Simple Unawareness in Dynamic Psychological Games

Nielsen, Carsten Søren; Sebald, Alexander Christopher

Publication date:
2014

Simple Unawareness in Dynamic Psychological Games

Carsten S. Nielsen[*] Alexander Sebald[†]

March 5, 2014

Abstract

Building on Battigalli and Dufwenberg (2009)'s framework of dynamic psychological games and the recent progress in the modeling of dynamic unawareness by Heifetz et al. (2013a), we model and analyze the impact of asymmetric awareness in the strategic interaction of players motivated by belief-dependent preferences like reciprocity and guilt. Specifically we characterize extensive-form games with belief-dependent preferences and simple unawareness, define extensive-form rationalizability and, using this, 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; Belief-dependent preferences; Extensive-form rationalizability.

JEL-Classifications: C72, C73, D80

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

[*] Department of Economics and Department of Psychology, University of Copenhagen, Øster Farimagsgade 2A, DK-1353, Copenhagen K, Denmark. Phone: (+45) 3532-4839. Fax: (+45) 3532-4932. E-mail: carsten.nielsen@gmail.com. Web: https://sites.google.com/site/carstennielsen/.

[†] Corresponding Author: 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.

psychological preferences [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 [e.g., Rabin (1993), Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006)] and guilt aversion [for example, Charness and Dufwenberg (2006), Battigalli and Dufwenberg (2007)]. 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 players 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.

A widely unspoken assumption that is underlying game-theoretic analyses, and therefore also the analyses of psychological games, is that players are aware of *all* facts characterizing the strategic environment they are in. However, in many real life situations this is not the case. People often have asymmetric awareness levels concerning their own as well as the feasible choices of others although they are part of the same strategic environment. People are frequently surprised in the sense that they become aware of new strategic alternatives by observing actions they had previously been unaware of. In recent years different models of unawareness have been proposed showing the importance of unawareness for individual decision making problems as well as the strategic interaction of players in standard (non-psychological) games [e.g., Fagin and Halpern (1988), Dekel et al. (1998), Modica and Rustichini (1999), Halpern (2001), Heifetz et al. (2006), Halpern and Rêgo (2006), Halpern and Rêgo (2008), Heifetz et al. (2008), Halpern and Rêgo (2009), Li (2009), Feinberg (2011), Grant and Quiggin (2012), Heifetz et al. (2013a) and Heifetz et al. (2013b)].

However, it is not only in standard games that unawareness is important. We show in our analysis here that unawareness has a profound and distinct 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 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 feeling guilty from letting down *Bob* if she canceled his party to take the exam the following morning. As a consequence, *Ann* might choose the second date to avoid letting *Bob* down. 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 *Bob*'s awareness. Only her belief-dependent feeling of guilt towards *Bob* creates the strong emotional incentive to leave him unaware.

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. We analyze the influence and importance of unawareness concerning feasible paths of play for the strategic interaction of players in psychological games. To simplify the analysis we concentrate on two-player strategic environments with simple unawareness. More specifically, building on Battigalli and Dufwenberg (2009)'s framework of dynamic psychological games and the recent progress in the modeling of unawareness [Heifetz et al. (2006, 2008, 2013a,b)], we define a two-player model in which players are motivated by belief-dependent preferences and one player is potentially unaware of certain feasible paths of play. Using this framework we 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 (2007). We limit ourselves to two-player environments and simple asymmetric awareness scenarios in order to intuitively introduce our model and clearly uncover the role of unawareness without burdening the analysis with technical issues arising in strategic environments allowing for more players and more complex unawareness.

Our examples demonstrate that the strategic behavior of players motivated by belief-dependent preferences crucially depends on their awareness concerning the strategic environment they are in, their perception concerning the awareness of others, their perception concerning the perception of others, and so on–a fact that implies both an opportunity as well as a challenge to analyses empirically investigating the strength and nature of belief-dependent preferences. On the one hand, in line with experimental evidence suggesting that people are more prone to selfish choices if they believe that others will remain unaware of them [e.g., Dana et al. (2006), Dana et al. (2007), Broberg et al. (2007), Tadelis (2008), Andreoni and Bernheim (2009), Lazear et al. (2009)], our examples show that varying the degree of awareness in the interaction of players that are motivated by belief-dependent preferences leads to intuitive and testable predictions distinct from predictions based on consequentialist preferences like selfishness and inequality aversion [Fehr and Schmidt (1999)]. On the other hand, it poses a challenge for experimental investigations in relatively uncontrolled environments like the field or the Internet. As also seen in our introductory example, not controlling for *Ann*'s perception concerning *Bob*'s awareness might lead to wrong inferences concerning *Ann*'s inclination to feel guilty towards *Bob*. Furthermore, our examples reveal that over and above the actual choices that are made, managing other people's awareness levels has to be understood as an integral and important part of any strategic interaction. By managing other's awareness levels, we influence the others' expectations and perceptions concerning our intentions, which in turn influences their behavior.

We start out by formulating a model concentrating on two-player extensive-forms with complete information, observable actions and no chance moves. To allow for unawareness we use a standard extensive-form representing the objective strategic environment and a subtree thereof, and define extensive-forms with simple unawareness with the help of a possibility function. This possibility function describes for each possible decision node in the objective extensive-form, and copy thereof in the subtree, the augmented history that players perceive to be at. Our two-player extensive-forms are in essence a special case of Heifetz et al. (2013a)'s generalized extensive-forms, and therefore embeddable in their setting.[1] Of course, our extensive-form with unawareness is not typically common knowledge among players, and therefore should be interpreted from the modeler's point of view. In fact, any game that does not explicitly distinguish between the players' description of the strategic environment and the modeler's will fail to capture unawareness [see Dekel et al. (1998)].

Having defined our class of two-player extensive-forms with unawareness, we formally characterize belief-dependent preferences in our setting. In synthesis, we define a player's

[1] As hinted at before, different models of unawareness have recently been presented in the literature. Although our analysis closely links to the setting of Heifetz et al. (2013a), we strongly believe that our idea and intuition can also be formalized extending one of the other frameworks for dynamic unawareness [Halpern and Rêgo (2006)].

strategies and conditional beliefs about the other player's pure strategies (first-order beliefs), beliefs about the other player's beliefs (second-order beliefs), and so on. The infinite hierarchy of conditional beliefs that we define takes player's awareness, players's perception regarding the other's awareness, and so forth, into account and is used for the general specification of our belief-dependent preferences and, hence, the characterization of our class of dynamic psychological games with simple unawareness. As mentioned above, specific types of belief-dependent preferences that can be embedded in our model are among others reciprocity and guilt aversion.

Dufwenberg and Kirchsteiger (2004), Battigalli and Dufwenberg (2007) and Sebald (2010) propose sequential equilibrium as a solution concept for their psychological games. However, assuming equilibrium play is very demanding in strategic environments involving unawareness. The implicit assumption made when imposing sequential equilibrium on strategic settings with unawareness is that if a player becomes aware of more during the game, he will compute new equilibrium beliefs not rationalizing, for example, why the other player made him aware. Sequential equilibrium only requires a player to reason about the other player's future behavior. For this reason, we impose extensive-form rationalizability [Pearce (1984)], which embodies forward induction, as a solution concept for our psychological games with simple unawareness. Extensive-form rationalizability implies, that along each feasible path of play, every active player is always certain that the other player sequential best responds, certain that the other player is certain that he sequential best responds, and so on. If a player finds himself at some augmented history, where the other player's strategies that could lead to that augmented history are inconsistent with the players previous certainty in the other player's best response, then the player seeks a *best rationalization* which could have led to that augmented history [Battigalli (1997), Battigalli and Siniscalchi (2002)]. That is, if the player is "surprised" by the other player's unexpected action, and cannot use Bayesian updating, then he forms new beliefs that justify this observed inconsistency. In its simplest form, forward-induction reasoning involves the assumption that, upon observing an unexpected (but undominated) action of the other player, a player maintains the working hypothesis that the latter is a sequential best response. The best rationalization principle captures precisely this type of argument.

After having defined our model, the solution concept and two prominent notions of belief-dependent preferences, reciprocity and guilt aversion, we describe two examples to highlight the role of unawareness in the interaction of agents motivated by reciprocity and guilt aversion. First, we consider a version of 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.[2] Different to Dufwenberg and Kirchsteiger (2004)'s analysis assuming full awareness, it is shown that as long as *Bob* is unaware of the fact that *Ann* could have defected, he defects independent of his sensitivity to reciprocity - even when *Ann* chooses to cooperate. 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*. *Ann* anticipates this and defects as long as she cannot cooperate and simultaneously make *Bob* aware of the fact that she could have defected. As a second example, we investigate a trust game with guilt aversion also analyzed in Battigalli and Dufwenberg (2009). We assume that the second mover, *Bob*, is guilt averse and aware of everything. Whereas the first mover, *Ann*, is unaware that *Bob* can actually 'share part of the pie'. Analogue to the intuition in our introductory example, in this scenario featuring asymmetric awareness *Bob* does not feel guilty when 'grabbing the entire pie', as he knows that *Ann* who does not expect him to share is not let down. Different to Battigalli and Dufwenberg (2009)'s analysis assuming full awareness, 'grabbing the entire pie' is Bob's unique equilibrium behavior independent of how guilt averse he is. 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.

The organization of the paper is as follows: In section 2 we introduce our two-player model. Following this, in section 3 we define psychological games with unawareness. Section 4 contains the definition of our solution concept: extensive-form rationalizability. Sections and 6 contain a formal definition of belief-dependent reciprocity and guilt aversion in our setting with unawareness and two examples. Finally, section 7 concludes.

2 Model

This section introduces most of the required game-theoretic notation and summarizes the features of unawareness that are relevant for our analysis. For simplicity, our analysis focuses on two-player, extensive-form games with observable actions and no chance

[2] Note that this awareness scenario is similar to the finitely repeated prisoner's dilemma with unawareness analyzed in Feinberg (2004).

moves. We allow for unawareness concerning feasible paths of play. An unaware player is unaware of his own unawareness and thinks that the other player is aware of the same as he is. An aware player, on the other hand, is certain what the other player is aware of.[3] The model relies on Heifetz et al. (2006, 2008, 2013a,b)'s class of games with unawareness.

