interaction) we only need to consider property 3-6. Property 7-8 concerns interactions with two or more stages (sequential interaction).

2.4 T-partial extensive forms

Extensive forms with subjective views $\langle N, \mathcal{T}, \phi \rangle$ constitute the modelers point of view. To describe what a player considers to be the strategic environment in any particular history we define T-partial extensive forms (henceforth; T-partial forms). That is, we divide the extensive form with subjective views into T-partial forms, which define the histories that players with a given frame of mind T subjectively perceive to be possible.

For any two subtrees in $T, T' \in \boldsymbol{T}$ we denote by $T \rightarrowtail T'$ whenever for some history in subtree T it is the case that the possibility function of some player maps into subtree T'. However, this is not enough because at some history in subtree T some player may perceive, that at a copy in some other subtree T' some other player might perceive to be at a copy

[5] The number following P in the last two properties corresponds to properties in Heifetz *et al.* (2011).

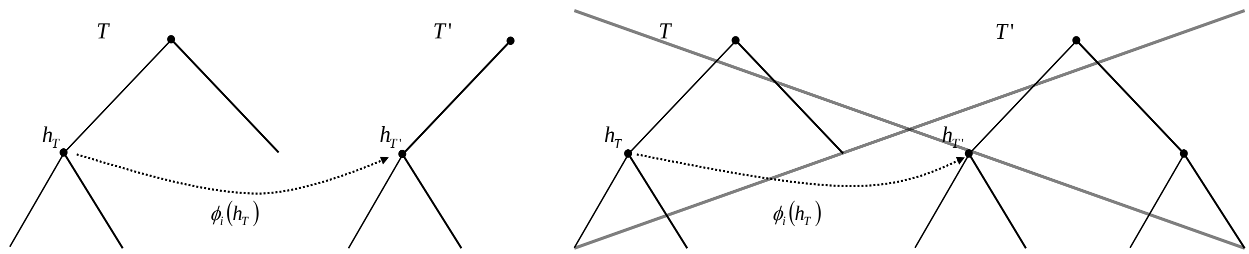

Figure 1: A correct and incorrect application of property 3

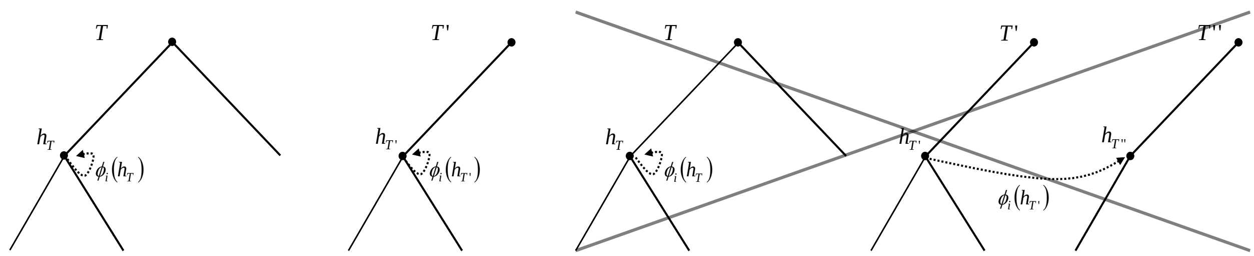

Figure 2: A correct and incorrect application of property 4

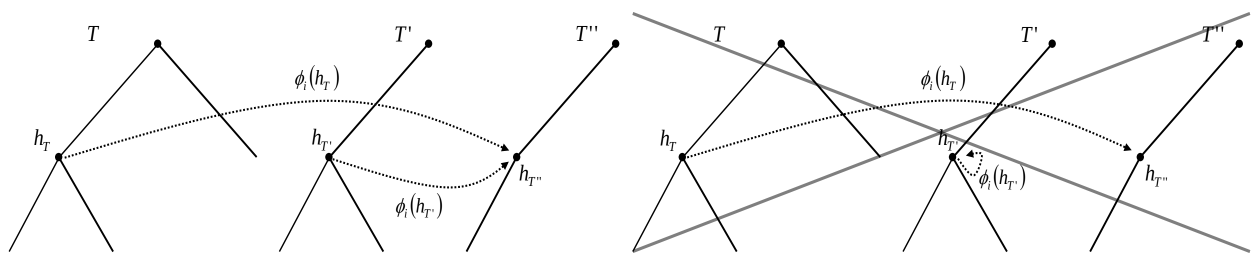

Figure 3: A correct and incorrect application of property 5

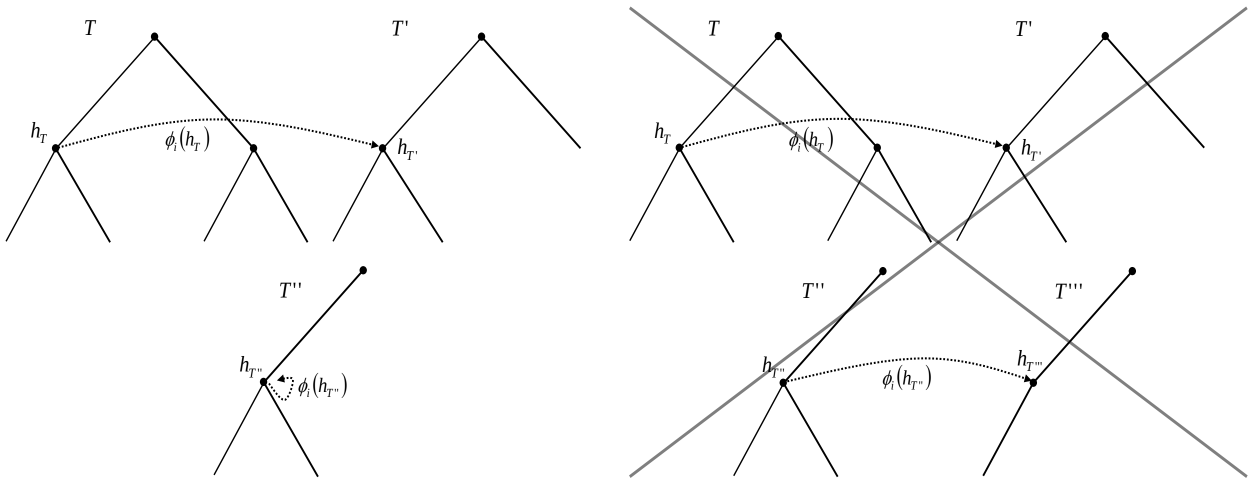

Figure 4: A correct and incorrect application of property 6

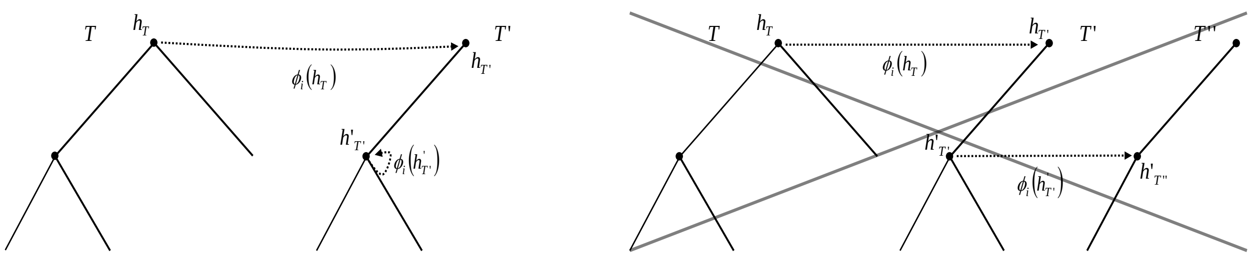

Figure 5: A correct and incorrect application of property 7

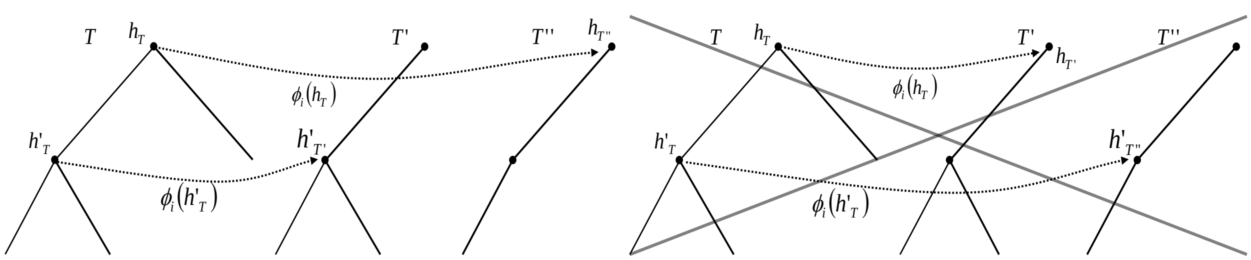

Figure 6: A correct and incorrect application of property 8

in yet another subtree T'', and so on. We therefore need to go beyond the binary relation and consider the transitive closure $\hookrightarrow$. Formally, a subtree T'' is in the transitive closure, denoted $T \hookrightarrow T''$, if and only if there is a sequence of subtrees $T, T', \ldots, T'' \in \boldsymbol{T}$ satisfying $T \rightarrowtail T' \rightarrowtail \cdots \rightarrowtail T''$. For every subtree $T \in \boldsymbol{T}$, the T-partial form is given by the partially ordered set of subtrees including T and all subtrees $T' \in \boldsymbol{T}$ satisfying the transitive closure $T \hookrightarrow T'$. Importantly, a T-partial form is itself an extensive form with subjective views. Denote the set of histories of player i in any T-partial form by Φ_i^T where $\Phi_i^T \subseteq \Phi_i$.[6]

The conception and interpretation of T-partial forms can be demonstrated by a simple example. Consider the extensive form underlying the sequential prisoners dilemma also analyzed by Dufwenberg and Kirchsteiger (2004).

[Figure 7]

Figure 7 shows two multi-player decision trees T and T' where $N = \{Ann, Bob\}$. Subtree T describes the objectively feasible extensive form underlying the sequential prisoners dilemma and T' a possible subtree thereof. Consider first subtree T. At the initial history h_T^0, *Ann* can choose between Cooperate (C) and Defect (D) and *Bob* is passive. While in history h_T^1 and h_T^2 *Bob* can choose between cooperate (c) and defect (d), and *Ann* is passive. In subtree T', on the other hand, *Ann* can only choose Cooperate (C) in history $h_{T'}^0$ and *Bob* can choose between cooperate (c) and defect (d) in history $h_{T'}^1$ following *Ann*'s choice Cooperate. Let the family of subtrees be given by $\boldsymbol{T} = \{T, T'\}$. Associated to this, a possible extensive form with subjective views is described in Figure 8.[7]

[Figure 8]

Intuitively, the possibility function drawn in Figure 8 indicates that e.g. at history h_T^0 in subtree T representing the objectively feasible paths of play *Ann* knows h_T^0, i.e. she perceives to be in h_T^0 and her initial frame of mind is given by T. In contrast, at history h_T^0 Bob perceives to be in history $h_{T'}^0$, i.e. $\phi_B(h_T^0) = h_{T'}^0$, and his initial frame of mind is given by T'. Remember, when $\boldsymbol{T}$ contains only two subtrees only property 3 is binding.

Given the family $\boldsymbol{T}$ and the set of possibility functions for both players, there are two partial forms in our example: (i) there is a T-partial form containing $\{T, T'\}$ and (ii) a T'-partial form containing $\{T'\}$. Rather than conceiving the whole extensive form with subjective views, players at each history perceive to be in a T-partial form. We will refer to the T-partial form that player i at ϕ_i is aware of as a T_{ϕ_i}-partial form. E.g., at history h_T^0 in

[6] I.e., $\Phi_i^T = \{\phi_i(h) : \text{ for all } h \in T, \text{ for all } h \in T' : T' \text{ satisfies } T \hookrightarrow T'\}$

[7] For the sake of clarity we have not labeled the terminal histories in Figure 7, nor drawn Bob's possibility function in histories $h_{T'}^0$ and $h_{T'}^1$ or the possibility function of both players in the terminal histories.

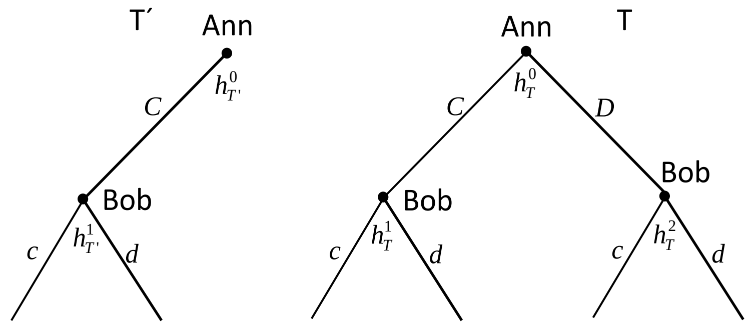

Figure 7: An extensive form and a subtree thereof

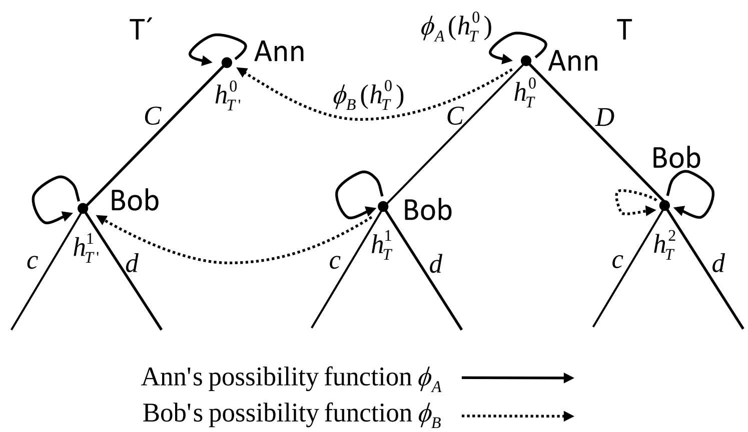

Figure 8: An extensive form with subjective views

Figure 8 *Ann* is aware of the $T_{\phi_A(h_T^0)}$-partial form containing $\{T, T'\}$, whereas *Bob* is aware of the $T_{\phi_B(h_T^0)}$-partial form containing only $\{T'\}$. Furthermore, as there is no uncertainty about the awareness of others and their knowledge about others' awareness, *Ann* knows that *Bob* is aware of the $T_{\phi_A(\phi_B(h_T^0))}$-partial form, the partial form at the outcome of the composite function $\phi_A(\phi_B(h_T^0))$, and *Bob* perceives *Ann* to be aware of the $T_{\phi_B(\phi_A(h_T^0))}$-partial form both containing $\{T'\}$.

The interpretation is the following. In the T_{ϕ_A}-partial form *Ann*'s frame of mind is T_{ϕ_A}. *Ann* knows *Bob*'s frame of mind $T_{\phi_A(\phi_B)}$ and knows what *Bob* thinks about her frame of mind $T_{\phi_A(\phi_B(\phi_A))}$, etc. Conversely, in the T_{ϕ_B}-partial form *Bob*'s frame of mind is T_{ϕ_B}, he has a perception about *Ann*'s frame of mind $T_{\phi_B(\phi_A)}$ and what *Ann*'s thinks about his frame of mind $T_{\phi_B(\phi_A(\phi_B))}$, and so on. Properties 3-6 imply that for two or more iterations, the perception function maps into the same history. More descriptively, let $\phi_A^0 = \phi_A$ and $\phi_A^{\iota+1} = \phi_A(\phi_B^\iota)$, then the properties imply that *Ann*'s perception at a history is such that $\phi_A(\phi_B^0) = \phi_A^{\iota+1}$ for all $\iota \geq 0$. Similar for *Bob*, or any other player. In the following we will thus without loss of generality only state the two first iterations.

