Forward Induction in a Backward Inductive Manner[*]

Martin Meier[†]and Andrés Perea[‡]

This version: January 2025

Abstract

We propose a new rationalizability concept for dynamic games with imperfect information, *forward and backward rationalizability,* that combines elements from forward and backward induction reasoning. It proceeds by applying the forward induction concept of *strong rationalizability* (also known as *extensive-form rationalizability*) in a backward inductive fashion. We argue that, compared to strong rationalizability, the new concept provides a more compelling theory for how players react to surprises. Moreover, we provide an epistemic characterization of the new concept, and show that (a) it always exists, (b) in terms of outcomes it is equivalent to strong rationalizability, (c) in terms of strategies it is a refinement of the pure backward induction concepts of *backward dominance* and *backwards rationalizability*, and (d) it satisfies *expansion monotonicity*: if a player learns that the game was actually preceded by some moves he was initially unaware of, then this new information will only refine, but never completely overthrow, his reasoning. Strong rationalizability violates this principle.

JEL Classification: C72

Keywords: Forward induction, backward induction, forward and backward rationalizability, strong rationalizability, backwards rationalizability, backward dominance, expansion monotonicity, dynamic games

1 Introduction

The main feature that distinguishes dynamic games from static games is the issue of *belief revision.* That is, how does a player revise his belief upon observing a surprising move by an opponent? Suppose you play a dynamic game, and you find yourself called to play at an information set that you initially believed not ever to be crossed, since rationality of each player and common certainty of rationality would prevent this set from being reached. Should you now rather believe that the choices of the other player are a true reflection of her reasoning (and hence this player is either not rational or believes that you are not rational) or should you believe that as far as reasoning and planning of strategies are concerned this other player is fully rational, but made a mistake when she implemented her strategies? Forward induction is in line with the first mode of reasoning, while backward induction is in line with the second.

[*] We thank Pierpaolo Battigalli, Adam Brandenburger, Emiliano Catonini, Satoshi Fukuda, Aviad Heifetz, David Pearce, Klaus Ritzberger and some anonymous referees for valuable comments. In addition, we are grateful to some anonymous referees at LOFT 2022 for their useful feedback. We also thank the audiences at the workshop "10 Years of Epicenter", the conference LOFT 2022, the seminar at Royal Holloway University of London, the seminar at NYU Abu Dhabi, and the SAET Conference 2023 for their comments and suggestions.

[†] *E-mail*: mmeier@unibz.it *Affiliation*: Free University of Bozen-Bolzano, Faculty of Economics and Management, Italy.

[‡] *E-mail*: a.perea@maastrichtuniversity.nl *Affiliation*: Maastricht University, EpiCenter and Department of Quantitative Economics, The Netherlands.

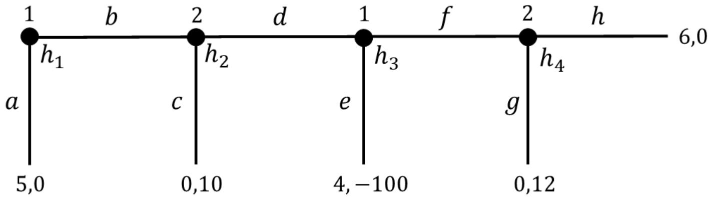

Figure 1: Strong rationalizability may lead to counterintuitive behavior

The older idea is backward induction, which dates at least back to von Neumann and Morgenstern (1953)[1], and has also been incorporated in concepts like *backwards rationalizability* (Penta (2015), Perea (2014), Catonini and Penta (2022)), *backward dominance* (Perea (2014)), *subgame perfect equilibrium* (Selten (1965)), *extensive-form perfect equilibrium* (Selten (1975)), *sequential equilibrium* (Kreps and Wilson (1982)) and *quasi-perfect equilibrium* (van Damme (1984)). According to these concepts, upon reaching an unexpected information set a player is free to believe that this is due to mistakes by his opponents in executing their planned strategies. Hence, a player is not required to try to learn from the past, even if doing so could refine his beliefs about the opponents' unobserved past, present and future moves.

In a sense, the forward induction concept of *strong rationalizability* (Pearce (1984), Battigalli (1997)), also known as *extensive-form rationalizability,* takes the other extreme by excluding, whenever possible, mistakes in the execution of planned strategies. However, to make this possible, a player may need to ascribe beliefs to the opponents that assume only a limited degree of rationality of their respective opponents. The example in Figure 1, which is a variant of Figure 3 in Reny (1992), will illustrate this.

Upon reaching h_2, player 2 is forced to believe that player 1 chooses strategy (b, f), as this is the only way for player 1 to get at least 5 – a payoff he could guarantee by choosing a. At the same time, player 2 must believe that player 1 ascribes a high probability to player 2 behaving irrationally at h_4. The unique best reply for player 2 is to choose strategy (d, g). However, if player 1 in fact believes that player 2 will choose rationally at h_4, and chooses rationally himself in the remainder of the game, then player 1 would choose e at h_3, yielding an extremely low payoff for player 2. Also this type of reasoning is therefore not free of problems.

Overall, we thus see some shortcomings with both backward and forward induction reasoning as discussed above. In response, we propose a rationalizability concept for dynamic games – different from both strong rationalizability and pure backward induction reasoning – which we call *forward and backward rationalizability.* In this new concept, we require players to learn from the past, but only as much as is consistent

[1] It is often claimed that backward induction first appeared in Zermelo (1913) in the proof of his famous theorem on chess. However, Zermelo did not assume a stopping rule for chess, and hence the game he considered did not have a finite horizon. Therefore, he could not use backward induction.

with “fully rational behavior” in the future.

More formally, the concept proceeds by applying the forward induction concept of strong rationalizability in a backward inductive fashion: We start by applying strong rationalizability to the last period in the game, which results in restrictions on strategies and beliefs there. Taking these restrictions as given, we then apply the strong rationalizability procedure to the parts of the game that start at the penultimate period, and so on, until we reach the beginning of the game.

We then ask: What does the new concept of forward and backward rationalizability entail in terms of reasoning? The epistemic characterization of strong rationalizability as given by Battigalli and Siniscalchi (2002) relies on *strong belief in rationality,* which means that a player, whenever possible, should believe that his opponents are choosing rationally in the whole game. On top of this, they require that a player must also believe, whenever possible, that his opponents choose rationally in the whole game while strongly believing in the other players' rationality. Iterating this argument leads to *rationality and common strong belief in rationality* – a set of epistemic conditions that characterizes extensive-form rationalizability.

In comparison, it is shown in Theorem 4.1 that our concept can be characterized epistemically by (a) first imposing common strong belief in rationality from the last period onwards, (b) then imposing common strong belief in rationality from the penultimate period onwards, keeping the restrictions from (a), and so on, until we reach the beginning of the game.

Intuitively, the reasoning process of a player choosing forward and backward rationalizable strategies can thus be described as follows: “Yesterday I believed that my opponents are rational from then (yesterday) on, believed that everybody else believed that everybody (else) is rational from then on, and so forth. If this is not falsified by today's observations, I should continue to believe not only that everybody else is rational from today on, and so forth, but also that everybody else *was rational from yesterday on*, and so forth. The same applies to the day before yesterday, the day before that, and so on.”

We next investigate how the new concept relates to existing concepts, such as strong rationalizability and backwards rationalizability. In the example of Figure 1, for instance, the new concept coincides with backward induction. However, there are other games where the concept is different, in terms of strategies, from both pure backward induction reasoning and strong rationalizability. Consider, for instance, the game in Figure 2.

Strong rationalizability would reason as follows: At h_2 player 2 must believe that player 1 chooses strategy (a, e), as this is the only strategy reaching h_2 that would yield player 1 at least 2 – a payoff he could guarantee by choosing b at the beginning. But then, player 2 would choose d, and player 1, anticipating this, would choose b.

Our concept of forward and backward rationalizability proceeds differently: In the last subgame, at h_3, it imposes no restrictions. Now consider the subgame starting at h_2, which is the classical Battle of the Sexes game with an outside option for player 2. Our concept uniquely selects the forward induction strategies (c, h) and f in this subgame. Finally, we turn to the whole game. Given the earlier restrictions, player 1 must believe that player 2 will choose (c, h), and therefore will choose b himself. In particular, it predicts that player 2 will choose (c, h) and not d, as strong rationalizability predicts.

Our new concept thus yields a different strategy for player 2 than strong rationalizability, but it induces the same outcome – player 1 choosing b at the beginning. In Theorem 6.3 we show that this is no coincidence:

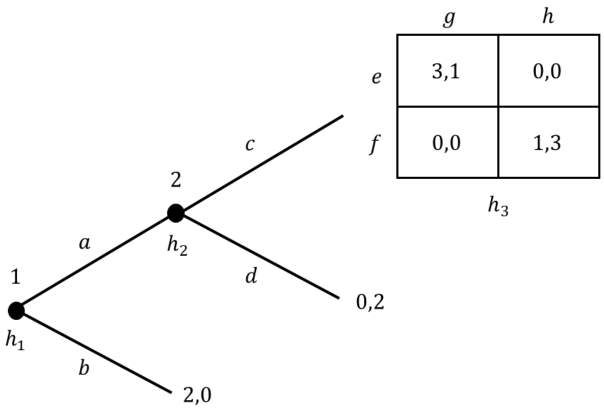

Figure 2: Battle of the sexes with double outside option

The two concepts will always yield the same outcomes, but may differ in terms of strategies.

When compared to the pure backward induction concept of backwards rationalizability, our concept always yields a refinement of the latter – not only in terms of outcomes but also in terms of strategies. See Theorem 5.1. In the game of Figure 2, for instance, backwards rationalizability yields the same strategies for player 1 as our concept, which is only strategy b, but allows for strategies (c, h) and d for player 2 whereas our concept only allows for (c, h). In fact, the new concept can be viewed as a backward looking strengthening of the purely forward looking concept of backwards rationalizability: On top of always believing that the opponent will choose rationally in the future, which is what backwards rationalizability entails, we require a player to also explain (some of) the opponent's *past* choices whenever possible.

Although our concept is equivalent to strong rationalizability in terms of outcomes, we believe that strategies matter. Surprises and mistakes *do* happen regularly in reality, and therefore we find it important to critically analyze how players react to surprises. Indeed, a satisfactory theory of rational behavior should also describe how rational players act and reason upon observing surprising, or even irrational, behavior by their opponents. Different from strong rationalizability, under forward and backward rationalizability a player, upon observing a surprising move by his opponent, will never believe that an opponent will choose irrationally in the future.

An important immediate consequence of the two results mentioned above is that in every dynamic game with imperfect information, every strongly rationalizable outcome is also induced by some profile of backwards rationalizable strategies. Catonini (2020) and Perea (2017) have already proven this result for dynamic games with observed past choices, but we show that this property even holds for games with imperfectly observed past choices. This may be viewed as a generalization of Battigalli's theorem, which states that in every dynamic game with perfect information and without relevant ties, the unique strongly rationalizable outcome is the backward induction outcome.

The paper finally investigates how players reason if they learn that the game was actually preceded by some earlier moves they were initially unaware of. Traditionally, we analyze a game by assuming that all players know that this is the game being played, and we may be interested in the players' behavior at "surprising" subgames, which were initially not expected to be reached. But instead of assuming that a player finds himself in a "surprising" subgame, it may also happen that a player initially views this subgame as the "whole game", and then learns that this game was actually preceded by some past moves.[2] If this happens, this could be a reason for the player to *refine* his reasoning, but, so we argue, it should never be a reason to completely *overthrow* the reasoning he did before. After all, the player reasoned himself towards a set of possible beliefs and strategies based on the accurate description of the game from now on, and the fact that the game actually started earlier does not change the accuracy of this description. If the game was in fact preceded by some past moves, this additional information should only lead to sharper predictions, but not to new beliefs and strategies which were considered irrational before receiving this new piece of information. Indeed, it is always possible that the moves preceding the game were chosen by mistake. This principle, that new information about past moves should only lead players to refine their reasoning, but not to overthrow their reasoning, is called *expansion monotonicity.*

Compare, for instance, Nash equilibrium and subgame perfect equilibrium in dynamic games. Nash equilibria in dynamic games that are not subgame perfect involve non-credible threats that are inconsistent with Nash equilibria in unreached subgames. Hence, Nash equilibrium violates expansion monotonicity, while subgame perfect equilibrium satisfies it.

As is easily seen, the game of Figure 1 shows that strong rationalizability violates expansion monotonicity. If the game were to start at h_2, then strong rationalizability would uniquely select the strategy c for player 2. However, if player 2 learns that the game in fact started at h_1, then strong rationalizability would uniquely select strategy (d, g) for player 2, and therefore the previous reasoning of player 2 would be completely overthrown by this new information. In contrast, the pure backward induction concept of backwards rationalizability *does* satisfy expansion monotonicity.

We show in Theorem 8.1 that also the new concept of forward and backward rationalizability always satisfies expansion monotonicity. Consider, for instance, the game from Figure 2. If the game were to start at h_3, our concept selects both e and f for player 1. However, if player 1 learns that the game started at h_2, this additional information will refine the set of possible choices for player 1 to only f.

The paper is organized as follows: In Section 2 we lay out the basic definitions and notation for dynamic games. In Section 3 we formally define the strong rationalizability procedure, the backwards rationalizability procedure and the new forward and backward rationalizability procedure, and prove the existence of the latter concept. In Section 4 we epistemically characterize forward and backward rationalizability. In Section 5 we show that the new concept constitutes a refinement, in terms of strategies, of backwards rationalizability, whereas we prove in Section 6 that in terms of outcomes it is equivalent to strong rationalizabity. In Section 7 we show how these results imply a generalization of Battigalli's theorem, by proving that in every

[2] This would be a special instance of an extensive-form game with unawareness. See, for example, Heifetz, Meier and Schipper (2013).

dynamic game with possibly imperfect information, all strongly rationalizable outcomes are also backwards rationalizable. In Section 8 we discuss the principle of expansion monotonicity. In Section 9 we provide some concluding discussions. The appendix contains all the proofs.

2 Preliminaries

In this section we introduce our model of dynamic games and define derived objects like strategies, conditional belief vectors, strong belief, and optimal choice at information sets.

2.1 Dynamic Games

In this paper we consider *finite dynamic games* that allow for *simultaneous moves* and *imperfect information*, and where every action and history can be *indexed by time.* We assume that every player, at each of his information sets, *knows the time.* Formally, a *dynamic game* is a tuple $\Gamma = (I, P, I^a, (A_i, H_i)_{i \in I}, Z, (u_i)_{i \in I})$, where

(a) I is the finite set of *players*;

(b) P is the finite set of *past action profiles*, or *histories;*

(c) the mapping I^a assigns to every history $p \in P$ the (possibly empty) set of *active players* $I^a(p) \subseteq I$ who must choose after history p. If $I^a(p)$ contains more than one player, there are simultaneous moves after p. We say that p is a *terminal history,* or an *outcome,* if $I^a(p) = \emptyset$, and p is called a *non-terminal history* otherwise. By P_i we denote the set of histories $p \in P$ with $i \in I^a(p)$;

(d) for every player i, the mapping A_i assigns to every history $p \in P_i$ the finite set of *actions* $A_i(p)$ from which player i can choose after history p. By $m_0 \in \mathbb{Z}$ we denote the first period of the game[3], whereas p_0 denotes the unique period m_0 history, marking the beginning of the game. For every $m \geq m_0 + 1$, the period m histories can then inductively be defined as the pairs $p' = (p, (a_i)_{i \in I^a(p)})$ where p is a non-terminal period $(m-1)$ history, and such that for every $i \in I^a(p)$ we have that $a_i \in A_i(p)$. By m_+ we denote the last period that contains a non-terminal history. We assume that the objects P, I^a and $(A_i)_{i \in I}$ are such that the histories in P are precisely those that are period m histories for some m. We say that a history p *precedes* a history p' (or p' *follows* p) if p' results by adding some action profiles after p;

(e) for every player i there is a partition H_i of the set of histories P_i where i is active. Every partition element $h_i \in H_i$ is called an *information set* for player i. In case h_i contains more than one history, the interpretation is that player i does not know at h_i which history in h_i has been realized. For every $h_i \in H_i$ there is a period m such that all histories in h_i are period m histories. In that case, h_i is called a period m information set. By H_i^m we denote the collection of period m information sets for player i. The objects A_i and H_i must be such that for every information set $h_i \in H_i$ and every two histories p, p' in h_i, we have that $A_i(p) = A_i(p')$. We can thus write $A_i(h_i)$ for the unique set of available actions at h_i. Moreover, it must be that $A_i(h_i) \cap A_i(h_i') = \emptyset$ for every two distinct information sets $h_i, h_i' \in H_i$. By $H := \cup_{i \in I} H_i$ we denote the collection of all information sets, whereas $H^m := \cup_{i \in I} H_i^m$ is the collection of all period m information sets;

[3] The reason we do not require $m_0 = 0$ is that in Section 8 we consider embedding Γ into a larger game that starts before Γ.

(f) $Z \subseteq P$ is the collection of *terminal histories* or *outcomes*;
(g) for every player i there is a utility function $u_i : Z \to \mathbb{R}$.

The dynamic game satisfies *perfect recall* (Kuhn (1953)) if every player always remembers which actions he chose in the past, and which information he had about the opponents' past actions. Formally, for every player i, information set $h_i \in H_i$, and histories $p, p' \in h_i$, the sequence of player i actions in p and p' must be the same (and consequently, the collection of player i information sets that precede p and p' must be the same). In the sequel we will always assume that the dynamic game under consideration satisfies perfect recall.

For every two information sets $h, h' \in H$, we say that h *precedes* h' (or h' *follows* h) if there is a history $p \in h$ and a history $p' \in h'$ such that p precedes p'.