2.1 The objectively feasible extensive-form

Consider two players i and j, their finite sets of actions A_i and A_j, and potentially feasible action profiles $(a_i, a_j) \in A_i \times A_j$. Let there be a finite set of decision nodes N including the initial node n^0. By convention, the game starts at the initial node n^0. As play unravels, each player is informed of the actions that have just occurred. Each subsequent node thus describes sequences of consecutive action profiles. That is, for some stage of the game $1 \leq l$, a subsequent node $n = (a^1, \dots, a^l)$ is represented by the action profile $a^k = (a_i^k, a_j^k)$ where $1 \leq k \leq l$.

In the continuation we take the view of player i. Analogous definitions apply for player j. The set of feasible actions for player i may depend on previous actions taken, and we therefore denote the set of potentially feasible actions at n by $A_{i,n}$. Player i is active at n if $A_{i,n}$ contains more than one element. There are simultaneous moves at n, if both players are active at n. A node is terminal, denoted $z = (a^1, \dots, a^l)$, if the set $A_{i,z} \times A_{j,z}$ is empty. Let Z denote the set of terminal nodes.[4]

2.2 Subjective Views

An aware player's subjective view of the feasible paths of play coincides with the objectively feasible extensive-form N. To model an unaware player's subjective view concerning the feasible paths of play, we make use of a derived extensive-form referred to as a *subtree*.

Definition 1. A subset of the objectively feasible extensive-form N is a subtree T if for some nonempty subset of terminal nodes $E \subseteq Z$:

$$T = \{n \in N : n \leq z \text{ for some } z \in E\}.$$

[3] An adaptation of our model and techniques to general extensive-forms with more complex awareness structures can be achieved at a notational cost.

[4] For a complete definition see Osborne and Rubinstein (1994).

Subtree T starts at the initial node and ends at one or more terminal nodes in Z. It thus represents a set of feasible paths of play. Importantly, subtree T is an unaware player's subjective view of the feasible paths of play, an aware player's view of the unaware player's subjective view, the unaware player's view of the aware player's view of the feasible paths of play, and so on. It may also represents an initially unaware player's view on his own subjective view of the feasible paths of play at an earlier stage of the game, after his awareness of the feasible paths of play has evolved. Definition 1 does, however, not imply that an unaware player necessarily needs to be aware of when the game starts. An unaware player may think that the game starts with the initial mover being passive effectively moving the start of the game, as he is aware of it, to the node at which either the aware player or he himself is active.

Nodes n that are in the subtree T also appear in the objectively feasible extensive-form N. When we jointly consider these nodes, we will need to explicitly differentiate these. We label by n_T the copy in subtree T of the node n in the objective extensive-form N whenever the copy of n is a part of subtree T, and assume that if the action profile taken at n leads to n', then it also leads copy n_T to copy n'_T. Formally, n_T is said to be a copy of n if $n = (a^1, \dots, a^l) = n_T$. Let $\mathcal{N}$ be the collection of *non-terminal* nodes in N and copies of these in subtree T and refer to elements of $\mathcal{N}$ as *decision nodes* and *decision copies*, respectively.

2.3 Extensive-forms with simple unawareness

In standard extensive-forms with observable actions, commonly known nodes describe all possible actions that can be taken throughout a game. However, this need not be the case when a player is unaware of certain feasible paths of play. At a decision node n in the objective extensive-form N, a player may only be aware of the possible actions that can be taken at decision copy n_T in subtree T.

To formulate such situations, we define augmented histories with the help of a possibility function. Augmented histories not only describe the sequence of actions taken before a player has to move, but also the player's awareness regarding the feasible set of actions he can choose from.

Definition 2. For player i there exists a one-to-one possibility function:

$$\boldsymbol{h}_i : \mathcal{N} \to \mathcal{N},$$

defining for each decision node/copy in $\mathcal{N}$ the decision node/copy player i perceives to be at.

We call the outcome of the possibility function an *augmented history*. A typical augmented history is denoted by h_i, as in $h_i := \boldsymbol{h}_i(n)$. Let H_i be the complete set of all augmented histories and H_i^T the subset thereof containing all mappings into decision copies in subtree T.

The mapping describes for each decision node in N, and possible decision copy thereof in T, the decision node/copy that player i perceives to be at. The mapping h_i adheres to properties that regulate which decision nodes/copies players are aware of.[5] Because actions are observable, a player knows the sequence of action profiles that has led to the decision node/copy he is at, knows that the other player knows, and so on. Thus, observable actions place two automatic restrictions on augmented histories. First, the decision node/copy that a player perceives to be at must be described by the sequence of action profiles observed. That is, augment history h_i at a decision node/copy must map into a decision node/copy that shares past actions (**Property 1**). Second, observable actions imply that the augmented history h_i maps into a single decision node/copy (**Property 2**). Since a player knows of the sequence of action profiles, and thus the decision node/copy, he does not need to make any inferences about its legitimacy.

We need two more properties to model simple unawareness. First, as we consider confined awareness it should hold that at decision copy $n_T \in T$ a player cannot perceive to be at an augmented history in the objective extensive-form:

Property 3. (*Confined awareness*): For a decision copy n_T in subtree T, the augmented history $\boldsymbol{h}_i(n_T)$ must map into an element in subtree T.

[Figure 1 and 2]

Second, at decision node n in the objective extensive-form N a player cannot anticipate to become unaware:

Property 4. (*No anticipation to become unaware*): For a decision node n and path $n, \ldots, n'$ in the objective extensive-form N, the augmented history $\boldsymbol{h}_i(n')$ must map into an element in the objective extensive-form N.

[5] As the simple unawareness structure that we consider here only requires one subtree T in addition to the objective extensive-form N, we do not have to consider all properties laid out by Heifetz et al. (2013a).

[Figure 3 and 4]

Given these properties, an augmented history can only map: (i) from a decision node/copy into itself, or (ii) from a decision node into its decision copy in the subtree T.[6]

Because the awareness of the two players may differ, we need to be explicit about their subjective view regarding each other's (perhaps more restricted) subjective view at each augmented history. For this purpose we make use of the composite of the possibility function. The composite possibility function of player i is $\boldsymbol{h}_j \circ \boldsymbol{h}_i(n)$, stating that player i: (i) considers the augmented history he perceives to be in, and (ii) from that point of view he considers the augmented history he thinks that player j perceives to be in. In a game without unawareness (i) and (ii) will coincide, but this need not to be the case in games with unawareness. A typical outcome of the composite possibility function, which is an augmented history, is denoted by h_{ji}, as in $h_{ji} := \boldsymbol{h}_j \circ \boldsymbol{h}_i(n)$.[7]

So what does it mean to be unaware? Whenever player i is at an augmented history $h_i \in H_i^T$, he is only aware of augmented histories in H_i^T. However, whenever player i finds himself at augmented history $h_i \in H_i \backslash H_i^T$ he is aware of augmented histories in H_i. In other words, at $h_i \in H_i^T$ player i is unaware of augmented histories in $H_i \backslash H_i^T$. For some augmented history $h_i \in H_i \backslash H_i^T$ it may also be the case that $h_{ji} \in H_j^T$, describing the situation where player i knows that player j at h_i is unaware of some feasible paths of play. Finally, for $h_i \in H_i^T$ we always have that $h_{ji} \in H_i^T$. That is, an unaware player i is unaware of his own unawareness.

We now demonstrate, by a simple example, the structure of extensive-forms with unawareness. Consider the extensive-form underlying the sequential prisoners dilemma also analyzed by Dufwenberg and Kirchsteiger (2004).

[Figure 5]

Figure 5 shows the objective extensive-form N and a subtree T. Consider first the objective extensive-form N. At the initial decision node n^0, *Ann* can choose between *Cooperate* (C) and *Defect* (D) while *Bob* is passive. In decision nodes n^1 and n^2, *Bob* can choose between *cooperate* (c) and *defect* (d) while *Ann* is passive. In subtree T, on

[6] The restrictions and properties imposed in this subsection are consistent with the relevant unawareness properties in Heifetz et al. (2013a, p. 59) (properties $U0$, $U1$, $U2$, and DA).

[7] Composite possibilities allow for more involved subjective views. For example, $\boldsymbol{h}_i \circ \boldsymbol{h}_j \circ \boldsymbol{h}_i(n)$ which is player i's perception at decision node n of player j's perception of what he perceives. However, in our setting with simple unawareness such involved iterative subjective views will be redundant.