This example also highlights that at $\phi_i \in \Phi_i$, player i is aware of subjective views in the T_{ϕ_i}-partial form. Conversely, player i is unaware of subjective views defined by any subtree $T' \in \boldsymbol{T}$ not in the transitive closure, $T_{\phi_i} \nrightarrow T'$. That is, player i does not know subtrees not in the transitive closure, he does not know that he does not know, and so on.

3 Dynamic psychological games with unawareness

Building on the class of extensive forms developed in the previous section we go on and define dynamic psychological games with unawareness. That is, we formally describe the strategies of players in extensive forms with subjective views (3.1), characterize hierarchies of conditional beliefs (3.2-3.3) and, finally, define psychological games with unawareness and belief dependent preferences (3.4). Lengthy mathematical proofs are relegated to Appendix (A).

3.1 Strategies

The set of feasible actions for player i at $\phi_i \in T$ is denoted $A_i(\phi_i) = \{a_i' : (\phi_i, a_i', a_{-i}) \in T\}$ and $A_i(\phi_i) \subseteq A_i(h)$ at $\phi_i(h)$. A player's strategy in our framework is a complete description of his disposition to act at different histories in the modeler's game. We let

$$S_i = \prod_{\phi_i \in \Phi_i \setminus \Phi_i^Z} A_i(\phi_i)$$

denote the set of (pure) strategies of player i from the modelers point of view. A typical strategy is denoted $s_i = (s_i(\phi_i))_{\phi_i \in \Phi_i \setminus \Phi_i^Z}$, where $s_i(\phi_i)$ is the action a_i that would be selected by strategy s_i if ϕ_i known by player i occurred. Define $S = \prod_{i \in N} S_i$ and $S_{-i} = \prod_{j \neq i} S_j$. The set of i's strategies that allow ϕ_i is denoted $S_i(\phi_i)$. Similar notation is used for strategy profiles consistent with ϕ_i: $S(\phi_i) = \prod_{j \in N} S_j(\phi_i)$ and $S_{-i}(\phi_i) = \prod_{j \neq i} S_j(\phi_i)$.

Strategies are conjunctions of behavioral conditionals of the form 'if ϕ_i occurred, player i would choose a_i.' We interpret such behavioral conditionals as objective descriptions: the strategy of player i at a known history ϕ_i is necessarily executed and every conditional 'if ϕ_i occurred, player i would choose a_i' is true if and only if $a_i = s_i(\phi_i)$, independently of whether or not a_i prescribed by $s_i(\phi_i)$ at ϕ_i is excluded by an earlier move of that very strategy.

For a strategy $s_i \in S_i$ and subtree $T \in \boldsymbol{T}$, we denote by s_i^T the realization equivalent strategy in the T-partial form. Two strategies s_i and s_i^T are realization equivalent if $s_i(\phi_i) = s_i^T(\phi_i)$ for every mapping ϕ_i of player i at histories in the T-partial form. We let $\psi(s_i^T) \in S_i^{T'}$ denote player i's realization equivalent strategy based on s_i^T in the T'-partial form.[8] Finally, we define $\zeta_i(s) \in \Phi_i^Z$ as player i's perceived terminal ϕ_i induced by $s^T = (s_i^T)_{i \in N}$.[9]

Define $S^T = \prod_{i \in N} S_i^T$ and $S_{-i}^T = \prod_{j \neq i} S_j^T$. The set of i's strategies that allow for history ϕ_i is denoted $S_i^T(\phi_i)$. Similar notation is used for strategy profiles consistent with ϕ_i: $S^T(\phi_i) = \prod_{j \in N} S_j^T(\phi_i)$ and $S_{-i}^T(\phi_i) = \prod_{j \neq i} S_j^T(\phi_i)$. Remember, each coordinate ϕ_i of $\phi = (\phi_1, \ldots, \phi_n)$ may map into copies of histories in different subtrees.

3.2 Conditional belief systems

Consider a player who is uncertain about which element in a set X is true. Assume X is a compact Polish space. Players assign probabilities to events $E, F, ...$ in the Borel sigma-algebra $\mathcal{B}_X$ of X according to some (countably additive) probability measure. Let $\Delta(X)$ denote the set of all probability measures on $(X, \mathcal{B}_X)$. As events unfold players update their beliefs. Let $\mathcal{C} \subseteq \mathcal{B}_X$ be a nonempty, finite or countable collection, such that $\varnothing \notin \mathcal{C}$. The interpretation is that player i is uncertain about the element $x \in X$, and $\mathcal{C}$ represents a collection of 'conditionals' or 'hypotheses'.

A conditional probability system (*cps*) on $(X, \mathcal{B}_X, \mathcal{C})$ is a mapping $\mu(\cdot|\cdot) : \mathcal{B}_X \times \mathcal{C} \to [0,1]$ such that, for all $E \in \mathcal{B}_X$ and $F', F \in \mathcal{C}$;

(i) $\mu(\cdot|\cdot) \in \Delta(X)$,

[8] The realization equivalent function $\psi(s_i^T)$ is defined by the $\psi : S_i^T \to S_i^{T'}$ such that $\psi(s_i^T) = (s_i^T(\phi_i))_{\phi_i \in \Phi_i^{T'} \cap \Phi_i^T} = s_i^{T'}$.

[9] The path function $\zeta_i : S^T \to \Phi_i^Z$ is defined such that $\phi_i^z = (a^1, \ldots, a^L) = \zeta_i(s^T)$ if and only if $a^1 = s^T(\phi_i^0)$ and $a^{t+1} = s^T(a^1, \ldots, a^t)$ for all $t \in \{1, \ldots, L-1\}$.

(ii) $\mu(F|F) = 1$, and

(iii) $E \subseteq F' \subseteq F$ implies $\mu(E|F) = \mu(E|F')\mu(F'|F)$.

We regard the set of cps' on $(X, \mathcal{B}_X, \mathcal{C})$ as a subset of the topological space $[\Delta(X)]^{\mathcal{C}}$ (the set of mappings from $\mathcal{C}$ to Δ(X)), where $\Delta(X)$ is endowed with the topology of weak convergence of measures which makes it Polish and $[\Delta(X)]^{\mathcal{C}}$ is endowed with the product topology. It is denoted by $\Delta^{\mathcal{C}}(X)$. Accordingly, we often write $\mu = (\mu(\cdot|F))_{F\in\mathcal{C}} \in \Delta^{\mathcal{C}}(X)$.

In our analysis, the family $\mathcal{C}$ corresponds to the collection of events that the players can observe in the game that they are aware of, namely, the elements of Φ_i.[10] If conditioning event F corresponds to ϕ_i, then we abbreviate as $\mu(\cdot|F) = \mu(\cdot|\phi_i)$.

3.3 Construction of hierarchies of conditional beliefs

A player's disposition to hold subjective beliefs about strategic choices is defined recursively by the hierarchy of conditional beliefs. The hierarchies of cps' in our framework are defined on extensive forms with subjective views. At ϕ_i each player must have a cps over each of the other player j's strategies $S_j^{T_{\phi_i}}$ in the T_{ϕ_i}-partial form player i's first-order cps', μ_i^1, are then elements of $\Delta^{\Phi_i}(S_{-i}^{T_{\phi_i}})$. Note that because i can become aware of more, the domain of the belief function can change depending on the conditioning event ϕ_i. Since each player may not know the belief of the others, each must have a second-order belief; i's second-order belief, μ_i^2, is thus an element of $\Delta^{\Phi_i}(S_{-i}^{T_{\phi_i}} \times \prod_{j\neq i}\Delta^{\Phi_j}(S_{-j}^{T_{\phi_i(\phi_j)}}))$. Similar for $-i$. Formally, define spaces

$$
\begin{aligned}
X_{-i}^0 &= S_{-i}^{T_{\phi_i}} \\
X_{-i}^1 &= S_{-i}^{T_{\phi_i}} \times \prod_{j\neq i} \Delta^{\Phi_j}(S_{-j}^{T_{\phi_i(\phi_j)}}) \\
&\vdots \\
X_{-i}^k &= X_{-i}^{k-1} \times \prod_{j\neq i} \Delta^{\Phi_j}(X_{-j}^{k-1}) \\
&\vdots
\end{aligned}
$$

A cps μ_i^k is called a k-order cps. For $k > 1$, μ_i^k is a joint cps on the other players' strategies and $(k-1)$-order cps' (possibly in different T-partial extensive forms). A hierarchy of cps' is

[10] Conditioning events corresponds to elements in Φ_i if $\mathcal{C} = \{F \subseteq S_{-i}^{T_{\phi_i}} : F = S_{-i}^{T_{\phi_i}}(\phi_{-i}), \phi_{-i} \in \Phi_{-i}\}$ if $X = S_{-i}^{T_{\phi_i}}$, or $\mathcal{C} = \{F \subseteq S_{-i}^{T_{\phi_i}} \times Y : F = S_{-i}^{T_{\phi_i}}(\phi_{-i}) \times Y, \phi_{-i} \in \Phi_{-i}\}$ if $X = S_{-i}^{T_{\phi_i}} \times Y$ where Y is a compact Polish space typically representing a set of other players' beliefs.

a countably infinite sequence of cps' $\boldsymbol{\mu}_i = (\mu_i^1, \mu_i^2, \dots) \in \prod_{k>0} \Delta^{\Phi_i}(X_{-i}^{k-1})$. A sequence of cps' $\boldsymbol{\mu}_i$ is coherent if the cps' of distinct orders cannot contradict each other: $\forall\ k > 0,\ \forall\ \phi_i \in \Phi_i$

$$\mu_i^k(\cdot|\phi_i) = \text{marg}_{X_{-i}^{k-1}} \mu_i^{k+1}(\cdot|\phi_i). \tag{1}$$

Let $\boldsymbol{M}_i$ denote the set of collectively coherent cps' $\boldsymbol{\mu}_i$ of player i in the extensive form with subjective views. We let M_i^k denote the set of collectively coherent k-order beliefs, that is, the projection of $\boldsymbol{M}_i$ on $\Delta^{\Phi_i}(X_{-i}^{k-1})$. Define $M_{-i}^k = \prod_{j\neq i} M_j^k$ and $\boldsymbol{M}_{-i} = \prod_{j\neq i} \boldsymbol{M}_j$.

The assumption that cps' are collective coherent closes the belief system; a hierarchy of cps' describes not only the strategic choices of other players, but also their hierarchy of beliefs.

Lemma 1. For each $i \in N$ there is a 1-to-1 and onto continuous function

$$f_i = (f_{\phi_i})_{\phi_i \in \Phi_i} : \boldsymbol{M}_i \to \Delta^{\Phi_i}(S_{-i}^{T_{\phi_i}} \times \boldsymbol{M}_{-i})$$

whose inverse is also continuous (i.e., f_i is a homeomorphism). Furthermore, each coordinate function f_{i,ϕ_i} is such that for all $\boldsymbol{\mu}_i = (\mu_i^1, \mu_i^2, \dots) \in \boldsymbol{M}_i, k \geq 1$,

$$\mu_i^k(\cdot|\phi_i) = \text{marg}_{S_{-i}^{T_{\phi_i}} \times M_{-i}^1 \times \cdots \times M_{-i}^{k-1}} f_{\phi_i}(\boldsymbol{\mu}_i).$$

Proof. See Appendix A.

Our specification of the conditioning events relies on interpreting each s_i^T as an objective description of how player i would behave at each history. However, we will also interpret each s_i^T as a plan of actions in the mind of player i. The implicit assumption is that a rational player assigns probability one to his own strategy whenever possible, i.e., conditional on each history consistent with the given strategy. This is the reason why we do not explicitly model player i's beliefs about his own behavior.

3.4 Psychological games with unawareness

We are now in a position to formally state the definition for our class of psychological game with unawareness:

Definition 3. A *dynamic psychological game with unawareness* based on the extensive form with subjective views $\langle N, \mathcal{T}, \phi \rangle$ is a structure $\Gamma = \langle N, \mathcal{T}, \phi, (u_i)_{i\in N} \rangle$ where $u_i : \Phi_i^Z \times \boldsymbol{M}_i \times \boldsymbol{M}_{-i} \times S_{-i}^{T_{\phi_i}} \to \mathbb{R}$ is i's (measurable and bounded) psychological payoff function.

In the following examples, a dynamic psychological game with unawareness is obtained from a material payoff game with unawareness $\langle N, \mathcal{T}, \phi, (\pi_i : \Phi_i^Z \to \mathbb{R})_{i\in N}\rangle$ according to some formula.

The two most prominent theories of belief-dependent preferences in the hitherto existing literature on dynamic psychological games are guilt aversion and reciprocity. Simple guilt aversion à la Battigalli and Dufwenberg (2007b), e.g., implies that players have a perception about the initial expectations of others concerning their material payoff and feel guilty whenever they do not live up to these expectations. More formally, consider a two-player situation. Let player i be in the T_{ϕ_i}-partial form with $s_i^T \in S_i^{T_{\phi_i}}$ and $s_j^T \in S_j^{T_{\phi_i}}$. Player i's perception of player j's initial expectations in the $T_{\phi_i^0(\phi_j)}$-partial form can be defined for $\psi(s_i^T) \in S_i^{T_{\phi_i^0(\phi_j)}}$ and $\psi(s_j^T) \in S_j^{T_{\phi_i^0(\phi_j)}}$ by

$$\mathbb{E}_{s_j^T,\mu_j^1}[\pi_j|\phi_i^0(\phi_j)] \coloneqq \sum_{s_i^T} \mu_j^1(\psi(s_i^T)|\phi_i^0(\phi_j)) \times \pi_j\left(\zeta_j\left(\psi(s_i^T), \psi(s_j^T)\right)\right) \tag{2}$$

Player i thinks he 'lets down' player j if j's actual material payoff is lower than the payoff i thinks j initially expects to get. This disappointment can be measured by the following expression:

$$\max\left\{0, \mathbb{E}_{s_j^T,\mu_j^1}[\pi_j|\phi_i^0(\phi_j)] - \pi_j\left(\zeta_i(s_i^T, s_j^T)\right)\right\} \tag{3}$$

Suppose that i likes his material payoff, but dislike disappointing j. By taking explicitly into account the material payoff game and player j's let down, we can model i's aversion to guilt by the following psychological payoff function:

$$u_i(\phi_i^z, \boldsymbol{\mu}, s_j^T) = \pi_i(\zeta_i(s_i^T, s_j^T)) - Y_i \times \max\left\{0, \mathbb{E}_{s_j^T,\mu_j^1}[\pi_j|\phi_i^0(\phi_j)] - \pi_j(\zeta_i(s_i^T, s_j^T))\right\}. \tag{4}$$

where $Y_i > 0$ is player i's sensitivity to guilt. Player i's expectations of u_i conditional on ϕ_i in the T_{ϕ_i}-partial form, given $s_i^T \in S_i^{T_{\phi_i}}$ and $\boldsymbol{\mu}_i$, can be defined for $s_j^T \in S_j^{T_{\phi_i}}$ by

$$\mathbb{E}_{s_i^T,\boldsymbol{\mu}_i}[u_i|\phi_i] \coloneqq \sum_{s_j^T} \mu_i^1(s_j^T|\phi_i) \times u_i\left(\zeta_i(s_i^T, s_j^T), \boldsymbol{\mu}, s_j^T\right). \tag{5}$$