2.2 Strategies

A *complete strategy* $\tilde{s}_i$ for player i assigns to every information set $h_i \in H_i$ an available action $\tilde{s}_i(h) \in A_i(h)$. Let $\tilde{S}_i$ be the set of complete strategies for player i, and $\tilde{S}_{-i} := \times_{j \neq i} \tilde{S}_j$ the set of opponents' complete strategy combinations. Every combination of complete strategies $(\tilde{s}_i)_{i \in I}$ induces an outcome $z((\tilde{s}_i)_{i \in I}) \in Z$. By

$$H_i(\tilde{s}_i) := \{h_i \in H_i \mid \text{there is } \tilde{s}_{-i} \in \tilde{S}_{-i} \text{ such that } z(\tilde{s}_i, \tilde{s}_{-i}) \text{ follows a history in } h_i\}$$

we denote the collection of player i information sets that can be reached by $\tilde{s}_i$. By $r_i(\tilde{s}_i)$ we denote the restriction of $\tilde{s}_i$ to information sets in $H_i(\tilde{s}_i)$, and it is called the *reduced strategy* induced by $\tilde{s}_i$. By $S_i := r_i(\tilde{S}_i)$ we denote the set of *reduced strategies* for player i. In the sequel, when we say *strategy* we always mean a reduced strategy. Every combination of strategies $(s_i)_{i \in I}$ reaches an outcome $z((s_i)_{i \in I}) \in Z$.

For a given player i and information set $h \in H$, we define the sets

$$\begin{aligned} S(h) &: = \{s \in \times_{i \in I} S_i \mid z(s) \text{ follows a history in } h\}, \\ S_i(h) &: = \{s_i \in S_i \mid \text{there is some } s_{-i} \in S_{-i} \text{ such that } (s_i, s_{-i}) \in S(h)\}, \text{ and} \\ S_{-i}(h) &: = \{s_{-i} \in S_{-i} \mid \text{there is some } s_i \in S_i \text{ such that } (s_i, s_{-i}) \in S(h)\}. \end{aligned}$$

Intuively, $S_i(h)$ is the set of strategies for player i that allow for information set h to be reached, whereas $S_{-i}(h)$ is the set of opponents' strategy combinations that allow for h to be reached. By perfect recall it holds, for every player i and every information set $h \in H_i$, that $S(h) = S_i(h) \times S_{-i}(h)$. For a given strategy $s_i \in S_i$ we define $H_i(s_i) := \{h \in H_i \mid s_i \in S_i(h)\}$.

2.3 Conditional Belief Vectors and Strong Belief

For a finite set X, let $\Delta(X)$ be the set of probability measures on X. A *conditional belief vector* for player i is a mapping b_i that assigns to every information set $h \in H_i$ a probabilistic belief $b_i(h) \in \Delta(S_{-i}(h))$. Moreover, a conditional belief vector b_i is required to satisfy *forward consistency*[4] (often called *Bayesian*

[4] We adopt this terminology from Battigalli, Catonini and Manili (2023).

updating). That is, for every $h, h' \in H_i$ where h precedes h' and $b_i(h)(S_{-i}(h')) > 0$ we have that

$$b_i(h')(s_{-i}) = \frac{b_i(h)(s_{-i})}{b_i(h)(S_{-i}(h'))}$$

for every $s_{-i} \in S_{-i}(h')$.[5] Let B_i be the set of all forward consistent conditional belief vectors for player i.

For a given player i, consider a set of opponents' strategy combinations $D_{-i} \subseteq S_{-i}$. Say that a conditional belief vector $b_i \in B_i$ *strongly believes* D_{-i} if for every information set $h \in H_i$ with $S_{-i}(h) \cap D_{-i} \neq \emptyset$ we have that $b_i(h)(D_{-i}) = 1$.

2.4 Optimal Choice at Information Sets

Recall that every strategy combination $s = (s_i)_{i \in I}$ induces a terminal history $z(s)$. For a strategy s_i, an information set $h \in H_i(s_i)$ and a conditional belief vector b_i, let

$$u_i(s_i, b_i(h)) := \sum_{s_{-i} \in S_{-i}(h)} b_i(h)(s_{-i}) \cdot u_i(z(s_i, s_{-i}))$$

be the *expected utility* induced by strategy s_i at information h under the conditional belief vector b_i. A strategy s_i is *optimal* for b_i at $h \in H_i(s_i)$ if

$$u_i(s_i, b_i(h)) \geq u_i(s_i', b_i(h)) \text{ for all } s_i' \in S_i(h).$$

For a given period m, strategy s_i is said to be *optimal* for b_i *from period m onwards* if for every period $\tau \geq m$ and every information set $h \in H_i(s_i) \cap H_i^\tau$, strategy s_i is optimal for b_i at h.

Note that if a strategy s_i does not allow any information set in H^τ with $\tau \geq m$ to be reached then, by definition, s_i is (vacuously) optimal from period m onwards for every $b_i \in B_i$. It can be shown that the following is true:

Remark 2.1 *For every conditional belief vector $b_i \in B_i$, every period m and every information set $h \in H^m$, there is a strategy $s_i \in S_i(h)$ that is optimal for b_i from period m onwards.*

3 Definition of the Concepts

In this section we recall the concepts of *strong rationalizability* and *backwards rationalizability,* and introduce the new concept of *forward and backward rationalizability.*

[5] By abuse of notation, we write $b_i(h)(s_{-i})$ instead of $b_i(h)(\{s_{-i}\})$.

3.1 Strong Rationalizability

The *strong rationalizability* procedure (Pearce (1984), Battigalli (1997)) is a forward induction procedure that recursively eliminates strategies and conditional belief vectors for every player. The main idea is that a player, whenever possible, must believe that his opponents are implementing strategies that are optimal from the beginning (that is, from period m_0 onwards).

Definition 3.1 (Strong rationalizability) ***Round*** 0***:*** *For every player i, set $B_i^0 := B_i$ and $S_i^0 := S_i$.*

Round $k \geq 1$***:*** *For every player i, set*

$$B_i^k := \{b_i \in B_i^{k-1} \mid b_i \text{ strongly believes } S_{-i}^{k-1}\}$$

and

$$S_i^k := \{s_i \in S_i^{k-1} \mid s_i \text{ is optimal for some } b_i \in B_i^k \text{ from period } m_0 \text{ onwards}\}.$$

Strategy $s_i \in S_i$ is ***strongly rationalizable*** *if $s_i \in S_i^k$ for all $k \geq 0$. Conditional belief vector b_i is* ***strongly rationalizable*** *if $b_i \in B_i^k$ for all $k \geq 0$.*

3.2 Backwards Rationalizability

The concept of *backwards rationalizability* (Perea (2014), Penta (2015), Catonini and Penta (2022)) is purely forward looking, and can thus be viewed as a genuine backward induction concept. This can also be seen from the fact that the resulting strategies can be characterized by *common belief in future rationality* (Perea (2014)), stating that a player always believes that his opponents will choose rationally now and in the future, always believes that his opponents always believe that the other players will choose rationally now and in the future, and so on. In the definition below, recall that m_+ is the last period that contains a non-terminal history.

Definition 3.2 (Backwards rationalizability) ***Period*** m_+, ***round*** 0. *Set $S_i^{m_+.0} := S_i$ and $B_i^{m_+.0} := B_i$ for all players i.*

Period m_+, ***round*** $k \geq 1$. *For every player i, set*

$$B_i^{m_+.k} := \{b_i \in B_i^{m_+.k-1} \mid b_i(h)(S_{-i}^{m_+.k-1}) = 1 \text{ for all } h \in H_i^{m_+}\},$$

and

$$S_i^{m_+.k} := \{s_i \in S_i^{m_+.k-1} \mid s_i \text{ is optimal for some } b_i \in B_i^{m_+.k} \text{ from period } m_+ \text{ onwards}\}.$$

Continue until $B_i^{m_+.K} = B_i^{m_+.K+1}$ and $S_i^{m_+.K} = S_i^{m_+.K+1}$ for some round K and all players i, and set $B_i^{m_+} := B_i^{m_+.K}$ and $S_i^{m_+} := S_i^{m_+.K}$.[6]

[6] Note that such a round K can always be found. Indeed, the set of strategies is finite, and there will thus be some K with $S_i^{m_+.K+1} = S_i^{m_+.K}$ for all players i. But then, by construction, $B_i^{m_+.K+2} = B_i^{m_+K+1}$ and $S_i^{m_+.K+2} = S_i^{m_+.K+1}$ for all players i.

Period $m \leq m_+ - 1$, *round* 0. *Set* $S_i^{m.0} := S_i^{m+1}$ *and* $B_i^{m.0} := B_i^{m+1}$ *for all players* i.

Period $m \leq m_+ - 1$, *round* $k \geq 1$. *For every player* i, *let*

$$B_i^{m.k} := \{b_i \in B_i^{m.k-1} \mid b_i(h)(S_{-i}^{m.k-1}) = 1 \text{ for all } h \in H_i^m\},$$

and

$$S_i^{m.k} := \{s_i \in S_i^{m.k-1} \mid s_i \text{ is optimal for some } b_i \in B_i^{m.k} \text{ from period } m \text{ onwards}\}.$$

Continue until $B_i^{m.K} = B_i^{m.K+1}$ *and* $S_i^{m.K} = S_i^{m.K+1}$ *for some round* K *and all players* i, *and set* $B_i^m := B_i^{m.K}$ *and* $S_i^m := S_i^{m.K}$.

A strategy s_i *is* ***backwards rationalizable*** *if* $s_i \in S_i^{m_0}$, *and a conditional belief vector* b_i *is* ***backwards rationalizable*** *if* $b_i \in B_i^{m_0}$.

The definition we have provided here uses the backwards order of elimination. That is, we start by eliminating at the ultimate period, subsequently we do the eliminations at the penultimate period, and so on, until we reach the beginning of the game. Perea (2014)'s definition is different, as in every round it (potentially) performs eliminations at each of the information sets in each of the periods. However, it is argued in Section 6.3 of Perea (2014) that the definition we provide here is equivalent, in terms of strategies and conditional belief vectors selected, to the definition in Perea (2014).

3.3 Forward and Backward Rationalizability

The new concept of *forward and backward rationalizability* combines elements from the strong rationalizability procedure and the backwards rationalizability procedure. Like the backwards rationalizability procedure, it proceeds in a backward inductive fashion by first performing eliminations at the ultimate period, and then proceeding backwards until we reach the beginning of the game. However, when we reach a certain period m in this way then, in line with strong rationalizability, we also require players at later periods to reason about the opponents' past moves at period m. This is fundamentally different from backwards rationalizability, where players at a given period are only required to reason about the opponents' moves at this period and *future* periods.

Definition 3.3 (Forward and backward rationalizability) ***Period*** m_+, ***round*** 0. *Set* $S_i^{m_+.0} := S_i$ *and* $B_i^{m_+.0} := B_i$ *for all players* i.

Period m_+, ***round*** $k \geq 1$. *For every player* i, *set*

$$B_i^{m_+.k} := \{b_i \in B_i^{m_+.k-1} \mid b_i \text{ strongly believes } S_{-i}^{m_+.k-1}\},$$

and

$$S_i^{m_+.k} := \{s_i \in S_i^{m_+.k-1} \mid s_i \text{ is optimal for some } b_i \in B_i^{m_+.k} \text{ from period } m_+ \text{ onwards}\}.$$

Continue until $B_i^{m_+.K} = B_i^{m_+.K+1}$ *and* $S_i^{m_+.K} = S_i^{m_+.K+1}$ *for some round* K*and all players* i, *and set* $B_i^{m_+} := B_i^{m_+.K}$ *and* $S_i^{m_+} := S_i^{m_+.K}$.

Period $m \leq m_+ - 1$, **round** 0. *Set* $S_i^{m.0} := S_i^{m+1}$ *and* $B_i^{m.0} := B_i^{m+1}$ *for all players* i.

Period $m \leq m_+ - 1$, **round** $k \geq 1$. *For every player* i, *let*

$$B_i^{m.k} := \{b_i \in B_i^{m.k-1} \mid b_i \textit{ strongly believes } S_{-i}^{m.k-1}\},$$

and

$$S_i^{m.k} := \{s_i \in S_i^{m.k-1} \mid s_i \textit{ is optimal for some } b_i \in B_i^{m.k} \textit{ from period } m \textit{ onwards}\}.$$

Continue until $B_i^{m.K} = B_i^{m.K+1}$ *and* $S_i^{m.K} = S_i^{m.K+1}$ *for some round* K *and all players* i, *and set* $B_i^m := B_i^{m.K}$ *and* $S_i^m := S_i^{m.K}$.

A strategy s_i *is* ***forward and backward rationalizable*** *if* $s_i \in S_i^{m_0}$, *and a conditional belief vector* b_i *is* ***forward and backward rationalizable*** *if* $b_i \in B_i^{m_0}$.

For a given player i and information set h in period m, let $\tau \leq m$ be the earliest period such that $S_i^{\tau.k} \cap S_i(h) \neq \emptyset$ for some round $k \geq 0$. For this period τ, let l be the highest such round with $S_i^{\tau.l} \cap S_i(h) \neq \emptyset$.[7] Then, we denote by $S_i^{fbr}(h) := S_i^{\tau.l} \cap S_i(h)$ the set of strategies that is predicted conditional on reaching information set h.

Note that according to our concept, a player $j \neq i$ with $h \in H_j$ may believe at information set h that player i chooses a strategy which is not in $S_i^{fbr}(h)$ in case there are more than two players. Indeed, it may be that $S_i^{fbr}(h) = S_i^{\tau.l} \cap S_i(h)$, whereas for a third player $n \neq i, j$ we have that $S_n^{fbr}(h) = S_n^{\tau'.l'} \cap S_n(h)$, where either $\tau' > \tau$ or ($\tau' = \tau$ and $l' < l$). In that case, a conditional belief vector b_j that is forward and backward rationalizable may at h assign positive probability to opponents' strategy combinations containing some $s_i \in S_i^{\tau'.l'} \backslash S_i^{\tau.l}$.

By construction, our procedure will refine the conditional beliefs of player i at some information set $h \in H_i^m$ until we reach a period $\tau < m$ where $S_{-i}^{\tau.k} \cap S_{-i}(h)$ becomes empty for some $k \geq 1$. In that case, player i's eventual conditional beliefs at h will thus be given by period τ and round $k-1$ of the procedure. This implies that at h, player i believes that his opponents were "level $(k-1)$-rational" from period τ onwards, and "fully rational" from period $\tau + 1$ onwards.

Nothing essential would change in the procedure if we would allow for randomized (that is, mixed or behavioral) strategies. In that case, the randomized strategies surviving the new procedure would exactly be the randomizations over pure strategies that survive our original procedure.

It is not difficult to prove that the concept of forward and backward rationalizability always yields at least one strategy and conditional belief vector for every player.

Theorem 3.1 (Existence) *For every player there are always at least one strategy and one conditional belief vector that are forward and backward rationalizable.*

Thus, it can never happen that all remaining strategies or conditional belief vectors for a given player are eliminated at a particular round.

[7] If $m > m_0$ then $\tau < m$, since h is always reachable by a strategy in $S_i^{m-1.0} = S_i^m$.

3.4 Examples

Example 1. Consider the example from Figure 2 in the introduction. We set $m_0 = 1$ and hence $m_+ = 3$. We will now run the forward and backward rationalizability procedure, starting at period 3.

Period 3. We have that $B_1^{3.0} = B_1^{3.1} = B_1$ and $B_2^{3.0} = B_2^{3.1} = B_2$. For player 1, both strategies (a, e) and (a, f) are optimal from period 3 onwards for some conditional belief vector in $B_1^{3.1}$, and similarly for player 2's strategies (c, g) and (c, h). Note that the strategies b and d are vacuously optimal from period 3 onwards for some conditional belief vector in $B_1^{3.1}$ and $B_2^{3.1}$, respectively. Thus,

$$S_1^{3.1} = S_1 = \{b, (a, e), (a, f)\} \text{ and } S_2^{3.1} = S_2 = \{d, (c, g), (c, h)\},$$

and this is where the procedure at Period 3 terminates.

Period 2. *Round 1.* We have that $B_1^{2.1} = B_1$ and $B_2^{2.1} = B_2$. For player 2, strategy (c, g) is not optimal from period 2 onwards for any conditional belief vector in $B_2^{2.1}$. In turn, strategies (c, h) and d *are* optimal from period 2 onwards for some conditional belief vector in $B_2^{2.1}$. Thus,

$$S_2^{2.1} = \{d, (c, h)\}.$$

Round 2. We then have

$$B_1^{2.2} = \{b_1 \in B_1 \mid b_1(h_1)(\{d, (c, h)\}) = 1 \text{ and } b_1(h_3)((c, h)) = 1\}.$$

Since for player 1 only strategies b and (a, f) are optimal from period 2 onwards for some conditional belief vector in $B_1^{2.2}$, it follows that

$$S_1^{2.2} = \{b, (a, f)\}.$$

Round 3. This implies that

$$B_2^{2.3} = \{b_2 \in B_2 \mid b_2(h_2)((a, f)) = 1 \text{ and } b_2(h_3)((a, f)) = 1\}.$$

For player 2, only strategy (c, h) is optimal from period 2 onwards for some conditional belief vector in $B_2^{2.3}$, and we thus conclude that

$$S_2^{2.3} = \{(c, h)\}.$$

Round 4. We then have that

$$B_1^{2.4} = \{b_1 \in B_1 \mid b_1(h_1)((c, h)) = 1 \text{ and } b_1(h_3)((c, h)) = 1\},$$

after which no further eliminations are possible in period 2.

Period 1. We start with the restrictions on the strategies and conditional belief vectors inherited from period 2. That is,

$$S_1^{1.0} = \{b, (a, f)\}, \; B_1^{1.0} = \{b_1 \in B_1 \mid b_1(h_1)((c, h)) = 1 \text{ and } b_1(h_3)((c, h)) = 1\},$$
$$S_2^{1.0} = \{(c, h)\} \text{ and } B_2^{1.0} = \{b_2 \in B_2 \mid b_2(h_2)((a, f)) = 1 \text{ and } b_2(h_3)((a, f)) = 1\}.$$

For player 1, the only strategy that is optimal from period 1 onwards for some conditional belief vector in $B_1^{1.0}$ is b. We therefore have

$$S_1^{1.1} = \{b\}.$$

Afterwards, no further eliminations are possible. We thus conclude that the strategies selected by the forward and backward rationalizability procedure are b for player 1 and (c, h) for player 2.