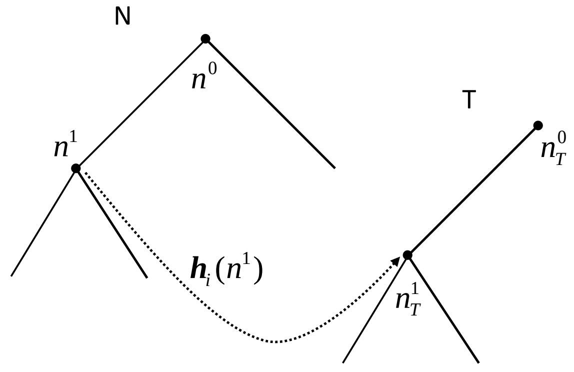


Figure 1: In line with Property 3

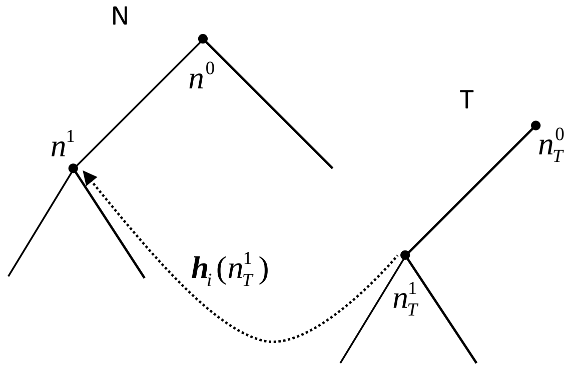


Figure 2: Violation of Property 3

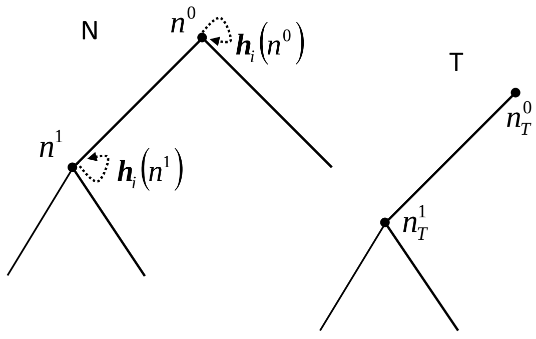


Figure 3: In line with Property 4

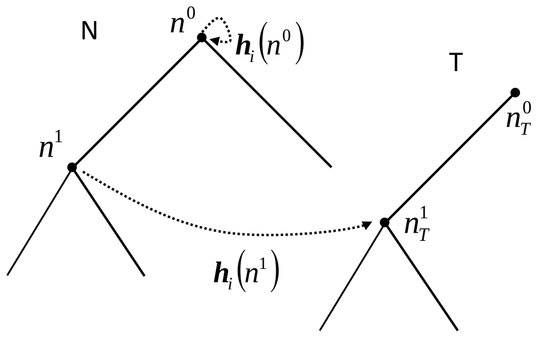


Figure 4: Violation of Property 4

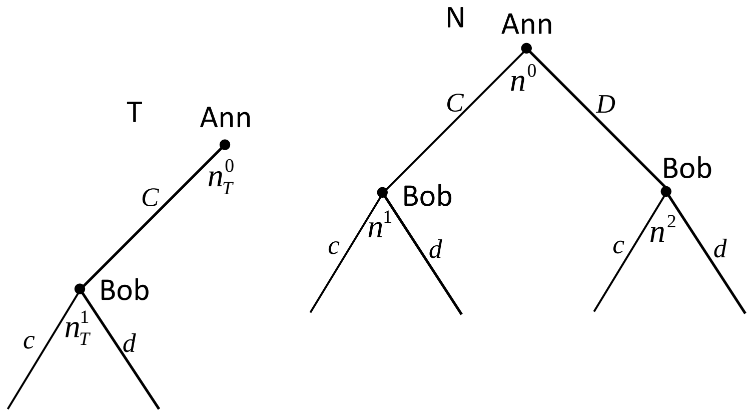
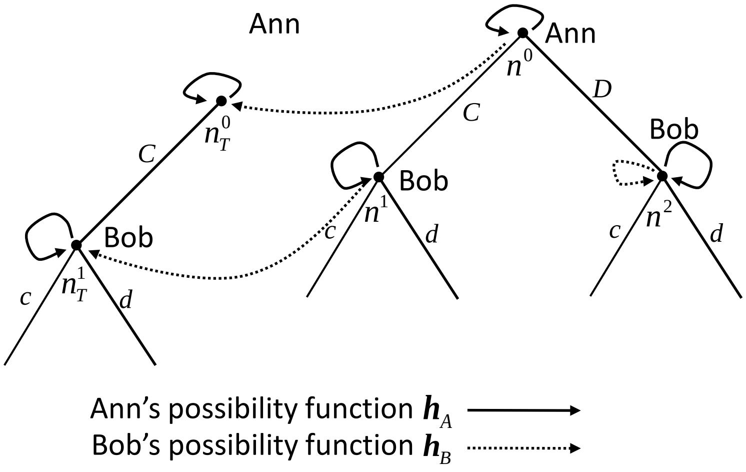


Figure 5: An extensive-form and a subtree thereof

the other hand, *Ann* can only choose *Cooperate* at decision copy n_T^0 and *Bob* can choose between *cooperate* and *defect* at decision copy n_T^1 following *Ann*'s action Cooperate.

Consider now a possible extensive-form with simple unawareness as depicted in Figure 6.[8]

[Figure 6]



Ann
Ann
n^0
C
D
n_T^0
C
Bob
Bob
n^1
Bob
n^2
c
d
c
d
c
n_T^1
d
Ann's possibility function h_A
Bob's possibility function h_B

Figure 6: An extensive-form with subjective views

The possibility functions shown in Figure 6 indicate that at the initial decision node n^0 of the objective extensive-form N, *Ann* is an aware player. She perceives the game to start at decision node $\boldsymbol{h}_A(n^0) = n^0$ and is aware of all augmented histories in H_A. Conversely, at decision node n^0, *Bob* is an unaware player. He perceives the game to start at decision copy $\boldsymbol{h}_B(n^0) = n_T^0$ and is, thus, only aware of augmented histories in H_B^T. At n^0 *Ann* knows that *Bob* is unaware, $\boldsymbol{h}_B \circ \boldsymbol{h}_A(n^0) = n_T^0$, and at n_T^0 *Bob* knows that *Ann* is of the same awareness as him, $\boldsymbol{h}_A \circ \boldsymbol{h}_B(n_T^0) = n_T^0$.

If *Ann* chooses *Cooperate*, then *Bob* stays unaware of the fact that *Ann* could have chosen *Defect*. *Bob* perceives to be at decision copy $\boldsymbol{h}_B(n^1)$ and is, thus, still only aware of the augmented histories in H_B^T. He perceives *Ann* to also be at decision copy $\boldsymbol{h}_A \circ \boldsymbol{h}_B(n_T^1) = n_T^1$ and share his subjective view with regard to the feasible paths of play. If *Ann* instead chooses *Defect*, then *Bob* perceives to be at decision node $\boldsymbol{h}_B(n^2) = n^2$ and is, thus, now aware of the augmented histories in H_B and knows that *Ann* has

[8] Because *Bob*'s augmented histories at decision nodes n_T^0 and n_T^1 are redundant, we omitted these for the sake of simplicity.

always been aware. Moreover, at this decision node *Bob* will be surprised. He realizes that had he observed *Ann* choosing *Cooperate*, then he would not have suspected that she could have chosen anything other than that action.

3 Psychological games with simple unawareness

3.1 Strategies

Let the set of feasible actions for player i at augmented history h_i be denoted by A_{i,h_i}. The structure of our extensive-form with simple unawareness implies that $A_{i,h_i} \subseteq A_{i,n}$. A player's strategy is a complete description of his disposition to act at different augmented histories in H_i. A pure strategy for player i thus specifies an action of player i at each of his augmented histories $h_i \in H_i$. A typical strategy is denoted $s_i = (s_{i,h_i})_{h_i \in H_i}$, where s_{i,h_i} is the action a_i that would be selected by strategy s_i at augmented history h_i. A strategy s_i specifies what player i does at each augmented history, both in the case when h_i maps into an element in the objective extensive-form N, and in the case when h_i maps into an element in the subtree T. Let

$$S_i = \prod_{h_i \in H_i} A_{i,h_i}$$

be the set of player i's strategies and $S = S_i \times S_j$ the set of strategy pairs.

In games with simple unawareness only the objective extensive-form N represents the physical paths of play in the game. The subtree T represents the restricted subjective view of the feasible paths in the mind of an unaware player, or the view of the feasible paths that an aware player assigns to the unaware player. Moreover, as the actual game in the objective extensive-form N evolves a player may become aware of paths of which he was unaware earlier. A strategy can therefore in games with simple unawareness *not* be conceived as an ex ante plan of action. Like (Heifetz et al., 2013a, p. 58), we interpret a strategy s_i of player i as a list of answers to the questions "what would player i do at each augmented history he considers as possible?" This list of answers should be interpreted as follows:

(i) For decision node n in the objective extensive-form N the action $s_{i,\boldsymbol{h}_i(n)}$ should be interpreted as the action that player i *actually* takes at n under strategy s_i, if and when n is reached.

(ii) For decision copy n_T in subtree T the action $s_{i,\boldsymbol{h}_i(n_T)}$ should be interpreted as the action that player i *would* take at n_T if he were unaware. This implies, for example, if player j perceives to be at $\boldsymbol{h}_j(n) = n_T$, then he knows that under strategy s_i player i would take the action $s_{i,\boldsymbol{h}_i \circ \boldsymbol{h}_j(n)}$ if and when n_T is reached.

To illustrate strategies in extensive-forms with simple unawareness we again consider Figure 6. *Ann* is active only after the initial augmented histories $\boldsymbol{h}_A(n^0)$ and $\boldsymbol{h}_A(n_T^0)$, so we can identify each of her strategies with the actions *Cooperate* or *Defect* at $\boldsymbol{h}_A(n^0)$ and *Corperate* at $\boldsymbol{h}_A(n_T^0)$. *Ann*'s set of strategies is thus $S_A = \{CC, DC\}$ (the first action is taken after $\boldsymbol{h}_A(n^0)$ and the second taken after $\boldsymbol{h}_A(n_T^0)$). *Bob* can take one of two possible actions (*cooperate* or *defect*) after each of the two augmented histories $\boldsymbol{h}_B(n^1)$ and $\boldsymbol{h}_B(n^2)$. Thus we identify each of *Bob*'s four strategies by: $S_B = \{cc, cd, dc, dd\}$. The set of strategy pairs in Figure 6 is $S = S_A \times S_B = \{(CC, cc), (CC, cd), \dots, (DC, dd)\}$.

The outcome function $\boldsymbol{z} : S \to Z$, associates each strategy profile $s \in S$ with the induced terminal node $\boldsymbol{z}(s) = (a^1, \dots, a^l)$, and is obtained as follows: at the initial decision node n^0 let $a^1 = (s_{i,\boldsymbol{h}_i(n^0)}, s_{j,\boldsymbol{h}_j(n^0)})$, and for each subsequent decision node n let $a^k = (s_{i,\boldsymbol{h}_i(a^1,\dots,a^{k-1})}, s_{j,\boldsymbol{h}_j(a^1,\dots,a^{k-1})})$ where $1 < k \leq l$. That is, for each strategy profile the outcome function maps out the associated path of actions that the player actually takes throughout the game. For example, if *Ann* chooses *Cooperate* at $\boldsymbol{h}_A(n^0)$, then the induced terminal node can either be given by the paths (C, c) or (C, d) depending on *Bob*'s chosen action $s_{B,\boldsymbol{h}_B(n^1)}$.

Strategy s_i reaches augmented history h_i if there is a strategy s_j of player j such that the path induced by $(s_i, s_j) \in S$ reaches h_i. Otherwise, we say that the augmented history is excluded by the strategy s_i. The set of strategy pairs that reaches h_i is denoted $S(h_i)$. In Figure 6 this implies that, for example, the strategy pairs that reach *Bob*'s augmented history $\boldsymbol{h}_B(n^1)$ are $S(\boldsymbol{h}_B(n^1)) = \{(CC, cc), (CC, cd), (CC, dc), (CC, dd)\}$. The set of player i's strategies that reaches h_i is denoted $S_i(h_i)$, which for *Bob* at $\boldsymbol{h}_B(n^1)$ implies the strategy set $S_B(\boldsymbol{h}_B(n^1)) = \{cc, cd, dc, dd\}$.