Different from guilt aversion, reciprocity implies that players have a perception about the kindness of others. Whenever players perceive others to be kind, they reciprocate by being kind themselves. Whenever players perceive others to be unkind, they act unkindly in return (Dufwenberg and Kirchsteiger (2004)). Hence, the essence of reciprocity theory concerns player i's perception of the actual kindness of player j towards him. We define i's perception of j's actual kindness for $\psi(s_i^T) \in S_i^{T_{\phi_i(\phi_j)}}$ and $\psi(s_j^T) \in S_j^{T_{\phi_i(\phi_j)}}$ by

$$K^j_{s_i^T, s_j^T} \coloneqq \pi_i \left(\zeta_i \left(\psi(s_i^T), \psi(s_j^T) \right) \right) - \pi^e_{s_i^T}, \tag{6}$$

where

$$\pi^e_{s_i^T} \coloneqq \frac{1}{2} \times \left[\max_{s_j^T} \pi_i \left(\zeta_i \left(\psi(s_i^T), \psi(s_j^T) \right) \right) + \min_{s_j^T} \pi_i \left(\zeta_i \left(\psi(s_i^T), \psi(s_j^T) \right) \right) \right] \tag{7}$$

is the 'equitable payoff', i.e. the average i thinks that j would be able to give based on his awareness. If $K^j_{s_i^T, s_j^T} > 0$, then player i thinks that he is treated kindly by player j. Conversely, if $K^j_{s_i^T, s_j^T} < 0$, then i thinks he is treated unkindly. Player j does not know the strategy of i meaning that his kindness towards player i is evaluate in terms of j' expectations:

$$\mathbb{E}_{s_j^T, \mu_j^1}[K^j_{s_i^T, s_j^T} | \phi_i(\phi_j)] \coloneqq \sum_{s_i^T} \mu_j^1(\psi(s_i^T) | \phi_i(\phi_j)) \times K^j_{s_i^T, s_j^T}. \tag{8}$$

Given this, player i's reciprocity payoff function can be formalized as:

$$u_i(\phi_i^z, \boldsymbol{\mu}, s_j^T) = \pi_i(\zeta_i(s_i^T, s_j^T)) + Y_i \times \mathbb{E}_{s_j^T, \mu_j^1}[K^j_{s_i^T, s_j^T} | \phi_i(\phi_j)] \times \pi_j(\zeta_j(s_i^T, s_j^T)), \tag{9}$$

where $Y_i > 0$ now is player i's sensitivity to reciprocity. Player i's expectations of u_i conditional on ϕ_i in the T_{ϕ_i}-partial form, given $s_i^T \in S_i^{T_{\phi_i}}$ and $\boldsymbol{\mu}_i$, can be defined for $s_j^T \in S_j^{T_{\phi_i}}$ by

$$\mathbb{E}_{s_i^T, \boldsymbol{\mu}_i}[u_i | \phi_i] \coloneqq \sum_{s_j^T} \mu_i^1(s_j^T | \phi_i) \times u_i \left(\zeta_i(s_i^T, s_j^T), \boldsymbol{\mu}, s_j^T \right). \tag{10}$$

Note that in both these examples we have used Lemma 1. This allows us to work with utilities of the form $u_i : \Phi_i^Z \times (S_j^{T_{\phi_i}} \times M_j^1) \to \mathbb{R}$, where i's second-order beliefs M_i^2 is not a factor of the domain.

Importantly, in both examples the belief-dependent psychological payoff of players depends on the T-partial form the players perceive to be in, the T-partial form the players perceive the other player to be in and so forth. This is, different to settings with full awareness, in our setting with unawareness e.g. the perceptions of players regarding the intentions and expectations of others, and hence, the influence of these on the players' behavior crucially depends on the awareness of players, the awareness players attribute to others, the awareness players belief other attribute to them etc.

4 Equilibrium analysis

In the following we propose a sequential equilibrium concept for dynamic psychological games with unawareness. We will define and interpret consistent assessments (4.1), state the main definition of equilibrium and provide an existence theorem (4.2). Lengthy mathematical proofs are relegated to Appendix (B-C).

4.1 Consistent assessments

Battigalli and Dufwenberg (2009) adapt Kreps and Wilson (1982)'s concept of sequential equilibrium to their class of dynamic psychological games. They do so by characterizing consistent assessments that do not only consist of first-, but also of higher-order beliefs and define sequential equilibria as sequential rational consistent assessments.

In turn, we adapt Battigalli and Dufwenberg (2009) concept to our framework. As in standard dynamic psychological games, assessments in our framework also refer to behavioral strategies. The interpretation of behavioral strategies we use excludes actual randomization. Rather, we assume that players do not know the pure strategies of others, and that the randomization represents players uncertainty, their first-order beliefs (or conjectures) about other players' pure strategies (Aumann and Brandenburger, 1995). We therefore interpret a behavioral strategy as a collection of independent first-order beliefs of other players about player i's actions at each history they are aware of. More formally,

$$\boldsymbol{\sigma}_i = ((\sigma_i(\cdot|\phi_j))_{\phi_j \in \Phi_j \setminus \Phi_j^Z})_{j \neq i} \in \prod_{j \neq i} \prod_{\phi_j \in \Phi_j \setminus \Phi_j^Z} \Delta(A_i(\phi_j)),$$

where $A_i(\phi_j)$ is the set of feasible actions for player i at the history as player j perceives it.

Each behavioral strategy $\boldsymbol{\sigma}_j$ induces an independent belief $\mathrm{Pr}_{\boldsymbol{\sigma}_j}$ about the continuation of play allowed for by ϕ_i:

$$\text{for all} \quad s_j^T \in S_j^{T_{\hat{\phi}_i}}(\hat{\phi}_i), \quad \mathrm{Pr}_{\boldsymbol{\sigma}_j} := \prod_{\phi_i \in \Phi_i \setminus \Phi_i^Z : \phi_i \nless \hat{\phi}_i} \sigma_j(s_j^T(\phi_i)|\phi_i),$$

where $\phi_i \nless \hat{\phi}_i$ means, that given the perception of player i, the history ϕ_i is not a predecessor, or prefix, of $\hat{\phi}_i$.

In their characterization Kreps and Wilson (1982) propose three conditions to ensure consistency of assessments: (i) beliefs must be derived using Bayes' rule, (ii) beliefs must reflect that players choose their strategies independently, and (iii) players with identical information have identical beliefs. In addition to these conditions, a further requirement for consistency is needed in psychological games: (iv) players hold correct beliefs about each others beliefs.

Condition (i) holds by the definition of cps'. That is, cps' are defined in such a way that they are consistent with Bayes' rule. Conditions (ii)-(iii) are ensured if we assume that the profile of first-order beliefs $\mu^1 = (\mu_i^1)_{i\in N}$ is derived from the independent behavioral strategy profile $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_i)_{i\in N}$. That is, if for all $i \in N$, $s_{-i}^T \in S_{-i}^{T_{\phi_i}}$, and $\phi_i \in \Phi_i$:

$$\mu_i^1(s_{-i}^T|\hat{\phi}_i) = \prod_{j\neq i} \mathrm{Pr}_{\sigma_j}(s_j^T|\hat{\phi}_i).$$

If a profile of first-order beliefs is derived from a profile of independent behavioral strategies, then the marginal first-order belief of any two players i, j about a third player k at history ϕ_i must coincide in the mind of player i. That is, for all $s_k^T \in S_k^{T_{\phi_i}}$ and $\phi_i \in \Phi_i$:

$$\mathrm{marg}_{S_k^{T_{\phi_i}}} \mu_i^1(s_k^T|\phi_i) = \mathrm{Pr}_{\sigma_k}(s_k^T|\phi_i) = \ \mathrm{marg}_{S_k^{T_{\phi_i}}} \mu_j^1(s_k^T|\phi_i).$$

Finally, condition (iv) follows from the second condition in the following definition of a consistent assessment:

Definition 4 (cf. Battigalli and Dufwenberg (2009))**.** An assessment $(\boldsymbol{\sigma}, \boldsymbol{\mu})$ in $\langle N, \mathcal{T}, \phi\rangle$ is consistent in the mind of *any* player if

(i) μ^1 is derived from $\boldsymbol{\sigma}$,

(ii) and higher order beliefs in $\boldsymbol{\mu}$ assign probability 1 to the lower order beliefs, such that for all $i \in N$, $k > 1$, $\phi_i \in \Phi_i$

$$\mu_i^k(\cdot|\phi_i) = \mu_{i,T}^{k-1}(\cdot|\phi_i) \times \delta_{\mu_{-i}^{k-1}}$$

where δ_x is the Dirac measure which assigns probability 1 to singleton $\{x\}$.

The first condition capture the assumption that first-order beliefs should be the end-product of a transparent reasoning process of players. Players reason about each others first-order beliefs at histories they know and do so in a consistent manner. However, unlike in a standard dynamic game, players in our framework may not know the same histories. E.g., player i may reason about beliefs at histories in Φ_i, while player j reasons about beliefs at histories in $\Phi_j \subset \Phi_i$. This would imply that player i thinks that player j has no beliefs at histories in $\Phi_i \backslash \Phi_j$, while player j would think that player i has beliefs at histories in $\Phi_i \cap \Phi_j$. The second condition is analog to Geanakoplos *et al.* (1989)'s condition requiring that players hold common and correct beliefs about each others' higher-order beliefs.

4.2 Sequential equilibrium assessments

We can now state this section's main definition: a consistent assessment is a sequential equilibrium if it satisfies sequential rationality. Formally, fix a hierarchy of cps' $\boldsymbol{\mu}_i$ a (non-terminal) ϕ_i that i perceives to be at and a strategy $s_i^T \in S_i^{T_{\phi_i}}(\phi_i)$. The expectation of u_i conditional on ϕ_i, given s_i^T and $\boldsymbol{\mu}_i$ for $s_{-i}^T \in S_{-i}^{T_{\phi_i}}(\phi_i)$ is:

$$\mathbb{E}_{s_i^T, \boldsymbol{\mu}_i}[u_i|\phi_i] := \sum_{s_{-i}^T} \mu_i^1\left(s_{-i}^T|\phi_i\right) u_i\left(\zeta_i\left(s_i^T, s_{-i}^T\right), \boldsymbol{\mu}, s_{-i}^T\right), \tag{11}$$

This gives the expected payoff from the strategies of others i is aware of. Player i does,in general, not know the strategies of the others and therefore evaluates his payoff with respect to his first-order belief.

Definition 5. An assessment $(\boldsymbol{\sigma}, \boldsymbol{\mu})$ is a sequential equilibrium (SE) if it is consistent and for all players $i \in N$, non-terminal histories $\phi_i \in \Phi_i \backslash \Phi_i^Z$, and optimal strategies $s_i^{T,*} \in S_i^{T_{\phi_i}}(\phi_i)$ consistent with ϕ_i for $s_i^T \in S_i^{T_{\phi_i}}(\phi_i)$:

$$\Pr_{\sigma_i}(s_i^{T,*}|\phi_i) > 0 \quad \Rightarrow \quad s_i^{T,*} \in \arg\max_{s_i^T} \mathbb{E}_{s_i^T, \boldsymbol{\mu}_i}[u_i|\phi_i]. \tag{12}$$

By consistency, σ_i represents the first-order beliefs of the other players about player i, and furthermore there is common certainty of the true belief profile $\boldsymbol{\mu}$ at every history ϕ_i. This clarifies that SE is a an equilibrium in beliefs.

The next result shows that it suffices to check whether there are any ϕ_i where player i can gain by deviating from the choices prescribed by $s_i^{T,*}$ at ϕ_i and conforming to $s_i^{T,*}$ thereafter. Since this 'one-stage-deviation principle' is essentially the principle of optimality in dynamic programming, which is based on backward induction, it also establishes that one can use backward induction to find optimal strategies.

If we take the point of view of an 'agent' $(i, \hat{\phi}_i)$ of player i in charge of a move at ϕ_i considering whether he should deviate by taking some action, need to take into account the uncertain continuation of play following his action. The induced probability measure of agent $(i, \hat{\phi}_i)$ following his action $a_i \in A_i(\hat{\phi}_i)$ is:

$$\text{for all } \; s_i^T \in S_i^{T_{\hat{\phi}_i}}(\hat{\phi}_i), \quad \Pr_{\sigma_i}(s_i^T|\hat{\phi}_i, a_i) := \prod_{\phi_i \in \Phi_i \backslash \Phi_i^Z : \phi_i \not\leq \hat{\phi}_i} \sigma_i(s_i^T(\phi_i)|\phi_i),$$

where $\phi_i \not\leq \hat{\phi}_i$ means, that given the perception of player i, the history ϕ_i is not a predecessor of $\hat{\phi}_i$.

The expected utility of i conditional on ϕ_i and $a_i \in A_i(\phi_i)$ given $(\boldsymbol{\sigma}, \boldsymbol{\mu})$ can be expressed for $s_i^T \in S_i^{T_{\phi_i}}(\phi_i, a_i)$ and $s_{-i}^T \in S_{-i}^{T_{\phi_i}}(\phi_i)$ as

$$\mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\mu}}[u_i|\phi_i, a_i] := \sum_{s_{-i}^T} \prod_{j \neq i} \mathrm{Pr}_{\sigma_j}\left(s_j^T|\phi_i\right) \times \sum_{s_i^T} \mathrm{Pr}_{\sigma_i}\left(s_i^T|\phi_i, a_i\right) u_i\left(\zeta_i\left(s_i^T, s_{-i}^T\right), \boldsymbol{\mu}, s_{-i}^T\right).$$

This specification presumes that (i, ϕ_i) assesses the probabilities of actions by other agents of player i in the same way as each player $j \neq i$; that is using the behavioral strategy $\boldsymbol{\sigma}_i$.

The following property formalizes the intuition of the one-stage-deviation principle: For a given combination of strategies of others, a player's strategy is optimal at any stage of the game if and only if there is no stage at which the player can gain by changing his strategy there, keeping it fixed at all other stages.

Proposition 1. An optimal strategy of any player satisfies the one-stage-deviation property since it holds for all $i \in N$, $\phi_i \in \Phi_i \backslash \Phi_i^Z$, that for $a_i \in A_i(\phi_i)$ and $s_i^T \in S_i^{T_{\phi_i}}(\phi_i)$

$$\max_{a_i} \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\mu}}[u_i|\phi_i, a_i] = \max_{s_i^T} \mathbb{E}_{s_i^T,\boldsymbol{\mu}_i}[u_i|\phi_i].$$

Proof. See Appendix (B).

The following existence theorem obtains:

Theorem 1. If the psychological payoff functions are continuous, then there exists at least one sequential equilibrium assessment.

Proof. See Appendix (B).