On the other hand, as we have seen in the introduction, strong rationalizability selects the strategies b for player 1 and d for player 2. The intuition for this difference is the following: According to forward and backward rationalizability, player 2 asks at h_2: What is the earliest period m such that player 1's past behavior – that is, player 1 choosing a – can be explained by "full rationality" from period m onwards? This must be period 2. Indeed, from period 2 onwards, player 1 expects player 2 to choose (c, h), which makes it optimal for player 1, from period 2 onwards, to choose (a, f). In turn, if player 2 expects player 1 to choose (a, f), then it is optimal for player 2 to choose (c, h). This is a plausible theory for the reasoning and play from period 2 onwards.

However, if player 1 anticipates player 2 choosing (c, h), then it can never be optimal for player 1 to choose a at h_1. As such, player 1 choosing a cannot be explained by "full rationality" from period 1 onwards.

According to strong rationalizability, player 2 asks at h_2: Is there a strategy for player 1 involving his observed past move a that is optimal for *some* belief, even if this belief attributes irrational future strategies to player 2? This reasoning leads player 2 to believe that player 1 chooses (a, e), as this is the only strategy involving a that can possibly yield him at least 2. As a consequence, player 2 will choose d. Note, however, that stratey (a, e) can only yield player 1 at least 2 if he believes that player 2 irrationally chooses the strategy (c, g) in the future. As such, believing that player 1 chooses (a, e) cannot be part of a "fully rational" theory from period 1 onwards. Therefore, our concept discards this type of reasoning by player 2.

Example 2. We now consider a larger example, with more information sets. Consider the dynamic game from Figure 3. Note that the information sets h_2, h_3 and h_5 are non-trivial. We set $m_0 = 1$, and hence $m_+ = 5$. Table 1 presents the sets of strategies that remain, for every period m and round k, in the forward and backward rationalizability procedure. We will now explain the elimination steps in the procedure.

In period 5, round 1, player 3's strategy (i, o) is suboptimal at h_7. In period 4, round 1, player 2 must at h_6 assign probability 1 to player 3's strategy (i, n), and hence player 2's strategies (d, j, l) and (d, k, l) become suboptimal at h_6. In period 3, round 1, player 3 must believe at h_4 that player 2 will either choose (d, j, m) or (d, k, m), and hence player 3's strategy g becomes suboptimal at h_4. In period 2 no eliminations can be made. In period 1, round 1, player 1's strategy b is always better than (a, e) at h_1, and hence player 1's strategy (a, e) is suboptimal at h_1. At period 1, round 2, player 2 must as h_5 assign probability zero to player 1 choosing (a, e) and probability zero to player 3 choosing g. But then, player 2 can at h_5 only assign positive probability to the second and fourth history, which makes player 2's strategy (d, j, m) suboptimal at h_5. At period 1, round 3, player 1 must believe at h_1 that player 2 chooses either c or (d, k, m), and hence player 1 expects 15 or 20 by choosing (a, f) at h_1. As b given player 1 the amount of 10 for sure, player 1's strategy b becomes suboptimal at h_1. At the same time, player 3 must believe at h_4 that player 2 chooses (d, k, m), which makes player 3's strategy (i, n) suboptimal at h_4. Finally, at period 1, round 4, player 2 must believe at h_2 that player 1 chooses (a, f), which makes player 2's strategy c suboptimal at h_2. The

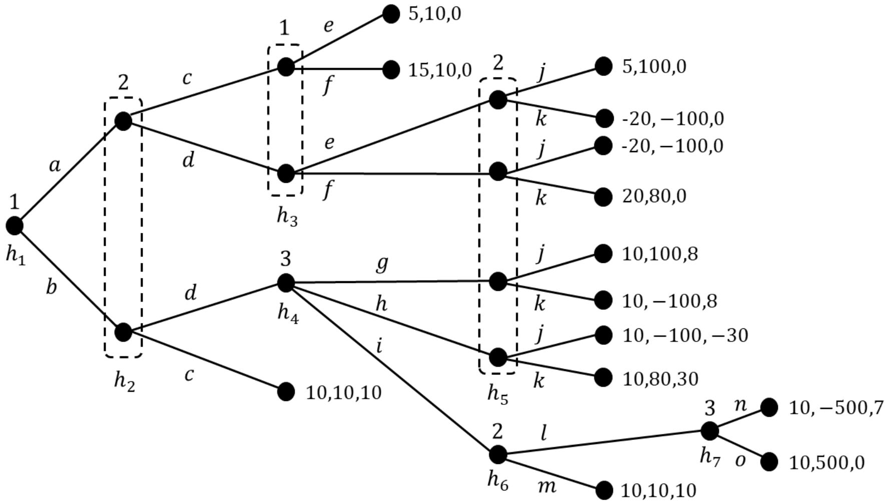

Figure 3: A dynamic game with non-trivial information sets

Period m	Round k	$S_1^{m.k}$	$S_2^{m.k}$	$S_3^{m.k}$
5	1	S_1	S_2	$\{g, h, (i, n)\}$
4	1	S_1	$\{(c, (d, j, m), (d, k, m)\}$	$\{g, h, (i, n)\}$
3	1	S_1	$\{(c, (d, j, m), (d, k, m)\}$	$\{h, (i, n)\}$
2	1	S_1	$\{(c, (d, j, m), (d, k, m)\}$	$\{h, (i, n)\}$
1	1	$\{(a, f), b\}$	$\{(c, (d, j, m), (d, k, m)\}$	$\{h, (i, n)\}$
1	2	$\{(a, f), b\}$	$\{(c, (d, k, m)\}$	$\{h, (i, n)\}$
1	3	$\{(a, f)\}$	$\{(c, (d, k, m)\}$	$\{h\}$
1	4	$\{(a, f)\}$	$\{(d, k, m)\}$	$\{h\}$

Table 1: Forward and backward rationalizability procedure in the game of Figure 3

unique forward and backward rationalizable strategies are thus (a, f) for player 1, (d, k, m) for player 2 and h for player 3.

4 Epistemic Characterization

In this section we investigate what the concept of forward and backward rationalizability entails in terms of reasoning. To this purpose, we offer epistemic conditions on the players' belief hierarchies such that the optimal strategies under these belief hierarchies are precisely the forward and backward rationalizable strategies. Before doing so, we first recall the definition of a (universal) type space for dynamic games, and subsequently formalize the notion of strong belief and optimal choice for types in a type space.

4.1 Type Space

The epistemic conditions we introduce will impose restrictions on the belief hierarchies that the players may have. Such belief hierarchies may conveniently be encoded by means of *types* in a type space. To formalize a type space, we need the following definition and pieces of notation. A topological space $(X, \mathcal{O})$ is called *Polish* if it is separable and completely metrizable. By $\Sigma(X)$ we denote the Borel σ-algebra on X, that is, the smallest σ-algebra that contains all open sets, whereas $\Delta(X)$ denotes the set of all probability measures on $(X, \Sigma(X))$. We endow $\Delta(X)$ with the smallest topology $\mathcal{O}_{\Delta(X)}$ such that each of the sets $\{\mu \in \Delta(X) | \int_X f d\mu \in O\}$ is open in $\Delta(X)$, where f runs over all bounded continuous functions $f : X \to \mathbb{R}$ and O runs over all open subsets of the reals. By Kechris (1995), Theorem 17.23, $(\Delta(X), \mathcal{O}_{\Delta(X)})$ is again a Polish space. We then consider $\Delta(X)$ as a measurable space that is endowed with the Borel σ-algebra (generated by $\mathcal{O}_{\Delta(X)}$). It is a well-known fact that a continuous map between two topological spaces is measurable if both of these spaces are endowed with their respective Borel σ-algebras.

Definition 4.1 (Type space) *A* ***type space*** $\mathcal{T} = ((T_i, \mathcal{O}_i), \beta_i)_{i \in I}$ *specifies, for every player* i,

(a) a Polish type space $(T_i, \mathcal{O}_i)$, *and*

(b) a continuous belief mapping β_i, *which assigns to every type* $t_i \in T_i$ *and information set* $h \in H_i$ *a probabilistic belief* $\beta_i(t_i, h) \in \Delta(S_{-i}(h) \times T_{-i})$.

Moreover, the types must satisfy forward consistency, that is, for every player i, *every type* $t_i \in T_i$, *and every two information sets* $h, h' \in H_i$ *where* h' *follows* h *and* $\beta_i(t_i, h)(S_{-i}(h') \times T_{-i}) > 0$, *we have that*

$$\beta_i(t_i, h')(\{s_{-i}\} \times E_{-i}) = \frac{\beta_i(t_i, h)(\{s_{-i}\} \times E_{-i})}{\beta_i(t_i, h)(S_{-i}(h') \times T_{-i})}$$

for every $s_{-i} \in S_{-i}(h')$ *and every* $E_{-i} \in \Sigma(T_{-i})$.

For our epistemic characterization we need to work with a *universal* type space. To explain what it is, we must first introduce the notion of a *type morphism.*

Definition 4.2 (Type morphism) *Consider two type spaces $\mathcal{T} = ((T_i, O_i), \beta_i)_{i \in I}$ and $\mathcal{T}' = ((T_i', O_i'), \beta_i')_{i \in I}$. A* ***type morphism*** *from $\mathcal{T}$ to $\mathcal{T}'$ is a tuple $(f_i)_{i \in I}$ of continuous functions $f_i : T_i \to T_i'$ such that, for every player i, every type $t_i \in T_i$, and every information set $h \in H_i$ we have that*

$$\beta_i'(f_i(t_i), h)(\times_{j \neq i}(\{s_j\} \times E_j')) = \beta_i(t_i, h)(\times_{j \neq i}(\{s_j\} \times f_j^{-1}(E_j')))$$

for every opponents' strategy combination $(s_j)_{j \neq i} \in S_{-i}(h)$ and every measurable set $\times_{j \neq i} E_j' \subseteq \times_{j \neq i} T_j'$ of opponents' type combinations.

A type space is then called *universal* if every other type space can be uniquely embedded into it by means of a type morphism.

Definition 4.3 (Universal type space) *A type space $\mathcal{T}$ is* ***universal***[8] *if for every other type space $\mathcal{T}'$ there is a unique type morphism from $\mathcal{T}'$ to $\mathcal{T}$.*

It turns out that every two universal type spaces are isomorphic. As such, we can speak about *the* universal type space. Battigalli and Siniscalchi (1999), Guarino (2024) and Fukuda (2024) have shown that we can always construct a universal type space in our setting.

4.2 Strong Belief

Consider a type space $\mathcal{T} = ((T_i, \mathcal{O}_i), \beta_i)_{i \in I}$, a type $t_i \in T_i$ and an event $E_{-i} \in \Sigma(S_{-i} \times T_{-i})$. The type t_i is said to *strongly believe* the event E_{-i} if it assigns probability 1 to the event whenever possible. That is,

$$\beta_i(t_i, h)(E_{-i}) = 1 \text{ at all } h \in H_i \text{ where } E_{-i} \cap (S_{-i}(h) \times T_{-i}) \neq \emptyset.$$

4.3 Optimal Choice

Consider a type $t_i \in T_i$, a strategy $s_i \in S_i$ and an information set $h \in H_i(s_i)$. Then, we denote by

$$u_i(s_i, t_i, h) := \sum_{s_{-i} \in S_{-i}(h)} \beta_i(t_i, h)(\{s_{-i}\} \times T_{-i}) \cdot u_i(z(s_i, s_{-i}))$$

the *expected utility* induced by s_i at h for the type t_i. The strategy s_i is *optimal* for the type t_i at information set h if $u_i(s_i, t_i, h) \geq u_i(s_i', t_i, h)$ for all $s_i' \in S_i(h)$. For a given period m, we say that strategy s_i is optimal for the type t_i *from period m onwards* if for every period $\tau \geq m$, and every information set $h \in H_i(s_i) \cap H^\tau$, the strategy s_i is optimal for type t_i at h.[9] For a given set of types $\hat{T}_i \in \Sigma(T_i)$, we denote by

$$(S_i \times \hat{T}_i)^{rat, \geq m} := \{(s_i, t_i) \in S_i \times \hat{T}_i \mid s_i \text{ optimal for } t_i \text{ from period } m \text{ onwards}\}$$

the event that player i chooses rationally from period m onwards and that i's type belongs to $\hat{T}_i$.

The following result states that the event of choosing rationally from a certain period onwards is always a "well-behaved" set.

[8] In the literature, such type spaces are sometimes called *terminal*.

[9] Note that if $H_i(s_i) \cap H^\tau = \emptyset$ for all $\tau \geq m$, then s_i is (vacuously) optimal for every type of player i from period m onwards.

Lemma 4.1 (Rationality is a measurable event) *Suppose that $\hat{T}_i$ is a closed (measurable) subset of T_i. Then, the set $(S_i \times \hat{T}_i)^{rat, \geq m}$ is a closed (measurable) subset of $S_i \times T_i$.*

This result will be important for guaranteeing that the epistemic conditions below are all well-defined. It will also play a key role in the proof of our epistemic characterization.

4.4 Epistemic Characterization

The epistemic conditions we impose on the players' types are as follows: First, we focus on the last period m_+ where players have to move. A player must $(m_+.1)$ strongly believe in the event that every opponent chooses rationally from period m_+ onwards, $(m_+.2)$ strongly believe in the event that every opponent chooses rationally from period m_+ onwards and that every opponent satisfies $(m_+.1)$, and so on. These conditions together yield *common backward strong belief in rationality from period m_+ onwards.* We refer to this event as (m_+).

We then move to period $m_+ - 1$. A player must $(m_+ - 1.1)$ strongly believe in the event that every opponent chooses rationally from period $m_+ - 1$ onwards and that every opponent satisfies (m_+). Moreover, a player must $(m_+ - 1.2)$ strongly believe in the event that every opponent chooses rationally from period $m_+ - 1$ onwards and that every opponent satisfies $(m_+ - 1.1)$, and so on. These conditions together yield *common backward strong belief in rationality from period $m_+ - 1$ onwards.*

We then continue in this fashion until we reach the beginning of the game. We thus give the highest epistemic priority to reasoning about the last period, the second-to-highest epistemic priority to reasoning about the last two periods, and so on.[10] The final restrictions on the types are called *common backward strong belief in rationality.*

Definition 4.4 (Common backward strong belief in rationality) *For every period m, number $k \in \{0, 1, ...\}$ and player i, we define the sets of types $SBR_i^{\geq m.k}$ that express k-fold backward strong belief in rationality from period m onwards. These sets $SBR_i^{\geq m.k}$ are inductively defined as follows.*

***Period** m_+. Set $SBR_i^{\geq m_+.0} := T_i$ for every player i. For every $k \geq 1$, inductively define*

$$SBR_i^{\geq m_+.k} := \{t_i \in SBR_i^{\geq m_+.k-1} \mid t_i \text{ strongly believes } \times_{j \neq i} (S_j \times SBR_j^{\geq m_+.k-1})^{rat, \geq m_+}\}.$$

Set $SBR_i^{\geq m_+} := \cap_{k \geq 0} SBR_i^{\geq m_+.k}$ for every player i.

***Period** $m \leq m_+ - 1$. Set $SBR_i^{\geq m.0} := SBR_i^{\geq m+1}$ for every player i. For every $k \geq 1$, inductively define*

$$SBR_i^{\geq m.k} := \{t_i \in SBR_i^{\geq m.k-1} \mid t_i \text{ strongly believes } \times_{j \neq i} (S_j \times SBR_j^{\geq m.k-1})^{rat, \geq m}\}.$$

Set $SBR_i^{\geq m} := \cap_{k \geq 0} SBR_i^{\geq m.k}$ for every player i.

For a given period m and round k, a type t_i is said to express up to k-fold backward strong belief in rationality

[10] For a detailed discussion of epistemic priority in rationalizability concepts, see Catonini (2019).

from period m onwards if $t_i \in SBR_i^{\geq m.k}$. The type is said to express common backward strong belief in rationality from period m onwards if $t_i \in SBR_i^{\geq m}$.The type t_i is said to express ***common backward strong belief in rationality*** *if $t_i \in SBR_i^{\geq m_0}$, where m_0 is the first period in the game.*

The following result guarantees that the epistemic conditions imposed above lead to "well-behaved" sets.

Lemma 4.2 (Epistemic conditions lead to closed sets) *Each of the sets $SBR_i^{\geq m.k}$ and $SBR_i^{\geq m}$ above is a closed subset of T_i.*

Let us now have a closer look at the epistemic conditions above. The conditions imply that at every information set where a player has to move, he looks for the earliest period m and the highest degree k such that it is possible to believe that (i) every player chooses rationally from period m onwards and expresses common backward strong belief in rationality from period m onwards, and (ii) every player chooses rationally from period $m-1$ onwards and expresses up to k-fold backward strong belief in rationality from period $m-1$ onwards. Moreover, he *will* then believe (i) and (ii). This may be viewed as a *best rationalization principle* for the epistemic concept above.

From this best rationalization principle it is clear that epistemic priority is given to backward induction reasoning: If a player is at an information set, he first looks for the earliest period m such that it is possible to believe that every player chooses rationally from period m onwards and expresses common backward strong belief in rationality from period m onwards. In that case, the player *will* express common backward strong belief in rationality from period m onwards, and hence will believe, in particular, that every opponent will choose rationally from period m onwards. Only afterwards will he think about period $m-1$, and look for the highest degree k such that it is possible to believe that, in addition, every player chooses rationally from period $m-1$ onwards and expresses up to k-fold backward strong belief in rationality from period $m-1$ onwards.

The following result shows that the epistemic conditions in *common backward strong belief in rationality* single out precisely those strategies that are *forward and backward rationalizable.*

Theorem 4.1 (Epistemic characterization) *Consider the universal type space $\mathcal{T} = ((T_i, \mathcal{O}_i), \beta_i)_{i \in I}$. Then, for every player i and strategy $s_i \in S_i$, the following holds:*

(a) strategy s_i is forward and backward rationalizable, if and only if, s_i is optimal from the first period onwards for a type $t_i \in T_i$ that expresses common backward strong belief in rationality,

(b) if $m \leq m_+ - 1$ then $s_i \in S_i^{m.0}$, if and only if, s_i is optimal from period $m+1$ onwards for a type $t_i \in SBR_i^{\geq m+1}$ that expresses common backward strong belief in rationality from period $m+1$ onwards, and

(c) if $k \geq 0$ then $s_i \in S_i^{m.k+1}$, if and only if, s_i is optimal from period m onwards for a type $t_i \in SBR_i^{\geq m.k}$ that expresses up to k-fold backward strong belief in rationality from period m onwards.