For a strategy s_i in the objective extensive-form, we denote by s_i^T the induced strategy in the subtree T. Strategy s_i is said to induce strategy s_i^T if $s_{i,h_i} = s_{i,h_i}^T$ for all $h_i \in H_i^T$. If $R_i \subseteq S_i$ is some set of strategies of player i, denote by R_i^T the set of strategies induced by R_i in the subtree T.

Observe that *Bob* at augmented history $\boldsymbol{h}_B(n^2)$ in Figure 6, after he has become aware, is not deluded to think that the strategic interaction at $\boldsymbol{h}_B(n^1)$ is described by paths of play in subtree T, nor does he think that *Ann* was ever unaware. Rather, *Bob* interprets the action C designated by *Ann*'s strategy $s_A = \{CC\}$ at $\boldsymbol{h}_A \circ \boldsymbol{h}_B(n^0)$ as describing his limited subjective view on *Ann*'s action that would have led to $\boldsymbol{h}_B(n_T^1)$ had she kept him unaware.

For any solution concept we need to analyze what *Ann* thinks of *Bob*'s action at $\boldsymbol{h}_B(n_T^1)$, and such actions are determined by a strategy s_B^T of *Bob* in subtree T. This is why a strategy s_i is defined at all the augmented histories of player i, including those in which he will never actually take an action. The induced strategy s_i^T becomes the object of contemplation and analysis of player j when he is unaware of parts of the actual game.

3.2 Conditional hierarchies of beliefs

Beliefs are modeled as in Battigalli and Dufwenberg (2009), but adapted to extensive-forms with simple unawareness. At an augmented history $h_i \in H_i \backslash H_i^T$, player i holds updated first-order beliefs about player j's disposition to act at augmented histories $h_i \in H_i$. Thus, conditional on each $h_i \in H_i \backslash H_i^T$, player i holds an updated, or revised, belief $\alpha_i(\cdot|h_i) \in \Delta(S_j(h_i))$ about the strategies s_j of player j that reach h_i.

Whereas, at augmented histories $h_i \in H_i^T$, player i's update first-order belief must be confined to player j's disposition to act at augmented histories in H_i^T. Therefore, at augmented histories $h_i \in H_i^T$ player i holds an updated first-order belief $\alpha_i^T(\cdot|h_i) \in \Delta(S_j^T(h_i))$ about the induced strategies s_j^T of player j that reach h_i.

The two systems of first-order beliefs of unaware and aware player i are

$$\alpha_i^T = (\alpha_i^T(\cdot|h_i))_{h_i \in H_i^T} \in \prod_{h_i \in H_i^T} \Delta(S_j^T(h_i))$$

and

$$\alpha_i = (\alpha_i(\cdot|h_i))_{h_i \in H_i \backslash H_i^T} \in \prod_{h_i \in H_i \backslash H_i^T} \Delta(S_j(h_i)),$$

respectively. Let $\boldsymbol{\alpha}_i = (\alpha_i^T, \alpha_i)$ denote player i's overt system of first-order beliefs.

Observe, for example, that *Bob* at $\boldsymbol{h}_B(n^0) = n_T^0$ in Figure 6 holds a conditional first-order belief $\alpha_B^T(\cdot|\boldsymbol{h}_B(n^0)) \in \Delta(S_A^T(\boldsymbol{h}_B(n^0)))$ about *Ann*'s induced strategies in subtree T (which is all of *Ann*'s induced strategies). If *Bob* then subsequently finds himself at $\boldsymbol{h}_B(n^2) = n^2$, then his belief system will change such that he now has a conditional first-order belief $\alpha_B(\cdot|\boldsymbol{h}_B(n^2)) \in \Delta(S_A(\boldsymbol{h}_B(n^2)))$ on *Ann*'s strategies S_A that reach *Bob*'s augmented history $\boldsymbol{h}_B(n^2)$. For example, at $\boldsymbol{h}_B(n^0)$, *Bob*'s conditional first-order belief is on *Ann*'s induced set of strategies $S_A^T(\boldsymbol{h}_B(n^0)) = \{C\}$ that reaches $\boldsymbol{h}_B(n^0)$. If *Bob* finds himself at $\boldsymbol{h}_B(n^2)$, he becomes aware and his first-order belief will be on *Ann*'s set of strategies $S_A(\boldsymbol{h}_B(n^2)) = \{CD\}$ that reaches $\boldsymbol{h}_B(n^2)$. That is, at $\boldsymbol{h}_B(n^2)$, *Bob* knows that *Ann* could have kept him unaware by using induced strategy $s_A^T = \{C\}$, but instead chose to make him aware by choosing $s_A = \{CD\}$.

Player i also holds an updated second-order belief about the first-order belief system of player j, a third-order belief about the second-order beliefs, and so on. For the purpose of this paper, we assume that higher-order beliefs are degenerate point beliefs. Thus, with slight abuse of notation we identify, conditional on each $h_i \in H_i$, $\beta_i(h_i)$ with a particular system of first-order beliefs $(\alpha_j^\lambda(\cdot|h_{ji}))_{h_{ji} \in H_j}$, where $\lambda = \varnothing$ if $h_{ji} \in H_j \backslash H_j^T$ and $\lambda = T$ if $h_{ji} \in H_j^T$. The overt system of second-order beliefs of player i is

$$\boldsymbol{\beta}_i = (\beta_i(h_i))_{h_i \in H_i}.$$

A similar notational convention applies to higher-order beliefs.

At *Ann*'s augmented history $\boldsymbol{h}_A(n^0) = n^0$ in Figure 6, she is aware and has first-order beliefs about *Bob*'s strategies in $S_B(\boldsymbol{h}_A(n^0)) = \{cc, cd, dc, dd\}$ that reach $\boldsymbol{h}_A(n^0)$, and she knows that *Bob* is unaware at $\boldsymbol{h}_B \circ \boldsymbol{h}_A(n^0) \in H_B^T$. Her belief about *Bob*'s first-order belief about her strategies that reach $\boldsymbol{h}_B \circ \boldsymbol{h}_A(n^0)$ must reflect this unawareness. *Ann* thus needs to take *Bob*'s point of view when considering his first-order beliefs about her strategies. To do so, she conditions *Bob*'s first-order beliefs on the composite possibility function $\boldsymbol{h}_B \circ \boldsymbol{h}_A(n^0) = n_T^0$, thereby restricting her second-order beliefs at $\boldsymbol{h}_A(n^0)$ to *Bob*'s first-order beliefs about her strategies in $S_A^T = \{C\}$, and so forth for higher-order beliefs.

Let $\boldsymbol{\mu}_i$ denote the overt (infinite) belief system of player i and $\boldsymbol{M}_i$ the (compact) set of all such beliefs. It is important to remember that at augmented histories H_i^T, the conditional belief system is restricted to first-order beliefs about the induced strategies S_j^T and second-order beliefs about these. Hence, a player at augmented history $h_i \in H_i^T$ does not have first- and second-order beliefs about actions selected by strategies at augmented histories $H_i \backslash H_i^T$. If, however, player i is at an augmented history $h_i \in H_i \backslash H_i^T$, then the conditional belief system is such that the player has first-order beliefs about the strategies S_j and second-order beliefs about (i) strategies S_i if $h_{ji} \in H_j \backslash H_j^T$ and (ii) induced strategies S_i^T if $h_{ji} \in H_j^T$. Players initial beliefs are those held at $\boldsymbol{h}_i(n^0)$ and $\boldsymbol{h}_i(n_T^0)$.

A player should not change his beliefs unless the play reaches an augmented history which falsifies it. We therefore assume that for awareness level $\lambda \in \{\varnothing, T\}$ the overt beliefs $\boldsymbol{\mu}_i$ of player i are consistent such that: (i) there is at least one strategy of player

j in the support of $\alpha_i^{\lambda}(\cdot|h_i)$ at some h_i, and (ii) that beliefs must satisfy Bayes' rule and common certainty of Bayes' rule whenever possible. Consistency of the updating system requires that $\alpha_i^{\lambda}(\cdot|h_i)$, $\beta_i^{\lambda}(h_i)$, and so on, at h_i are consistent with h_i being reached and that no beliefs are abandoned unless falsified. Thus if *Bob*, in Figure 6, finds himself at $\boldsymbol{h}_B(n^2)$ and becomes aware, then he must change his beliefs such that they are consistent at his new awareness level.

3.3 Psychological games with simple unawareness

We are now in a position to formally state the definition of psychological games with simple unawareness:

Definition 3. *A psychological game with simple unawareness* based on the extensive-form with subjective views $\langle\{i,j\},\mathcal{N},\boldsymbol{h}_i,\boldsymbol{h}_j\rangle$ is a structure $\Gamma = \langle\{i,j\},\mathcal{N},\boldsymbol{h}_i,\boldsymbol{h}_j,u_i,u_j\rangle$ where $u_i : Z \times \boldsymbol{M}_i \to \mathbb{R}$ is player i's (measurable and bounded) psychological payoff function.

A psychological game with simple unawareness is obtained from a material payoff game with simple unawareness $\langle\{i,j\},\mathcal{N},\boldsymbol{h}_i,\boldsymbol{h}_j,\pi_i,\pi_j\rangle$, where $\pi_i : Z \to \mathbb{R}$, according to some formula. Following our definition of extensive-form rationalizability in the subsequent section, we will give two examples of specific psychological payoff functions in our model.

4 Extensive-form rationalizability

As a solution concept we use extensive-form rationalizability [Pearce (1984), Battigalli (1997), Battigalli and Siniscalchi (2002)] which embodies forward inductive reasoning. In what follows we extend this definition to psychological games with simple unawareness.

4.1 Sequential Rationality

Our basic behavioral assumption is that player i chooses and carries out a strategy s_i that reaches augmented history h_i and is optimal given his overt belief $\boldsymbol{\mu}_i$, conditional upon any history consistent with s_i. It is thus not required that a strategy specifies behavior at augmented histories that cannot be reached by s_i.