In a dynamic psychological game with unawareness players will be aware of different T-partial forms. This implies that solving for a sequential equilibrium in the smallest T-partial form where players are only aware of one subtree, corresponds to solving a dynamic psychological game à la Battigalli and Dufwenberg (2009). However, for players aware of two or more subtrees the game is such that we must solve for a sequential equilibrium by first considering the smallest T-partial form, and then extending the equilibrium step-by-step by taking the equilibria of other players in embeddable T-partial forms as given.

This observation suggests a procedure for finding an equilibrium in our structure. Consider player i. Start at the last stage: any history ϕ_i for which the feasible actions terminate the game. Then look for an equilibrium in the T_{ϕ_i}-partial form that a player is aware of, by: (i) calculating the best responses of other players at the copy in the last stage in the smallest embeddable partial form, and (ii) extend the equilibrium step-by-step to copies in the last stage in larger embeddable partial forms by finding a fixed point given the optimal actions of others. Now go backward and look at a history ϕ_i' in the second-to-last stage. The best responses has already been calculated for the history (ϕ_i', a), because such a history correspond to a history ϕ_i in the last stage of the game. We assume that each active player at the second-to-last stage makes feasible choices that maximizes his expected payoff given the best responses in the last stage, because he expects that the other players will also best respond in the last stage. Again, extend the equilibrium in the second-to-last stage step-by-step from the smallest embeddable partial form to the $T_{\phi_i'}$-partial form. We continue to go backwards in this ways until we reach the initial stage.

4.3 Two examples: reciprocity and guilt aversion

In the following we present two examples to demonstrate the impact and importance of unawareness in strategic interactions of players with belief-dependent preferences. First, we analyze a sequential prisoners dilemma featuring unawareness and reciprocity. Second, we investigate a trust game with guilt aversion. A full description of the strategic interaction with all possible awareness levels and equilibria is beyond the scope of this paper. Therefore, we limit the analysis to specific awareness scenarios and the respective characterization of only one equilibrium.

A sequential prisoners dilemma with reciprocity: To start with consider the following awareness scenario already depicted in Figure 8 augmented by *Ann* and *Bob*'s material payoffs:[11]

[Figure 9]

In the strategic setting depicted in Figure 9 *Ann* is initially aware of everything, whereas *Bob* is initially only aware of histories in T'. *Ann* knows that *Bob* is initially only aware of histories in T' and that he only becomes aware of more when she chooses Defect (D), in which case he will become aware of everything. Furthermore, *Ann* knows that *Bob* is certain that *Ann* is aware of what he is aware of.

For simplicity assume that only *Bob* is motivated by belief-dependent reciprocity. That is, *Bob*'s utility is described by equations 6 to 10. *Ann* is only interested in her own monetary payoff.

As suggested by the afore-described procedure for finding equilibria in our structure, we start by analyzing *Bob*'s optimal behavior. *Bob*'s optimal behavior following *Ann*'s choice

[11] For the sake of clarity we only depict the function ϕ for *Ann* and *Bob* in the non-terminal histories relevant for solving the game.

Cooperate (C) and his choice following *Ann*'s choice Defect (D).[12] Note that when *Ann* chooses to cooperate *Bob* perceives to be in the T'-partial form. On the other hand, when *Ann* defects *Bob* perceives to be in the T-partial form. In this awareness scenario *Bob*'s strategy in the T-partial form defines an action in histories $h^1_{T'}$ and h^2_T, whereas *Bob*'s strategy in the T'-partial form only defines an action in history $h^1_{T'}$. We therefore have to analyze *Bob*'s optimal behavior in histories $h^1_{T'}$ and h^2_T.[13]

Result 1. *Bob* chooses defect (d) in histories $h^1_{T'}$ and h^2_T irrespective of his sensitivity to reciprocity Y_B in all sequential equilibria.

Different to Dufwenberg and Kirchsteiger (2004, p. 282) our awareness scenario implies *Bob*'s behavior following *Ann*'s choice Cooperate is independent of his sensitivity to reciprocity Y_B. If *Bob* is unaware of *Ann*'s action Defect and all of his own subsequent actions, *Bob* evaluates the kindness of *Ann* on the basis of the T'-partial form (which in this example is congruent with histories in T'). In the T'-partial form *Ann*'s action set at history $h^0_{T'}$, Ann's initial history as perceived by *Bob*, is a singleton. I.e., what *Ann* intends to give *Bob* in the T'-partial form is equal to the average that *Ann* is able to give independent of her belief about *Bob*'s choice in $h^1_{T'}$, and, hence, $K^A = 0$. Consequently, *Bob* only takes into account his own monetary payoff when optimizing his own choice in history $h^1_{T'}$ (i.e. given *Ann*'s choice Cooperate and, hence, $K^A = 0$, *Bob*'s utility as defined in equation 9 reduces to Bob's material payoff in history $h^1_{T'}$). On the other hand, *Bob* re-evaluates *Ann*'s kindness towards him on the basis of the T-partial form following *Ann*'s action Defect. As can easily be seen, independent of his choice in history h^2_T, *Bob*'s material payoff in the T-partial form is lowest following *Ann*'s choice Defect. Hence, what she intends to give *Bob* is for sure lower than the average that *Ann* is able to give and, hence, $K^A < 0$. Given this, *Bob* has an incentive to reciprocate by choosing defect in history h^2_T. Furthermore, *Bob* also prefers to defect in history h^2_T out of own monetary reasons. I.e., *Bob* chooses defect in history h^2_T out of reciprocity as well as own monetary considerations. (A result which is analog to Dufwenberg and Kirchsteiger (2004, p. 282)'s Observation 1.)

Given *Bob*'s behavior in histories $h^1_{T'}$ and h^2_T, *Ann*'s behavior can be described as follows

Result 2. Independent of Bob's sensitivity to reciprocity *Ann* chooses Defect in history h^0_T in all sequential equilibria.

[12] We abstract from the behavior of players in histories in which they are passive, i.e. in which their action sets are singletons.

[13] Note we do not have to analyze *Bob*'s optimal behavior in history h^1_T as neither Bob's strategy in the T-partial nor his strategy in the T'-partial form defines an action in $h^1_{T'}$. Intuitively, no matter what happens in the game *Bob* never has to make a choice in $h^1_{T'}$, nor does he have to reason about what he would do in that history.

Intuitively, *Ann* anticipates that given his awareness *Bob* will not perceive her as kind and thus defect independent of what she does. Consequently, since *Ann* is only interested in her own material payoff, she chooses to defect herself in order to get a 0 material payoff instead of -2.

In synthesis: although *Bob* is reciprocal, his equilibrium behavior in this awareness scenario stands in contrast to the result with full awareness described in Dufwenberg and Kirchsteiger (2004). With full awareness *Bob*'s equilibrium behavior following *Ann*'s action Cooperate depends on *Bob*'s sensitivity to reciprocity. For sufficiently high levels of sensitivity *Bob* reciprocates by choosing cooperate (c). With unawareness as described above *Bob*'s behavior is independent of his sensitivity to reciprocity. *Bob* simply defects as he perceives *Ann*'s action as unkind no matter what she does. As a consequence, also *Ann*'s equilibrium behavior is qualified. She defects as well.

Of course, if *Ann* could choose Cooperate and by this make *Bob* simultaneously aware of what he is unaware of, the situation would be different. This scenario could, e.g., be depicted by the following variant of the strategic setting presented above.

[Figure 10]

As before, *Bob* is initially unaware of *Ann*'s action Defect and all of his own subsequent actions. However, *Ann* knows now that *Bob* will become aware if she chooses to Cooperate.

In this scenario, *Bob*'s optimal behavior in histories $h^1_{T'}$ and h^2_T remain the same as before, i.e., *Bob* chooses defect out of monetary and reciprocal reasons. However, different from before, *Bob*'s strategy in the T-partial form now also defines an optimal choice in history h^1_T following *Ann*'s action Cooperate in history h^0_T. Interestingly, if *Ann* chooses to cooperate in history h^0_T, *Bob*'s behavior in history h^1_T is analog to Dufwenberg and Kirchsteiger (2004, p. 282)'s Observation 2. By cooperating *Ann* can influence the basis upon which *Bob* evaluates her kindness. We saw before, that *Bob* evaluates the kindness of *Ann* following her choice Cooperate on the basis of the T'-partial form as long as he is unaware of *Ann*'s action Defect and all of his own subsequent actions. However, when *Ann* cooperates in this variant scenario, *Bob*'s awareness increases to the T-partial form and, hence, the partial form he plays as well as the basis upon which he evaluates *Ann*'s kindness changes. Aware of *Ann*'s action Defect and *Bob*'s actions cooperate and defect following it, *Bob* realizes that *Ann*'s action Cooperate was actually kind. This is something he would not have realized had he remained unaware.

This also has consequences for *Ann*'s optimal behavior. If *Bob*'s sensitivity to reciprocity is low *Ann* defects. However, if *Bob* is sensitive enough to *Ann*'s (un)kindness, *Ann* cooperates so as to induce *Bob* to reciprocate by choosing cooperate, something she would not do

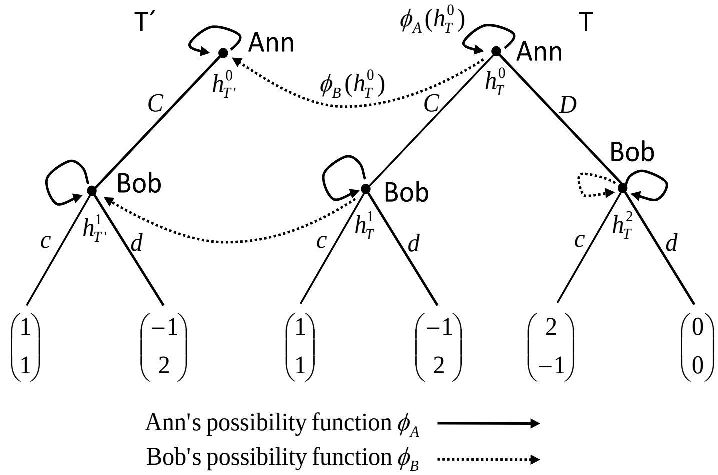

Figure 9: Sequential Prisoners Dilemma with unawareness

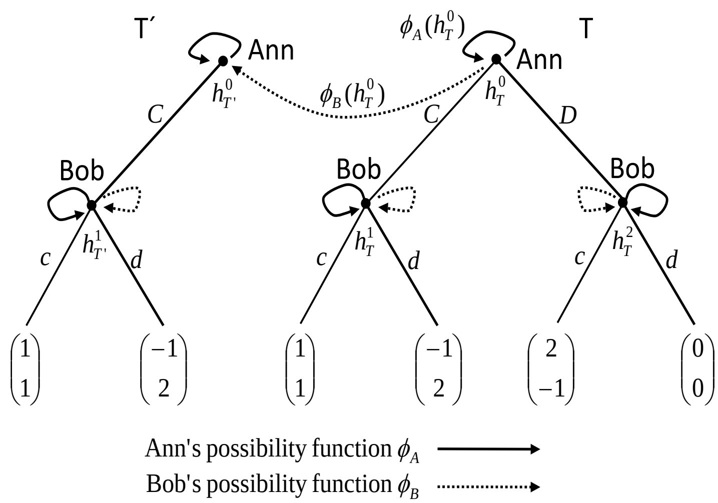

Figure 10: Sequential Prisoners Dilemma with subjective views

were she sure that *Bob* would not become aware of more.

It is really at the intersection of these two scenarios that the implications of unawareness for the behavior of people motivated by belief-dependent preferences become most visible. It is easy to see that, if *Bob* were only interested in his own material payoff, his behavior in the two awareness scenarios would be the same. Most importantly, being only interested in his own material payoff means *Bob* would choose defect following *Ann*'s choice Cooperate independent of whether he is only aware of the T'-partial form (as in the first scenario) or the T-partial form (as in the second scenario). It is only his belief-dependent utility which explains the above described difference in behavior between the first and second awareness scenario.

A trust game: Consider the following trust game also analyzed by Battigalli and Dufwenberg (2009). However, different to them assume that *Bob* is aware of everything, but *Ann* is not aware of *Bob*'s action *Share*:[14]

[Figure 11]

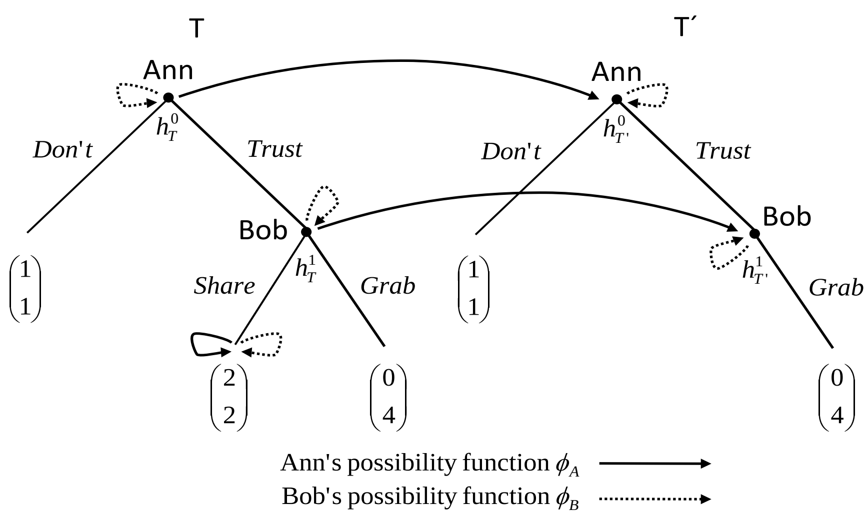

Figure 11: Trust game with unawareness

Assume *Bob* is motivated by simple guilt aversion as described by eq. 2 to 5 and *Ann* is only interested in her own material payoff. Given this and assuming full awareness of both *Ann* and *Bob*, Battigalli and Dufwenberg (2009, p. 19) demonstrate that there exist

[14]Again, for the sake of clarity we only depict the function ϕ for *Ann* and *Bob* in the non-terminal histories relevant for solving the game.

three sequential psychological equilibria in this game: (i) a 'trust' equilibrium in which *Ann* chooses *Trust* and *Bob* chooses *Share*, (ii) a 'no-trust' equilibrium in which *Ann* chooses *Don't* and *Bob* chooses *Grab* and (iii) and an 'insufficient' trust equilibrium in which *Ann* chooses *Don't* and *Bob* chooses *Share* with probability $\frac{2}{5}$.

In our strategic situation with unawareness as depicted in Figure 11 *Bob* knows that *Ann* is not 'let down', if he chooses *Grab* following her choice *Trust*. I.e., he simply knows that *Ann* does not expect him to choose *Share* following *Trust* because she is unaware of the possibility that he could share the pie. Hence, *Bob* does not feel any guilt towards *Ann* from choosing *Grab* following *Trust*. Of course, *Ann* correctly anticipates that *Bob* chooses *Grab* following *Trust* and chooses *Don't* in the history $h^0_{T'}$ she initially perceives to be in.

Result 3. If *Bob* is guilt averse and *Ann* unaware of *Bob*'s action *Share* the unique sequential psychological equilibrium is: *Ann* chooses *Don't* and *Bob* chooses *Grab* independent of *Bob*'s sensitivity to guilt Y_B.