In particular, since we know from Theorem 3.1 that forward and backward rationalizable strategies always exist, it follows that there is always a type that expresses common backward strong belief in rationality. That is, the system of epistemic conditions we offer never leads to logical contradictions.

A major difference with strong rationalizability is that forward and backward rationalizability requires players to do forward induction reasoning from a certain period onwards, in a backward inductive fashion. Strong rationalizability, in contrast, always requires players to do the forward induction reasoning in the whole game, that is, from the first period onwards.

As such, we can also consider a *bounded rationality* version of forward and backward rationalizability in which players only do the forward induction reasoning from period m_+ onwards, from period $m_+ - 1$ onwards, until we reach period m. Players would thus not actively reason about choices that are made before period m. Parts (b) and (c) in Theorem 4.1 reveal what has to be imposed, in terms of reasoning, to establish such a bounded rationality variant.

5 Relation with Backwards Rationalizability

In this section we start by showing that our concept is a refinement of backwards rationalizability in terms of strategies, and link this to the epistemic conditions of *common belief in future rationality* (Perea (2014)).

5.1 Refinement of Backwards Rationalizability

In the game of Figure 2 we saw that forward and backward rationalizability selects a different strategy for player 2 than strong rationalizability. The reason was that according to the former concept, player 2, at a given information set h, only interprets player 1's past move as a rational move if this is compatible with the completed reasoning from h onwards. This shows that forward and backward rationalizability is, above all, a forward looking concept, and thus gives priority to backward induction reasoning. This intuition is confirmed by the following result.

Theorem 5.1 (Relation with backwards rationalizability) *Every strategy and conditional belief vector that is forward and backward rationalizable, is also backwards rationalizable.*

Hence, one can argue that the new concept gives epistemic priority to backward induction reasoning compared to forward induction reasoning. At the same time, as the game in Figure 2 shows, our concept may be more restrictive than backwards rationalizability since it additionally imposes some forward induction reasoning.

5.2 Belief in Future Rationality

In Perea (2014) it is shown that backwards rationalizability can be epistemically characterized by the conditions of *common belief in future rationality,* stating that a player always believes that his opponents

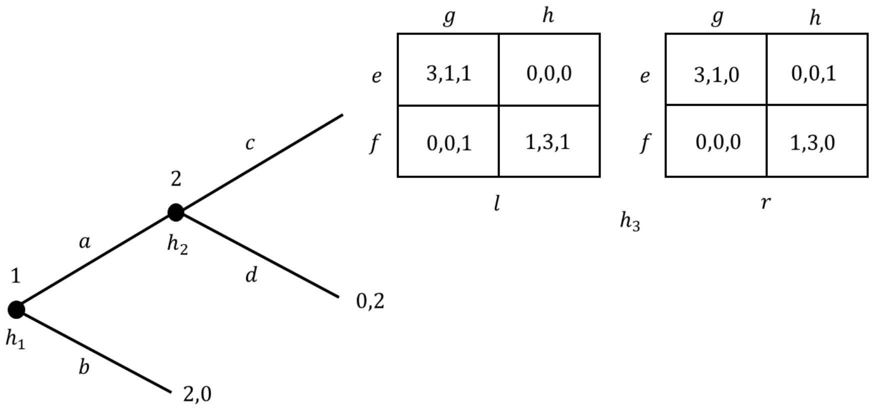

Figure 4: Double outside option game with a bet for the third player

will choose rationally now and in the future, that a player always believes that his opponents always believe that the other players will chooses rationally now and in the future, and so on. Hence, even if a player is suprised by an opponent's past move, which may possibly lead him to conclude that his opponent has chosen irrationally in the past, he will still believe that the same opponent will "recover" and choose rationally from now on.

As, by Theorem 5.1, forward and backward rationalizability refines the reasoning of backwards rationalizability, it follows that the former concept always reasons within the bounds set by common belief in future rationality. In particular, a player will only interpret an opponent's past move as a signal for the opponent's future behavior – a typical forward inductive inference – if this is in accordance with common belief in future rationality.

In this sense, our concept is fundamentally different from strong rationalizability. To further illustrate this, consider the game in Figure 4. This is a three-player double outside option game, where the third player has to "bet" on the behavior of the first two players after they have both foregone the outside option. In that case, player 1 must choose between e and f, player 2 must choose between g and h, whereas player 3 must choose between l and r (left matrix or right matrix). Note that r will only be optimal for player 3 if he believes, with sufficiently high probability, that players 1 and 2 miscoordinate on e and h at information set h_3.

According to strong rationalizability, player 3 will conclude at h_3 that players 1 and 2 will go for e and h, respectively, and therefore player 3 will choose r. To see this, note that a for player 1 can only be optimal at h_1 if he chooses e at h_3, and that c for player 2 can only be optimal at h_2 if he chooses h at h_3.

But assume now that player 1 would choose optimally at h_3 given his beliefs there. According to strong rationalizability, player 1 will believe at h_3 that player 2 chooses h, for the same reasons as outlined above.

If player 1 chooses optimally at h_3, then he would go for f and not e, rendering r suboptimal for player 3. In that sense, the choice r selected by strong rationalizability is a risky choice for player 3, as this bet is based on the assumption that player 1 will not choose optimally given his belief at h_3.

The concept of forward and backward rationalizability prescribes a completely different line of reasoning for player 3 here. If player 3 finds himself at h_3, then he first asks whether there is a plausible theory from information set h_2 onwards that could explain the event of reaching h_3. Such a theory can indeed be found: For player 2 it is only optimal to choose c at h_2 if he would choose h at h_3. Therefore, player 3 will believe at h_3 that player 2 will choose h, and will believe that player 1 also believes at h_3 that player 2 chooses h. Assuming that player 1 chooses optimally at h_3, player 3 then believes that player 1 chooses f at h_3.

Clearly, this belief about the strategies of players 1 and 2 cannot be refined any further by subsequently analyzing the game from h_1 onwards, and therefore forward and backward rationalizability leads player 3 to believe that players 1 and 2 will choose f and h at h_3, and player 3 will thus choose l. Note that player 1's choice f is optimal at h_3 if he believes that player 2 chooses h there, and *vice versa.* As such, player 3's reasoning about the behavior of players 1 and 2 is in line with common belief in future rationality, contrary to player 3's reasoning prescribed by strong rationalizability.

6 Relation with Strong Rationalizability

In this section we show that the forward and backward rationalizability procedure is equivalent, in terms of outcomes, to the strong rationalizability procedure. To prove this result we use the *strong belief reduction operator* from Perea (2018), and rely on results and proof techniques from Perea (2018), observing that these can be extended to games with *unobserved past choices.*

6.1 Strong Belief Reduction Operator

A *product of strategy sets* is a set $D = \times_{i \in I} D_i$ where $D_i \subseteq S_i$ for every player i. A *reduction operator* r assigns to every product of strategy sets D a subset $r(D) \subseteq D$, where $r(D)$ is again a product of strategy sets. In the sequel, we always assume that we are talking about products of strategy sets. A set E with $r(D) \subseteq E \subseteq D$ is a *partial reduction* of D. For a given $k \in \mathbf{N}$, let $r^k(D)$ be the k-fold application of the operator r to D.

For a given product of strategy sets D, let $H(D)$ be the collection of information sets that are reached by some strategy profile in D.

Definition 6.1 (Strong belief reduction operator) *The* ***strong belief reduction operator*** *sb assigns to every product of strategy sets $D = \times_{i \in I} D_i$ the set $sb(D) = \times_{i \in I} sb_i(D)$, where*

$$\begin{aligned} sb_i(D) := \{ s_i \in D_i \mid & \text{ there is some } b_i \in B_i \text{ that strongly believes } D_{-i} \\ & \text{such that } s_i \text{ is optimal for } b_i \text{ at every } h \in H(D) \cap H_i(s_i) \}. \end{aligned}$$

Recall that $H_i(s_i)$ is the collection of information sets for player i that can be reached by s_i. In Perea (2018), Theorem 2.1, it is shown that for dynamic games with observed past choices, the strongly rationalizable strategies are obtained by iteratedly applying the strong belief reduction operator to the full set of strategies. As the proof does not rely on the property of observed past choices, this result holds for all games in our setting as well.

Theorem 6.1 (Characterization of strong rationalizability) *For every $k \in \mathbf{N}$ and every player i, let $S_i^{sr,k}$ be the set of strategies for player i that survive round k of the strong rationalizability procedure. Let $S^{sr,k} = \times_{i \in I} S_i^{sr,k}$ and $S = \times_{i \in I} S_i$. Then, $S^{sr,k} = sb^k(S)$ for every $k \in \mathbf{N}$.*

6.2 Order Independence with Respect to Outcomes

Consider a reduction operator r. An *elimination order* for r is a finite sequence $(D^0, D^1, ..., D^K)$ of products of strategy sets such that (a) $D^0 = S$, (b) $r(D^k) \subseteq D^{k+1} \subseteq D^k$ for every $k \in \{0, ..., K-1\}$, and (c) $r(D^K) = D^K$.

For a product of strategy sets D, let $Z(D)$ be the collection of terminal histories reached by strategy profiles in D.

Definition 6.2 (Order independence with respect to outcomes) *A reduction operator r is **order independent with respect to outcomes** if for every two elimination orders $(D^0, D^1, ..., D^K)$ and $(E^0, E^1, ..., E^L)$ we have that $Z(D^K) = Z(E^L)$.*

Corollary 3.1 in Perea (2018) states that for all dynamic games with observed past choices, the strong belief reduction operator is order independent with respect to outcomes. As it turns out, the proof in Perea (2018) does not rely on the property of observed past choices, and holds for our class of dynamic games as well.

Theorem 6.2 (Order independence with respect to outcomes) *The strong belief reduction operator sb is order independent with respect to outcomes.*

6.3 Outcome Equivalence with Strong Rationalizability

The following result states that the reduction steps in the forward and backward rationalizability procedure correspond to a specific elimination order of the strong belief reduction operator.

Lemma 6.1 (Procedure as elimination order) *Let $(D^0, D^1, ..., D^K)$ be the products of strategy sets generated by the forward and backward rationalizability procedure in every round.*[11] *Then, $(D^0, D^1, ..., D^K)$ is an elimination order of the strong belief reduction operator.*

[11] Here, the rounds $0, 1, ..., K$ actually correspond to the rounds $m.k$ in the definition of the forward and backward rationalizability procedure.

By combining Theorem 6.1, Lemma 6.1 and Theorem 6.2, we conclude that strong rationalizability leads to the same set of outcomes as forward and backward rationalizability.[12]

Theorem 6.3 (Outcome equivalence with strong rationalizability) *Let S^{sr} and S^{fbr} be the products of strategy sets induced by the strong rationalizability procedure, and the forward and backward rationalizability procedure, respectively. Then, $Z(S^{sr}) = Z(S^{fbr})$.*

That is, if one would be only interested in the induced outcomes, it makes no difference whether strong rationalizability is used, or forward and backward rationalizability. However, as we have argued before, we believe that the latter concept provides a more compelling theory for how players react to surprises at information sets to which the players initially assign probability zero.

7 Generalization of Battigalli's Theorem

Battigalli (1997) has shown that in every dynamic game with perfect information but without relevant ties, strong rationalizability leads to the unique backward induction outcome. Alternative proofs can be found in Catonini (2020), Chen and Micali (2013), Heifetz and Perea (2015) and Perea (2018).

Catonini (2020) and Perea (2017) provide generalizations of this result, by showing that in every dynamic game with *observable past choices* (but allowing for simultaneous moves), strong rationalizability refines, in terms of outcomes, the concepts of backwards rationalizability and backward dominance, respectively.

However, one of the most attractive properties of strong rationalizability is that in games with *imperfect* information (that is, where some past choices are unobservable), it allows an active player at an information set to exclude some of the histories at this information set from consideration. This is so, since reaching those histories would imply a lower degree of rationality for some players that moved in the past than the remaining "more rational" histories. Therefore, it is of importance to understand the relationship between forward and backward induction especially for games with *imperfect* information. As is well-known, there are games that allow for more backwards rationalizable outcomes than strongly rationalizable outcomes.[13] To the best of our knowledge, it was – up to now – unknown whether for games with *imperfect* information strongly rationalizable outcomes are always backwards rationalizable outcomes.

It turns out that in every dynamic game considered in our setup, that is, also in games with imperfect information, every strongly rationalizable outcome is indeed backwards rationalizable. This follows immediately from our Theorems 5.1 and 6.3: Take an outcome induced by strong rationalizability. Then, by

[12] This result could also be established by using tools from Chen and Micali (2013): It can be shown that forward and backward rationalizability corresponds to a particular elimination order of the iterated conditional dominance procedure by Shimoji and Watson (1998). Since Shimoji and Watson (1998) prove that the latter procedure is equivalent to strong rationalizability, and Chen and Micali (2013) show that this procedure is order independent with respect to outcomes, the statement from Theorem 6.3 follows. Perea (2024) shows the outcome equivalence result for the version of forward and backward rationalizability that does not insist on forward consistency of the conditional belief vectors.

[13] See, for instance, the classical Battle-of-the-Sexes game with an outside option, which is the game that starts at h_2 in Figure 2. In that game, strong rationalizability uniquely yields the forward induction outcome $(c, (f, h))$, whereas backwards rationalizability also allows for the outcomes $(c, (e, h))$ and d.

Theorem 6.3, this outcome will also be induced by forward and backward rationalizability. As, by Theorem 5.1, forward and backward rationalizability refines backwards rationalizability in terms of strategies, it follows that this outcome is also induced by backwards rationalizability. We thus obtain the following result.[14]

Corollary 7.1 (Generalization of Battigalli's theorem) *Let S^{sr} and S^{br} denote the products of strategy sets induced by the strong rationalizability procedure and the backwards rationalizability procedure, respectively. Then, $Z(S^{sr}) \subseteq Z(S^{br})$.*

This result, in turn, implies Battigalli's theorem, as in every dynamic game with perfect information but without relevant ties, the concept of backwards rationalizability leads to the unique backward induction strategies, and thus, in particular, to the unique backward induction outcome. The existence of strongly rationalizable outcomes now implies that this must also be the unique strongly rationalizable outcome.

8 Expansion Monotonicity

In this section we introduce the principle of *expansion monotonicity* and show that the concepts of *forward and backward rationalizability* and *backwards rationalizability* satisfy the principle.

Consider two dynamic games Γ and $\hat{\Gamma}$. Then, Γ is a *subgame* of $\hat{\Gamma}$ if Γ is the restriction of $\hat{\Gamma}$ to those histories in $\hat{\Gamma}$ that weakly follow the initial history in Γ, and such that every information set in Γ is either contained in, or disjoint from, $\hat{\Gamma}$. Whenever Γ is a subgame of $\hat{\Gamma}$, we say that $\hat{\Gamma}$ is an *expansion* of Γ.

Intuitively, expansion monotonicity states that if a player learns that the game Γ was actually preceded by some earlier moves, resulting in an expansion $\hat{\Gamma}$, then this new information should only refine, but never overthrow, his previous reasoning. But what do we mean by the reasoning of a player in the games Γ and $\hat{\Gamma}$?

Formally, we model the output of the players' reasoning processes by conditional belief vectors, which specify at every information set where this player is active the belief that he holds about the other players' strategies. In general, a *solution concept* φ specifies for every game Γ and every player i in Γ a set $B_i^\varphi(\Gamma)$ of conditional belief vectors, representing the possible beliefs that player i can end up with if he reasons according to the standards of φ.

Now suppose that the players in Γ learn that the game was actually preceded by earlier moves, resulting in the expansion $\hat{\Gamma}$. Then, the reasoning of the players in the new, larger game $\hat{\Gamma}$ is represented by the new sets of conditional belief vectors $B_i^\varphi(\hat{\Gamma})$. How do we connect this to the reasoning and strategy choices in the smaller game Γ?

Let us denote by $S_i^{\hat{\Gamma}}$ and S_i^{Γ} the sets of strategies for player i in the games $\hat{\Gamma}$ and Γ, respectively. Similarly, we denote by $H_i^{\hat{\Gamma}}$ and H_i^{Γ} the collections of information sets in $\hat{\Gamma}$ and Γ, respectively, where player i is active.

[14] Corollary 7.1 could also be established by using tools from Chen and Micali (2013). Indeed, it can be shown that the backwards rationalizability procedure corresponds to a particular *non-finished* elimination order of the iterated conditional dominance procedure. We thank Pierpaolo Battigalli and Emiliano Catonini for pointing this out to us.

Consider a strategy $s_i \in S_i^\Gamma(h)$ in the subgame and an opponents' strategy combination $\hat{s}_{-i} \in S_{-i}^{\hat{\Gamma}}(h)$ in the expansion. Then, $\hat{s}_{-i}$ induces an opponents' strategy combination in Γ, by restricting $\hat{s}_{-i}$ to the information sets in Γ. Consequently, $(s_i, \hat{s}_{-i})$ induces a terminal history $z(s_i, \hat{s}_{-i})$ in Γ.

Consider a conditional belief vector $\hat{b}_i \in B_i^\varphi(\hat{\Gamma})$ in the new game $\hat{\Gamma}$, an information set $h \in H_i^\Gamma$ in the subgame, and a strategy $s_i \in S_i^\Gamma(h)$ for player i in the subgame. Then, the expected utility at h, given s_i and $\hat{b}_i(h)$, is then given by

$$u_i(s_i, \hat{b}_i(h)) := \sum_{\hat{s}_{-i} \in S_{-i}^{\hat{\Gamma}}(h)} \hat{b}_i(h)(\hat{s}_{-i}) \cdot u_i(z(s_i, \hat{s}_{-i})).$$

We say that the strategy s_i is *optimal* for $\hat{b}_i$ at h if

$$u_i(s_i, \hat{b}_i(h)) \geq u_i(s_i', \hat{b}_i(h)) \text{ for all } s_i' \in S_i^\Gamma(h).$$

Then, we denote by

$$S_i^\varphi(\Gamma \mid \hat{\Gamma}) := \{s_i \in S_i^\Gamma \mid s_i \text{ is optimal for some } \hat{b}_i \in B_i^\varphi(\hat{\Gamma}) \text{ at all information sets } h \in H_i^\Gamma(s_i)\}$$

the set of strategies for player i that are optimal in the subgame Γ if the players learn that the actual game is $\hat{\Gamma}$. We call $S_i^\varphi(\Gamma \mid \hat{\Gamma})$ the set of strategies that is *predicted for the subgame* Γ if the solution concept φ is applied to the expansion $\hat{\Gamma}$. In particular, $S_i^\varphi(\Gamma \mid \Gamma)$ contains those strategies that the solution concept induces for player i in the game Γ if the players do not learn any new information there.