Fix an overt belief $\boldsymbol{\mu}_i$, an augmented history h_i, and a strategy s_i. The expectation of the psychological payoff u_i, given s_i and $\boldsymbol{\mu}_i$ is

$$\mathbb{E}_{s_i,\boldsymbol{\mu}_i}[u_i|h_i] := \sum_{s_j \in S_j} \alpha_i^{\lambda}(s_j^{\lambda}|h_i) \times u_i(\boldsymbol{z}(s_i, s_j), \boldsymbol{\mu}_i), \tag{1}$$

where $\lambda = \{\varnothing\}$ if $h_i \in H_i \backslash H_i^T$ and $\lambda = T$ if $h_i \in H_i^T$. At an augmented history in the subtree, a player's expected psychological payoff is thus restricted to the part of the strategy that describes the actions of an unaware player.

Definition 4. Fix an overt belief $\boldsymbol{\mu}_i \in \boldsymbol{M}_i$. Strategy s_i is a *sequential best response* to $\boldsymbol{\mu}_i$ if for all $h_i \in H_i$

$$s_i \in \arg\max_{s_i \in S_i(h_i)} \mathbb{E}_{s_i,\boldsymbol{\mu_i}}[u_i|h_i].$$

Similar for player j.

For any overt belief $\boldsymbol{\mu}_i$, let $\mathrm{BR}(\boldsymbol{\mu}_i)$ denote the set of strategies s_i that are sequential best responses to $\boldsymbol{\mu}_i$ in accordance with Definition 4. The set of best responses thus consists of strategies s_i of player i that, for a given $\boldsymbol{\mu}_i$, are undominated at every augmented history $h_i \in H_i$ given his awareness level λ at h_i.

For example, for *Bob*'s strategy $s_B = \{cd\}$ to be a sequential best response in Figure 6, it must be that it is undominated by his strategies $\hat{s}_B \in \{cc, dc, dd\}$ at every $h_B \in H_B$, given his awareness at h_B. The augmented history at which *Bob* is passive is omitted in the following. At $\boldsymbol{h}_B(n^1) = n_T^1$, *Bob* is unaware and knows that *Ann*'s strategy is $s_A^T = \{C\}$. Strategy $s_B = \{cd\}$ undominated if its induced strategy $s_B^T = \{c\}$ gives him an equal or higher expected psychological payoff at $\boldsymbol{h}_B(n^1)$ than induced strategy $\hat{s}_B^T = \{d\}$, giventhat *Ann* chooses $s_A = \{CC\}$. At $\boldsymbol{h}_B(n^2) = n^2$ *Bob* has observed that *Ann* has chosen to make him aware by choosing $s_A = \{CD\}$. He therefore, at $\boldsymbol{h}_B(n^2)$, believes with certainty that *Ann*'s strategy is $s_A = \{CD\}$. *Bob* considers strategy $s_B = \{cd\}$ undominated if it gives him an equal or higher expected psychological payoff at $\boldsymbol{h}_B(n^2)$ than any of the strategies $s_B \in \{cc, dc, dd\}$, given he observes that *Ann* has chosen $s_A = \{CD\}$. Thus, *Bob*'s strategy $s_B = \{cd\}$ is a sequential best response if, and only if, it is undominated at both $\boldsymbol{h}_B(n^1)$ and $\boldsymbol{h}_B(n^2)$.

Clearly, BR_i is nonempty valued. The conditional first-order belief $\alpha_i^{\lambda}(\cdot|h_i)$ is continuous. Since u_i is also continuous, we have that $\alpha_i^{\lambda}(\cdot|h_i)u_i(\cdot)$ is continuous, which implies that BR_i is an upper hemicontinuous correspondence.

4.2 Best-Rationalization Principle

The best-rationalization principle [Battigalli (1997), Battigalli and Siniscalchi (2002)] requires that players' beliefs conditional upon observing augmented history h_i be consistent with the highest degree of *strategic sophistication* of other players. Our analysis clarifies what is meant by strategic sophistication in terms of psychological games with unawareness.

Definition 5. Consider the following *extensive-form rationalization* procedure for player i:

$$\boldsymbol{M}_i[1] = \boldsymbol{M}_i,$$

$$R_i[1] = \begin{cases} s_i \in S_i & \text{such that there exists an overt belief } \boldsymbol{\mu}_i \in \boldsymbol{M}_i[1] \\ & \text{for which } s_i \in \mathrm{BR}(\boldsymbol{\mu}_i). \end{cases}$$

$$\vdots$$

$$\boldsymbol{M}_i[k] = \begin{cases} \boldsymbol{\mu}_i \in \boldsymbol{M}_i[k-1] & \text{such that for all augmented histories } h_i \in H_i, \text{ if} \\ & R_j[k-1](h_i) \neq \varnothing, \text{ then } \alpha_i^{\lambda}(R_j^{\lambda}[k-1]|h_i) = 1, \text{ where} \\ & \lambda = \{\varnothing\} \text{ if } h_i \in H_i \backslash H_i^T \text{ and } \lambda = T \text{ if } h_i \in H_i^T. \end{cases}$$

$$R_i[k] = \begin{cases} s_i \in S_i & \text{such that there exists an overt belief } \boldsymbol{\mu}_i \in \boldsymbol{M}_i[k] \\ & \text{for which } s_i \in \mathrm{BR}(\boldsymbol{\mu}_i). \end{cases}$$

Let $\mathrm{R}_i[\infty] = \bigcap_{k\geq 0} \mathrm{R}_i[k]$. Player i's strategies in $\mathrm{R}_i[\infty]$ are said to be extensive-form (correlated) rationalizable in a psychological game with simple unawareness. Similar for player j.

We start our definition of extensive-form rationalizability at the level of strategic thinking of player i, whose *first level* rationalizable strategies are sequential best responses to some nonrestricted overt belief. Strategies that are not sequential best responses to any nonrestricted overt belief are not first level rationalizable. Next, player i restricts his overt beliefs to those for which he, at each augmented history h_i, is certain (given the awareness level at h_i) of those first level rationalizable strategies of player j that reach h_i. Player i then chooses *second level* rationalizable strategies that are sequential best responses to these restricted overt beliefs. Strategies that are not sequential best responses to these restricted overt beliefs, on player j's first level rationalizable strategies, are not second level rationalizable. Furthermore, player i must restrict his restricted overt beliefs to those for which he, at each augmented history h_i, is certain (given the awareness level at h_i) of those second level rationalizable strategies of player j that reach h_i. Player i then chooses *third level* rationalizable strategies that are sequential best responses to these restricted-restricted overt beliefs. Strategies that are not sequential best responses to these restricted-restricted overt beliefs, on player j's second level rationalizable strategies, are not third level rationalizable, and so on.

Remark 1. $\{\mathrm{R}_i[k] : k \geq 0\}$ is a weakly decreasing sequence, that is, $\mathrm{R}_i[k+1] \subseteq \mathrm{R}_i[k]$ for all k. Since R_i is a closed set (because the correspondences BR_i is upper hemicontinuous), the sequence converges in countably many steps. The limit is given by the first integer K such that $\mathrm{R}_i[K] = \mathrm{R}_i[K+1]$. Similar for player j.

Definition 5 and Remark 1 can be interpreted as follows. Consider the limit set $\mathrm{R}_i[K]$ of player i. The sequence $\mathrm{R}_j[0], \mathrm{R}_j[1], \ldots, \mathrm{R}_j[K-1]$ represents a hierarchy of increasingly strong hypotheses of player i about the behavior of player j. When player i implements a strategy $s_i \in \mathrm{R}_i[K]$, he always optimize accordingly. At the beginning of the game, it is the common belief that all players update and behave in this way.

The set $\boldsymbol{M}_k$ (for $k > 1$) implies that along each feasible path of play, at an augmented history an active player is certain that the other player sequential best responds, certain that the other player is certain he sequential best responds, and so on. If a player finds himself at some succeeding augmented history, where the other player's strategies that could lead to that augmented history are inconsistent with the player's previous certainty in the other player's best response, then the player seeks a *best rationalization* which could have led to that augmented history. That is, if the player is "surprised" by the other player's unexpected action, and cannot use Bayesian updating, then he forms new beliefs that justify this observed inconsistency. In its simplest form, forward-induction reasoning involves the assumption that, upon observing an unexpected (but undominated) action of the other player, a player maintains the working hypothesis that the latter is a sequential best response. The best-rationalization principle captures precisely this type of argument.

Forward-induction reasoning implies that at $\boldsymbol{h}_B(n_T^0)$ and onwards, unaware *Bob* is certain that *Ann*'s sequential best response is $s_A^T = \{CC\}$. However, if augmented history $\boldsymbol{h}_B(n^2)$ is reached and *Bob* becomes aware, then he is certain given his newly found awareness that *Ann*'s action $s_A = \{CD\}$ is a sequential best response to some overt belief of hers. At $\boldsymbol{h}_B(n^2)$, *Bob* has no choice but to revert to being certain that

Ann would not choose the strategy $s_A = \{CC\}$ rationally, and excludes *Ann*'s overt beliefs for which $s_A = \{CC\}$ is a sequential best response.

Using extensive-form rationalizability as a solution concept highlights the need to define strategies by the actions taken not only at decision nodes, which represent actual paths of play, but also at decision copies. However, at each decision node, the augmented history h_i that is considered possible could be in the subtree describing a restrictive view of the feasible paths of play, and at that point the player can only perceive his strategy in terms of these paths. Furthermore, players can only rank the other player's strategies according to their perhaps restricted awareness. An aware player must, for example, consider an unaware player's strategies in terms of how he perceives the game, that is, in the subtree which represents an unaware players' subjective view.