Like the previous example featuring reciprocity, also this example with guilt aversion demonstrates the impact of unawareness on the behavior of players with belief-dependent preferences. Furthermore, also this scenario highlights that 'managing others awareness levels' concerning feasible paths of play is an important and integral part of strategic interactions when players are motivated by belief-dependent preferences. The fact that *Ann* is unaware of *Bob*'s action *Share* implies that he would not feel any guilt towards *Ann* for choosing *Grab*, since she would not be let down. However, if he could, *Bob* would like to make *Ann* aware of his option *Share* before she chooses between *Don't* and *Trust*, an option not considered in our example in Figure 11. He would like to make her aware in order to signal to her that they could *Share* the pie. Of course, if he were to do this the equilibrium analysis would mirror Battigalli and Dufwenberg (2009, p. 19)'s analysis.

5 Adding uncertainty to the framework

In the previous sections, we focused on games with observable actions and no chance moves. However, our concepts and results carry over to the more general setting where past actions need not be perfectly observed (5.1) and chance may play a role (5.2). Doing so implies extending our class of extensive forms further towards Heifetz *et al.* (2011)'s generalized extensive forms.

5.1 Imperfectly observable actions

Let $\mathbf{\Phi}_i$ be the partition of the finite set Φ_i into information sets of player i, and $A_i(\boldsymbol{\phi}_i)$ be the set of feasible actions for player i at an information set $\boldsymbol{\phi}_i \in \mathbf{\Phi}_i$. Since actions are no longer observable, we will need to explicitly consider the properties that (i) a player never excludes the true history from the set of histories he regards as feasible (*Generalized reflexivity*), and (ii) that a player uses the consistency or inconsistency of histories with his information to make inference about the true history (*Introspection*).[15] Formally,

Property 9 (*P*1 *Generalized reflexivity*)**.** At a history h_T in subtree T, if an information set $\boldsymbol{\phi}_i(h_T)$ is a subset of an embeddable subtree T' and T' contains a copy $h_{T'}$ of h_T, then it must be that $\boldsymbol{\phi}_i(h_T)$ also contains $h_{T'}$.

and

Property 10 (*P*2 *Introspection*)**.** If a history h'_T is in the information set $\boldsymbol{\phi}_i(h_T)$, then it must be that the information set $\boldsymbol{\phi}_i(h'_T)$ is equal to $\boldsymbol{\phi}_i(h_T)$.

Finally, assume perfect recall. I.e., players remember whatever they knew previously, including their past actions.

Property 11 (*P*6 *Perfect recall*)**.** For any two histories $h_T = (a^1, \ldots, a^l)$ and $h'_T = (a^1, \ldots, a^l)$ the sequences of actions are the same whenever h_T and h'_T are in the same information set $\boldsymbol{\phi}_i$ of player i.

Together with the straight forward modifications of property 3-8, these additional properties give sufficient conditions for a dynamic game with imperfect information and unawareness (see Heifetz *et al.* (2011)).

In such a game the set of strategy profiles consistent with any information set $\boldsymbol{\phi}_i$ of player i must have the form $S^T(\boldsymbol{\phi}_i) = S_i^T(\boldsymbol{\phi}_i) \times S_{-i}^T(\boldsymbol{\phi}_i)$, where $S_i^T(\boldsymbol{\phi}_i) := \bigcup_{\phi_i \in \boldsymbol{\phi}_i} S_i^T(\phi_i)$ and $S_{-i}^T(\boldsymbol{\phi}_i) := \bigcup_{\phi_i \in \boldsymbol{\phi}_i} S_{-i}^T(\phi_i)$.

The collection of conditioning events for player i's first-order beliefs is now given by $\{F_i : F_i = S_i^{T_{\boldsymbol{\phi}_i}}(\boldsymbol{\phi}_i), \boldsymbol{\phi}_i \in \mathbf{\Phi_i}\}$. Furthermore, let X_{-i}^{k-1} be space of $(k-1)$-order uncertainty of player i; we obtain the set of k-order cps' for the extended game with imperfectly observable actions $\Delta^{\mathbf{\Phi}_i}(X_{-i}^{k-1})$, and the k-order uncertainty space $X_{-i}^k = X_{-i}^{k-1} \times \prod_{j \neq i} \Delta^{\mathbf{\Phi}_j}(X_{-j}^{k-1})$. The resulting set of infinite hierarchies of cps' $\boldsymbol{M}_i$ is homeomorphic to $\Delta^{\mathbf{\Phi}_i}(S^{T_{\boldsymbol{\phi}_i}} \times \boldsymbol{M}_{-i})$ via a belief mapping $f_i = (f_{\boldsymbol{\phi}_i})_{\boldsymbol{\phi}_i \in \mathbf{\Phi}_i}$.

[15] A graphical illustration of each property can be found on page 32

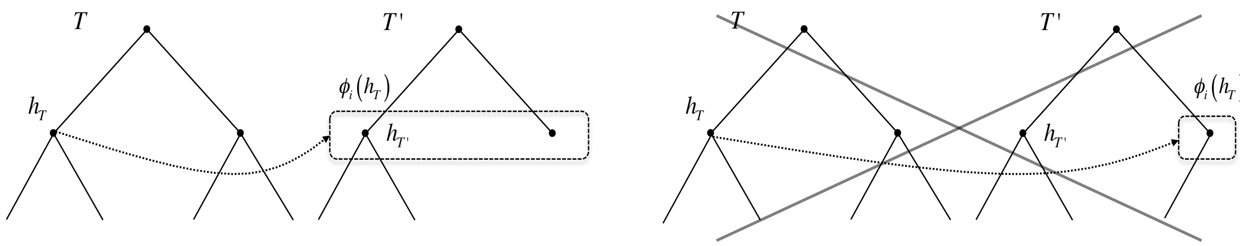

Figure 12: A correct and incorrect application of property 9

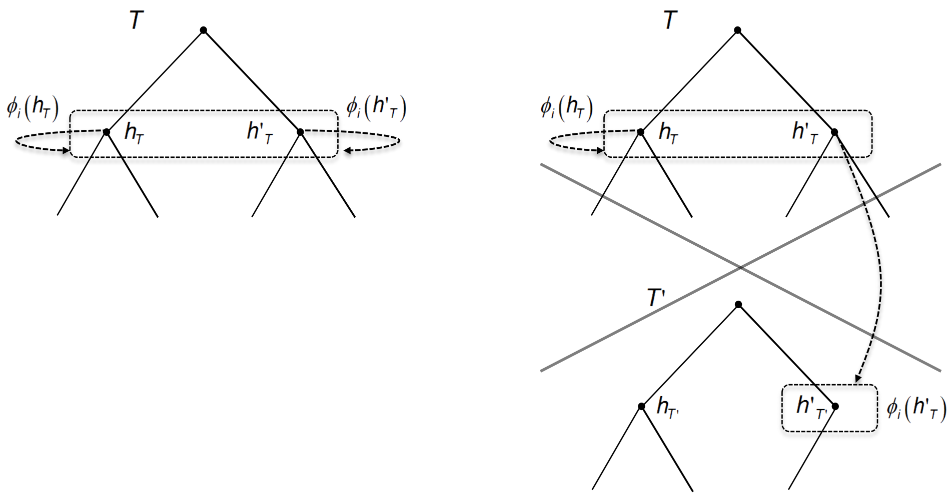

Figure 13: A correct and incorrect application of property 10

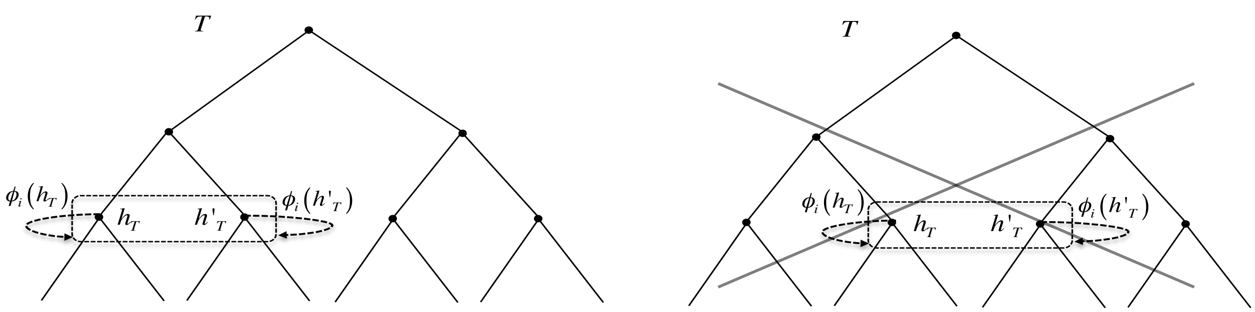

Figure 14: A correct and incorrect application of property 11

5.2 Chance moves and incomplete information

First we consider observable chance moves. Let $N^c = \{c, 1, \dots, n\}$ be the set of players where index c denotes the chance player, and $\boldsymbol{\sigma}_c := \sigma_c(\cdot|h_T) \in \prod_{h_T \in \bigcup_{T \in \boldsymbol{T}}[T \setminus Z_T]} \Delta^0(A_c(h_T))$ the strictly positive objective probabilities of chance moves. It is routine to define closed and compact subsets of hierarchies $\boldsymbol{M}_i$ for real players $i \in N$, reflecting the assumption that each players' beliefs about the chance player c are determined by $\boldsymbol{\sigma}_c$ and there is common certainty about this.

An assessment $(\boldsymbol{\sigma}, \boldsymbol{\mu}) = (\boldsymbol{\sigma}_i, \boldsymbol{\mu}_i)_{i \in N^c}$ is consistent if there is a sequence of strictly positive behavioral strategy profiles $\boldsymbol{\sigma}^k \to \boldsymbol{\sigma}$ such that for all real players $i \in N$, $s^T_{-i} \in S^{T\boldsymbol{\phi}_i}_{-i}$, $\boldsymbol{\phi}_i \in \boldsymbol{\Phi}_i$,

$$\mu^1_i(s^T_{-i}|\boldsymbol{\phi}_i) = \lim_{k \to \infty} \frac{\Pr_{\boldsymbol{\sigma}_c}(s^T_c) \prod_{j \neq i,c} \Pr_{\boldsymbol{\sigma}^k_j}(s^T_j|\boldsymbol{\phi}_i)}{\sum_{\tilde{s}^T_{-i} \in S^{T\boldsymbol{\phi}_i}_{-i}} \Pr_{\boldsymbol{\sigma}_c}(s^T_c|\boldsymbol{\phi}_i) \prod_{j \neq i,c} \Pr_{\boldsymbol{\sigma}^k_j}(\tilde{s}^T_j|\boldsymbol{\phi}_i)}.$$

Kreps and Wilson (1982, Section 5) have a similar condition that refers to cps' of histories (or nodes). In psychological games we furthermore assume that for all $k > 1$, $\boldsymbol{\mu}^k_i$ assigns probability 1 to $\boldsymbol{\mu}^{k-1}_{-i}$. $(\boldsymbol{\sigma}, \boldsymbol{\mu})$ is a sequential equilibrium if it is consistent and for all real players $i \in N$, $\boldsymbol{\phi}_i \in \boldsymbol{\Phi}_i \setminus \boldsymbol{\Phi}^Z_i$, $s^{T,*}_i \in S^{T\boldsymbol{\phi}_i}_i(\boldsymbol{\phi}_i)$,

$$\Pr_{\boldsymbol{\sigma}_i}(s^{T,*}_i|\boldsymbol{\phi}_i) > 0 \quad \Rightarrow \quad s^{T,*}_i \in \arg \max_{s^T_i \in S^{T\boldsymbol{\phi}_i}_i(\boldsymbol{\phi}_i)} \mathbb{E}_{s^T_i, \boldsymbol{\mu}_i}[u_i|\boldsymbol{\phi}_i].$$

where $\mathbb{E}_{s^T_i, \boldsymbol{\mu}_i}[u_i|\boldsymbol{\phi}_i]$ is the obvious modification of equation 12. It can easily be proven that the existence theorem also holds when we add chance as a player (assuming that the belief-dependent payoff functions are continuous).

This extended framework also allows for the analysis of situations with incomplete information, modeling them as games with asymmetric information about an initial chance move: at an empty history h^0_T the only active player is c (chance), $A_c(h^0_T) = \Theta^T$, where $\Theta^T \subseteq \Theta \subseteq \Theta_c \times \Theta_1 \times \cdots \times \Theta_n$ is the non-empty and finite set of conceivable parameter values in the subtree T. For every real player $i \in N$, each coordinate $\theta_i \in \Theta^T_i$ represents player i's private information about the unknown payoff-relevant aspect of the game; we call it player i's information type. Each coordinate $\theta_c \in \Theta^T_c$ (the information type of the chance player) represents any residual uncertainty about payoffs that remains after pooling the real player's private information. We typically refer to profile $\theta = \theta_c \times \theta_1 \times \cdots \times \theta_n$ as the state of nature. Dynamic psychological games with incomplete information are easily represented by means of the histories that players perceive and information sets: the set of histories in a subtree is $\Theta^T \times T$. Information sets for real players $i \in N$ have the following form:

$$\boldsymbol{\Phi}_i(\theta_i, h_T) = \{(\theta_i, \theta'_{-i}, h_T) : \theta'_{-i} \in \Theta^T_{-i}\}.$$

5.3 A simple principal-agent example with reciprocity

Imagine that a selfish principal p and a reciprocal agent a are entering a new venture (e.g. as manager-worker, investor-entrepreneur, or manufacture-distributor). The venture can have a big potential, B, or a small potential, S, decided by chance. We will assume that the principal quantifies the success of the venture in terms of profit, which depends on the potential of the venture and the effort exerted by the agent. Conversely, as the agent is reciprocal, the effort of the agent will depend on his wage (profit forgone by the principal) and how kind he perceives the principal to be (the relative profit forgone by the principal).

If the venture has a small potential, then by exerting high effort e_H the agent generates a profit of 20, while if he exerts low effort e_L the profit will be 15. Instead, if the venture has a big potential the agent generates either a profit of 30 or 20 by exerting high or low effort, respectively. Exerting effort is not costless to the agent. Assume that the agent incurs a cost of 4 by exerting high effort, the cost of exerting low effort is normalized to zero. The cost incurred by the principal is the profit forgone as wage to the agent.

The timing of the game is as follows: First chance chooses the state of nature $\theta \in \{B, S\}$, thereafter the principal decides on how much he wants to pay the agent as a share s of the profit, with $s \in \{0, 0.01, 0.02, ..., 1\}$, and finally the reciprocal agent (see eq. 6 to 10) ends the game by choosing how much effort he wants to exert.

As a benchmark, consider first the case of full awareness and perfect information (observable chance moves), where both, the principal and the agent, learn the state of nature $\theta \in \{S, B\}$ chosen by chance. Subsequently, we extent this framework to study unawareness in the presence of incomplete information.

[Figure 15]

Given this, the following result obtains:

Result 4. Independent of the state of nature and the principal's belief concerning the agent's effort choice, the principal's kindness towards the agent is negative, $K^p < 0$, if the agent receives a profit share $s < 0.5$ and positive, $K^p > 0$, if the received profit share is $s > 0.5$.

Proof. See Appendix (C).