Expansion monotonicity then states that every strategy in $S_i^\varphi(\Gamma \mid \hat{\Gamma})$, which is allowed by the new reasoning after learning that the actual game is $\hat{\Gamma}$, must also be allowed by the original reasoning which took place before receiving this new information – that is, it must be in $S_i^\varphi(\Gamma \mid \Gamma)$.

Definition 8.1 (Expansion monotonicity) *A solution concept φ satisfies **expansion monotonicity** if for every dynamic game Γ, every expansion $\hat{\Gamma}$, and every player i, it holds that $S_i^\varphi(\Gamma \mid \hat{\Gamma}) \subseteq S_i^\varphi(\Gamma \mid \Gamma)$.*

This property corresponds to requirement BI1 in Kohlberg and Mertens (1986), which states that a solution of a game should always induce a solution in each of its subgames. Note that we use set inclusion – and not set equality – here. The reason is that, upon learning that the actual game is $\hat{\Gamma}$, the player may receive new information that helps him to *refine* his reasoning in Γ.

It can be shown that the new concept proposed in this paper, as well as the backwards rationalizability concept, satisfy expansion monotonicity.

Theorem 8.1 (Expansion monotonicity) *The concept of forward and backward rationalizability and the concept of backwards rationalizability satisfy expansion monotonicity.*

In the introduction we have seen that *strong rationalizability* violates the principle of expansion monotonicity.

9 Discussion

Role of surprises. On a conceptual level, an important difference between *forward and backward rationalizability* and *strong rationalizability* lies in the role of surprises. In strong rationalizability, a player never believes his opponents to make mistakes in the execution of their strategies, only in the planning of their strategies. Indeed, when a player is surprised by another player's move, he believes the other player consciously made this move because he attributed a lower level of iterated strong belief in rationality to his opponents. Forward and backward rationalizability is agnostic about how players explain surprising opponents' moves: a player may, but need not, explain surprising opponents' moves by believing that these are due to past mistakes in the execution of strategies. However, a player will always believe that these same opponents will reason and behave in accordance with forward and backward rationalizability directly after the occurence of the surprise, and from then onwards. This is in line with Catonini and Penta (2022), in which it is argued that backwards rationalizability remains agnostic about how players explain surprising past moves. In this light, it would be interesting to embed the concepts of strong rationalizability and forward and backward rationalizability within the framework of Battigalli and de Vito (2021), which explicitly distinguishes between plans and actual behavior.

Computational tractability. In comparison with strong rationalizability the procedure of forward and backward rationalizability is often easier to implement. By applying strong rationalizability at the last information sets and then working its way backwards, the latter procedure typically keeps the decision problems at these information sets rather small.

Expansion monotonicity. We have shown that forward and backward rationalizability satisfies expansion monotonicity, unlike strong rationalizability. The intuitive reason is that the former concept applies strong rationalizability to *every subgame*, in a backward inductive fashion. In that sense, the relation between these two concepts is analogous to the relation between Nash equilibrium and subgame perfect equilibrium.

10 Appendix

10.1 Proof of Theorem 3.1

To prove Theorem 3.1 we need the following result.

Lemma 10.1 (Strong belief of nested sets) *For a given player i let*

$$\emptyset \neq D_k \subseteq D_{k-1} \subseteq ... \subseteq D_0 = S_{-i}$$

be a sequence of nested subsets of opponents' strategy combinations. Then, there is a conditional belief vector $b_i \in B_i$ that strongly believes each of the sets $D_k, D_{k-1}, \ldots, D_0$.

Proof. Take some arbitrary probability distribution $p_i \in \Delta(S_{-i})$ such that $p_i(s_{-i}) > 0$ for all $s_{-i} \in S_{-i}$. For a given information set $h \in H_i$ let $m(h)$ be the highest number in $\{0, ..., k\}$ such that $S_{-i}(h) \cap D_{m(h)} \neq \emptyset$.

Define the conditional belief $b_i(h)$ by

$$b_i(h)(s_{-i}) := \begin{cases} \dfrac{p_i(s_{-i})}{p_i(S_{-i}(h) \cap D_{m(h)})}, & \text{if } s_{-i} \in S_{-i}(h) \cap D_{m(h)} \\ 0, & \text{otherwise.} \end{cases} .$$

Then, it may be verified that the conditional belief vector $b_i = (b_i(h))_{h \in H_i}$ so constructed satisfies forward consistency, and strongly believes each of the sets $D_k, D_{k-1}, \ldots, D_0$. ■

Proof of Theorem 3.1. Recall that $B_i^{m.k}$ and $S_i^{m.k}$ are the sets of conditional belief vectors and strategies, respectively, that survive round k at period m of the forward and backward rationalizability procedure. We show, by induction on $m.k$, that $B_i^{m.k}$ and $S_i^{m.k}$ are always non-empty, starting with $m_+.0$.

By definition, $B_i^{m_+.0} = B_i$ and $S_i^{m_+.0} = S_i$. Applying Lemma 10.1 to the case $D_k = D_0 = S_{-i}$ shows that B_i is non-empty.

Now, suppose that m and k are such that $(m.k) \neq (m_+.0)$. We distinguish two cases: (1) $k \geq 1$, and (2) $k = 0$.

Case 1. Suppose that $k \geq 1$. Then, by definition, $B_i^{m.k} = \{b_i \in B_i^{m.k-1} \mid b_i \text{ strongly believes } S_{-i}^{m.k-1}\}$. By construction,

$$S_{-i}^{m.k-1} \subseteq S_{-i}^{m.k-2} \subseteq ... \subseteq S_{-i}^{m_+.0},$$

and $B_i^{m.k}$ consists of all those conditional belief vectors that strongly believe each of these nested sets. As, by the induction assumption, each of these nested sets is non-empty, it follows by Lemma 10.1 that $B_i^{m.k}$ is non-empty as well.

Now, take some $b_i \in B_i^{m.k}$. Then, b_i is forward consistent. It follows from Lemma 8.13.2 in Perea (2012) that there is a strategy s_i that is optimal for b_i at all $h \in H_i(s_i)$. In particular, s_i is optimal for b_i from period m onwards, and hence s_i is in $S_i^{m,k}$. Thus, $S_i^{m.k}$ is non-empty.

Case 2. Suppose that $k = 0$. Then, by construction, $B_i^{m.0} = B_i^{m+1.K}$ and $S_i^{m.0} = S_i^{m+1.K}$, where K is the round at which the procedure at period $m + 1$ terminates.[15] By the induction assumption, $B_i^{m+1.K}$ and $S_i^{m+1.K}$ are non-empty, and thus $B_i^{m.0}$ and $S_i^{m.0}$ are non-empty as well.

It thus follows, by induction on m and k, that $S_i^{m.k}$ and $B_i^{m.k}$ are non-empty for every m and k, and for every player i. As the procedure terminates after finitely many rounds, we conclude that every player i has at least one strategy and one conditional belief vector that are forward and backward rationalizable. ■

10.2 Proofs of Section 4

Proof of Lemma 4.1. Consider some information set $h \in H_i$ of player i and let $s_i \in S_i(h)$. We first show that the set of types $T_i(s_i, h)$ of player i for whom playing s_i is optimal at h is a closed set of types. To this

[15] The fact that such a round K exists follows from arguments similar to those outlined in Footnote 5.

purpose, we consider for any alternative strategy $r_i \in S_i(h)$, any opponents' strategy combination s_{-i} and any opponents' type combination t_{-i} the utility difference

$$u_i(s_i, s_{-i}, t_{-i}) - u_i(r_i, s_{-i}, t_{-i}) := u_i(z(s_i, s_{-i})) - u_i(z(r_i, s_{-i})).$$

By fixing s_i, r_i, and h, and varying s_{-i} and t_{-i}, we obtain a bounded continuous function

$$u_i(s_i, \cdot, \cdot) - u_i(r_i, \cdot, \cdot) : S_{-i}(h) \times T_{-i} \to \mathbb{R}.$$

This is indeed the case, since we endow $S_{-i}(h)$ with the discrete topology and $u_i(s_i, s_{-i}, t_{-i}) - u_i(r_i, s_{-i}, t_{-i})$ only depends on the S_{-i}-dimension of $S_{-i} \times T_{-i}$. Hence, the set of measures $\mu_i \in \Delta(S_{-i}(h) \times T_{-i})$ such that integrating over this function with respect to μ_i is non-negative is a closed set of measures, which we call $\Delta(S_{-i}(h) \times T_{-i})^{s_i \geq r_i}$. The set of measures $\mu_i \in \Delta(S_{-i}(h) \times T_{-i})$ such that s_i is optimal at h is the intersection

$$\Delta(S_{-i}(h) \times T_{-i})^{s_i,h} := \bigcap_{r_i \in S_i(h)} \Delta(S_{-i}(h) \times T_{-i})^{s_i \geq r_i},$$

which is closed as well. Note that, by construction,

$$T_i(s_i, h) = \{t_i \in T_i \mid \beta_i(t_i, h) \in \Delta(S_{-i}(h) \times T_{-i})^{s_i,h}\}.$$

Since the mapping $\beta_i(\cdot, h) : T_i \to \Delta(S_{-i}(h) \times T_{-i})$ is continuous, we conclude that the set $T_i(s_i, h)$ is closed.

Recall that the set of types t_i such that s_i is optimal at h is precisely $T_i(s_i, h)$. For a given period m, let $H^{\geq m} := H^m \cup H^{m+1} \cup \ldots \cup H^{m_+}$ be the collection of information sets from period m onwards. Then, the set of types for which s_i is optimal from period m onwards is

$$T_i^{\geq m}(s_i) := \bigcap_{h \in H_i(s_i) \cap H^{\geq m}} T_i(s_i, h),$$

which is closed in T_i. Note that if s_i does not reach any information set in $H^{\geq m}$, then s_i is automatically optimal from period m onwards for all types in T_i. For each of the finitely many strategies $s_i \in S_i$ of player i, the set $\{s_i\} \times T_i^{\geq m}(s_i)$ is closed in the product topology of $S_i \times T_i$, since it is the product of two closed sets. The set

$$(S_i \times T_i)^{rat, \geq m} = \bigcup_{s_i \in S_i} (\{s_i\} \times T_i^{\geq m}(s_i))$$

is closed in $S_i \times T_i$ since it is the union of finitely many closed sets. If $\hat{T}_i$ is a measurable subset of T_i then $(S_i \times \hat{T}_i)^{rat, \geq m} = (S_i \times T_i)^{rat, \geq m} \cap (S_i \times \hat{T}_i)$ is measurable since it is an intersection of a closed and a measurable set. If $\hat{T}_i$ is closed, then $S_i \times \hat{T}_i$ is closed and hence $(S_i \times \hat{T}_i)^{rat, \geq m}$ is closed, being the intersection of two closed sets. ■

Proof of Lemma 4.2. We start by proving the following result.

Claim. Let E be a closed subset of $S_{-i} \times T_{-i}$. Then, the set $\{ t_i \mid t_i$ strongly believes $E \}$ is a closed subset of T_i.

Proof of claim. Let $h \in H_i$ be such that $(S_{-i}(h) \times T_{-i}) \cap E \neq \emptyset$. We show that the set of measures in $\Delta(S_{-i}(h) \times T_{-i})$ that assign probability 1 to E is closed set. To this end, let $(\mu_n)_{n\in\mathbb{N}} \to \mu$ be a sequence of probability measures in $\Delta(S_{-i}(h) \times T_{-i})$ converging to $\mu \in \Delta(S_{-i}(h) \times T_{-i})$ such that $\mu_n(E) = 1$ for all $n \in \mathbb{N}$. We have to show that $\mu(E) = 1$. But this follows immediately from the Portemanteau Theorem (Kechris (1995), Theorem 17.20). By continuity, the set $\{t_i \in T_i \mid \beta_i(t_i, h)(E) = 1\}$ is a closed set of types. The set of types t_i that strongly believe E is the finite intersection of such sets of types over all $h \in H_i$ such that $(S_{-i}(h) \times T_{-i}) \cap E \neq \emptyset$. Hence, this is a closed set of types. $\Diamond$

The lemma now follows immediately by iteratively applying the claim and Lemma 4.1. ■

Proof of Theorem 4.1. Recall that, for every player i, period m, and round k, the sets $B_i^{m.k}$ and $S_i^{m.k}$ are the collections of conditional belief vectors and strategies, respectively, selected by the forward and backward rationalizability procedure at round k of period m. Moreover, $B_i^{m_0}$ and $S_i^{m_0}$ are the sets of conditional belief hierarchies and strategies, respectively, that survive all rounds at all periods.

For every player i and strategy $s_i \in S_i^{m_0}$ choose a conditional belief vector $b_i[s_i] \in B_i^{m_0}$ such that s_i is optimal for $b_i[s_i]$ from the first period onwards. For all other strategies s_i there is a period $m \in \{m_0, ..., m_+\}$ and a round k such that $s_i \in S_i^{m.k} \backslash S_i^{m.k+1}$. For such a strategy $s_i \in S_i^{m.k} \backslash S_i^{m.k+1}$ we can then choose a conditional belief vector $b_i[s_i] \in B_i^{m.k}$ such that s_i is optimal for $b_i[s_i]$ from period m onwards if $k \geq 1$, and s_i is optimal from period $m+1$ onwards if $k = 0$. If $m.k = m_+.0$, then optimality from period $m_+ + 1$ onwards means that s_i need not be optimal for $b_i[s_i]$ at all.

Based on these conditional belief vectors $b_i[s_i]$ we will now construct a finite type space $\hat{\mathcal{T}} = ((T_i, \mathcal{O}_i), \beta_i)$ where the sets of types are given by $T_i = \{t_i^{b_i[s_i]} \mid s_i \in S_i\}$, and the belief mappings β_i are such that

$$\beta_i(t_i^{b_i[s_i]}, h)((s_j, t_j)_{j\neq i}) = \begin{cases} b_i[s_i](h)((s_j)_{j\neq i}), & \text{if } t_j = t_j^{b_j[s_j]} \text{ for all } j \neq i \\ 0, & \text{otherwise} \end{cases} \tag{10.1}$$

for all players i, all strategies s_i, all information sets $h \in H_i$, and all opponents' strategy-type combinations $(s_j, t_j)_{j\neq i} \in S_{-i} \times T_{-i}$. Hence, every type $t_i^{b_i[s_i]}$ has the belief $b_i[s_i](h)$ about the opponents' strategy combinations at every information set $h \in H_i$, and matches, in its belief, every opponent's strategy s_j with the associated type $t_j^{b_j[s_j]}$. It is easy to see that every type in this model is forward consistent. Note that $b_i[s_i] = b_i[\hat{s}_i]$ implies that $t_i^{b_i[s_i]} = t_i^{b_i[\hat{s}_i]}$, and hence the type space $\hat{\mathcal{T}}$ is non-redundant by construction.

For every player i and conditional belief vector $b_i \in B_i \backslash \{b_i[s_i] \mid s_i \in S_i\}$ not present in $\hat{\mathcal{T}}$, we add a new type $t_i^{b_i}$ to $\hat{\mathcal{T}}$ whose conditional beliefs are given by

$$\beta_i(t_i^{b_i}, h)((s_j, t_j)_{j\neq i}) = \begin{cases} b_i(h)((s_j)_{j\neq i}), & \text{if } t_j = t_j^{b_j[s_j]} \text{ for all } j \neq i \\ 0, & \text{otherwise} \end{cases} \tag{10.2}$$

The new type space obtained after adding the type $t_i^{b_i}$ to $\hat{\mathcal{T}}$ is denoted by $\hat{\mathcal{T}} \cup \{t_i^{b_i}\}$.

Let $\mathcal{T}$ be a universal type space. Then, by definition, each of the finite type spaces $\hat{\mathcal{T}}$ and $\hat{\mathcal{T}} \cup \{t_i^{b_i}\}$ maps in a unique way to the universal type space $\mathcal{T}$ by a type morphism. Note that for every type $t_i^{b_i[s_i]}$ in $\hat{\mathcal{T}}$ the induced conditional belief hierarchy is the same, no matter whether it is regarded as a type in $\hat{\mathcal{T}}$ or a type in $\hat{\mathcal{T}} \cup \{t_i^{b_i}\}$. Since a type morphism always preserves the induced conditional belief hierarchy, the type $t_i^{b_i[s_i]}$ will be mapped to the same type in the universal type space $\mathcal{T}$, no matter whether it is regarded as a type in $\hat{\mathcal{T}}$ or a type in $\hat{\mathcal{T}} \cup \{t_i^{b_i}\}$. As each of the type spaces $\hat{\mathcal{T}}$ and $\hat{\mathcal{T}} \cup \{t_i^{b_i}\}$ is non-redundant, every type in these type spaces may be uniquely identified with a type in the universal type space $\mathcal{T}$.

For every player i, period m and number $k \in \{0, 1, ...\}$, we denote by $SBR_i^{\geq m.k}$ the set of types for player i in the universal type space $\mathcal{T}$ that express k-fold backward strong belief in rationality from period m onwards. Define

$$\hat{B}_i^{m.k} := \{b_i \in B_i \mid \text{there is some } t_i \in SBR_i^{\geq m.k} \text{ that induces the conditional belief vector } b_i\}$$

and

$$\hat{S}_i^{m.k} := \{s_i \in S_i \mid \text{there is some } t_i \in SBR_i^{\geq m.k} \text{ with } (s_i, t_i) \in (S_i \times SBR_i^{\geq m.k})^{rat, \geq m}\}.$$

Here, when we say that "t_i induces the conditional belief vector b_i", we mean that $\text{marg}_{S_{-i}(h)} \beta_i(t_i, h) = b_i(h)$ for every $h \in H_i$. We prove the following result.