5 Guilt Aversion and Reciprocity

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 (2007), for example, implies that player i judges the initial expectations of player j concerning his material payoff and feels guilty whenever he does not live up to these expectations. More formally, consider a psychological game with simple unawareness as defined above. Given his strategy s_j and the overt first-order belief system $\boldsymbol{\alpha}_j$, player j forms an initial expectation about his material payoff π_j:

$$\mathbb{E}_{s_j,\boldsymbol{\alpha}_j}[\pi_j|\boldsymbol{h}_j(n^0)] = \sum_{s_i} \alpha_j^{\lambda}(s_i^{\lambda}|\boldsymbol{h}_j(n^0))\pi_j(\boldsymbol{z}(s_j,s_i)),$$

with $\lambda = \{\varnothing\}$ if $\boldsymbol{h}_j(n^0) \in H_j \backslash H_j^T$ (i.e. if player j is initially aware) and $\lambda = T$ if $\boldsymbol{h}_j(n^0) \in H_j^T$ (i.e. if player j is initially unaware). For every terminal node z the function

$$D_j(z,s_j,\boldsymbol{\alpha}_j) = max\{0, E_{s_j,\boldsymbol{\alpha}_j}[\pi_j|\boldsymbol{h}_j(n^0)] - \pi_j(z)\}$$

measures how much player j is let down relative to his initial expectation. Of course, player i does not know player j's strategy and first-order beliefs, but holds a belief about these. Denote player i's belief about player j's let-down by $D_{ij}(z,s_j,\boldsymbol{\beta}_i)$, where $\boldsymbol{\beta}_i$ is player i's overt second-order belief. Given this, player i is motivated by simply guilt if he has belief-dependent preferences represented by a utility function of the following form:

$$u_i(\boldsymbol{z}(s_i, s_j), \boldsymbol{\mu}_i) = \pi_i(z) - \theta_{ij} D_{ij}(z, s_j, \boldsymbol{\beta}_i) \tag{2}$$

where θ_{ij} is player i's sensitivity to guilt. A guilt averse player i tries to maximize the expected value of equation 2 (see equation 1 in section 4.1). He senses a psychological cost connected to his feeling of guilt in case he does not live up to his belief about player j's expectation and takes this into account when deciding on his optimal behavior. Different to Battigalli and Dufwenberg (2007), players' feelings of guilt in our setting with simple unawareness depend on their awareness level and the awareness level of the other player. If, for example, player j is unaware of certain feasible paths of play, player i's belief about how much he let's down player j takes player j's unawareness into account.

Different from guilt aversion, reciprocity assumes that players judge the kindness of others. Whenever player i judges player j to be kind, he reciprocates by being kind himself. Whenever player i judges player j to be unkind, he acts unkindly in return (see e.g. Dufwenberg and Kirchsteiger (2004)). More formally, given an augmented history h_j, a strategy s_j that reaches h_j and the overt first-order belief system $\boldsymbol{\alpha}_j$, player j forms an expectation about player i's material payoff π_i:

$$\mathbb{E}_{s_j, \boldsymbol{\alpha}_j}[\pi_i | h_j] = \sum_{s_i} \alpha_j^{\lambda}(s_i^{\lambda} | h_j) \pi_i(\boldsymbol{z}(s_i, s_j))$$

with $\lambda = \{\varnothing\}$ if $h_j \in H_j \backslash H_j^T$ and $\lambda = T$ if $h_j \in H_j^T$. Player j's kindness towards player i is described by how much player j expects to give player i relative to some equitable payoff $\pi_i^e(h_j)$ at augmented history h_j:

$$K_{ji}(h_j) = \mathbb{E}_{s_j, \boldsymbol{\alpha}_j}[\pi_i | h_j] - \pi_i^e(h_j).$$

The equitable payoff is the threshold or neutral payoff above (below) which player j treats player i kindly (unkindly). In other words, if $K_{ji}(\cdot) > 0$, then player j treats player i kindly. Conversely, if $K_{ji}(\cdot) < 0$, then player j treats player i unkindly. Let the equitable payoff be

$$\pi_i^e(h_j) = \frac{1}{2} \times \left[\max_{s_j} \left(E_{s_j, \boldsymbol{\alpha}_j}[\pi_i | h_j] \right) + \min_{s_j} \left(E_{s_j, \boldsymbol{\alpha}_j}[\pi_i | h_j] \right) \right].$$

The equitable payoff is the average player j is able to give to player i in material terms based on his awareness and first-order belief. Of course, as in the case of guilt aversion, player i does not know strategy s_j and first-order beliefs $\boldsymbol{\alpha}_j$, but holds a first- and second-order belief about them. Denote player i's judgment of player j's kindness at augmented history h_i by $K_{iji}(h_i) = \mathbb{E}_{s_j,\boldsymbol{\beta}_i}[\pi_i|h_{ji}] - \pi_i^e(h_{ji})$, where $\boldsymbol{\beta}_i$ is player i's overt second-order belief. We say player i is motivated by reciprocity if he has belief-dependent preferences represented by a utility function of the form:

$$u_i(z, \boldsymbol{\mu}_i|h_i) = \pi_i(z) + Y_i \times K_{iji}(h_i) \times \pi_j(z), \tag{3}$$

where $Y_i > 0$ is player i's sensitivity to reciprocity. A reciprocal player i tries to maximize the expected value of equation 3. Whenever player i perceives player j to be kind, player i is motivated to also maximize player j's material payoff. In case player i judges player j to be unkind, player i is motivated to reduce player j's material payoff. This definition of reciprocity implies that if player i is unaware of certain feasible paths of play, he judges the kindness of player j based on the feasible paths that he is aware of.

With unawareness the judgments of players regarding the intentions and expectations of others, and hence, the influence of these on the players' behavior crucially depend on the awareness of players, the awareness players attribute to others, the awareness players belief other attribute to them, and so on. In the following section we apply these belief-dependent preferences in two examples highlighting the role of unawareness in the interaction of agents motivated by reciprocity and guilt aversion.

6 Two examples

In the following we present two examples to highlight the impact and importance of simple unawareness in the strategic interaction of players with belief-dependent preferences. In particular, our examples demonstrate that the strategic behavior of players motivated by belief-dependent preferences crucially depends on their awareness concerning the strategic environment they are in. As a consequence, players' awareness levels are an important and integral part of the strategic environment. 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 is beyond the scope of this paper. Therefore, we limit the analysis to specific awareness scenarios.

A sequential prisoners dilemma with reciprocity: Consider the following awareness scenario already depicted in Figure 6 with *Ann*'s and *Bob*'s material payoffs added:[9]

[Figure 7]

In the strategic setting depicted in Figure 7, *Ann* is initially aware of everything, whereas *Bob* is initially unaware. *Ann* knows this, and knows that he only becomes aware of everything if she chooses *Defect.* Also, *Ann* knows that *Bob* perceive her to be unaware.

We assume that *Bob* is motivated by belief-dependent reciprocity and *Ann* is selfish. *Bob*'s psychological utility is thus described by equations 3, while *Ann*'s is equal to her monetary payoff ($u_A(z) = \pi_A(z)$). *Ann*'s optimal strategy depends on her first-order belief about *Bob* strategy (conditional on her behavior). Her strategy $s_A = \{CC\}$ is a sequential best response as long as her expected material payoff from strategy $s_A = \{CC\}$ exceeds her expected material payoff from strategy $s_A = \{CD\}$. Her expected material payoff from strategy $s_A = \{CC\}$ is:

$$\mathbb{E}_{\{CC\},\boldsymbol{\alpha}_A}[\pi_A|\boldsymbol{h}_A(n^0)] = \alpha_A(\{c\cdot\})|\boldsymbol{h}_A(n^0))(1) + (1 - \alpha_A(\{c\cdot\}|\boldsymbol{h}_A(n^0)))(-1),$$

where $\alpha_A(\{c\cdot\}|\boldsymbol{h}_A(n^0)) \coloneqq \sum_{s_B \in \{cc,cd\}} \alpha_A(s_B|\boldsymbol{h}_A(n^0))$ is a shorthand notation for *Ann*'s first-order belief about the strategies of *Bob* that select the action *cooperate* at augmented history $\boldsymbol{h}_B(n^1) = n^1_T$. *Ann*'s expected material payoff from following strategy $s_A = \{CD\}$ is:

$$\mathbb{E}_{\{CD\},\boldsymbol{\alpha}_A}[\pi_A|\boldsymbol{h}_A(n^0)] = \alpha_A(\{\cdot c\}|\boldsymbol{h}_A(n^0))(2) + (1 - \alpha_A(\{\cdot c\}|\boldsymbol{h}_A(n^0)))(0),$$

where $\alpha_A(\{\cdot c\}|\boldsymbol{h}_A(n^0)) \coloneqq \sum_{s_B \in \{cc,dc\}} \alpha_A(s_B|\boldsymbol{h}_A(n^0))$ is an akin shorthand notation. The first level rationalizable strategies of *Ann* are thus

$$R_A[1] = \{s_A : \alpha_A(\{c\cdot\}|\boldsymbol{h}_A(n^0)) - \alpha_A(\{\cdot c\}|\boldsymbol{h}_A(n^0)) \geq \frac{1}{2} \Rightarrow s_A = \{CC\},$$
$$\text{otherwise} \Rightarrow s_A = \{CD\}\}.$$

Bob, on the other hand, is initially passive and only becomes active at augmented histories $\boldsymbol{h}_B(n^1) = n^1_T$ and $\boldsymbol{h}_B(n^2) = n^2$. Remember that at augmented history $\boldsymbol{h}_B(n^1)$, *Bob* is unaware and holds a first-order belief $\alpha^T_B(s^T_A|\boldsymbol{h}_B(n^1))$ and a second-order point belief $\beta_B(\boldsymbol{h}_B(n^1))$ concerning *Ann*'s restricted first-order belief $\alpha^T_A(s^T_B|\boldsymbol{h}_B(n^0_T))$. Given his unawareness at $\boldsymbol{h}_B(n^1)$, *Bob* thinks that *Ann*'s only action is *Cooperate.* Independent

[9] For the sake of clarity we only depict augmented histories h_i for *Ann* and *Bob* that are relevant for solving the game.

of his second-order belief and his sensitivity to reciprocity Y_B he thus judges *Ann* as neither kind nor unkind ($K_{BAB}(\boldsymbol{h}_B(n^1)) = 0$). Consequently, whenever *Bob* finds himself at $\boldsymbol{h}_B(n^1)$ he will simply maximize his own material payoff by choosing *defect*.