I.e., given that both learn the true state of nature, paying the agent a profit share of less than 0.5 is considered unkind and might be reciprocated by low effort independent of whether the potential of the venture is big or small. On the other hand, a profit share of more than 0.5 is considered kind and, hence, might lead to high effort by the agent. Note that are profit share of $s = 0.5$ implies a payoff of the agent of 6 $(0.5 \cdot 20 - 4 = 6)$ in material terms, if the state of nature is S and the agent chooses high effort and a payoff of 7.5 ($0.5 \cdot 15 = 7.5$) in material terms, if he chooses low effort. Furthermore, it implies a payoff of 11 ($0.5 \cdot 30 - 4 = 11$) in material terms for the agent, if the state of nature is B and the agent chooses high effort and a payoff of 10 ($0.5 \cdot 20 = 10$) in material terms, if he chooses low effort. Interestingly, this shows that at a profit share $s = 0.5$ the agent has a material self-interest to choose low effort when the potential is small, and a material self-interest to choose high effort when the potential is big.

At profit shares above or below, this incentive might, however, be mitigated by the agent's inclination to behave reciprocal. Remember, reciprocity implies that the agent has an incentive to behave kind towards the principal if the principal is kind to him and vice versa. The following can be concluded regarding the agent's effort choice when the potential is small:

Result 5. If the state of nature is S, the agent chooses high effort e_H in equilibrium out of reciprocity reasons, if he receives a share $0.5 < s \leq 0.8$ of the profit to be made given that his sensitivity to reciprocity is

$$Y \geq \frac{4 - 5s}{5(1 - s)} \frac{1}{20s - 10}. \tag{13}$$

If $0.5 < s \leq 0.8$ and $Y < \frac{4-5s}{5(1-s)} \frac{1}{20s-10}$ the agent chooses low effort. At profit shares above and below, the agent respectively chooses high and low effort out of reciprocity and material self-interest.

Proof. See Appendix (C).

This shows that there exists a range of profit shares in which the agent perceives the principal as kind and reciprocates by choosing high effort, although it is not in his material self-interest. Importantly, if the agent were only interested in his own monetary payoff, then given that the potential is small he would only choose high effort at profit shares $s > 0.8$, i.e. at payoffs above 12 ($0.8 \cdot 20 - 4 = 12$ where 20 is the profit when the agent chooses high effort given S and 4 is the cost of effort.). Hence, Result 5 shows that if the potential is small and the agent is sufficiently motivated by reciprocity, the principal can induce high effort at a lower cost compared to a situation in which the agent is only driven by his own material payoff.

An opposing result obtains for an enterprize with big potential:

Result 6. If the state of nature is B, the agent chooses high effort e_H in equilibrium out of reciprocity reasons, if he receives a share $0.4 < s \leq 0.5$ of the profit to be made given that his sensitivity to reciprocity is

$$Y \leq \frac{4 - 10s}{10(1 - s)} \frac{1}{30s - 15}. \tag{14}$$

If $0.4 < s \leq 0.5$ and $Y > \frac{4-10s}{10(1-s)} \frac{1}{30s-15}$ the agent chooses low effort. At profit shares above and below, the agent respectively chooses high and low effort out of reciprocity and material self-interest.

Proof. See Appendix (C).

At profit shares $0.4 < s \leq 0.5$ of the profit to be made, given that the potential is big, the agent chooses high effort only if his sensitivity to reciprocity $Y \leq \frac{4-10s}{10(1-s)} \frac{1}{30s-15}$. Contrary to the case where the potential is small, there now exists a range of profit shares of the profit to be made in which it is in the agent's material self interest to choose high effort, but as the agent perceives these profit shares as an unkind offer by the principal, he reciprocates by choosing low effort. If the potential is big and the agent is only interested in his own monetary payoff, he would choose high effort at profit shares $s > 0.4$, i.e., at a payoff above 8 ($0.4 \cdot 30 - 4 = 8$ where 30 is the profit when the agent chooses high effort given B and 4 is the cost of effort.). However, if the agent is motivated sufficiently by reciprocity, the principal can only induce high effort by the agent at higher costs.

Given the above, consider now a situation with private information concerning the state of nature chosen by chance. When information is private players do not observe the whole state of nature, but only their own coordinate–their information type. Assume, e.g., that chance can choose between two states of nature $\theta \in \{(B,S),(S,S)\}$, where the first coordinate refers to the principal's information type and second refers to the agent's information type. Furthermore assume that the awareness of the principal and agent is as given in Figure 16 and that the agent's sensitivity to reciprocity $Y = 0.25$.

[Figure 16]

I.e., in this awareness scenario we assume that the principal and the agent are respectively confined to the T-partial and T'-partial form throughout the whole game. This implies that the principal is aware that chance can draw either of the two states of nature (big and small potential), while the agent is only aware of one state of nature since he only observes coordinate S, and therefore thinks that chance draws a small potential with certainty. Formally, the path of play observed by the agent is the same no matter which state of nature is drawn. Furthermore, the agent evaluates a payment offered by the principal always relative to the profit to be made given the small potential, i.e., 20 and 15. Observe also that the only source of uncertainty is chance. After chance selects information types, the actions are observed by all.

This scenario can be analyzed using Results 5 and 6. Interestingly, in this scenario featuring unawareness, the principal has a material incentive to leave the agent unaware.

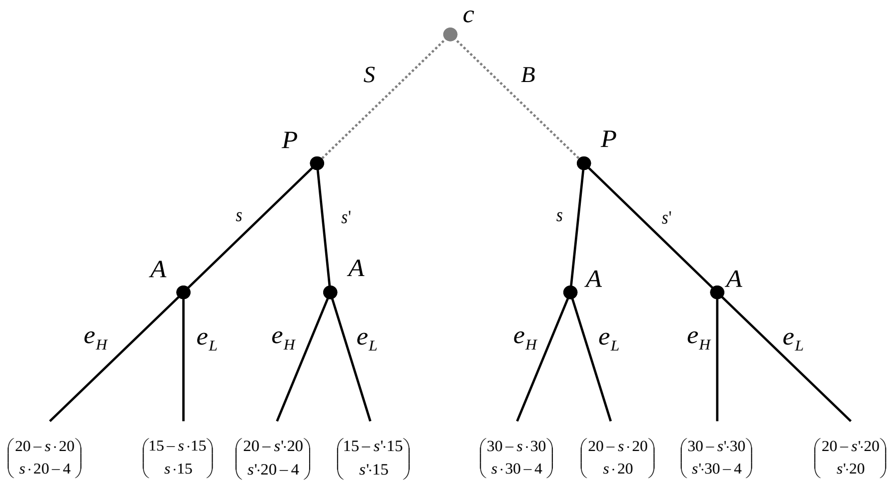

Figure 15: Simplified Principal-Agent Example with two possible profit shares s and s'

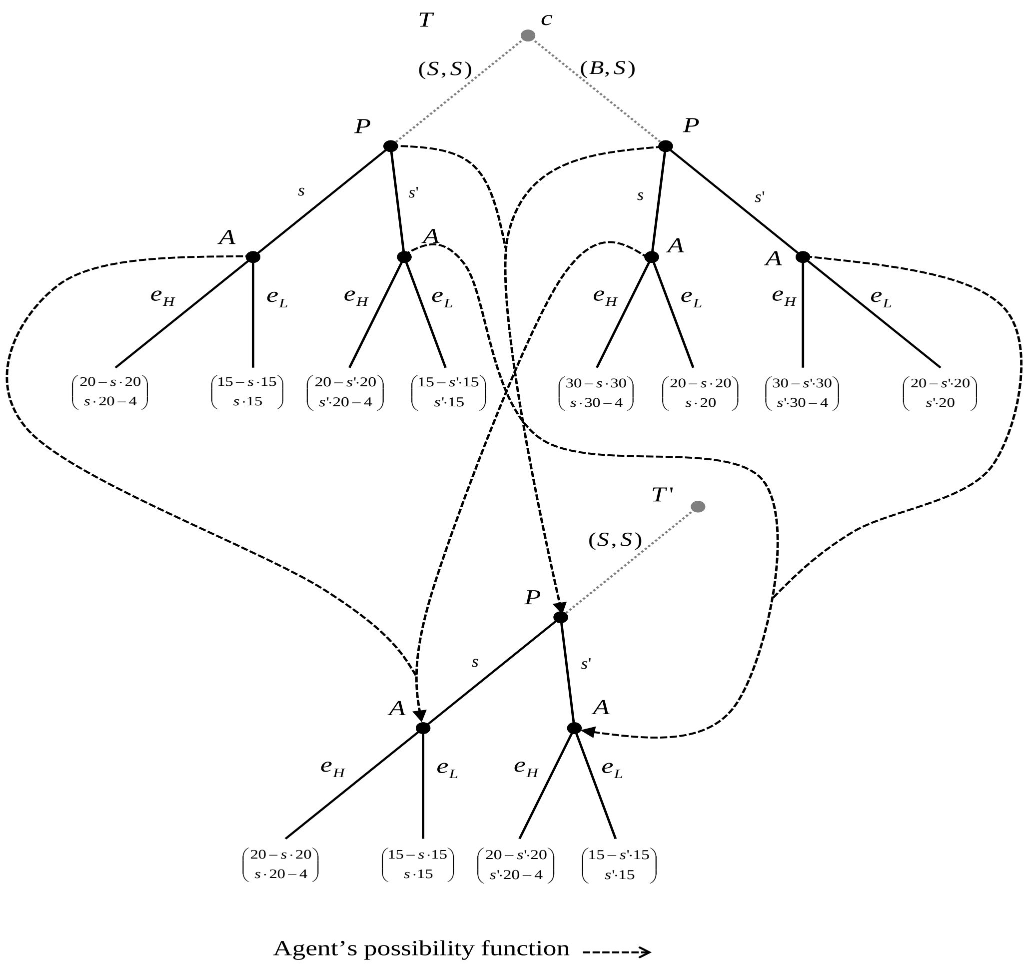

Figure 16: Simplified Principal-Agent Example with unawareness and two possible profit shares s and s'

Result 7. If the agent's sensitivity is $Y = 0.25$, the potential of the project is big B and the agent is unaware of this, the principal offers the agent a share $s = 0.6$ of the profit to be earned given the state of nature S (i.e. which translates into a share $s = 0.4$ of the profit to be earned given the state of nature is B) and in this way incentivizes the agent to choose e_h.

Intuitively, a profit share of $s = 0.6$ as perceived by the agent is kind and given his sensitivity, he reciprocates by choosing high effort, although it is not in his material self-interest to do so (see result 5). On the other hand, the offered compensation would be perceived as unkind were the agent aware of the fact that the business venture has a big potential (see result 6). Were the agent aware of the fact that the state of nature is B, he would choose low effort given $s = 0.4$ and his sensitivity to reciprocity. Hence, the principal benefits from the agent's unawareness. The reason is that being unaware of B and, hence, the actual profit to be made, the agent evaluates the principal's offer on the basis of the profit to be made given S. I.e., he reacts to his perception concerning the kindness of the principal given that the venture has a small potential, rather than the true state of nature. In this way he perceives the offer of the principal as kind, although it is actually not.

In synthesis, this principal-agent situation with asymmetric awareness concerning an initial move of chance represents yet another example highlighting the importance of unawareness in the interaction of players with belief-dependent preferences. Again, unawareness influences the frame of mind of players which constitute the background against which they e.g. judge the their own as well as the other's (un)kindness which, in turn, influences their strategic interaction.

6 Conclusion

We have provided a general framework for dynamic psychological games with unawareness, presented a solution concept and showed that any game in our class of games with unawareness has at least one sequential equilibrium provided that players have continuous belief-dependent utility functions. Furthermore, we have analyzed different examples demonstrating the influence of unawareness in the strategic interaction of players with belief-dependent preferences like reciprocity and guilt.

Our analysis has shown that unawareness has a profound impact on the strategic interaction of players with belief-dependent psychological preferences. Unawareness in the interaction of players with belief-dependent preferences leads to very intuitive behavioral predictions distinct from predictions using standard (non-psychological) preferences or no unawareness. Hence, it is an issue which should not simply be neglected and assumed away, but rather taken into account as an integral and important part of strategic interactions.

A Appendix

A.1 Proof of Lemma 1

The following Lemma 2 and proof of Lemma 1 relies heavily on Battigalli and Siniscalchi (1999)'s extension of Proposition 1 and 2 in Brandenburger and Dekel (1993) to conditional probability systems.

Let the infinite (not necessarily coherent) hierarchies of cps' be $\bar{\boldsymbol{M}}_i = \prod_{k\geq 0} \Delta^{\Phi_i}(X^k_{-i})$, and $\hat{\boldsymbol{M}}_i \subset \bar{\boldsymbol{M}}_i$ be the subset satisfying eq. (1).

Lemma 2. Consider the following set:

$$\hat{M}_i = \{(\mu^1_i(\cdot|\phi_i), \mu^2_i(\cdot|\phi_i), \dots) : k \geq 1,\ \mu^k_i(\cdot|\phi_i) \in \Delta(Z^0_{-i} \times \cdots \times Z^{k-1}_{-i}),$$
$$\operatorname{marg}_{Z^0_{-i}\times\cdots\times Z^{k-1}_{-i}} \mu^{k+1}_i(\cdot|\phi_i) = \mu^k_i(\cdot|\phi_i)\}.$$

There is a 1-to-1 and onto continuous function $h_{\phi_i} : \hat{M}_i \to \Delta(S^{T_{\phi_i}}_{-i} \times \bar{\boldsymbol{M}}_{-i})$ whose inverse is also continuous.

Proof. In Lemma 2 set $Z^0_{-i} = X^0_{-i}$, $Z^k_{-i} = \Delta^{\Phi_i}(X^k_{-i})$ for $k \geq 1$. So $Z^0_{-i} \times \cdots \times Z^{k-1}_{-i} = X^k_{-i}$ and $\prod_{k\geq 0} Z^k_{-i} = S^{T_{\phi_i}}_{-i} \times \bar{M}_{-i}$. Since $S^{T_{\phi_i}}_{-i}$ is a finite set then the weak topology on $\Delta^{\Phi_i}(S^{T_{\phi_i}}_{-i})$ is compact Polish (Aliprantis and Kim (1999, Theorem 14.15)), hence the Z^k_{-i}'s will be compact Polish spaces. A countable product of compact Polish spaces, endowed with the product topology, is itself compact Polish. The proof then follows from Lemma 1 in Brandenburger and Dekel (1993). ∎

Proof. (Proof of Lemma 1) For each $\phi_i \in \Phi_i$, let $\gamma_{\phi_i} : \hat{\boldsymbol{M}}_i \to \hat{M}_i$ be the following function:

$$\gamma_{\phi_i}(\mu^1_i, \mu^2_i, \dots) = (\mu^1_i(\cdot|\phi_i), \mu^2_i(\cdot|\phi_i), \dots).$$

γ_{ϕ_i} is clearly continuous. By Lemma 2 the composite function

$$g_{i,\phi_i} = h_{\phi_i} \circ \gamma_{\phi_i} : \hat{\boldsymbol{M}}_i \to \Delta(S^{T_{\phi_i}}_{-i} \times \bar{\boldsymbol{M}}_{-i})$$

is also continuous. Let $(\mu^1_i(\cdot|\phi_i), \mu^2_i(\cdot|\phi_i), \dots) = \gamma_{\phi_i}(\mu^1_i, \mu^2_i, \dots)$. Clearly, $(\mu^1_i(\phi_i|\phi_i), \mu^2_i(\phi_i|\phi_i), \dots) = 1$ and for all $k \geq 1$ eq. (1) is satisfied. Thus the mapping

$$g_i = (g_{\phi_i})_{\phi_i \in \Phi_i} : \hat{\boldsymbol{M}}_i \to \left[\Delta(S^{T_{\phi_i}}_{-i} \times \bar{\boldsymbol{M}}_{-i})\right]^{\Phi_i}$$

is continuous and satisfies eq. (1). The latter fact implies that g_i is 1-to-1 and the restriction of g_i^{-1} to $g_i(\hat{\boldsymbol{M}}_i)$ is continuous (i.e., it is homeomorphic).