Claim. For every period m and number $k \in \{0, 1, ...\}$ it holds that (i) $\hat{B}_i^{m.k} \subseteq B_i^{m.k+1}$, (ii) $B_i^{m.k+1} \subseteq \hat{B}_i^{m.k}$ and for every $b_i \in B_i^{m.k+1}$ we have that $t_i^{b_i} \in SBR_i^{\geq m.k}$, (iii) $\hat{S}_i^{m.k} \subseteq S_i^{m.k+1}$ and (iv) $S_i^{m.k+1} \subseteq \hat{S}_i^{m.k}$.

Proof of claim. We show the four statements by induction on $m.k$.

We start with $m_+.0$. Then, $\hat{B}_i^{m_+.0}$ is, by definition, the set of conditional belief vectors induced by the types in $SBR_i^{\geq m_+.0}$. As $SBR_i^{\geq m_+.0} = T_i$, this is the set of all conditional belief vectors, and hence $\hat{B}_i^{m_+.0} = B_i$. As, by construction, $B_i^{m_+.1} = B_i$ as well, it follows that $\hat{B}_i^{m_+.0} = B_i^{m_+.1}$. Moreover, as $SBR_i^{\geq m_+.0} = T_i$, for every $b_i \in B_i^{m_+.1}$ we have that $t_i^{b_i} \in SBR_i^{\geq m_+.0}$. This establishes (i) and (ii).

Moreover, $\hat{S}_i^{m_+.0}$ contains precisely those strategies that are optimal from period m_+ onwards for some $t_i \in SBR_i^{\geq m_+.0}$. As $SBR_i^{\geq m_+.0} = T_i$, these are precisely the strategies that are optimal from period m_+ onwards for some conditional belief vector $b_i \in B_i$. By definition, these are precisely the strategies in $S_i^{m_+.1}$. Hence, we conclude that $\hat{S}_i^{m_+.0} = S_i^{m_+.1}$. This establishes (iii) and (iv).

Next, take some $m.k \neq m_+.0$, and assume that the claim holds for $m.k - 1$ if $k \geq 1$, and that the claim holds for any $m + 1.k'$ if $k = 0$. We distinguish two cases: (1) $k = 0$, and (2) $k \geq 1$.

Case 1. Suppose that $k = 0$. Then, by definition, there is some round K such that $\hat{B}_i^{m.0} = \hat{B}_i^{m+1.K}$ and $B_i^{m.1} = B_i^{m+1.K+1}$. As, by the induction assumption, $\hat{B}_i^{m+1.K} = B_i^{m+1.K+1}$, we conclude that $\hat{B}_i^{m.0} = B_i^{m.1}$. Moreover, by construction, $\hat{S}_i^{m.0}$ is the set of strategies that are optimal, from period m onwards, for some $b_i \in \hat{B}_i^{m.0}$, whereas $S_i^{m.1}$ is the set of strategies that are optimal, from period m onwards, for some $b_i \in B_i^{m.1}$. Since $\hat{B}_i^{m.0} = B_i^{m.1}$, it follows that $\hat{S}_i^{m.0} = S_i^{m.1}$.

Case 2. Suppose that $k \geq 1$.

(i) We show that $\hat{B}_i^{m.k} \subseteq B_i^{m.k+1}$. Take some $b_i \in \hat{B}_i^{m.k}$. Then, there is some $t_i \in SBR_i^{\geq m.k}$ that induces b_i. By definition, $SBR_i^{\geq m.k} \subseteq SBR_i^{\geq m.k-1}$, and hence $b_i \in \hat{B}_i^{m.k-1}$. By the induction assumption on (i) it follows that $b_i \in B_i^{m.k}$. Hence, we only need to show that b_i strongly believes $S_{-i}^{m.k}$. Let $h \in H_i$ be such that $S_{-i}^{m.k} \cap S_{-i}(h) \neq \emptyset$. We must show that $b_i(h)(S_{-i}^{m.k}) = 1$. By the induction assumption applied to (iii) and (iv) we know that $S_{-i}^{m.k} = \hat{S}_{-i}^{m.k-1}$. Hence, by the definition of $\hat{S}_j^{m.k-1}$ for every $j \neq i$, we know that $\times_{j\neq i}(S_j \times SBR_j^{\geq m.k-1})^{rat,\geq m} \cap (S_{-i}(h) \times T_{-i}) \neq \emptyset$. Therefore, since $t_i \in SBR_i^{\geq m.k}$, we conclude that $\beta_i(t_i, h)(\times_{j\neq i}(S_j \times SBR_j^{\geq m.k-1})^{rat,\geq m}) = 1$. This, in turn, implies that $b_i(h)(\hat{S}_{-i}^{m.k-1}) = 1$. As, by the induction assumption on (iii) and (iv), $S_{-i}^{m.k} = \hat{S}_{-i}^{m.k-1}$, we conclude that $b_i(h)(S_{-i}^{m.k}) = 1$. Hence, b_i strongly believes $S_{-i}^{m.k}$. Since $b_i \in B_i^{m.k}$, it follows that $b_i \in B_i^{m.k+1}$. As such, $\hat{B}_i^{m.k} \subseteq B_i^{m.k+1}$.

(ii) We show that $B_i^{m.k+1} \subseteq \hat{B}_i^{m.k}$ and for every $b_i \in B_i^{m.k+1}$ we have that $t_i^{b_i} \in SBR_i^{\geq m.k}$. Take some $b_i \in B_i^{m.k+1}$. Then, in particular, $b_i \in B_i^{m.k}$ and hence we know, by the induction assumption on (ii), that $t_i^{b_i} \in SBR_i^{\geq m.k-1}$. Thus, to prove that $t_i^{b_i} \in SBR_i^{\geq m.k}$ it only remains to show that $t_i^{b_i}$ strongly believes $\times_{j\neq i}$ $(S_j \times SBR_j^{\geq m.k-1})^{rat.\geq m}$. To this end, let $h \in H_i$ be such that $(\times_{j\neq i}(S_j \times SBR_j^{\geq m.k-1})^{rat,\geq m}) \cap (S_{-i}(h) \times T_{-i}) \neq \emptyset$. We must show that $\beta_i(t_i^{b_i}, h)(\times_{j\neq i}(S_j \times SBR_j^{\geq m.k-1})^{rat,\geq m}) = 1$.

By definition, $\hat{S}_{-i}^{m.k-1} = \text{proj}_{S_{-i}}(\times_{j\neq i}(S_j \times SBR_j^{\geq m.k-1})^{rat,\geq m})$. Note that by the induction assumption of (iii) and (iv) we have that $\hat{S}_{-i}^{m.k-1} = S_{-i}^{m.k}$. Therefore, we have $S_{-i}(h) \cap S_{-i}^{m.k} \neq \emptyset$. Since $b_i \in B_i^{m.k+1}$ it follows that $b_i(h)(S_{-i}^{m.k}) = 1$, and hence $b_i(h)(\hat{S}_{-i}^{m.k-1}) = 1$. By the definition of $t_i^{b_i}$ in (10.2) we have $\beta_i(t_i^{b_i}, h)((\hat{S}_{-i}^{m.k-1} \cap S_{-i}(h)) \times T_{-i}) = 1$, and that $\beta_i(t_i^{b_i}, h)$ assigns probability 1 to $\{(s_j, t_j^{b_j[s_j]})_{j\neq i} \mid s_j \in \hat{S}_j^{m.k-1}$ for all $j \neq i\}$. As $\hat{S}_{-i}^{m.k-1} = S_{-i}^{m.k}$ we conclude that $\beta_i(t_i^{b_i}, h)$ assigns probability 1 to $\{(s_j, t_j^{b_j[s_j]})_{j\neq i} \mid s_j \in S_j^{m.k}$ for all $j \neq i\}$.

Consider a type $t_j^{b_j[s_j]}$ where $s_j \in S_j^{m.k}$. Then, we know by the definition of type $t_j^{b_j[s_j]}$ in (10.1) that $t_j^{b_j[s_j]}$ induces the conditional belief vector $b_j[s_j] \in B_j^{m.k}$, and that s_j is optimal for $b_j[s_j]$ from period m onwards. Hence, s_j is optimal for $t_j^{b_j[s_j]}$ from period m onwards. As $b_j[s_j] \in B_j^{m.k}$ we conclude by the induction assumption of (ii) that $t_j^{b_j[s_j]} \in SBR_j^{\geq m.k-1}$. As s_j is optimal for type $t_j^{b_j[s_j]}$ from period m onwards, it follows that $(s_j, t_j^{b_j[s_j]}) \in (S_j \times SBR_j^{\geq m.k-1})^{rat,\geq m}$. Recall that $\beta_i(t_i^{b_i}, h)$ assigns probability 1 to $\{(s_j, t_j^{b_j[s_j]})_{j\neq i} \mid s_j \in S_j^{m.k}$ for all $j \neq i\}$. Hence, it follows that $\beta_i(t_i^{b_i}, h)(\times_{j\neq i}(S_j \times SBR_j^{\geq m.k-1})^{rat,\geq m}) = 1$. As such, we conclude that $t_i^{b_i}$ strongly believes $\times_{j\neq i}(S_j \times SBR_j^{\geq m.k-1})^{rat.\geq m}$.

Since $t_i^{b_i} \in SBR_i^{\geq m.k-1}$ it follows that $t_i^{b_i} \in SBR_i^{\geq m.k}$. We thus conclude that for every $b_i \in B_i^{m.k+1}$ we have that $t_i^{b_i} \in SBR_i^{\geq m.k}$. Since, by (10.2), $t_i^{b_i}$ induces the conditional belief vector b_i, it follows that $b_i \in \hat{B}_i^{m.k}$. Hence, $B_i^{m.k+1} \subseteq \hat{B}_i^{m.k}$.

(iii) We show that $\hat{S}_i^{m.k} \subseteq S_i^{m.k+1}$. Let $s_i \in \hat{S}_i^{m.k}$. Then, in particular, $s_i \in \hat{S}_i^{m.k-1}$. By the induction assumption of (iii) it follows that $s_i \in S_i^{m.k}$. Since $s_i \in \hat{S}_i^{m.k}$, there is a $t_i \in SBR_i^{\geq m.k}$ such that s_i is optimal for t_i from period m onwards. Let b_i be the conditional belief vector induced by t_i. As the expected utility depends only on first-order beliefs, s_i is optimal for b_i from period m onwards. Since $t_i \in SBR_i^{\geq m.k}$ it follows, by definition, that $b_i \in \hat{B}_i^{m.k}$. By (i) it then follows that $b_i \in B_i^{m.k+1}$. Hence, s_i is optimal for some $b_i \in B_i^{m.k+1}$ from period m onwards. As we have seen above that $s_i \in S_i^{m.k}$, we conclude that $s_i \in S_i^{m.k+1}$. Hence, $\hat{S}_i^{m.k} \subseteq S_i^{m.k+1}$.

(iv) We finally show that $S_i^{m.k+1} \subseteq \hat{S}_i^{m.k}$. Let $s_i \in S_i^{m.k+1}$. Then, by construction, $b_i[s_i] \in B_i^{m.k+1}$ and s_i is optimal for $b_i[s_i]$ from period m onwards. By (ii) we know that $t_i^{b_i[s_i]} \in SBR_i^{\geq m.k}$. Moreover, $t_i^{b_i[s_i]}$ induces the conditional belief vector $b_i[s_i]$. Since the expected utility depends only on first-order beliefs, we conclude that s_i is optimal for $t_i^{b_i[s_i]}$ from period m onwards. This implies that $(s_i, t_i^{b_i[s_i]}) \in (S_i \times SBR_i^{\geq m.k})^{rat.\geq m}$, and hence $s_i \in \hat{S}_i^{m.k}$. Thus, $S_i^{m.k+1} \subseteq \hat{S}_i^{m.k}$. This completes the proof of the claim. $\Diamond$

We are now able to prove the theorem.

(a) Take first a strategy s_i that is forward and backward rationalizable. Then, there is a conditional belief vector $b_i \in B_i^{m_0}$ such that s_i is optimal for b_i from the first period onwards. Note that $b_i \in B_i^{m_0.k+1}$ for all k and hence, by part (ii) of the claim, $t_i^{b_i} \in SBR_i^{\geq m_0.k}$ for all k. Therefore, $t_i^{b_i} \in SBR_i^{\geq m_0}$, and hence $t_i^{b_i}$ expresses common backward strong belief in rationality. As $t_i^{b_i}$ induces the conditional belief vector b_i, and s_i is optimal for b_i from the first period onwards, it follows that s_i is optimal for $t_i^{b_i}$ from the first period onwards. As such, s_i is optimal, from the first period onwards, for a type that expresses common backward strong belief in rationality.

Conversely, suppose that s_i is optimal, from the the first period onwards, for a type t_i that expresses common backward strong belief in rationality. Hence, $t_i \in SBR_i^{\geq m_0}$. Suppose that t_i induces the conditional belief vector b_i. Then, s_i is optimal, from the first period onwards, for b_i. Since $t_i \in SBR_i^{\geq m_0.k}$ for all k, and t_i induces the conditional belief vector b_i, it follows that $b_i \in \hat{B}_i^{m_0.k}$ for all k. By part (i) of the claim it follows that $b_i \in B_i^{m_0.k+1}$ for all k, and hence b_i is forward and backward rationalizable. Since s_i is optimal for b_i from the first period onwards, we conclude that s_i is forward and backward rationalizable.

(b) Take first a strategy $s_i \in S_i^{m.0}$. Then, $s_i \in S_i^{m+1}$. Hence, there is a conditional belief vector $b_i \in B_i^{m+1}$ such that s_i is optimal for b_i from period $m+1$ onwards. Note that $b_i \in B_i^{m+1.k+1}$ for all k and hence, by part (ii) of the claim, $t_i^{b_i} \in SBR_i^{\geq m+1.k}$ for all k. Therefore, $t_i^{b_i} \in SBR_i^{\geq m+1}$, and hence $t_i^{b_i}$ expresses common backward strong belief in rationality from period $m+1$ onwards. As $t_i^{b_i}$ induces the conditional belief vector b_i, and s_i is optimal for b_i from period $m+1$ onwards, it follows that s_i is optimal for $t_i^{b_i}$ from period $m+1$ onwards. As such, s_i is optimal, from period $m+1$ onwards, for a type that expresses common backward strong belief in rationality from period $m+1$ onwards.

Conversely, suppose that s_i is optimal, from period $m+1$ onwards, for a type t_i that expresses common backward strong belief in rationality from period $m+1$ onwards. Hence, $t_i \in SBR_i^{\geq m+1}$. Suppose that t_i induces the conditional belief vector b_i. Then, s_i is optimal, from period $m+1$ onwards, for b_i. Since $t_i \in SBR_i^{\geq m+1.k}$ for all k, and t_i induces the conditional belief vector b_i, it follows that $b_i \in \hat{B}_i^{m+1.k}$ for all k. By part (i) of the claim it follows that $b_i \in B_i^{m+1.k+1}$ for all k, and hence $b_i \in B_i^{m.0}$. Since s_i is optimal for b_i from period $m+1$ onwards, we conclude that $s_i \in S_i^{m.0}$.

(c) Take first a strategy $s_i \in S_i^{m.k+1}$. Hence, there is a conditional belief vector $b_i \in B_i^{m.k+1}$ such that s_i is optimal for b_i from period m onwards. By part (ii) of the claim we conclude that $t_i^{b_i} \in SBR_i^{\geq m.k}$. As

$t_i^{b_i}$ induces the conditional belief vector b_i, and s_i is optimal for b_i from period m onwards, it follows that s_i is optimal for $t_i^{b_i}$ from period m onwards. As such, s_i is optimal, from period m onwards, for a type in $SBR_i^{\geq m.k}$ that expresses k-fold backward strong belief in rationality from period m onwards.

Conversely, suppose that s_i is optimal, from period m onwards, for a type $t_i \in SBR_i^{\geq m.k}$ that expresses k-fold backward strong belief in rationality from period m onwards. Suppose that t_i induces the conditional belief vector b_i. Then, s_i is optimal, from period m onwards, for b_i. Since $t_i \in SBR_i^{\geq m.k}$ and t_i induces the conditional belief vector b_i, it follows that $b_i \in \hat{B}_i^{m.k}$. By part (i) of the claim it follows that $b_i \in B_i^{m.k+1}$. Since s_i is optimal for b_i from period m onwards, we conclude that $s_i \in S_i^{m.k+1}$. This completes the proof. ■

10.3 Proof of Theorem 5.1

Let $S_i^{fbr,m.k}$ and $B_i^{fbr,m.k}$ be the sets of strategies and conditional belief vectors for player i that result from period m, round k, of the forward and backward rationalizability procedure. Similarly, we define the sets $S_i^{br,m.k}$ and $B_i^{br,m.k}$ for the backwards rationalizability procedure. For every period m, let K_m be the first round where both procedures terminate at period m. We show, by induction on $m.k$, that for all players i,

$$B_i^{fbr,m.k} \subseteq B_i^{br,m.k} \text{ and } S_i^{fbr,m.k} \subseteq S_i^{br,m.k}. \tag{10.3}$$

For $m.k = m_+.0$ this is true, since $B_i^{fbr,m_+.0} = B_i^{br,m_+.0} = B_i$ and $S_i^{fbr,m_+.0} = S_i^{br,m_+.0} = S_i$.

Now, take some $m.k \neq m_+.0$, and assume that (10.3) holds for $m{+}1.K_{m+1}$ if $k = 0$ or for $m.k{-}1$ if $k \geq 1$. If $k = 0$ then, by definition, $B_i^{fbr,m.0} = B_i^{fbr,m+1.K_{m+1}}$, $B_i^{br,m.0} = B_i^{br,m+1.K_{m+1}}$, $S_i^{fbr,m.0} = S_i^{fbr,m+1.K_{m+1}}$ and $S_i^{br,m.0} = S_i^{br,m+1.K_{m+1}}$. Thus, (10.3) would hold by the induction assumption.