At augmented history $\boldsymbol{h}_B(n^2) = n^2$, *Bob* is aware of everything and knows that he could have earned a monetary payoff of 2 had *Ann* initially chosen to keep him unaware by choosing *Cooperate*. *Bob*'s judgment of *Ann*'s intention towards him at $\boldsymbol{h}_B(n^2)$ is

$$\mathbb{E}_{s_A,\boldsymbol{\beta}_B}[\pi_B|\boldsymbol{h}_B(n^2)] = \beta_B(\{\cdot c\}|\boldsymbol{h}_B(n^2)) \cdot (-1) + (1 - \beta_B(\{\cdot c\}|\boldsymbol{h}_B(n^2))) \cdot (0),$$

where $\beta_B(\{\cdot c\}|\boldsymbol{h}_B(n^2)) \coloneqq \sum_{s_B \in \{cc,dc\}} \beta_B(\boldsymbol{h}_B(n^2))$ is a shorthand notation for *Bob*'s second-order belief about the likelihood with which *Ann* believes he chooses *cooperate* at augmented history $\boldsymbol{h}_B(n^2)$. Clearly, $2 > \mathbb{E}_{s_A,\boldsymbol{\beta}_B}[\pi_B|\boldsymbol{h}_B(n^2)]$ independent of second-order belief $\beta_B(\boldsymbol{h}_B(n^2))$. Hence, *Bob* judges *Ann* as unkind when finding himself at $\boldsymbol{h}_B(n^2)$. The first level rationalizable strategies of *Bob* are thus

$$R_B[1] = \{s_B : \text{ for all } \boldsymbol{\beta}_B, Y_B \Rightarrow s_B = \{dd\}\}.$$

Although *Bob* is motivated by belief-dependent reciprocity as defined in equation 3, his behavior in our sequential prisoners dilemma with unawareness is independent of his (second-order) beliefs and independent of his sensitivity to reciprocity Y_B.

Bob's set of first level rationalizable strategies is a singleton set $R_B[1] = \{dd\}$. *Ann* is thus at all her augmented histories certain that *Bob* follows strategy $s_B = \{dd\}$. That is, *Ann*'s overt beliefs $\boldsymbol{\mu}_A \in \boldsymbol{M}_A[2]$ are all such that $\alpha_A(\{dd\}|\cdot) = 1$. Being certain that *Bob* chooses *defect* no matter what, *Ann*'s sequential best response strategy must also select Defect as an action (since $0 > -1$). The second level rationalizable strategies of *Ann* are thus

$$R_A[2] = \{s_A : \text{ for all } \boldsymbol{\mu}_A \in \boldsymbol{M}_A[2] \Rightarrow s_A = \{CD\}\}.$$

Ann anticipates that given his awareness, *Bob* judges her as unkind and chooses *defect* independent of what she does. Consequently, since *Ann* is only interested in her own material payoff, she chooses *Defect* herself to get a material payoff of 0 instead of −1.

To study the impact of unawareness, we compare this awareness scenario to the rationalizable solution of the sequential prisoners dilemma with reciprocity and full awareness (see figure).

[Figure 8]

Now *Ann* chooses to *Cooperate* in the initial augmented history $\boldsymbol{h}_A(n^0) = n^0$ as long as she believes sufficiently strongly that *Bob* will *cooperate* given that she chooses strategy $s_A = \{CC\}$. Her expected payoff from choosing either $s_A = \{CC\}$ or $s_A = \{CD\}$ at the augmented history $\boldsymbol{h}_A(n^0)$ is the same as before, and her first level rationalizable strategies are again:

$$R_A[1] = \{s_A : \alpha_A(\{c\cdot\}|\boldsymbol{h}_A(n^0)) - \alpha_A(\{\cdot c\}|\boldsymbol{h}_A(n^0)) \geq \frac{1}{2} \Rightarrow s_A = \{CC\},$$
$$\text{otherwise } \Rightarrow s_A = \{CD\}\}.$$

Bob's optimal behavior at $\boldsymbol{h}_B(n^2) = n^2$ remains the same as before. That is, *Bob* chooses *defect* out of monetary and reciprocal reasons. However, *Bob*'s optimal behavior at $\boldsymbol{h}_B(n^1) = n^1$ now depends on his sensitivity to reciprocity Y_B. Let *Bob*'s (second-order) belief about *Ann*'s belief concerning the likelihood with which he chooses *cooperate* at $\boldsymbol{h}_B(n^1)$, be $\beta_B(\{c\cdot\}|\boldsymbol{h}_B(n^1)) \coloneqq \sum_{s_B \in \{cc,cd\}} \beta_B(\boldsymbol{h}_B(n^1))$. The first level rationalizable strategies of *Bob* are thus:

$$R_B[1] = \{s_B : \beta_B(\{c\cdot\}|\boldsymbol{h}_B(n^1)) \leq 2 - \frac{1}{Y_B} \Rightarrow s_B = \{cd\}, \tag{4}$$
$$\text{otherwise } \Rightarrow s_B = \{dd\}\}.$$

The lower *Bob*'s second-order belief is, the kinder he perceives *Ann*'s strategy $s_A = \{CC\}$ ($K_{BAB} = 1 - \frac{1}{2} \cdot \beta_B(\{c\cdot\}|\boldsymbol{h}_B(n^1))$), which provides him with a payoff which is higher than if she had chosen strategy $s_A = \{CD\}$. At $\boldsymbol{h}_B(n^1)$, *Bob* never actually thinks *Ann* is unkind. The question at this augmented history simply is whether he thinks she is kind enough, given his sensitivity to reciprocity Y_B, such that he prefers to reciprocate her kindness.

If *Bob*'s sensitivity to reciprocity is low ($Y_B \leq \frac{1}{2}$), such that he for sure chooses *defect* if *Ann* chooses *Cooperate*, then *Ann* chooses strategy $s_A = \{CD\}$ as this provides her with a higher expected material payoff. Conversely, if *Bob*'s sensitivity to reciprocity is high ($Y_B \geq 1$), *Ann* is certain that *Bob* chooses *cooperate* if she chooses *Cooperate*. Given this, she chooses $s_A = \{CC\}$ as this provides her with a higher expected material payoff. Notice, if *Bob* is sensitive enough to *Ann*'s kindness, then he chooses *cooperate* at $\boldsymbol{h}_B(n^1)$ independent of his second-order belief. Given this *Ann* also chooses *Cooperate*, something she would not do were she sure that *Bob* would be unaware. Based on $R_B[1]$,

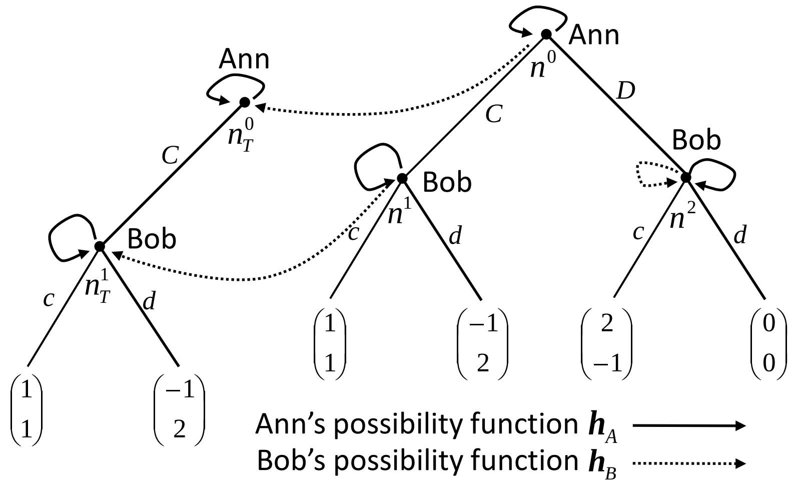


Figure 7: Sequential Prisoners Dilemma with unawareness

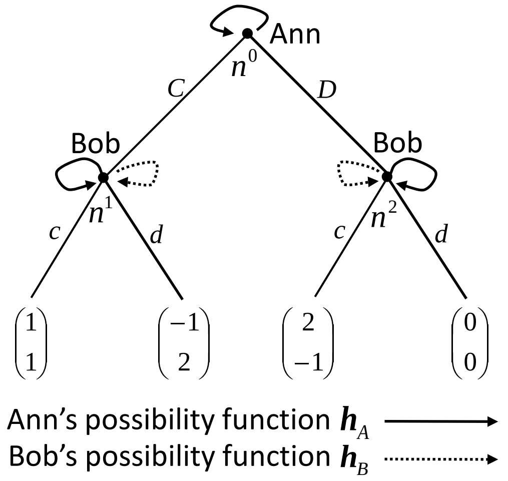

Ann
n^0
C
D
Bob
Bob
n^1
n^2
c
d
c
d
(1, 1)
(−1, 2)
(2, −1)
(0, 0)
Ann's possibility function h_A
Bob's possibility function h_B

Figure 8: Sequential Prisoners Dilemma with full awareness

Ann's overt beliefs are

$$\boldsymbol{M}_A[2] = \{\boldsymbol{\mu}_A : \text{for all } Y_B \geq 1 \Rightarrow \alpha_A(\{c\cdot\}|\boldsymbol{h}_A(n^0)) = 1, \\ \text{for all } Y_B < \frac{1}{2} \Rightarrow \alpha_A(\{c\cdot\}|\boldsymbol{h}_A(n^0)) = 0\}.$$

At the second level of reasoning, *Ann* is certain that a very sensitive *Bob* will always *cooperate* if she also cooperates, whereas a very insensitive *Bob* will always *defect* no matter what. For intermediate sensitivity levels, $\frac{1}{2} \leq Y_B < 1$, her overt beliefs are $\boldsymbol{M}_A[2] = \boldsymbol{M}_A[1]$. *Ann*'s second level rationalizable strategies are thus:

$$R_A[2] = \{s_A : \text{ for } Y_B \geq 1 \Rightarrow s_A = \{CC\}, \\ \text{for } \frac{1}{2} \leq Y_B < 1, \gamma_A(\{c\cdot\}|\boldsymbol{h}_A(n^0)) \leq 2 - \frac{1}{Y_B} \Rightarrow s_A = \{CC\}, \\ \text{otherwise } \Rightarrow s_A = \{CD\}\}.$$

where γ_A denotes *Ann*'s (third-order) point belief about *Bob*'s second-order belief. Based on $R_A[2]$, *Bob*'s second level overt beliefs are

$$\boldsymbol{M}_B[3] = \{\boldsymbol{\mu}_B : \text{ for } Y_B \geq 1 \Rightarrow \beta_B(\{c\cdot\}|\boldsymbol{h}_b(n^1)) = 1, \\ \text{for } Y_B < \frac{1}{2} \Rightarrow \beta_B(\{c\cdot\}|\boldsymbol{h}_B(n^1)) = 0, \\ \text{otherwise } \Rightarrow \delta_B(\{c\cdot\}|\boldsymbol{h}_B(n^1)) \leq 2 - \frac{1}{Y_B}, \beta_B(\{c\cdot\}|\boldsymbol{h}_B(n^1)) = 1\}.$$

where δ_B denotes *Bob*'s (fourth-order) belief about *Ann*'s third-order beliefs. $\boldsymbol{M}_B[3]$ implies that *Bob* believes that *Ann* expects him to choose $s_B = \{cd\}$ whenever he finds himself at augmented history $\boldsymbol{h}_B(n^1)$. *Bob*'s third-level rationalizable strategies are thus:

$$R_B[3] = \{s_B : \text{ for } Y_B \geq 1 \Rightarrow s_B = \{cd\}, \\ \text{otherwise } Y_B < 1 \Rightarrow s_B = \{dd\}\}.$$