The next step is to show that $g_i(\hat{\boldsymbol{M}}_i) = \Delta^{\Phi_i}(S_{-i}^{T_{\phi_i}} \times \bar{\boldsymbol{M}}_{-i})$.

First, $\Delta^{\Phi_i}(S_{-i}^{T_{\phi_i}} \times \bar{\boldsymbol{M}}_{-i}) \subset g_i(\hat{\boldsymbol{M}}_i)$: Take $\mu_i = (\mu_i(\cdot|\phi_i))_{\phi_i \in \Phi_i} \in \Delta^{\Phi_i}(S_{-i}^{T_{\phi_i}} \times \bar{\boldsymbol{M}}_{-i})$ and for all $\phi_i \in \Phi_i$ and $k \geq 1$, define $\mu_i^k(\cdot|\phi_i)$ using eq. (1). If $\boldsymbol{\mu}_i = (\mu_i^1, \mu_i^2, \dots) \in \hat{\boldsymbol{M}}_i$, then $g_i(\boldsymbol{\mu}_i) = \mu_i \in g_i(\hat{\boldsymbol{M}}_i)$. It is thus sufficient to show that $\boldsymbol{\mu}_i \in \hat{\boldsymbol{M}}_i$; in order to do this we have to verify that each μ_i^k satisfies property (iii) of the cps (coherency of $\boldsymbol{\mu}_i$ is satisfied by construction).

Second, $g_i(\hat{\boldsymbol{M}}_i) \subset \Delta^{\Phi_i}(S_{-i}^{T_{\phi_i}} \times \bar{\boldsymbol{M}}_{-i})$: Take $\boldsymbol{\mu}_i \in \hat{\boldsymbol{M}}_i$ and let $\mu_i = g_i(\boldsymbol{\mu}_i)$. We must verify that property (iii) of the cps holds for μ_i.

In the proof of their Proposition 1, Battigalli and Siniscalchi (1999) establishes (at a general level) that each μ_i^k in the first case and μ_i in the second case satisfies property (iii) of the cps. This again implies that $g_i(\hat{\boldsymbol{M}}_i) = \Delta^{\Phi_i}(S_{-i}^{T_{\phi_i}} \times \bar{\boldsymbol{M}}_{-i})$.

Even if i's hierarchy of cps' $\boldsymbol{\mu}_i$ is coherent, some elements of $g_i(\boldsymbol{\mu}_i)$ may assign positive probability to sets of incoherent hierarchies of players $-i$. Player i is certain that the other players have coherent beliefs if for some $\phi_i \in \Phi_i$ if $g_{\phi_i}(\boldsymbol{\mu}_i)(S_{-i}^{T_{\phi_i}} \times \hat{\boldsymbol{M}}_{-i}) = 1$. Common certainty of coherency for some $\phi_i \in \Phi_i$ can thus be inductively defined as follows:

$$
\begin{aligned}
&\boldsymbol{M}_i^1 = \hat{\boldsymbol{M}}_i \\
&\text{for all } n \geq 2, \\
&\boldsymbol{M}_i^n = \left\{ \boldsymbol{\mu}_i \in \boldsymbol{M}_i^{n-1} : \; g_{\phi_i}(\boldsymbol{\mu}_i)(S_{-i}^{T_{\phi_i}} \times \boldsymbol{M}_{-i}^{n-1}) = 1 \right\}, \\
&\boldsymbol{M}_i = \bigcap_{n \geq 1} \boldsymbol{M}_i^n.
\end{aligned}
$$

$\boldsymbol{M}_i \subset \hat{\boldsymbol{M}}_i$ is thus the set of collectively coherent cps' $\boldsymbol{\mu}_i$ of player i.

The final step in the proof is to show that the restriction of $g_i = (g_{\phi_i})_{\phi \in \Phi_i}$ to $\boldsymbol{M}_i \subset \hat{\boldsymbol{M}}_i$ induces a homeomorphism

$$f_i = (f_{\phi_i})_{\phi_i \in \Phi_i} : \boldsymbol{M}_i \to \Delta^{\Phi_i}(S_{-i}^{T_{\phi_i}} \times \boldsymbol{M}_{-i}).$$

It is easy to check that $\boldsymbol{M}_i = \{\boldsymbol{\mu_i} \in \hat{\boldsymbol{M}}_i : g_i(\boldsymbol{\mu}_i)(S_{-i}^{T_{\phi_i}} \times \boldsymbol{M}_{-i}) = 1\}$, so $g_i(\boldsymbol{M}_i) = \left\{\mu_i \in \Delta^{\Phi_i}(S^{T_{\phi_i}} \times \bar{\boldsymbol{M}}_{-i}) : \mu_i(S^{T_{\phi_i}} \times \boldsymbol{M}_{-i}) = 1\right\}$, since g_i is onto. But $g_i(\boldsymbol{M}_i)$ is homeomorphic to $\boldsymbol{M}_i$ and $\{\boldsymbol{\mu}_i \in \bar{\boldsymbol{M}}_i : g_i(\boldsymbol{\mu}_i)(S^{T_{\phi_i}} \times \boldsymbol{M}_{-i}) = 1\}$ is homeomorphic to $\Delta^{\Phi_i}(S^{T_{\phi_i}} \times \boldsymbol{M}_{-i})$ (for any Polish space Z and measurable subset W of Z, $\{\mu_i \in \Delta(Z) : \mu_i(W) = 1\}$ is homeomorphic to $\Delta(W)$). So $\boldsymbol{M}_i$ is homeomorphic to $\Delta^{\Phi_i}(S^{T_{\phi_i}} \times \boldsymbol{M}_{-i})$. ■

B Appendix

B.1 Proof of Proposition 1

The Proof follows naturally from the following Lemma, which itself is essentially an adaptation of the dynamic programming approach due to Battigalli and Dufwenberg (2007a, Section 3). We want to relate the problem $\max_{s_i^T} \mathbb{E}_{s_i^T, \boldsymbol{\mu}_i}[u_i|\phi_i]$ to a dynamic programming problem on the decision tree induced by $\boldsymbol{\mu}_i$, i.e., the decision tree player i thinks he is in. Important for the following analysis is our assumption that player i knows his own belief and assigns probability one to the strategy he intends to carry out. However, first we develop some notation needed for the Lemma.

Depth of the decision tree:

- For each k with $0 \leq k \leq l(\phi_i)$ (recall that $l(\phi_i)$ denotes the length of history ϕ_i). Let a_i^k be the action taken by some $i \in N$ in ϕ_i at the predecessor of ϕ_i of length k. Thus, by definition $\phi_i = (a^0, a^1, \ldots, a^{l(\phi_i)-1})$ where $a^k = (a_1^k, \ldots, a_n^k)$.
- Let $d(\phi_i) = \max_{\phi_i \leq \phi_i^z} [l(\phi_i^z) - l(\phi_i)]$ denote the depth of the decision tree with root ϕ_i.

Strategies:

- $(s_i^T|\phi_i)$ denotes the strategy that takes all the actions of player i in history ϕ_i and behaves as s_i^T otherwise:

$$(s_i^T|\phi_i)_{\tilde{\phi}_i} = \begin{cases} s_i^T(\tilde{\phi}_i) & \text{if } \tilde{\phi}_i \nless \phi_i, \\ a_i^{l(\tilde{\phi}_i)} & \text{if } \tilde{\phi}_i < \phi_i. \end{cases}$$

  Intuitively, $(s_i^T|\phi_i)$ is a strategy that takes on the observed actions made prior to the history ϕ_i, and then agrees with strategy s_i^T at ϕ_i and in what follows.

- Now change $(s_i^T|\phi_i)$ at ϕ_i so that it is the strategy obtained from $(s_i^T|\phi_i)$ by replacing $s_i^T(\tilde{h}_T)$ with $a_i \in A_i(\phi_i)$. The resulting strategy is denoted $(s_i^T|\phi_i, a_i)$. I.e.,

$$(s_i^T|\phi_i, a_i)_{\tilde{\phi}_i} = \begin{cases} (s_i^T|\phi_i)_{\tilde{\phi}_i} & \text{if } \tilde{\phi}_i \neq \phi_i, \\ a_i & \text{if } \tilde{\phi}_i = \phi_i. \end{cases}$$

  In words, $(s_i^T|\phi_i, a_i)$ is the strategy consistent with ϕ_i that chooses a_i at ϕ_i and behaves as $(s_i^T|\phi_i)$ in all other histories $\tilde{\phi}_i$. I.e., $(s_i^T|\phi_i)$ takes an ex ante (before player i makes his choice at ϕ_i) point of view of the strategy $s_i^T \in S_i^{T_{\phi_i}}(\phi_i)$ which is consistent with ϕ_i, while $(s_i^T|\phi_i, a_i)$ takes on an ex post (after player i makes his choice at ϕ_i) view of the strategy $s_i^T \in S_i^{T_{\phi_i}}(\phi_i, a_i)$ which is consistent with ϕ_i and the choice a_i he is about to make.

Value functions on the decision tree:

- Define the two value functions $V_{\boldsymbol{\mu}_i} : \Phi_i \to \mathbb{R}$ and $\overline{V}_{\boldsymbol{\mu}_i} : (\Phi_i \backslash \Phi_i^Z) \times A_i(\phi_i) \to \mathbb{R}$ induced by $\boldsymbol{\mu}_i$.

- For terminal histories $\phi_i^z \in \Phi_i^Z$, let

$$V_{\boldsymbol{\mu}_i}(\phi_i^z) = \mu_i^1(S_{-i}^{T_{\phi_i^z}}(\phi_i^z)|\phi_i^z) u_i(\phi_i^z, \boldsymbol{\mu}_i, \boldsymbol{\mu}_{-i}, s_{-i}^T).$$

- Assuming that the value function $V_{\boldsymbol{\mu}_i}(\phi_i, a)$ has been defined for all immediate successors (ϕ_i, a), let

$$\overline{V}_{\boldsymbol{\mu}_i}(\phi_i, a_i) = \sum_{a_{-i} \in A_{-i}(\phi_i)} \mu_i^1(S_{-i}^{T_{\phi_i}}(\phi_i, a_{-i})|\phi_i) V_{\boldsymbol{\mu}_i}(\phi_i, a_i, a_{-i}). \tag{i}$$

  For each $a_i \in A_i(\phi_i)$, $V_{\boldsymbol{\mu}_i}(\phi_i)$ is defined as

$$V_{\boldsymbol{\mu}_i}(\phi_i) = \max_{a_i \in A_i(\phi_i)} \overline{V}_{\boldsymbol{\mu}_i}(\phi_i, a_i).$$

Next we state the dynamic programming problem:

Lemma 3 (Dynamic Programming)**.** Suppose that for all $\phi_i \in \Phi_i \backslash \Phi_i^Z$,

$$s_i^{T,*}(\phi_i) \in \arg \max_{a_i \in A_i(\phi_i)} \overline{V}_{\boldsymbol{\mu}_i}(\phi_i, a_i).$$

Then for all $\phi_i \in \Phi_i \backslash \Phi_i^Z$,

$$\mathbb{E}_{(s_i^{T,*}|\phi_i), \boldsymbol{\mu}_i}[u_i|\phi_i] = V_{\boldsymbol{\mu}_i}(\phi_i) = \max_{s_i^T \in S_i^{T_{\phi_i}}(\phi_i)} \mathbb{E}_{s_i^T, \boldsymbol{\mu}_i}[u_i|\phi_i]. \tag{DP}$$

Proof of Lemma. The proof is by induction on $d(\phi_i)$.

Basic step: We start from the last stage: ϕ_i is such that all feasible actions following histories ϕ_i terminate the game, i.e. $d(\phi_i) = 1$. Clearly (DP) holds for all ϕ_i for which $d(\phi_i) = 1$.

Inductive step: We now fix some stage $k \geq 1$, which is not the last stage, and look at the stage just preceding it. Suppose (DP) holds for all ϕ_i such that $1 \leq d(\phi_i) \leq k$. Let $d(\phi_i) = k + 1$.

By the law of iterated expectations for all $a_i \in A_i(\phi_i)$:

$$\mathbb{E}_{(s_i^{T,*}|\phi_i,a_i),\boldsymbol{\mu}_i}[u_i|\phi_i] = \sum_{a_{-i}\in A_{-i}(\phi_i)} \mu_i^1(S_{-i}^{T_{\phi_i}}(\phi_i, a_{-i})|\phi_i)\mathbb{E}_{(s_i^{T,*}|\phi_i,a_i),\boldsymbol{\mu}_i}[u_i|\phi_i, a_i, a_{-i}]. \tag{ii}$$

By the inductive hypothesis, for all $a_i \in A_i(\phi_i)$ and $a_{-i} \in A_{-i}(\phi_i)$:

$$\mathbb{E}_{(s_i^{T,*}|\phi_i,a_i),\boldsymbol{\mu}_i}[u_i|\phi_i, a_i, a_{-i}] = V_{\boldsymbol{\mu}_i}(\phi_i, a_i, a_{-i}) = \max_{s_i^T \in S_i^{T_{\phi_i}}(\phi_i,a_i,a_{-i})} \mathbb{E}_{s_i^T,\boldsymbol{\mu}_i}[u_i|\phi_i, a_i, a_{-i}]. \tag{iii}$$

If we plug (iii) into (ii) and compare with (i), we get:

$$\mathbb{E}_{(s_i^{T,*}|\phi_i,a_i),\boldsymbol{\mu}_i}[u_i|\phi_i] = \overline{V}_{\boldsymbol{\mu}_i}(\phi_i, a_i).$$

Therefore,

$$\mathbb{E}_{(s_i^{T,*}|\phi_i),\boldsymbol{\mu}_i}[u_i|\phi_i] = V_{\boldsymbol{\mu}_i}(\phi_i) = \max_{s_i^T \in S_i^{T_{\phi_i}}(\phi_i)} \mathbb{E}_{s_i^T,\boldsymbol{\mu}_i}[u_i|\phi_i]$$
$$\textit{if and only if}$$
$$s_i^{T,*}(\phi_i) \in \arg\max_{a_i\in A_i(\phi_i)} \mathbb{E}_{(s_i^{T,*}|\phi_i,a_i),\mu_i}[u_i|\phi_i]$$
$$\textit{if and only if}$$
$$s_i^{T,*}(\phi_i) \in \arg\max_{a_i\in A_i(\phi_i)} \overline{V}_{\boldsymbol{\mu}_i}(\phi_i, a_i).$$

The latter condition holds by assumption and the inductive step is hereby proven. ∎

Proof of Proposition. Let $(\boldsymbol{\sigma}, \boldsymbol{\mu})$ be a consistent assessment. Then for each $\phi_i^z \in \Phi_i^Z$,

$$V_{\boldsymbol{\mu}_i}(\phi_i^z) = \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\mu}}[u_i|\phi_i],$$

and for all ϕ_i with $d(\phi_i) = 1$ we have

$$V_{\boldsymbol{\mu}_i}(\phi_i) = \max_{a_i\in A_i(\phi_i)} \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\mu}}[u_i|\phi_i, a_i]. \tag{BI}$$

Then a straightforward backwards induction argument shows (BI) holds for all $\phi_i \in \Phi_i \backslash \Phi_i^Z$. Therefore the Lemma implies that the Proposition holds. ■

B.2 Proof of Theorem 1

First let $\beta^1(\boldsymbol{\sigma}) = (\beta^1(\boldsymbol{\sigma}))_{i\in N}$ denote the profile of first-order beliefs derived from $\boldsymbol{\sigma}$ according to condition (i) in Definition 4. The profile of infinite belief hierarchies $\boldsymbol{\mu} = \beta(\boldsymbol{\sigma})$ is obtained by condition (ii) in Definition 4. By construction, the assessment $(\boldsymbol{\sigma}, \beta(\boldsymbol{\sigma}))$ is consistent. It follows that $\beta(\cdot)$ is a continuous function.