Suppose now that $k \geq 1$ and that (10.3) holds for $m.k - 1$. We first show that $B_i^{fbr,m.k} \subseteq B_i^{br,m.k}$. Take some $b_i \in B_i^{fbr,m.k}$. Then, in particular, $b_i \in B_i^{fbr,m.k-1}$. By the induction assumption it follows that $b_i \in B_i^{br,m.k-1}$. As $b_i \in B_i^{fbr,m.k}$ we know that b_i strongly believes $S_{-i}^{fbr,m.k-1}$. Take some $h \in H_i^m$. Since so far no restrictions have been imposed on the optimality of strategies at information sets preceding h, strongly believing $S_{-i}^{fbr,m.k-1}$ implies that $b_i(h)(S_{-i}^{fbr,m.k-1}) = 1$. As, by the induction assumption, $S_{-i}^{fbr,m.k-1} \subseteq S_{-i}^{br,m.k-1}$, it follows that $b_i(h)(S_{-i}^{br,m.k-1}) = 1$. As $b_i \in B_i^{br,m.k-1}$, we conclude that $b_i \in B_i^{br,m.k}$. Hence, $B_i^{fbr,m.k} \subseteq B_i^{br,m.k}$.

We next show that $S_i^{fbr,m.k} \subseteq S_i^{br,m.k}$. Take some $s_i \in S_i^{fbr,m.k}$. Then, s_i is optimal, from period m onwards, for some $b_i \in B_i^{fbr,m.k}$. As we have seen above that $B_i^{fbr,m.k} \subseteq B_i^{br,m.k}$, we conclude that s_i is optimal, from period m onwards, for some $b_i \in B_i^{br,m.k}$. Hence, $s_i \in S_i^{br,m.k}$. We have thus shown that $S_i^{fbr,m.k} \subseteq S_i^{br,m.k}$.

By induction, it follows that $B_i^{fbr,m_0.K_{m_0}} \subseteq B_i^{br,m_0.K_{m_0}}$ and $S_i^{fbr,m_0.K_{m_0}} \subseteq S_i^{br,m_0.K_{m_0}}$, which completes the proof. ■

10.4 Proof of Lemma 6.1

Without loss of generality, suppose that the first time period is 1. For every period $m \in \{1, ..., m_+\}$ and $k \in \{0, 1, 2, ..., K_m\}$, let $m.k$ be the elimination step in period m, round k, of the forward and backward rationalizability procedure. Here, K_m denotes the last round of the procedure at period m. This leads to the sequence

$$(D^{m_+.0}, D^{m_+.1}, ..., D^{m_+.K_{m_+}},\ D^{m_+-1.0}, D^{m_+-1.1}, ..., D^{m_+-1.K_{m_+-1}},\ ...\ , D^{1.0}, D^{1.1}, ..., D^{1,K_1})$$

of products of strategy sets.

We show that this sequence is an elimination order for sb. By definition, $D^{m_+.0} = \times_{i \in I} S_i$ and $sb(D^{1.K_1}) = D^{1.K_1}$. It remains to show condition (b) in the definition of an elimination order.

Consider first a step $m.k$ with $k \leq K_m - 1$. Then, by definition, $D^{m.k+1} = \times_{i \in I} D_i^{m.k+1}$, where

$$\begin{aligned} D_i^{m.k+1} = \{s_i \in D_i^{m.k} \mid & \text{ there is some } b_i \in B_i \text{ that strongly believes } D_{-i}^{m_+.0}, D_{-i}^{m_+.1}, ..., D_{-i}^{m.k} \\ & \text{such that } s_i \text{ is optimal for } b_i \text{ from period } m \text{ onwards}\} \end{aligned} \qquad (10.4)$$

for every player i. Define, for every player i, the set

$$\begin{aligned} E_i^{m.k+1} = \{s_i \in D_i^{m.k} \mid & \text{ there is some } b_i \in B_i \text{ that strongly believes } D_{-i}^{m.k} \\ & \text{such that } s_i \text{ is optimal for } b_i \text{ at every } h \in H(D^{m.k}) \cap H_i(s_i) \\ & \text{that belongs to } H^\tau \text{ for some } \tau \geq m\}. \end{aligned} \qquad (10.5)$$

Claim. $E_i^{m.k+1} = D_i^{m.k+1}$.

Proof of claim. Clearly, $D_i^{m.k+1} \subseteq E_i^{m.k+1}$. To prove that $E_i^{m.k+1} \subseteq D_i^{m.k+1}$, take some $s_i \in E_i^{m.k+1}$. We will show that $s_i \in D_i^{m.k+1}$. As $s_i \in E_i^{m.k+1}$, there is some $b_i^{m.k} \in B_i$ that strongly believes $D_{-i}^{m.k}$ such that s_i is optimal for $b_i^{m.k}$ at every $h \in H(D^{m.k}) \cap H_i(s_i)$ that belongs to H^τ for some $\tau \geq m$. To show that $s_i \in D_i^{m.k+1}$ we distinguish two cases: (1) $k \geq 1$ and (2) $k = 0$.

Case 1. Suppose that $k \geq 1$. As $s_i \in E_i^{m.k+1}$ we know, by definition, that $s_i \in D_i^{m.k}$. Hence, there is some $b_i' \in B_i$ that strongly believes $D_{-i}^{m_+.0}, ..., D_{-i}^{m.k-1}$ such that s_i is optimal for b_i' from period m onwards. Define the conditional belief vector b_i where

$$b_i(h) := \begin{cases} b_i^{m.k}(h), & \text{if } D_{-i}^{m.k} \cap S_{-i}(h) \neq \emptyset \\ b_i'(h), & \text{otherwise} \end{cases}$$

for every $h \in H_i$.

Then, by construction, b_i strongly believes $D_{-i}^{m_+.0}, D_{-i}^{m_+.1}, ..., D_{-i}^{m.k}$. To show that s_i is optimal for b_i from period m onwards, take some $h \in H_i(s_i) \cap H^\tau$ for some $\tau \geq m$. We distinguish two cases.

If $D_{-i}^{m.k} \cap S_{-i}(h) \neq \emptyset$, then $h \in H(D^{m.k}) \cap H_i(s_i)$ since $s_i \in D_i^{m.k}$. Moreover, $b_i(h) = b_i^{m.k}(h)$. As, by construction, s_i is optimal for $b_i^{m.k}$ at h, it follows that s_i is optimal for b_i at h.

If $D_{-i}^{m.k} \cap S_{-i}(h) = \emptyset$ then, by construction, $b_i(h) = b_i'(h)$. As s_i is optimal for b_i' at h, it follows that s_i is optimal for b_i at h. Hence, s_i is optimal for b_i from period m onwards.

We finally show that b_i is in B_i, by proving that it is forward consistent. Consider some information sets $h, h' \in H_i$, where h precedes h'. We distinguish three cases:

(i) If $D_{-i}^{m.k} \cap S_{-i}(h) \neq \emptyset$ and $D_{-i}^{m.k} \cap S_{-i}(h') \neq \emptyset$, then b_i coincides with $b_i^{m.k} \in B_i$ at h and h'. Since $b_i^{m.k}$ is forward consistent, it follows that b_i is forward consistent if the game moves from h to h'.

(ii) If $D_{-i}^{m.k} \cap S_{-i}(h) = \emptyset$ and $D_{-i}^{m.k} \cap S_{-i}(h') = \emptyset$, then b_i coincides with $b_i' \in B_i$ at h and h'. Since b_i' is forward consistent, it follows that b_i is forward consistent if the game moves from h to h'.

(iii) Suppose, finally, that $D_{-i}^{m.k} \cap S_{-i}(h) \neq \emptyset$ and $D_{-i}^{m.k} \cap S_{-i}(h') = \emptyset$. Then, $b_i(h) = b_i^{m.k}(h)$ and $b_i(h') = b_i'(h')$. As $D_{-i}^{m.k} \cap S_{-i}(h') = \emptyset$ and $b_i^{m.k}$ strongly believes $D_{-i}^{m.k}$, we have that $b_i^{m.k}(h)(D_{-i}^{m.k}) = 1$. Since $D_{-i}^{m.k} \cap S_{-i}(h') = \emptyset$, it thus follows that $b_i^{m.k}(h)(S_{-i}(h')) = 0$. Thus, b_i is trivially forward consistent if the game moves from h to h'.

By combining the cases (i), (ii) and (iii), we see that b_i is forward consistent.

Hence, for the strategy $s_i \in E_i^{m.k+1}$ there is some $b_i \in B_i$ that strongly believes $D_{-i}^{m_+.0}, D_{-i}^{m_+.1}, ..., D_{-i}^{m.k}$ such that s_i is optimal for b_i from period m onwards. That is, $s_i \in D_i^{m.k+1}$, which completes Case 1.

Case 2. Suppose that $k = 0$. If $m = m_+$, then $D^{m.k} = D^{m_+.0} = \times_{j \in I} S_j$. In that case, it would follow immediately that $s_i \in D_i^{m.k+1} = D_i^{m_+.1}$.

Suppose now that $k = 0$ and $m \leq m_+ - 1$. Then, $D^{m.k} = D^{m.0} = D^{m+1.K_{m+1}}$. Since $s_i \in D_i^{m.0} = D^{m+1.K_{m+1}}$, there is some $b_i' \in B_i$ that strongly believes $D_{-i}^{m_+.0}, ..., D_{-i}^{m+1.K_{m+1}-1}$ such that s_i is optimal for b_i' from period $m+1$ onwards. Define the conditional belief vector b_i where

$$b_i(h) := \begin{cases} b_i^{m.0}(h), & \text{if } D_{-i}^{m.0} \cap S_{-i}(h) \neq \emptyset \\ b_i'(h), & \text{otherwise} \end{cases}$$

for every $h \in H_i$.

To show that $s_i \in D_i^{m.k+1} = D_i^{m.1}$, we show that $b_i \in B_i$, that b_i strongly believes $D_{-i}^{m_+.0}, ..., D_{-i}^{m+1.K_{m+1}-1}, D_{-i}^{m.0}$, and that s_i is optimal for b_i from period m onwards.

Note that, by construction, b_i strongly believes $D_{-i}^{m_+.0}, ..., D_{-i}^{m+1.K_{m+1}-1}$ and strongly believes $D_{-i}^{m.0}$. In the same way as for Case 1, it can be shown that b_i is forward consistent, and hence $b_i \in B_i$.

We finally show that s_i is optimal for b_i from period m onwards. Take some information set $h \in H_i(s_i) \cap H^\tau$, for some $\tau \geq m$.

If $D_{-i}^{m.0} \cap S_{-i}(h) \neq \emptyset$, then $h \in H(D^{m.0}) \cap H_i(s_i)$ since $s_i \in D_i^{m.0}$. Thus, by construction, s_i is optimal for $b_i^{m.0}$ at h. Since $b_i^{m.0}(h) = b_i(h)$, we know that s_i is optimal for b_i at h.

If $D_{-i}^{m.0} \cap S_{-i}(h) = \emptyset$, then, by definition, $b_i(h) = b_i'(h)$. Since $D_{-i}^{m.0} \cap S_{-i}(h) = \emptyset$, it must necessarily be that $h \in H^\tau$ with $\tau \geq m+1$. Indeed, since $D_{-i}^{m.0}$ only imposes restrictions on actions at period $m+1$ and later, it follows that $D_{-i}^{m.0} \cap S_{-i}(h) \neq \emptyset$ for every $h \in H_i \cap H^m$. As s_i is optimal for b_i' from period $m+1$ onwards, we conclude that s_i is optimal for $b_i'(h)$ at h, and thus also for $b_i(h)$ at h.

Hence, s_i is optimal for b_i from period m onwards. Altogether, we see that $b_i \in B_i$, that b_i strongly believes $D_{-i}^{m_+.0}, ..., D_{-i}^{m+1.K_{m+1}-1}, D_{-i}^{m.0}$, and that s_i is optimal for b_i from period m onwards. Hence, $s_i \in D_i^{m.1} = D_i^{m.k+1}$, which completes Case 2.

By Cases 1 and 2 we thus conclude that $E_i^{m.k+1} \subseteq D_i^{m.k+1}$, and hence $E_i^{m.k+1} = D_i^{m.k+1}$. This completes the proof of the claim. $\Diamond$

Since, by construction, $sb(D^{m.k}) \subseteq E^{m.k+1}$, it follows by the claim that

$$sb(D^{m.k}) \subseteq E^{m.k+1} = D^{m.k+1} \subseteq D^{m.k}. \tag{10.6}$$

Consider now the case where $k = K_m$. Then the next step is $m - 1.0$ where, by definition, $D^{m-1.0} = D^{m.K_m}$. Hence, we trivially have that

$$sb(D^{m.K_m}) \subseteq D^{m-1.0} \subseteq D^{m.K_m}. \tag{10.7}$$

By (10.6) and (10.7) we conclude that $(D^{m_+.0}, ..., D^{1.K_1})$ is an elimination order for sb. This completes the proof. ■

10.5 Proof of Theorem 8.1

(a) We first prove the statement concerning forward and backward rationalizability. Let h^* be the singleton information set containing only the initial history of Γ. Then, every strategy $\hat{s}_i \in S_i^{\hat{\Gamma}}(h^*)$ in the expansion that allows for h^* to be reached naturally induces a strategy $f_i(\hat{s}_i) \in S_i^{\Gamma}$ in the subgame, such that $\hat{s}_i$ and $f_i(\hat{s}_i)$ prescribe the same actions at all information sets in $H_i^{\hat{\Gamma}}(\hat{s}_i) \cap H_i^{\Gamma}$. Note that the mapping f_i is onto.

Now, consider a conditional belief vector $\hat{b}_i$ for player i in the expansion $\hat{\Gamma}$. Then, $\hat{b}_i$ naturally induces the conditional belief vector $g_i(\hat{b}_i)$ in the subgame Γ, where for every information set $h \in H_i^{\Gamma}$ and every opponents' strategy combination $(s_j)_{j\neq i} \in S_{-i}(h)$,

$$(g_i(\hat{b}_i))(h)((s_j)_{j\neq i}) := \sum_{(\hat{s}_j)_{j\neq i} \in S_{-i}^{\hat{\Gamma}}(h) : f_j(\hat{s}_j) = s_j \text{ for all } j\neq i} \hat{b}_i(h)((\hat{s}_j)_{j\neq i}). \tag{10.8}$$

It may easily be verified that $g_i(\hat{b}_i)$ is forward consistent and that the mapping $g_i : B_i^{\hat{\Gamma}} \to B_i^{\Gamma}$ is onto.

Suppose, without loss of generality, that the expansion $\hat{\Gamma}$ starts at period 1, that the subgame Γ starts at the singleton information set h^* in period $m_0 \geq 1$, and that the last period in which players are active is m_+ in both $\hat{\Gamma}$ and Γ. It is of course possible that some terminal histories in $\hat{\Gamma}$ are longer than in Γ, but every terminal history in $\hat{\Gamma}$ that passes through information sets in Γ will be in Γ as well. This will therefore not affect $Z(S^{fbr}(\Gamma \mid \hat{\Gamma}))$.

For every player i, period $m \in \{1, ..., m_+\}$ and round k, let $\hat{B}_i^{m.k}$ and $\hat{S}_i^{m.k}$ be the set of conditional belief vectors and the set of strategies for player i that survive round k at period m in the forward and backward rationalizability procedure for $\hat{\Gamma}$. Similarly, we denote by $B_i^{m.k}$ and $S_i^{m.k}$ the set of conditional belief vectors and the set of strategies for player i that survive round k at period m in the forward and backward rationalizability procedure for Γ. For every period $m \in \{m_0, ..., m_+\}$, let K_m be the earliest round in this period at which both procedures terminate.

Claim. For every player i, period $m \in \{m_0, ..., m_+\}$ and round $k \in \{1, ..., K_m\}$, we have that

$$g_i(\hat{B}_i^{m.k}) = B_i^{m.k} \text{ and } f_i(\hat{S}_i^{m.k}) = S_i^{m.k}.$$

Proof of claim. By induction on $m.k$. We start by with $m.k = m_+.0$. Then, $\hat{B}_i^{m_+.0} = B_i^{\hat{\Gamma}}$, $B_i^{m_+.0} = B_i^{\Gamma}$, $\hat{S}_i^{m_+.0} = S_i^{\hat{\Gamma}}$ and $S_i^{m_+.0} = S_i^{\Gamma}$, which implies that

$$g_i(\hat{B}_i^{m_+.0}) = g_i(B_i^{\hat{\Gamma}}) = B_i^{\Gamma} = B_i^{m_+.0} \text{ and } f_i(\hat{S}_i^{m_+.0}) = f_i(S_i^{\hat{\Gamma}}) = S_i^{\Gamma} = S_i^{m_+.0},$$

since both g_i and f_i are onto.

Now, take some $m.k \neq m_+.0$, and assume that the claim holds for $m+1.K_{m+1}$ if $k = 0$ or for $m.k-1$ if $k \geq 1$. If $k = 0$ then, by definition, $\hat{B}_i^{m.0} = \hat{B}_i^{m+1.K_{m+1}}$, $B_i^{m.0} = B_i^{m+1.K_{m+1}}$, $\hat{S}_i^{m.0} = \hat{S}_i^{m+1.K_{m+1}}$ and $S_i^{m.0} = S_i^{m+1.K_{m+1}}$. Thus, the claim would hold by the induction assumption.