If *Bob* is sensitive enough to reciprocity ($Y_B \geq 1$), he always chooses *cooperate* if *Ann* chooses *Cooperate*, and chooses *defect* if *Ann* chooses *Defect.* If he is insensitive enough to reciprocity ($Y_B < \frac{1}{2}$), then he always chooses *defect* no matter what *Ann* does. However, if his sensitivity to reciprocity is $\frac{1}{2} \leq Y_B < 1$ and he finds himself at $\boldsymbol{h}_B(n^1)$, then he is certain that *Ann* believes that he will *cooperate* ($\beta_B(\{c\cdot\}|\boldsymbol{h}_B(n^1)) = 1$). That is, $1 > 2 - \frac{1}{Y_B}$ and he defects although he believes that *Ann* believes that he will *cooperate*. Finally, *Ann*'s fourth-level rationalizable strategies are:

$$\begin{aligned} R_A[4] = \{s_A : \ & \text{for all } Y_B \geq 1 \Rightarrow s_A = \{CC\}, \\ & \text{for all } Y_B < 1 \Rightarrow s_B = \{CD\}\} \end{aligned}$$

With full awareness, if *Bob* is sufficiently sensitive to reciprocity ($Y_B \geq 1$), then *Ann* chooses to *Cooperate* since this induces *Bob* to choose *cooperate* as well. This stands in contrast to the result of the first awareness scenario in which *Bob* was unaware of *Ann*'s possibility to defect even after her choice *Cooperate*. This example highlights that, although belief-dependent reciprocity is only based on first- and second-order beliefs, the recursive nature of extensive-form rationalizability requires the specification of higher (potentially infinite) orders of beliefs.

In synthesis: although *Bob*'s sensitivity to reciprocity might be very high, his behavior in the first awareness scenario stands in contrast to the result with full awareness. With full awareness *Bob*'s behavior following *Ann*'s action *Cooperate* depends on *Bob*'s sensitivity to reciprocity. For sufficiently high levels of sensitivity *Bob* reciprocates by choosing *cooperate*. With unawareness as in the previous scenario *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 behavior is qualified. She defects as well. Interestingly, not controlling for his awareness, *Bob* behaves *as if* he is selfish, although he is not. It is only his subjective perception concerning the strategic environment which drives his optimal behavior in our sequential prisoners dilemma with simple unawareness.

It is 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 decision to *Cooperate* independent of whether he is only aware of augmented histories in subtree T (as in the first scenario) or everything (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 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* (see figure):[10]

[Figure 9]

Assume *Bob* is motivated by simple guilt aversion as described by equation 2, and that *Ann* is only interested in her own material payoff. If there is full awareness, Battigalli and Dufwenberg (2009, p. 21) demonstrate that the rationalizable solution is such that *Ann* chooses *Trust* and *Bob* chooses *Share*.

Less formally than before: With unawareness as depicted in Figure 9, *Bob* knows that *Ann* is not 'let down', if he chooses *Grab* following her decision to *Trust*. That is, 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*. Hence, *Bob* does not feel any guilt towards *Ann* from choosing *Grab* following *Trust* ($D_{BA} = 0$). Of course, *Ann* (correctly) anticipates that *Bob* chooses *Grab* following *Trust* and thus chooses *Don't* in the augmented history $\boldsymbol{h}_A(n^0) = n^0_T$ she initially perceives to be in.

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. In particular, it 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 9. He would like to make her aware in order to signal to her that they could *Share*. Of course, if he were to do this the analysis would mirror Battigalli and Dufwenberg (2009, p. 21)'s analysis. Interestingly, were *Bob* only interested in his own monetary payoff, he would not be concerned about *Ann* being or not being aware of his option *Share*, as Ann would in any case choose *Don't* anticipating his selfish behavior.

7 Conclusion

We have analyzed the influence and importance of unawareness concerning feasible paths of play for the strategic interaction of players in psychological games, and defined a two-

[10] Again, for the sake of clarity we only depict the function $\boldsymbol{h}_i$ for *Ann* and *Bob* in the non-terminal histories relevant for solving the game.

player model in which players are motivated by belief-dependent preferences and simple unawareness of certain feasible paths of play. Using this model we 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 (2007).

Our examples demonstrate that the strategic behavior of players motivated by belief-dependent preferences crucially depends on their awareness concerning the strategic environment they are in, their perception concerning the awareness of others, their perception concerning the perception of others etc.

In other words, unawareness has a profound and intuitive impact on the strategic interaction of players with belief-dependent psychological preferences - a fact that creates both an opportunity as well as a challenge to empirically investigations analyzing the strength and nature of belief-dependent preferences.

Concentrating on two-player environments and simple awareness scenarios obviously puts limits to the strategic situations that can be analyzed with our model. Nevertheless our simple model has allowed us to uncover intriguing effects. More general strategic environments with more complex awareness scenarios are left for future research.

8 Acknowledgements

We would like to thank Geir Asheim, Pierpaolo Battigalli, Mie la Cour Sonne, 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), the Seminar in Microeconomics 2011 (Lund), the CNEE Workshop 2011 (Copenhagen) and the CES seminar 2011 (Leuven) for helpful comments and suggestions. All errors are our own.

References

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

Battigalli, P., 1997. On rationalizability in extensive games. Journal of Economic Theory 74 (1), 40–61.

Battigalli, P., Dufwenberg, M., 2007. Guilt in games. The American Economic Review, Papers and Proceedings 97 (2), 170–176.

Battigalli, P., Dufwenberg, M., 2009. Dynamic psychological games. Journal of Economic Theory 144, 1–35.

Battigalli, P., Siniscalchi, M., 2002. Strong belief and forward induction reasoning. Journal of Economic Theory 106 (2), 356–391.

Bellemare, C., Sebald, A., Strobel, M., 2010. Measuring the willingness to pay to avoid guilt: Estimation using equilibrium and stated belief models. Journal of Applied Econometrics, forthcomming.

Broberg, T., Ellingsen, T., Johannesson, M., 2007. Is generosity involuntary? Economics Letters 94 (1), 32–37.

Charness, G., Dufwenberg, M., 2006. Promises and partnership. Econometrica 74 (6), 1579–1601.

Dana, J., Cain, D., 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.

Dana, J., Weber, R., Kuang, J., 2007. Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. Economic Theory 33 (1), 67–80.

Dekel, E., Lipman, B., Rustichini, A., 1998. Standard state-space models preclude unawareness. Econometrica 66, 159–173.

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

Fagin, R., Halpern, J., 1988. Beliefs, awareness, and limited reasoning. Artificial Intelligence 34, 39–76.

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

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

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

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

Feinberg, Y., 2004. Subjective reasoninggames with unawareness. Stanford University Research Paper (1875).

Feinberg, Y., 2011. Games with unawareness. Mimeo.

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

Grant, S., Quiggin, J., 2012. Inductive reasoning about unawareness. Risk and Uncertainty Working Papers.

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

Halpern, J., Rêgo, L., 2006. Extensive games with possibly unaware players. In: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems. ACM, pp. 744–751.

Halpern, J., Rêgo, L., 2008. Interactive unawareness revisited. Games and Economic Behavior 62, 323–262.

Halpern, J., Rêgo, L., 2009. Reasoning about knowledge of unawareness. Games and Economic Behavior 67 (2), 503–525.

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

Heifetz, A., Meier, M., Schipper, B., 2008. A cannonical model for interactive unawareness. Games and Economic Behavior 62, 304–324.

Heifetz, A., Meier, M., Schipper, B., 2013a. Dynamic unawareness and rationalizable behavior. Games and Economic Behavior 81, 50–68.

Heifetz, A., Meier, M., Schipper, B., 2013b. Unawareness, beliefs, and speculative trade. Games and Economic Behavior 77, 100–121.

Lazear, E., Malmendier, U., Weber, R., 2009. Sorting and social preferences. Mimeo, 97.

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

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

Osborne, M., Rubinstein, A., 1994. A course in game theory. MIT press.

Pearce, D., 1984. Rationalizable strategic behavior and the problem of perfection. Econometrica, 1029–1050.

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

Sebald, A., 2010. Attribution and reciprocity. Games and Economic Behavior 68 (1), 339–352.

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

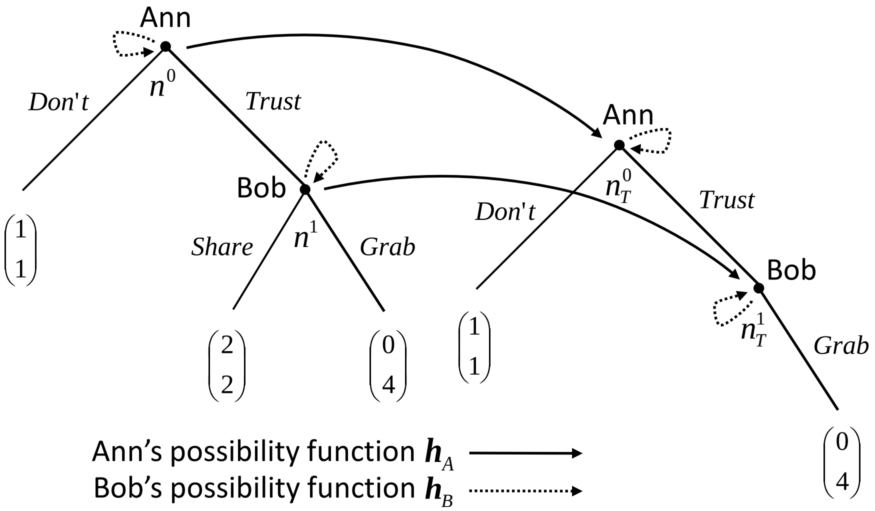


Figure 9: Trust game with unawareness