Suppose that each player i is subject to a slight imperfection of rationality (tremble) of the following kind. At every history ϕ_i there is a small positive probability ϵ_{i,ϕ_i} for the breakdown of rationality. Whenever rationality breaks down, every action a_i will be selected with some positive probability $\sigma_i(a_i|\phi_i) = \epsilon_{i,\phi_i}(a_i)$. Formally, fix a strictly positive vector $\boldsymbol{\epsilon} = ((\epsilon_{i,\phi_i}(a_i))_{a_i\in A_i(\phi_i)})_{i\in N,\phi_i\in\Phi_i\backslash\Phi_i^Z}$ such that for all $\phi_i \in \Phi_i\backslash\Phi_i^Z$, $\sum_{a_i\in A_i(\phi_i)} \epsilon_{i,\phi_i}(a_i) < 1$.

Definition 6 ($\boldsymbol{\epsilon}$-constrained equilibrium)**.** An $\boldsymbol{\epsilon}$-constrained equilibrium is a set of behavioral strategies profiles $\boldsymbol{\sigma}$ such that for all $i \in N$, $\phi_i \in \Phi_i$, $a_i \in A_i(\phi_i)$:

(i) $\sigma_i(a_i|\phi_i) \geq \epsilon_{i,\phi_i}(a_i)$,

(ii) $a_i \notin \arg\max_{a_i\in A_i(\phi_i)} \mathbb{E}_{\boldsymbol{\sigma},\beta(\boldsymbol{\sigma})}[u_i|\phi_i, a_i] \Rightarrow \sigma_i(a_i|\phi_i) = \epsilon_{i,\phi_i}(a_i)$.

Let $\Sigma_{\boldsymbol{\epsilon}} = \prod_{i\in N} \Sigma_{\boldsymbol{\epsilon},i}$ be the set of behavioral strategy profiles satisfying condition (i) in Definition 6, and let $\text{BR}_{\boldsymbol{\epsilon}} : \Sigma_{\boldsymbol{\epsilon}} \to \Sigma_{\boldsymbol{\epsilon}}$ be the $\boldsymbol{\epsilon}$-best response correspondence that assigns to each profile $\boldsymbol{\sigma}$ the subset of profiles in $\Sigma_{\boldsymbol{\epsilon}}$ satisfying condition (ii) of the definition,

$$\begin{aligned}\text{BR}_{\boldsymbol{\epsilon},i}(\boldsymbol{\sigma}) = \{\boldsymbol{\sigma}_i \in \Sigma_{\boldsymbol{\epsilon},i} : & a_i \notin \arg\max_{\tilde{a}_i\in A_i(\phi_i)} \mathbb{E}_{\boldsymbol{\sigma},\beta(\boldsymbol{\sigma})}[u_i|\phi_i, \tilde{a}_i] \\ & \Rightarrow \sigma_i(a_i|\phi_i) = \epsilon_{i,\phi_i}(a_i), \text{ for all } \phi_i \in \Phi_i, \text{ for all } a_i \in A_i(\phi_i),\end{aligned}$$

$$\text{BR}_{\boldsymbol{\epsilon}}(\boldsymbol{\sigma}) = \prod_{i\in N} \text{BR}_{\boldsymbol{\epsilon},i}(\boldsymbol{\sigma}).$$

$\text{BR}_{\boldsymbol{\epsilon},i}(\boldsymbol{\sigma})$ is a nonempty convex subset of Euclidean space $\Delta(A_i(\phi_i))$. Since $\mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\mu}}[u_i|\phi_i, a_i]$ is continuous in $(\boldsymbol{\sigma}, \boldsymbol{\mu})$ and $\boldsymbol{\mu} = \beta(\boldsymbol{\sigma})$ is a continuous function, $\mathbb{E}_{\boldsymbol{\sigma},\beta(\boldsymbol{\sigma})}[u_i|\phi_i, a_i]$ is continuous in $\boldsymbol{\sigma}$.

We now have enough structure to apply Kakutani's fixed point theorem to the best response correspondence. $\text{BR}_{\boldsymbol{\epsilon}}(\boldsymbol{\sigma})$ is upper hemicontinuous because $\mathbb{E}_{\boldsymbol{\sigma},\beta(\boldsymbol{\sigma})}[u_i|\phi_i, a_i]$ is continuous for each (finite) $\phi_i \in \Phi_i$ and $a_i \in A_i(\phi_i)$, nonempty since each $\mathbb{E}_{\boldsymbol{\sigma},\beta(\boldsymbol{\sigma})}[u_i|\phi_i, a_i]$ is continuous and $\Sigma_{\boldsymbol{\epsilon}}$ is compact, and convex valued because each $\mathbb{E}_{\boldsymbol{\sigma},\beta(\boldsymbol{\sigma})}[u_i|\phi_i, a_i]$ is quasi-concave on $\Sigma_{\boldsymbol{\epsilon}}$. Therefore $\mathrm{BR}_{\boldsymbol{\epsilon}}(\boldsymbol{\sigma})$ has a fixed point, which is an $\boldsymbol{\epsilon}$-constrained equilibrium.

Fix a sequence $\boldsymbol{\epsilon}^k \to 0$ and a corresponding sequence of $\boldsymbol{\epsilon}^k$-constraint equilibrium strategies $\boldsymbol{\sigma}^k$. By compactness, the sequence $(\boldsymbol{\sigma}^k)$ has a limit point $\boldsymbol{\sigma}^*$. A trembling-hand perfect equilibrium is any limit of $\boldsymbol{\epsilon}$-constraint equilibria as $\boldsymbol{\epsilon}^k \to 0$. We will now prove that the trembling-hand perfect equilibrium $(\boldsymbol{\sigma}^*, \beta(\boldsymbol{\sigma}^*))$ is a sequential equilibrium.

Assessment $(\boldsymbol{\sigma}^*, \beta(\boldsymbol{\sigma}^*))$ is continuous: to see this note that, by continuity, $\beta(\boldsymbol{\sigma}^*)$ is a limit point of $\beta(\boldsymbol{\sigma}^k)$, and that the set of consistent assessment is closed. By continuity of $\mathbb{E}_{\boldsymbol{\sigma},\beta(\boldsymbol{\sigma})}[u_i|\phi_i, a_i]$ in $\boldsymbol{\sigma}$ (and fitness of $A_i(\phi_i)$), for k sufficiently large

$$\arg\max_{a_i \in A_i(\phi_i)} \mathbb{E}_{\boldsymbol{\sigma}^*,\beta(\boldsymbol{\sigma}^*)}[u_i|\phi_i, a_i] = \arg\max_{a_i \in A_i(\phi_i)} \mathbb{E}_{\boldsymbol{\sigma}^k,\beta(\boldsymbol{\sigma}^k)}[u_i|\phi_i, a_i].$$

By Definition 5 and Proposition 1 each $(\boldsymbol{\sigma}^*, \beta(\boldsymbol{\sigma}^*))$ is a sequential equilibrium assessment.
■

C Appendix

C.1 Proof of Result 4

Independent of the state of nature, what the principal intends to give to the agent is $\alpha_p(\theta)[s\pi(\theta,h)-4]+(1-\alpha_p(\theta))[s\pi(\theta,l)]$ where $\alpha_p(\theta)$ is the principal's belief about the agent's effort choice given a certain state of nature $\theta \in \{B,S\}$ (a feature of the principal's first order belief) and $\pi(\cdot)$ is the principal's profit depending on the effort choice of the agent and the state of nature. The equitable payoff π^e that the principal could give to the agent given the state of nature and first order belief of the principal is

$$\pi^e = \frac{1}{2}[\alpha_p(\theta)(\pi(\theta,h)-4)+(1-\alpha_p(\theta))(\pi(\theta,l))-4\alpha_p(\theta)]$$

which reduces to

$$\pi^e = \frac{1}{2}[\alpha_p(\theta)(\pi(\theta,h)-\pi(\theta,l))+(\pi(\theta,l))-8\alpha_p(\theta)]$$

Given these two terms

$$K^p = \alpha_p(\theta)[s(\pi(\theta,h)-\pi(\theta,l))]-4\alpha_p(\theta)+[s\pi(\theta,l)] - \frac{1}{2}[\alpha_p(\theta)(\pi(\theta,h)-\pi(\theta,l))+(\pi(\theta,l))-8\alpha_p(\theta)]$$

and $K^p < 0$ if $s < \frac{1}{2}$.

C.2 Proof of Result 5

The agent chooses high effort in equilibrium as long as

$$20s-4+Y(20s-10)(20-20s) \geq 15s+Y(20s-10)(15-15s) \tag{15}$$

which reduces to

$$Y \geq \frac{4-5s}{5(1-s)}\frac{1}{20s-10} \tag{16}$$

For values $0.5 \leq s \leq 0.8$ the minimum sensitivity to reciprocity Y required to induce high effort e_h is weakly positive. This means, high effort can be implemented at $0.5 < s \leq 0.8$ if $Y \geq \frac{4-5s}{5(1-s)}\frac{1}{20s-10}$, otherwise the agent chooses low effort l. At profit shares above $K^p > 0$ and the agent materially prefers high effort. Hence, the agent chooses high effort independent of his sensitivity to reciprocity. The analog argument holds for profit shares $s \leq 0.5$.

C.3 Proof of Result 6

The agent chooses high effort in equilibrium as long as

$$30s - 4 + Y(30s - 15)(30 - 30s) \geq 20s + Y(30s - 15)(20 - 20s) \tag{17}$$

which reduces to

$$Y \leq \frac{4 - 10s}{10(1 - s)} \frac{1}{30s - 15} \tag{18}$$

For values $0.4 < s \leq 0.5$ the maximum sensitivity to reciprocity Y required to induce high effort h is weakly positive. This means, high effort can be implemented at $0.4 < s \leq 0.5$ if $Y \leq \frac{4-10s}{10(1-s)} \frac{1}{30s-15}$, otherwise the agent chooses low effort l. At profit shares above $K^p > 0$ and the agent materially prefers high effort. Hence, the agent chooses high effort independent of his sensitivity to reciprocity. The analog argument holds for profit shares $s \leq 0.4$.

References

ALIPRANTIS, C. and KIM, C. (1999). *Infinite Dimensional Analysis: A Hitchhikers Guide*. Springer-Verlag, New York.

ANDREONI, J. and BERNHEIM, B. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, **77** (5), 1607–1636.

AUMANN, R. and BRANDENBURGER, A. (1995). Epistemic conditions for nash equilibrium. *Econometrica*, **63**, 1161–1180.

BATTIGALLI, P. and DUFWENBERG, M. (2007a). Dynamic psychological games: Omitted material.

— and — (2007b). Guilt in games. *The American Economic Review, Papers and Proceedings*, **97** (2), 170–176.

— and — (2009). Dynamic psychological games. *Journal of Economic Theory*, **144**, 1–35.

— and SINISCALCHI, M. (1999). Hierarchies of conditional beliefs and interactive epistemology in dynamic games. *Journal of Economic Theory*, **88**, 188–230.

BELLEMARE, C., SEBALD, A. and STROBEL, M. (2010). Measuring the willingness to pay to avoid guilt: Estimation using equilibrium and stated belief models. *Journal of Applied Econometrics, forthcomming.*

BRANDENBURGER, A. and DEKEL, E. (1993). Hierarchies of beliefs and common knowledge. *Journal of Economic Theory*, **59**, 189–198.

BROBERG, T., ELLINGSEN, T. and JOHANNESSON, M. (2007). Is generosity involuntary? *Economics Letters*, **94** (1), 32–37.

CHARNESS, G. and DUFWENBERG, M. (2006). Promises and partnership. *Econometrica*, **74** (6), 1579–1601.

DANA, J., CAIN, D. and DAWES, R. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, **100** (2), 193–201.

—, WEBER, R. and KUANG, J. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, **33** (1), 67–80.

DEKEL, E., LIPMAN, B. and RUSTICHINI, A. (1998). Standard state-space models preclude unawareness. *Econometrica*, **66**, 159–173.

DUFWENBERG, M. and KIRCHSTEIGER, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, **47** (2), 268–298.

FAGIN, R. and HALPERN, J. (1988). Beliefs, awareness, and limited reasoning. *Artificial Intelligence*, **34**, 39–76.

Falk, A., Fehr, E. and Fischbacher, U. (2008). Testing theories of fairness–intentions matter. *Games and Economic Behavior*, **62** (1), 287–303.

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

Fehr, E., Kirchsteiger, G. and Riedl, A. (1993). Does fairness prevent market clearing? an experimental investigation. *The Quarterly Journal of Economics*, **108** (2), 437–459.

Feinberg, Y. (2009). Games with unawareness. *Mimeo*.

Geanakoplos, J., Pearce, D. and Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, **1** (1), 60–79.

Grant, S. and Quiggin, J. (2009). Inductive reasoning about unawareness. *Risk and Uncertainty Working Papers*.

Halpern, J. (2001). Alternative semantics for unawareness. *Games and Economic Behavior*, **37**, 321–339.

— and Rêgo (2008). Interactive unawareness revisited. *Games and Economic Behavior*, **62**, 323–262.

Heifetz, A., Meier, M. and Schipper, B. (2006). Interactive unawareness. *Journal of Economic Theory*, **130**, 78–94.

—, — and — (2008). A cannonical model for interactive unawareness. *Games and Economic Behavior*, **62**, 304–324.

—, — and — (2011). Dynamic unawareness and rationalizable behavior. *Mimeo*.

Kreps, D. and Wilson, R. (1982). Sequential equilibrium. *Econometrica*, **50**, 863–894.

Lazear, E., Malmendier, U. and Weber, R. (2009). Sorting and social preferences. *Mimeo*, p. 97.

Li, J. (2009). Information structures with unawareness. *Journal of Economic Theory*, **144** (3), 977–993.

Modica, S. and Rustichini, A. (1999). Unawareness and partitional information structures. *Games and Economic Behavior*, **27**, 265–298.

Nielsen, C. and Sebald, A. (2011). Forward induction, strong beliefs, and unawareness in dynaminc psychological games. *Working Paper, University of Copenhagen*.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, pp. 1281–1302.

Tadelis, S. (2008). The power of shame and the rationality of trust. *Mimeo*.