Assume now that $k \geq 1$ and that the claim holds for $m.k-1$. To show that $g_i(\hat{B}_i^{m.k}) \subseteq B_i^{m.k}$, take some $b_i \in g_i(\hat{B}_i^{m.k})$. Then, by definition, there is some $\hat{b}_i \in \hat{B}_i^{m.k}$ such that $b_i = g_i(\hat{b}_i)$. By definition of $\hat{B}_i^{m.k}$, it must be that $\hat{b}_i$ strongly believes $\hat{S}_{-i}^{m.k-1}$. In particular, for every $h \in H_i^{\Gamma}$ it must be that

$$\hat{b}_i(h)(\hat{S}_{-i}^{m.k-1}) = 1 \text{ whenever } \hat{S}_{-i}^{m.k-1} \cap S_{-i}^{\hat{\Gamma}}(h) \neq \emptyset. \tag{10.9}$$

Take such $h \in H_i^{\Gamma}$. We will show that

$$(g_i(\hat{b}_i))(h)(S_{-i}^{m.k-1}) = 1 \text{ whenever } S_{-i}^{m.k-1} \cap S_{-i}^{\Gamma}(h) \neq \emptyset. \tag{10.10}$$

Suppose that $S_{-i}^{m.k-1} \cap S_{-i}^{\Gamma}(h) \neq \emptyset$. By the induction assumption we know that $S_{-i}^{m.k-1} = f_{-i}(\hat{S}_{-i}^{m.k-1})$, where $f_{-i}(\hat{S}_{-i}^{m.k-1}) := \times_{j \neq i} f_j(\hat{S}_j^{m.k-1})$. Moreover, by construction, the sets $\hat{S}_{-i}^{m.k-1}$ and $f_{-i}(\hat{S}_{-i}^{m.k-1})$ induce the same actions in Γ. Since $m \geq m_0$, it follows that $\hat{S}_{-i}^{m.k-1} \cap S_{-i}^{\hat{\Gamma}}(h) \neq \emptyset$. By (10.9) we then conclude that $\hat{b}_i(h)(\hat{S}_{-i}^{m.k-1}) = 1$. It then follows by (10.8) that

$$(g_i(\hat{b}_i))(h)(f_{-i}(\hat{S}_{-i}^{m.k-1})) = 1.$$

Since, by the induction assumption, $S_{-i}^{m.k-1} = f_{-i}(\hat{S}_{-i}^{m.k-1})$, we conclude that

$$(g_i(\hat{b}_i))(h)(S_{-i}^{m.k-1}) = 1,$$

which establishes (10.10).

As this holds for every $h \in H_i^{\Gamma}$ with $S_{-i}^{m.k-1} \cap S_{-i}^{\Gamma}(h) \neq \emptyset$, it follows that $g_i(\hat{b}_i)$ strongly believes $S_{-i}^{m.k-1}$. Moreover, as $\hat{b}_i \in \hat{B}_i^{m.k-1}$ and, by the induction assumption, $g_i(\hat{B}_i^{m.k-1}) = B_i^{m.k-1}$, we conclude that

$g_i(\hat{b}_i) \in B_i^{m.k-1}$. Hence, by definition, $b_i = g_i(\hat{b}_i) \in B_i^{m.k}$. Since this applies to every $b_i \in g_i(\hat{B}_i^{m.k})$, it follows that $g_i(\hat{B}_i^{m.k}) \subseteq B_i^{m.k}$.

To show that $B_i^{m.k} \subseteq g_i(\hat{B}_i^{m.k})$, take some $b_i \in B_i^{m.k}$. Then, in particular, $b_i \in B_i^{m.k-1}$. As, by the induction assumption, $B_i^{m.k-1} = g_i(\hat{B}_i^{m.k-1})$, there is some $\hat{b}_i \in \hat{B}_i^{m.k-1}$ such that $b_i = g_i(\hat{b}_i)$.

Moreover, as $b_i \in B_i^{m.k}$ we know, by definition, that b_i strongly believes $S_{-i}^{m.k-1}$. Hence, for every $h \in H_i^\Gamma$ we have that

$$b_i(h)(S_{-i}^{m.k-1}) = 1 \text{ whenever } S_{-i}^{m.k-1} \cap S_{-i}^\Gamma(h) \neq \emptyset.$$

As, by the induction assumption, $S_{-i}^{m.k-1} = f_{-i}(\hat{S}_{-i}^{m.k-1})$, it follows that

$$b_i(h)(f_{-i}(\hat{S}_{-i}^{m.k-1})) = 1 \text{ whenever } f_{-i}(\hat{S}_{-i}^{m.k-1}) \cap S_{-i}^\Gamma(h) \neq \emptyset. \tag{10.11}$$

By construction, $f_{-i}(\hat{S}_{-i}^{m.k-1})$ prescribes the same actions in the subgame Γ as $\hat{S}_{-i}^{m.k-1}$. Moreover, as $m \geq m_0$ and the first history of Γ is in period m_0, the set $\hat{S}_{-i}^{m.k-1}$ imposes no restrictions on actions before Γ starts. We therefore have that

$$f_{-i}(\hat{S}_{-i}^{m.k-1}) \cap S_{-i}^\Gamma(h) \neq \emptyset \text{ if and only if } \hat{S}_{-i}^{m.k-1} \cap S_{-i}^\Gamma(h) \neq \emptyset. \tag{10.12}$$

By combining (10.11) and (10.12), we see that for every $h \in H_i^\Gamma$,

$$b_i(h)(f_{-i}(\hat{S}_{-i}^{m.k-1})) = 1 \text{ whenever } \hat{S}_{-i}^{m.k-1} \cap S_{-i}^\Gamma(h) \neq \emptyset. \tag{10.13}$$

Since $b_i = g_i(\hat{b}_i)$ for some $\hat{b}_i \in \hat{B}_i^{m.k-1}$, it follows by (10.13) and (10.8) that we can choose $\hat{b}_i \in \hat{B}_i^{m.k-1}$ such that

$$\text{for every } h \in H_i^{\hat{\Gamma}}, \text{ we have } \hat{b}_i(h)(\hat{S}_{-i}^{m.k-1}) = 1 \text{ whenever } \hat{S}_{-i}^{m.k-1} \cap S_{-i}^\Gamma(h) \neq \emptyset.$$

This means that $\hat{b}_i$ strongly believes $\hat{S}_{-i}^{m.k-1}$. As $\hat{b}_i \in \hat{B}_i^{m.k-1}$, it follows that $\hat{b}_i \in \hat{B}_i^{m.k}$. Since $g_i(\hat{b}_i) = b_i$ we conclude that $b_i \in g_i(\hat{B}_i^{m.k})$. As this holds for every $b_i \in B_i^{m.k}$ we conclude that $B_i^{m.k} \subseteq g_i(\hat{B}_i^{m.k})$. Together with our insight above that $g_i(\hat{B}_i^{m.k}) \subseteq B_i^{m.k}$, it follows that $g_i(\hat{B}_i^{m.k}) = B_i^{m.k}$.

We next show that $f_i(\hat{S}_i^{m.k}) = S_i^{m.k}$ for every player i. To prove that $f_i(\hat{S}_i^{m.k}) \subseteq S_i^{m.k}$, take some $s_i \in f_i(\hat{S}_i^{m.k})$. Then, there is some $\hat{s}_i \in \hat{S}_i^{m.k}$ such that $s_i = f_i(\hat{s}_i)$. By definition of $\hat{S}_i^{m.k}$, there is some $\hat{b}_i \in \hat{B}_i^{m.k}$ such that $\hat{s}_i$ is optimal for $\hat{b}_i$ at every $h \in H_i^{\hat{\Gamma}}(\hat{s}_i)$ from period m onwards.

Take some $h \in H_i^\Gamma(s_i)$ from period m onwards. We will show that s_i is optimal for $(g_i(\hat{b}_i))(h)$ at h. Recall that, for every $\hat{s}_i' \in \hat{S}_i$, the transformed strategy $f_i(\hat{s}_i')$ induces the same actions as $\hat{s}_i$ at information sets in Γ. This insight, together with (10.8), leads to the conclusion that

$$u_i(\hat{s}_i', \hat{b}_i(h)) = u_i(f_i(\hat{s}_i'), (g_i(\hat{b}_i))(h) \text{ for all } \hat{s}_i' \in S_i^{\hat{\Gamma}}(h). \tag{10.14}$$

As $\hat{s}_i$ is optimal for $\hat{b}_i$ at h, we know that

$$u_i(\hat{s}_i, \hat{b}_i(h)) \geq u_i(\hat{s}_i', \hat{b}_i(h)) \text{ for all } \hat{s}_i' \in S_i^{\hat{\Gamma}}(h).$$

Together with (10.14), this yields

$$u_i(f_i(\hat{s}_i), (g_i(\hat{b}_i))(h) \geq u_i(f_i(\hat{s}_i'), (g_i(\hat{b}_i))(h) \text{ for all } \hat{s}_i' \in S_i^{\hat{\Gamma}}(h).$$

As $s_i = f_i(\hat{s}_i)$ and $f_i(S_i^{\hat{\Gamma}}(h)) = S_i^{\Gamma}(h)$, it follows that

$$u_i(s_i, (g_i(\hat{b}_i))(h) \geq u_i(s_i', (g_i(\hat{b}_i))(h) \text{ for all } s_i' \in S_i^{\Gamma}(h).$$

Since this holds for every $h \in H_i^{\Gamma}(s_i)$ from period m onwards, we conclude that s_i is optimal for $g_i(\hat{b}_i)$ at all $h \in H_i^{\Gamma}(s_i)$ from period m onwards.

As $\hat{b}_i \in \hat{B}_i^{m.k}$, we know from above that $g_i(\hat{b}_i) \in B_i^{m.k}$. Since s_i is optimal for $g_i(\hat{b}_i)$ at all $h \in H_i^{\Gamma}(s_i)$ from period m onwards, it follows, by definition, that $s_i = f_i(\hat{s}_i) \in S_i^{m.k}$. As this holds for every $s_i \in f_i(\hat{S}_i^{m.k})$, we conclude that $f_i(\hat{S}_i^{m.k}) \subseteq S_i^{m.k}$.

We next prove that $S_i^{m.k} \subseteq f_i(\hat{S}_i^{m.k})$. Take some $s_i \in S_i^{m.k}$. Then, by definition, there is some $b_i \in B_i^{m.k}$ such that s_i is optimal for b_i at every $h \in H_i^{\Gamma}(s_i)$ from period m onwards.

From above we know that $B_i^{m.k} = g_i(\hat{B}_i^{m.k})$, and hence there is some $\hat{b}_i \in \hat{B}_i^{m.k}$ with $b_i = g_i(\hat{b}_i)$. Choose some strategy $\hat{s}_i \in S_i^{\hat{\Gamma}}$ such that (i) $f_i(\hat{s}_i) = s_i$, and (ii) $\hat{s}_i$ is optimal for $\hat{b}_i$ at every $h \in H_i^{\hat{\Gamma}}(\hat{s}_i) \backslash H_i^{\Gamma}$ from period m onwards.

Now, take some $h \in H_i^{\Gamma}(s_i)$ from period m onwards. As s_i is optimal for b_i at h, we have that

$$u_i(s_i, b_i(h)) \geq u_i(s_i', b_i(h)) \text{ for all } s_i' \in S_i^{\Gamma}(h).$$

Since $s_i = f_i(\hat{s}_i)$ and $b_i = g_i(\hat{b}_i)$, we know that

$$u_i(f_i(\hat{s}_i), (g_i(\hat{b}_i))(h)) \geq u_i(f_i(\hat{s}_i'), (g_i(\hat{b}_i))(h)) \text{ for all } \hat{s}_i' \in S_i^{\hat{\Gamma}}(h).$$

Together with (10.14) we then conclude that

$$u_i(\hat{s}_i, \hat{b}_i(h)) \geq u_i(\hat{s}_i', \hat{b}_i(h)) \text{ for all } \hat{s}_i' \in S_i^{\hat{\Gamma}}(h),$$

which means that $\hat{s}_i$ is optimal for $\hat{b}_i$ at h. As this holds for every $h \in H_i^{\Gamma}(\hat{s}_i)$ from period m onwards, and since we know from above that $\hat{s}_i$ is optimal for $\hat{b}_i$ at every $h \in H_i^{\hat{\Gamma}}(\hat{s}_i) \backslash H_i^{\Gamma}$ from period m onwards, we conclude that $\hat{s}_i$ is optimal for $\hat{b}_i$ at every $h \in H_i^{\hat{\Gamma}}(\hat{s}_i)$ from period m onwards. This, together with the fact that $\hat{b}_i \in \hat{B}_i^{m.k}$, implies that $\hat{s}_i \in \hat{S}_i^{m.k}$.

As $s_i = f_i(\hat{s}_i)$, we conclude that $s_i \in f_i(\hat{S}_i^{m.k})$. Since this holds for every $s_i \in S_i^{m.k}$, we conclude that $S_i^{m.k} \subseteq f_i(\hat{S}_i^{m.k})$. Together with the insight above that $f_i(\hat{S}_i^{m.k}) \subseteq S_i^{m.k}$, it follows that $f_i(\hat{S}_i^{m.k}) = S_i^{m.k}$.

By induction on $m.k$, the proof of the claim is complete. $\diamondsuit$

To prove expansion monotonicity, take some $s_i \in S_i^{fbr}(\Gamma \mid \hat{\Gamma})$. Then, there is some conditional belief vector $\hat{b}_i \in \hat{B}_i^{1.K_1}$ that survives the forward and backward rationalizability procedure in $\hat{\Gamma}$ such that s_i is optimal for $\hat{b}_i$ at all $h \in H_i^{\Gamma}(s_i)$. In particular, we then know that $\hat{b}_i \in \hat{B}_i^{m_0.K_{m_0}}$, where m_0 is the period where the subgame Γ starts. By the claim it follows that $g_i(\hat{b}_i) \in B_i^{m_0.K_{m_0}}$, which means that $g_i(\hat{b}_i) \in B_i^{fbr}(\Gamma)$.

Since s_i is optimal for $\hat{b}_i$ at all $h \in H_i^\Gamma(s_i)$ it follows by (10.14) that s_i is optimal for $g_i(\hat{b}_i)$ at all $h \in H_i^\Gamma(s_i)$. Since $g_i(\hat{b}_i) \in B_i^{fbr}(\Gamma)$ we conclude that $s_i \in S_i^{fbr}(\Gamma \mid \Gamma)$. As this holds for every $s_i \in S_i^{fbr}(\Gamma \mid \hat{\Gamma})$, we conclude that $S_i^{fbr}(\Gamma \mid \hat{\Gamma}) \subseteq S_i^{fbr}(\Gamma \mid \Gamma)$, and hence expansion monotonicity holds for the forward and backward rationalizability procedure.

(b) The proof for backwards rationalizability is very similar, since both *backwards rationalizability* and *forward and backward rationalizability* proceed in a backward inductive fashion. The proof is therefore left to the reader. ■

References

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

[2] Battigalli, P., Catonini, E. and J. Manili (2023), Belief change, rationality, and strategic reasoning in sequential games, *Games and Economic Behavior* **142,** 527–551.

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

[4] Battigalli, P. and N. de Vito (2021), Beliefs, plans, and perceived intentions in dynamic games, *Journal of Economic Theory* **195,** 105283.

[5] Catonini, E. (2019), Rationalizability and epistemic priority orderings, *Games and Economic Behavior* **114,** 101–117.

[6] Catonini, E. (2020), On non-monotonic strategic reasoning, *Games and Economic Behavior* **120,** 209–224.

[7] Catonini, E. and A. Penta (2022), Backward induction reasoning beyond backward induction, *Barcelona School of Economics Working Paper 1315.*

[8] Chen, J. and S. Micali (2013), The order independence of iterated dominance in extensive games, *Theoretical Economics* **8,** 125–163.

[9] Fukuda, S. (2024), The existence of universal qualitative belief spaces, *Journal of Economic Theory* **216,** 105784.

[10] Guarino, P. (2024), Topology-free type structures with conditioning events, Forthcoming at *Economic Theory.*

[11] Heifetz, A., Meier, M. and B.C. Schipper (2013), Dynamic unawareness and rationalizable behavior, *Games and Economic Behavior* **81,** 50–68.

[12] Heifetz, A. and A. Perea (2015), On the outcome equivalence of backward induction and extensive form rationalizability, *International Journal of Game Theory* **44,** 37–59.

[13] Kechris, A. (1995), *Classical Descriptive Set Theory,* Springer, Graduate Texts in Mathematics.

[14] Kohlberg, E. and J.-F. Mertens (1986), On the strategic stability of equilibria, *Econometrica* **54,** 1003–1037.

[15] Kreps, D.M. and R. Wilson (1982), Sequential equilibria, *Econometrica* **50**, 863–894.

[16] Kuhn, H.W. (1953), Extensive games and the problem of information, in H.W. Kuhn and A.W. Tucker (eds.), *Contributions to the Theory of Games,* Volume II (Princeton University Press, Princeton, NJ), pp. 193–216 (*Annals of Mathematics Studies* **28**).

[17] Pearce, D.G. (1984), Rationalizable strategic behavior and the problem of perfection, *Econometrica* **52,** 1029–1050.

[18] Penta, A. (2015), Robust dynamic implementation, *Journal of Economic Theory* **160,** 280–316.

[19] Perea, A. (2012), *Epistemic Game Theory: Reasoning and Choice,* Cambridge University Press.

[20] Perea, A. (2014), Belief in the opponents' future rationality, *Games and Economic Behavior* **83,** 231–254.

[21] Perea, A. (2017), Order independence in dynamic games, *Epicenter Working Paper No. 8.*

[22] Perea, A. (2018), Why forward induction leads to the backward induction outcome: A new proof for Battigalli's theorem, *Games and Economic Behavior* **110,** 120–138.

[23] Perea, A. (2024), More reasoning, less outcomes: A monotonicity result for reasoning in dynamic games, *Epicenter Working Paper No. 32.*

[24] Reny, P. (1992), Backward induction, normal form perfection and explicable equilibria, *Econometrica* **60,** 627–649.

[25] Selten, R. (1965), Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit, *Zeitschrift für die Gesammte Staatswissenschaft* **121,** 301–324, 667–689.

[26] Selten, R. (1975), Reexamination of the perfectness concept for equilibrium points in extensive games, *International Journal of Game Theory* **4,** 25–55.

[27] Shimoji, M. and J. Watson (1998), Conditional dominance, rationalizability, and game forms, *Journal of Economic Theory* **83,** 161–195.

[28] van Damme, E. (1984), A relation between perfect equilibria in extensive form games and proper equilibria in normal form games, *International Journal of Game Theory* **13**, 1–13.

[29] von Neumann, J. and O. Morgenstern (1953), *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ.

[30] Zermelo, E. (1913), Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels, *Proceedings Fifth International Congress of Mathematicians* **2,** 501–504.